ივანე ჯავახიშვილის სახელობის თბილისის
სახელმწიფო უნივერსიტეტი

ო. ფურთუხია

აღწერითი სტატისტიკა, ალბათობა, სტატისტიკური დასკვნების თეორია

თბილისი
2008

წინამდებარე სახელმძღვანელო წარმოადგენს ავტორის მიერ წლების განმავლობაში ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტში ალბათობის თეორიასა და მათემატიკურ სტატისტიკაში წაკითხული ლექციების კურსის გადამუშავებულ ვარიანტს. იგი შედგება სამი ნაწილისაგან: აღწერითი (დესკრიფციული) სტატისტიკა, ალბათობის თეორია და სტატისტიკური დასკვნების თეორია. მასში თვალსაჩინოდაა გადმოცემული ის ძირითადი საკითხები, რომლებიც ყველაზე უფრო ხშირად გვხვდება თანამედროვე გამოყენებით კვლევებში. წიგნი არაა გადატვირთული გრძელი დამტკიცებით, ძირითადი აქცენტი გადატანილია შინაარსობრივ მხარეზე და სამოტივაციო და საილუსტრაციო მაგალითებზე. წიგნით პირველადი სარგებლობისათვის სრულიად საკმარისია საშუალო სკოლის მათემატიკის ცოდნა. სასახელმძღვანელო შედგენილია თბილისის სახელმწიფო უნივერსიტეტში ამჟამად მოქმედი სასწავლო პროგრამებისა და სილაბუსების მიხედვით "სტატისტიკასა" (სოციალურ და პოლიტიკურ მეცნიერებათა ფაკულტეტისათვის) და "ალბათობის თეორიასა და მათემატიკურ სტატისტიკაში" (ეკონომიკისა და ბიზნესის ფაკულტეტისათვის).

სახელმძღვანელო განკუთვნილია სოციალურ-პოლიტიკური და ეკონომიკური პროფილის სტუდენტებისათვის. იგი სასარგებლო იქნება აგრეთვე როგორც მაღალი კლასების მათემატიკის პედაგოგებისათვის, ისე სტატისტიკის გამოყენებითი ასპექტებით დაინტერესებული ნებისმიერი მკითხველისათვის.

რედაქტორი: საქართველოს მეცნიერებათა ეროვნული აკადემიის წევრ-კორესპონდენტი, თსუ სრული პროფესორი ე. ნადარაია

რეცენზენტები: თსუ ასოცირებული პროფესორი, ფიზიკა-მათემატიკის მეცნიერებათა დოქტორი ო. ღლონტი,
თსუ მათემატიკის ლაბორატორიის უფრ. ლაბორანტი,
ფიზ.-მათ. მეცნიერებათა კანდიდატი ზ. ხეჩინაშვილი

განხილულია და მოწონებულია ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტის ალბათობის თეორიისა და მათემატიკური სტატისტიკის მიმართულების სემინარზე

ISBN 978-9941-0-0693-7

ს ა რ ჩ ე ვ ი

აღწერითი (დესკრიფციული) სტატისტიკა

ალბათობის თეორია

სტატისტიკური დასკვნების თეორია

აღწერითი (დესკრიფციული) სტატისტიკა

შესავალი

ისევე როგორც გეომეტრია, რომელმაც დაიწყო არსებობა კითხვიდან: "რამდენად დიდია მიწის ამ ნაკვეთის ფართობი", ალბათობის თეორია დაიბადა კითხვიდან: "რამდენად ხშირად" გვხვდება ესა თუ ის ხდომილება (შედეგი) ცდების დიდ სერიაში. კლასიკური ალბათობის თეორიის მიზანს წარმოადგენს ხდომილებების ფარდობითი სიხშირეების გამოთვლის მეთოდების დადგენა მასთან დაკავშირებული სხვა ხდომილებების მოცემული ფარდობითი სიხშირეების საშუალებით. ალბათობა განუზღვრელობისა და მასთან დაკავშირებული რისკის საზომია. თანამედროვე გაგებით, ალბათობის თეორიის საგანს წარმოადგენს "შემთხვევითი მოვლენების" მათემატიკური ანალიზი.

მათემატიკური სტატისტიკა – ეს არის მათემატიკის დარგი, რომელიც შეისწავლის სტატისტიკური მონაცემების სისტემატიზაციის, ანალიზისა და გამოყენების მეთოდებს თეორიული და პრაქტიკული დასკვნების მიღების მიზნით. იგი შედგება ორი ნაწილისაგან: აღწერითი (დესკრიფციული) სტატისტიკა და სტატისტიკური დასკვნების თეორია. მისი მეთოდები ალბათობის თეორიაზეა დაფუძნებული. ალბათობის თეორია წარმოადგენს ხიდს აღწერით სტატისტიკასა და სტატისტიკური დასკვნების თეორიას შორის, იგი საშუალებას იძლევა გავიაზროთ, თუ როგორ იქმნება და გამოიყენება სტატისტიკური პროცედურები, როგორ შეიძლება ავიცილოთ თავიდან ამ მეთოდების არასწორი გამოყენება. გარკვეული აზრით, მათემატიკური სტატისტიკის ამოცანები ალბათობის თეორიის ამოცანების საწინააღმდეგოა. კერძოდ, თუ ალბათობის თეორიის ძირითადი მიზანია რთული ხდომილებების ალბათობების გამოთვლა მოცემული ალბათური მოდელისათვის, მათემატიკური სტატისტიკა მიზნად ისახავს შებრუნებულ ამოცანას – ამა თუ იმ რთულ ხდომილებაზე დაკვირვებების შედეგად გამოვავლინოთ ალბათურ-სტატისტიკური მოდელის სტრუქტურა.

ალბათობის თეორიის როგორც მეცნიერების განვითარება მიეკუთვნება მე-17 საუკუნის მეორე ნახევარს და დაკავშირებულია პასკალის (1623-1662), ფერმას (1601-1665) და ჰიუგენსის (1629-1695) სახელებთან, თუმცა ცალკეული ამოცანები, რომლებიც დაკავშირებული იყო აზარტულ თამაშებში შანსების დათვლასთან, XV-XVI საუკუნეებშიც განიხილებოდა იტალიელი მათემატიკოსების მიერ (კარდანო, პაჩოლი, ტარტალია და სხვა). ამ მეცნიერების ჭეშმარიტი ისტორია კი იწყება იაკობ ბერნულის (1654-1705), პიერ-სიმონ ლაპლასისა (1749-1827) და სიმონ-დენის პუასონის (1781-1840) შრომებიდან. ალბათობის თეორიის განვითარების თანამედროვე პერიოდი დაკავშირებულია ა. ნ. კოლმოგოროვის სახელთან (1903-1987).

მათემატიკური სტატისტიკა როგორც მეცნიერება იწყება კარლ ფრიდრიხ გაუსის (1777-1855) შრომებიდან. მათემატიკური სტატისტიკის განვითარებაში დიდი წვლილი მიუძღვით კ. პირსონს (1857-1936), რ. ა. ფიშერს (1860-1962), ე. ნეიმანს (1894-1977), ა. ნ. კოლმოგოროვს, ნ. ვ. სმირნოვს (1900-1966), ა. ვალდს (1902-1950) და სხვებს. გასული საუკუნის 20-იან

წლებს განეკუთვნება მათემატიკური სტატისტიკის როგორც დამოუკიდებელი მათემატიკური დისციპლინის ჩამოყალიბება.

საქართველოში ალბათობის თეორიისა და მათემატიკური სტატისტიკის სწავლების ტრადიციას საფუძველი დაუდო ანდრია რაზმაძემ (1889-1929), რომელიც კითხულობდა ლექციების კურსს ახლად დაარსებულ თბილისის უნივერსიტეტში, ხოლო მათემატიკის ამ დარგის განვითარებას საფუძველი ჩაუყარა გვანჯი მანიამ (1918-1985), მასვე ეკუთვნის პირველი ქართული სახელმძღვანელოები და მონოგრაფიები ამ მიმართულებით. აღსანიშნავია, რომ საქართველოში ალბათობის თეორიისა და მათემატიკური სტატისტიკის განვითარებას არა მარტო წმინდა მათემატიკური გამოკვლევებით, არამედ გამოყენებითი ხასიათის კვლევებითაც ღრმა კვალი დაამჩნია რევაზ ჩიტაშვილმა (1942-1995), მანვე დაუდო სათავე სტოქასტური ფინანსური ანალიზის განვითარებას საქართველოში.

მათემატიკური სტატისტიკა ფართოდ გამოიყენება ტექნიკურ გამოკვლევებში, ეკონომიკაში, საბანკო სფეროში, მართვის (მენეჯმენტის) თეორიასა და პრაქტიკაში, სოციოლოგიაში, მედიცინაში, გეოლოგიაში, ისტორიაში და ა. შ. დაკვირვებების, გაზომვების, ექსპერიმენტების შედეგების დამუშავებასა და ანალიზთან საქმე გვაქვს როგორც ადამიანის პრაქტიკული მოღვაწეობის ყველა სფეროში, ისე სამეცნიერო კვლევების ყველა მიმართულებაში. ალბათურ-სტატისტიკური მეთოდების გამოყენების გარეშე წარმოუდგენელია რისკის, დაზღვევის, ფინანსური საქმიანობის შეფასება და ანალიზი. ინჟინრები, მენეჯერები, ეკონომისტები, სოციოლოგები, ექიმები, ფსიქოლოგები, ისტორიკოსები, ფილოლოგები, ლინგვისტები წარმატებით იყენებენ ალბათობასა და მათემატიკურ სტატისტიკაზე დაფუძნებულ გადაწყვეტილებების მიღების ინტელექტუალურ ინსტრუმენტებს.

სტატისტიკის აღწერილობითი მხარიდან ნათელი ხდება, თუ ადამიანის მოღვაწეობის რომელ მხარეს ეხება ესა თუ ის ამოცანა: ეკონომიკას, დემოგრაფიას, ბიოლოგიას, მედიცინას, სოციოლოგიას და ა.შ. შესაბამისად, შეიძლება ვისაუბროთ ეკონომეტრიკაზე, დემოგრაფიულ სტატისტიკაზე, ბიომეტრიაზე, ბიო-სამედიცინო სტატისტიკაზე, სტატისტიკაზე სოციალურ მეცნიერებებში და ა.შ., როგორც სტატისტიკის ერთი შეხედვით სხვადასხვა სფეროებზე. სინამდვილეში, ყოველი მათგანი წარმოადგენს მონაცემთა სხვადასხვა ბაზას, რომელთა დამუშავება და ანალიზი ხდება მათემატიკური სტატისტიკის მეთოდების გამოყენებით.

სტატისტიკური მეთოდები და მიდგომები, ისევე როგორც მისი თეორიული ბაზა – ალბათობის თეორია ინტენსიურად ვითარდება მთელს მსოფლიოში. ამ მიმართულებით გამოდის ხუთასზე მეტი სამეცნიერო ჟურნალი. ამერიკის სტატისტიკური ასოციაცია ითვლის 20000-ზე მეტ წევრს, ხოლო ბრიტანეთის სამეფო სტატისტიკური ასოციაცია – 10000-ზე მეტს. აშშ-ს უნივერსიტეტებში სტატისტიკური ფაკულტეტები მეტია ვიდრე მათემატიკური და ფიზიკური. ეკონომეტრიკის სპეციალისტებს (სტატისტიკური მეთოდების სპეციალისტებს ეკონომიკაში) მიღებული აქვთ ექვსი ნობელის პრემია.

როგორ შეიძლება ვირწმუნოთ, რომ საბუნებისმეტყველო მეცნიერებების მიერ შექმნილი სამყაროს სურათი რეალური სამყაროს ადექვატუ-

რია? საქმე იმაშია, რომ საბუნებისმეტყველო მეცნიერებების მიერ დადგენილი კანონზომიერებები საშუალებას გვაძლევს ვიწინასწარმეტყველოთ (წინასწარ განვჭვრიტოთ) ზოგიერთი მოვლენა, ხოლო ეს წინასწარმეტყველება შეიძლება შემოწმდეს პრაქტიკით. ამ შემოწმებების შედეგებით ჩვენ ვასკვნით რამდენად ზუსტად და სრულად აღვიქვამთ ჩვენ სამყაროს. მაგალითად, ნიუტონის მიერ დადგენილი კლასიკური მექანიკის კანონებით შესაძლებელია განვჭვრიტოთ როგორ იმოძრავებს კოსმოსური ხომალდი დედამიწის გარშემო ორბიტაზე, სად დაეცემა ჰორიზონტის მიმართ გარკვეული კუთხით ნასროლი ქვა, როგორ ირხევა ქანქარა.

მაგრამ, სამწუხაროდ ნებისმიერი კანონზომიერება არ იძლევა საშუალებას ზუსტად (ცალსახად) ვიწინასწარმეტყველოთ ამა თუ იმ მოვლენის შედეგი. არსებობს ორი ტიპის კანონზომიერებები. ერთნი საშუალებას იძლევიან გავაკეთოთ ცალსახა წინასწარმეტყველება, მაშინ როდესაც დანარჩენების საფუძველზე კეთდება არა ცალსახა წინასწარმეტყველება, არამედ მხოლოდ ალბათური. პირველი ტიპის კანონზომიერებებს უწოდებენ *დინამიკურს*, ხოლო მეორე ტიპისას – *სტატისტიკურს ან ალბათურს*. კანონზომიერებების ეს დაყოფა შემოთავაზებულ იქნა მე-19 საუკუნის 60-იან წლებში ცნობილი ინგლისელი ფიზიკოსის – ჯეიმს მაქსველის მიერ.

მოვიყვანოთ სტატისტიკური კანონზომიერების მარტივი მაგალითი. ავიღოთ მონეტა და დავიწყოთ მისი აგდება. მონეტის ნებისმიერი აგდება – არის ხდომილება, რომლის შედეგის წინასწარ გამოცნობა შეუძლებელია. მაგრამ შეიძლება დავამტკიცოთ (სხვა სიტყვებით, შეიძლება ვიწინასწარმეტყველოთ), რომ თუ ავაგდებთ მონეტას საკმაოდ ბევრჯერ, მაშინ დაახლოებით ნახევარ შემთხვევაში მოვა “გერბი”, ხოლო მეორე ნახევარ შემთხვევაში – “საფასური”. ეს კანონზომიერება შეიძლება შემოწმდეს პრაქტიკულად, თუ ავაგდებთ მონეტას, ვთქვათ, 500-ჯერ. ცხადია, მონეტის აგდების ესა თუ ის შედეგი შემთხვევითია, მიუხედავად ამისა ადგილი აქვს სტატისტიკურ კანონზომიერებას, რომელიც მდგომარეობს “გერბის” მოსვლის ფარდობითი სიხშირის სტატისტიკურ მდგრადობაში.

ალბათობის თეორიის, როგორც მათემატიკური მეცნიერების, ჩამოყალიბებაში საწყისი ეტაპი იყო სწორედ მრავალრიცხოვანი რეალური მოვლენების სიხშირეების სტატისტიკურ მდგრადობაზე დაკვირვებები. ბ. გნედენკოს მიხედვით ალბათობის თეორია შეიძლება განვმარტოთ როგორც მათემატიკის დარგი, რომელიც იკვლევს იმ შემთხვევითი მოვლენების (ექსპერიმენტების) მათემატიკურ მოდელებს, რომლებსაც ახასიათებს სიხშირეების მდგრადობა. რაც არ უნდა გასაკვირი იყოს სწორედ სტატისტიკური (და არა დინამიკური!) კანონზომიერებები უფრო ადექვატურად აღწერენ რეალურ სამყაროს. ჩვენ ვცხოვრობთ სამყაროში, სადაც ყოველ ფეხის ნაბიჯზე ხდება შემთხვევითი ხდომილებები: შემთხვევითი შეხვედრა, შემთხვევითი პოვნა, შემთხვევითი დაზიანება, შემთხვევითი შეცდომა, შემთხვევითი დაკარგვა, შემთხვევითი აღმოჩენა და ა. შ. უსასრულოდ.

ალბათობის თეორია აგებს სხვადასხვა ხდომილებების ემპირიულად დაკვირვებადი სიხშირეების მათემატიკურ მოდელებს. გამონათქვამი “A ხდომილების ალბათობა ტოლია მაგალითად ნახევრის” გვესმის შემდეგნაირად: დაკვირვებების გარკვეული რაოდენობისას A ხდომილება ხდება

50% შემთხვევაში. ასეთი გაგებით ალბათობის თეორიას გააჩნია ფართე გამოყენებები ადამიანის მოღვაწეობის ყველა სფეროში. მათემატიკური სტატისტიკა გაიგება როგორც მათემატიკური მეცნიერება (საკუთრივ როგორც ალბათობის თეორიის ნაწილი), რომელიც ამოდელირებს მოვლენას, რომელთაც ჯ. ნეიმანი უწოდებს "ინდუქტიურ ქცევას". ადამიანის ქცევაში შეცდომები გარდაუვალია. ადამიანები და სხვა გონიერი არსებები ინტუიციურად ცდილობენ გამოიყენონ მათ მეხსიერებაში დაგროვილი ცოდნის მარაგი და ბუნებაში მიმდინარე მოვლენებზე დაკვირვებები, რათა მათ ქმედებებში რაც შეიძლება იშვიათად იყოს შეცდომები. სწორედ ამ მოქმედებების არჩევას, რომელიც შეესაბამება მეხსიერების მარაგს და ფაქტებზე დაკვირვებებს, უწოდებენ ინდუქტიურ ქცევას. მათემატიკური სტატისტიკა წარმოადგენს ინდუქტიური ქცევის მოდელს, როდესაც დაკვირვებადი ფაქტები "შემთხვევითი" ხასიათისაა. მათემატიკური სტატისტიკის ტიპიური ამოცანა მდგომარეობს შემდეგში: სიხშირეების ტერმინებში ჩამოყალიბებული დაკვირვებების შედეგებიდან გამოვიყვანოთ მოქმედებების ამორჩევის ისეთ მეთოდები, როცა შეცდომების სიხშირე იქნება მინიმალური.

კაცობრიობის განვითარების პროგრესი დაფუძნებულია ადამიანის უნარზე შეამჩნიოს პერმანენტულობა (პერმანენტულობა არის ის რაც არის, ან გვეჩვენება, უცვლელი) როგორც თავის გარშემო საგნებში, ისე ამ საგნებში მიმდინარე ცვლილებებში, რომლებსაც ჩვენ მოვლენებს ვუწოდებთ. პერმანენტულობის შენიშვნა და ინდუქტიური ქცევა ახასიათებს არა მარტო ადამიანებს, არამედ დაბალი განვითარების ორგანიზმებსაც. ხშირად ინდუქტიურ ქცევას მივყავართ კარგ შედეგებამდე, მაგრამ, ბუნებრივია, არა ყოველთვის.

ადამიანების მიერ ერთ-ერთი პირველი შენიშნული პერმანენტულობები იყო ზომები, მანძილები, წონები და ა. შ. შედარებით გვიან აზარტული თამაშების მოთამაშეებმა შენიშნეს ახალი სახის "პერმანენტულობა". ეს იყო ე. წ. ფარდობითი სიხშირე, რომლითაც გვხვდება კონკრეტული შედეგი განმეორებით ცდებში, რომლებშიც ცალკეული ცდის შედეგის განჭვრეტა შეუძლებელია. გაცილებით გვიან იქნა შენიშნული პერმანენტულობა ადამიანის ცხოვრებისეულ მოვლენებში. მე-17 საუკუნეში ინგლისელმა მასონებმა შენიშნეს, რომ სიკვდილიანობის სიხშირე ერთი და იგივე ასაკის ცალკე კაცებში და ცალკე ქალებში, რომლებიც ცხოვრობდნენ ერთი და იგივე პირობებში, წლიდან წლამდე მერყეობდა ძალიან ვიწრო საზღვრებში. იმის დადგენა თუ ვინ დაიღუპება კონკრეტულად 50 წლიანი მამაკაცების მოცემული ჯგუფიდან მომავალ წელს შეუძლებელია, მაგრამ იმის მითითება თუ რამდენი იქნება სიკვდილიანობის რიცხვი ამ ჯგუფში შესაძლებელია დიდი სიზუსტით. როგორც ცნობილია, სწორედ ეს პერმანენტულობა წარმოადგენს სიცოცხლის დაზღვევის საფუძველს, რომელსაც სწავლობს ე.წ. აქტუარული (სადაზღვევო) მათემატიკა.

თავი I
სტატისტიკის საგანი

რატომ უნდა ვისწავლოთ სტატისტიკა?

ჩვენს ინფორმაციის საუკუნეში, მსოფლიო გადავსებულია სხვადასხვა სახის მონაცემებით. საგაზეთო სტატიები და სატელევიზიო ახალი ამბები აჭრელებულია ისეთი ტიპის განცხადებებითა და მტკიცებულებებით, როგორიცაა: "დოუ-ჯონსის საშუალო დაეცა 6 პუნქტით", "სამომხმარებლო საქონელზე ფასების ინდექსი გასული თვის განმავლობაში გაიზარდა 8%-ით", "უკანასკნელი გამოკითხვა გვიჩვენებს, რომ პრეზიდენტის რეიტინგი ამჟამად შეადგენს 63%-ს", "პაციენტთა 98%-ს, რომელიც იმყოფებოდა კლინიკურ გამოკვლევაზე, არ განუცდია რაიმე გვერდითი ეფექტი მკერდის კიბოს ახალი პრეპარატისაგან". ხშირია შემთხვევები, როცა იმისათვის რომ მიმდინარე მოვლენებს მიეცეს გონივრული (ინტელექტუალური) შეფასება, ჩვენ გვესაჭიროება (გვიხდება) გადავამუშაოთ მონაცემების არსებითად დიდი რაოდენობა. მთავრობა, საფინანსო წრეები და მეცნიერ-მკვლევარები ხარჯავენ მილიონობით დოლარს, რათა მოაგროვონ მონაცემები. ფედერალური მთავრობები ხელს უწყობენ მონაცემების შეგროვებას, როგორც საკუთარი ძალისხმევით, ისე კორპორაციების მიმართ მოთხოვნით – რათა მათ მისაწვდომი გახადონ ინფორმაციები. კერძო სექტორი აგრეთვე მონაწილეობს ამ პროცესში. შეგროვებულ მონაცემთა რაოდენობა სულ უფრო და უფრო იზრდება უკანასკნელ პერიოდში.

ჩვენი მიზანია თითოეულ მონაცემს მივცეთ თავისი შინაარსი. კომპიუტერის ერა საშუალებას გვაძლევს ერთის მხრივ, მოვახდინოთ მონაცემების სწრაფი შეგროვება, გადამუშავება, დაჯამება და ანალიზი, ხოლო მეორეს მხრივ, კი – მოვახდინოთ სულ უფრო და უფრო მეტი მონაცემის მიღება და შენახვა. კომპიუტერის საშუალებით ჩვენ შეგვიძლია მივიღოთ მონაცემები, მაგალითად, აქციათა კოტირებების შესახებ, კომპიუტერის კლავიშზე თითის დაჭერით. ჩვენ სწორად უნდა გავაანალიზოთ და მივცეთ შესაბამისი ინტერპრეტაცია ყოველ მონაცემს.

გადაწყვეტილებები ხშირად დაფუძნებულია არასრულ ინფორმაციაზე. მაგალითად, საგამოცდო კომისიის მიერ უნივერსიტეტში მიღებული პირველი კურსის სტუდენტები ისე ირჩევენ თავიანთ ძირითად სპეციალობას, რომ არა აქვთ ნათელი წარმოდგენა მომავალი კარიერის შესახებ. ან პაციენტები, რომლებიც დაავადებული არიან კიბოთი, თანხმდებიან მონაწილეობა მიიღონ ახალი პრეპარატის კლინიკურ გამოკვლევაზე ისე, რომ მათთვის უცნობია ამ პრეპარატის მოქმედების ყველა შესაძლო შედეგი. ანალოგიურად, ის ადამიანები, რომლებიც რეგულარულად ღებულობენ საქმიან გადაწყვეტილებებს გარემო პირობების გათვალისწინებით, არ შეიძლება დარწმუნებულები იყვნენ იმ ფაქტორების მომავალ ყოფაქცევაზე, რომლებიც საბოლოო გავლენას ახდენენ განსახილველი მოვლენების საბოლოო შედეგზე.

მეწარმის მიერ შემოთავაზებული კონტრაქტის საფასური არ შეიძლება იყოს სრულად გარკვეული, ის ვერ მოიცავს მოსალოდნელ ფასებს მომავალში და ვერ იქნება მასში გათვალისწინებული კონკურენტების მიერ შემოთავაზებული ფასები. ამ დასანანი გაურკვევლობის მიუხედავად, მეწარმემ

უნდა დააწესოს თავისი ფასი. ინვესტორმა გარანტირებულად არ იცის ფინანსური ბაზარი იქნება აღმავალი, გაწონასწორებული თუ დაღმავალი. მიუხედავად ამისა, ინვესტორმა უნდა გადაწყვიტოს როგორ დააბალანსოს თავის ფინანსური პორტფელი აქციებით, ობლიგაციებით და ფინანსური ბაზრის სხვა ინსტრუმენტებით, მაშინ როცა ფინასური ბაზრის მომავალი განვითარება მისთვის უცნობია.

განვიხილოთ შემდეგი განცხადებები:

- "IBM-ის აქციების ფასი 6 თვე იქნება უფრო მაღალი, ვიდრე ის არის ამჟამად."
- "თუ სახელმწიფო ბიუჯეტის დეფიციტი ისეთი მაღალია როგორსაც აცხადებენ, მაშინ საპროცენტო განაკვეთი დარჩება მაღალი წლის ბოლომდე."
- "კოლეჯის კურსდამთავრებულის წლიური შემოსავალი იქნება უფრო მეტი, ვიდრე ინდივიდისა საკოლეჯო განათლების გარეშე."

ნებისმიერი ეს წინადადება შეიცავს გაუმართლებელ ენობრივ დაშვებებს (გაურკვევლობას, ილუზორულ, არაგარანტირებულ მტკიცებულებას). რა დაპირებებიც არ უნდა გაკეთდეს, შეუძლებელია დარწმუნებული ვიყოთ მათ სიმართლეში. მიუხედავად ამისა, მკვლევარს შეიძლება ჯეროდეს, რომ მოვლენების განვითარება მომავალი რამოდენიმე თვის განმავლობაში შეიძლება იყოს ისეთი, რომ IBM-ის აქციების ფასი მოსალოდნელია გაიზარდოს, თუმცა ის ვერ იქნება ამაში დარწმუნებული. შესაბამისად, ზემოთმოყვანილი განცხადებები შეიძლება მოდიფიცირებული იქნეს შემდეგნაირად:

- "IBM-ის აქციების ფასი 6 თვე შესაძლებელია (სავარაუდოა) იყოს უფრო მაღალი, ვიდრე ის არის ამჟამად."
- "თუ სახელმწიფო ბიუჯეტის დეფიციტი ისეთი მაღალია როგორსაც აცხადებენ, მაშინ საპროცენტო განაკვეთი ალბათ დარჩება მაღალი წლის ბოლომდე."
- "კოლეჯის კურსდამთავრებულის წლიური შემოსავალი სავარაუდოდ იქნება უფრო მეტი, ვიდრე ინდივიდისა საკოლეჯო განათლების გარეშე."

მნიშვნელოვანია სიტყვების ზუსტი და ფრთხილი შერჩევა. არასწორი (გაუმართლებელი) სიტყვები, არააუცილებლად გასაგები და არაადექვატური, უნდა შეცვალოს. ყოველივე ამის შემდეგ ბუნებრივია ისმის კითხვა თუ რას ნიშნავს: "სავარაუდოა, რომ", "მოსალოდნელია, რომ" ან "ალბათურია, რომ"? ამ კითხვებზე პასუხს იძლევა სწორედ ალბათობის თეორია და მათემატიკური სტატისტიკა.

როგორ უნდა გაეცეს პასუხი კითხვაზე: როგორ იკვებებიან აშშ-ის მოქალაქეები? აშშ-ში დადგენილ იქნა, რომ დიეტა, რომელიც მდიდარია ხილითა და ბოსტნეულით ასოცირდება ისეთი ქრონიკული დაავადებების დაბალ რისკთან, როგორიცაა კიბო. დიეტოლოგები რეკომენდაციას უწევენ ამერიკელებს, რომ სასურველია მათ მიირთვან ხილის ან ბოსტნეულის ხუთი ან მეტი პორცია ყოველდღიურად. რადენიმე მკვლევარმა დიეტოლოგიის დეპარტამენტიდან, ქრონიკული დაავადებების კონტროლისა და პრევენციის ეროვნული ცენტრიდან, კიბოს ეროვნული ინსტიტუტიდან და ჯანმრთელობის ეროვნული ინსტიტუტიდან გადაწყვიტეს გამოეყენებინათ სტატისტიკური პროცედურები, რათა ენახათ როგორი პროგრესია ამ მიმართულებით. ქვემ-

ოთ ახსნილი იქნება მათ მიერ გამოყენებული პროცედურები და კვლევის შედეგები.

აღსანიშნავია, რომ ადამიანების უმრავლესობისთვის ალბათობა და სტატისტიკა ახლობელი გახდა რადიოს, ტელევიზიის, გაზეთებისა და ჟურნალების საშუალებით. მაგალითად, გაზეთებში ნაპოვნი იყო შემდეგი განცხადებები: ა). ჩვეულებრივ, დღეში ერთი აბი ვიტამინი და მინერალი ასაკოვან ადამიანებში ზრდის გარკვეულ იმუნურ რეაქციებს 64 პროცენტით, იმ კვლევების თანახმად, რომელიც ჩატარებული იქნა ნიუ ჯერსის მედიცინისა და სტომატოლოგიის უნივერსიტეტში (USA Weekend, January 6-8, 1995); ბ). ქვეყნის მასშტაბით გამოკითხული 1 000 ოჯახიდან, 40% განაცხადა, რომ ფლობენ სულ ცოტა ერთ უკაბელო ტელეფონს, 9% ფლობს ორს ან მეტს (Tribune-Rev-iew. Greensburg, PA, Jenuary 8, 1995, p.B3); გ). 1994 წელს საცალო ვაჭრობის სპეციალურმა სათხილამურო სპორტის მაღაზიებმა გაყიდეს 760 000 თხილამური საშუალო ფასად 153$ და 176 200 სათხილამურო კოსტუმი საშუალო ფასად 280$ (Tribune-Review, Greensburg, PA, Jenuary 8, 1995, p. G1); დ). ღორტმონდში (პენსილვანია) გაყიდული უძრავი ქონების საშუალო ფასი ბოლო 12 თვის განმავ-ლობაში იყო 64 304$ (Tribune-Review, Greensburg, PA, Jenuary 8, 1995, p.B3); ე). აშშ-ის მოქალაქეები წლის განმავლობაში სპორტულ ტანსაცმელზე საშუალოდ ხარჯავენ 193$ (USA Today, Jenuary 10, 1995).

სტატისტიკა გამოიყენება ადამიანი მოღვაწეობის თითქმის ყველა სფეროში. მაგალითად, სპორტში სტატისტიკოსს შეუძლია აწარმოოს აღრიცხვა იარდების რიცხვისა რაც გაირბინეს ფეხბურთელებმა ფეხბურთის მატჩის განმავლობაში ან დაითვალოს რაოდენობა დარტყმებისა რაც ბეისბოლისტმა გააკეთა სეზონის განმავლობაში. სხვა სფეროებში, როგორიცაა ჯანდაცვა, ადმინისტრატორი დაინტერესებულია იმ მოქალაქეების რაოდენობით, რომლებიც ავადმყოფობდნენ ახალი შტამის ვირუსით გარკვეული პერიოდის განმავლობაში. განათლებაში, მკვლევარს შეიძლება აინტერესებდეს სწავლების ახალი მეთოდი უკეთესია თუ არა, ვიდრე ძველი. აქ ჩამოთვლილია მხოლოდ რამოდენიმე მაგალითი, თუ როგორ შეიძლება იქნეს გამოყენებული სტატისტიკა სხვადასხვა მიმართულებებით.

გარდა ამისა, სტატისტიკა გამოიყენება გამოკვლევის შედეგების გასაანალიზებლად და როგორც ინსტრუმენტი სამეცნიერო კვლევებში გადაწყვეტილების მისაღებად, რომელიც დაფუძნებულია ექსპერიმენტებზე. სტატისტიკის სხვა გამოყენებები მოიცავს ოპერაციების სტატისტიკურ გამოკვლევას, პროდუქციის ხარისხის შემოწმებას, შეფასებასა და პროგნოზირებას.

სტუდენტები სწავლობენ სტატისტიკას რამდენიმე მიზეზის გამო:

1. სტუდენტებს, ისევე როგორც პროფესიონალ ადამიანებს, უნდა შეეძლოთ წაკითხვა და გაგება სხვადასხვა სტატისტიკური გამოკვლევების, რომლებიც ჩატარებულია მათ შესაბამის სფეროებში. ისინი უნდა იყვნენ გათვითცნობიერებული ლექსიკაში, სიმბოლოებში, ცნებებში და სტატისტიკურ პროცედურებში, რომლებიც გამოიყენება ამ კვლევებში.

2. სტუდენტები და პროფესიონალი ადამიანები შესაძლებელია იყვნენ მიწვეული, რათა მათ ჩაატარონ კვლევები შესაბამის სფეროებში, ვინაიდან სტატისტიკური პროცედურები არის კვლევის საფუძველი. იმისათვის, რომ მათ მიაღწიონ შედეგებს, მათ უნდა შეძლონ ექსპერიმენტის დაპროექტება

(დაგეგმვა); უნდა შეძლონ მონაცემების შეგროვება, ორგანიზება, გაანალიზება და შეჯამება; გარდა ამისა, შესაძლებელია, შეძლონ საიმედო წინასწარმეტყველება და გონივრული პროგნოზი სამომავლო გამოყენების შესახებ. მათ ასევე უნდა შეეძლოთ გადმოსცენ კვლევის შედეგები საკუთარი სიტყვებით.

3. სტუდენტებს და პროფესიონალ ადამიანებს აგრეთვე შეუძლიათ გამოიყენონ სტატისტიკის შესწავლიდან მიღებული ცოდნა, რათა გახდნენ უკეთესი მომხმარებლები და მოქალაქეები. მაგალითად, მათ შეუძლიათ მიიღონ ინტელექტუალური (გონივრული) გადაწყვეტილება, თუ რომელი პროდუქტი შეიძინონ სამომხმარებლო კვლევებზე დაყრდნობით, განსაზღვრონ სამთავრობო ხარჯები სტატისტიკური კვლევების მიხედვით და ასე შემდეგ.

ეს მიზეზები შეიძლება ჩაითვალოს მიზნებად იმ სტუდენტებისთვის და პროფესიონალებისთვის, რომლებიც სწავლობენ სტატისტიკას.

სტატისტიკის მიმართულებები.

ამ შემთხვევაში ჩვენი მიზანია გავაცნოთ სტუდენტებს ალბათობისა და სტატისტიკის საწყისი ცნებები, მაგალითად, შემდეგი ტიპის კითხვებზე პასუხის გაცემით:

რა მიმართულებები გააჩნია სტატისტიკას?

რა არის მონაცემი?

როგორაა მონაცემები შერჩეული?

იმისათვის, რომ სტატისტიკოსებმა მიიღონ ცოდნა ერთი შეხედვით უსისტემო ხდომილებებიდან (მოვლენებიდან), ისინი აგროვებენ ინფორმაციას იმ სიდიდეებზე (ცვლადებზე), რომლებიც აღწერენ ხდომილებებს.

სიდიდე *არის მახასიათებელი ან ატრიბუტი, რომელსაც შეუძლია მიიღოს სხვადასხვა მნიშვნელობა.*

მონაცემი *არის გაზომვების ან დაკვირვებების ის მნიშვნელობები, რომლებსაც ღებულობენ სიდიდეები. სიდიდეს, რომლის მნიშვნელობები განისაზღვრება შემთხვევით (დამოკიდებულია შემთხვევითობაზე), უწოდებენ* ***შემთხვევით სიდიდეს.***

დავუშვათ, რომ სადაზღვევო კომპანიამ გამოიკვლია თავისი ჩანაწერები უკანასკნელი რამოდენიმე წლის განმავლობაში და დაადგინა, რომ საშუალოდ 3 ავტომობილი ყოველი 100 ავტომობილიდან, რომელიც მოყვა ავტოსაგზაო შემთხვევაში, დაზღვეული იყო ერთი წლის განმავლობაში. მიუხედავად იმისა, რომ არ არსებობს ხერხი, რომლის საშუალებითაც შეიძლება ვიწინასწარმეტყველოთ რომელი კონკრეტული ავტომობილი მოყვება ავტოსაგზაო შემთხვევაში (ავტო-საგზაო შემთხვევა შემთხვევით ხდება), სადაზღვევო კომპანიას შეუძლია შეიტანოს გარკვეული კორექტივები თავის გადასახადებში, რამდენადაც კომპანიამ იცის ზოგადი ტენდენცია ხანგრძლივი დროის განმავლობაში (კერძოდ, საშუალოდ, დაზღვეული ავტომობილების 3% შესაძლებელია მოხვდეს ავტო-საგზაო შემთხვევაში).

მონაცემის მნიშვნელობების ერთობლიობა ქმნის ***მონაცემების სიმრავლეს.*** მონაცემების სიმრავლის ცალკეულ წევრს ***დაკვირვება, მონაცემი ანუ მონაცემი წერტილი*** ეწოდება.

მონაცემები შეიძლება გამოყენებულ იქნეს სხვადასხვანაირად. ცოდნის ის სისტემა, რომელსაც სტატსიტიკას უწოდებენ, იმის მიხედვით თუ როგორ გამოიყენება მონაცემები, ზოგჯერ იყოფა ორ ძირითად მიმართულებად. ეს ორი მიმართულებაა: 1. *აღწერითი სტატისტიკა* და 2. *სტატისტიკური დასკვნები.*

აღწერით სტატისტიკაში სტატისტიკოსი ცდილობს აღწეროს სიტუაცია. განვიხილოთ მოსახლეობის ნაციონალური აღწერის საკითხი, რომელსაც აშშ მთავრობა ატარებს ყოველ ათ წელიწადში. ამ აღწერის შედეგები გვაძლევს მოსახლეობის საშუალო ასაკს, შემოსავალს და სხვა განსაკუთრებულ მახასიათებლებს აშშ-ის მოსახლეობის. ამ ინფორმაციის მისაღებად, მოსახლეობის აღწერის ბიუროს უნდა გააჩნდეს გარკვეული საშუალებები შესაბამისი მონაცემების შესაგროვებლად. მას შემდეგ რაც მიღებულია (შეგროვებულია) მონაცემები, აღწერის ბიურომ უნდა მოახდინოს მათი ორგანიზება და დაჯამება. საბოლოოდ, აღწერის ბიუროს ესაჭიროება საშუალებები (მეთოდები) რათა მიღებული მონაცემები წარმოადგინოს აზრიანი (ან თვალსაჩინო) ფორმით, მაგალითად, ცხრილების, გრაფიკების ან დიაგრამების ფორმით.

ისტორიული შენიშვნა. აღწერითი სტატისტიკის საწყისები სათავეს იღებს მონაცემთა შეგროვების მეთოდებით, რომლებიც გამოიყენებოდა მოსახლეობის აღწერისას ბაბილონსა და ეგვიპტეში ძვ. წ. აღრიცხვის 4500 – 3000 წლებში. გარდა ამისა, რომის იმპერატორმა ავგუსტმა (ძვ. წ. აღრიცხვის 27–17 წლები) ჩაატარა გამოკვლევა რომის იმპერიის მოქალაქეების როგორც დაბადებისა და სიკვდილიანობის შესახებ, ისე ყოველი მოქალაქის კუთვნილებაში მყოფი საქონლის რაოდენობისა და მათ მიერ წლის განმავლობაში მოყვანილი სასოფლო-სამეურნეო კულტურების შესახებ.

აღწერითი სტატისტიკა გულისხმობს მონაცემების შეგროვებას, ორგანიზებას, დაჯამებასა და წარმოდგენას.

სტატისტიკის მეორე განშტოებას უწოდებენ *სტატისტიკური დასკვნების თეორიას.* აქ, სტატისტიკოსი ცდილობს გააკეთოს დასკვნები *პოპულაციის შერჩევებიდან (ამონარჩევებიდან).* სტატისტიკური დასკვნები იყენებს *ალბათობას,* ე. ი. ხდომილების მოხდენის შანსს. ადამიანების დიდი ნაწილისათვის ცნობილია ალბათობის პრინციპიალური სქემები აზარტული თამაშების სხვადასხვა ფორმებიდან. ადამიანი რომელიც თამაშობს კარტს, სათამაშო კამათელს, ბინგოსა და ლატარეას, იგებს ან კარგავს (აგებს) ალბათური განაწილების კანონის მიხედვით. ალბათობის თეორია ასევე გამოიყენება სადაზღვევო საქმიანობაში და ადამიანის მოღვაწეობის სხვა მრავალ სფეროში.

ისტორიული შენიშვნა. სტატისტიკური დასკვნების თეორია სათავეს იღებს 1600 წლიდან, როდესაც ჯონ გრანტმა გამოაქვეყნა წიგნი მოსახლეობის ზრდის შესახებ, სადაც სიკვდილიანობის ცხრილზე დაყრდნობით გაკეთებულია ბუნებრივი და პოლიტიკური ხასიათის შენიშვნები. თითქმის იმავე დროს, მათემატიკოსმა და ასტრონომმა ედმუნდ ჰალიმ გამოაქვეყნა სიკვდილიანობის პირველი სრულყოფილი ცხრილი (აღსანიშნავია, რომ სადაზღვევო კომპანია იყენებს სიკვდილიანობის ცხრილებს დაზღვევის ტარიფის დასადგენად).

შემოვიღოთ ახლა პოპულაციისა და შერჩევის ცნებები და ვნახოთ რა განსხვავებაა მათ შორის.

პოპულაცია შედგება ყველა იმ ობიექტისაგან, რომელიც შეისწავლება (ის არის დაკვირვების ყველა შესაძლო შედეგთა სიმრავლე).

უმეტეს შემთხვევაში, რიგი მიზეზების გამო (მაგალითად, დანახარჯების სიძვირე, დროის სიმცირე, პოპულაციის მოცულობის სიდიდე, სამედიცინო პრობლემები და ა. შ.) მკვლევარს არა აქვს შესაძლებლობა სტატისტიკური კვლევისათვის გამოიყენოს მთლიანი პოპულაცია. ამიტომ მკვლევარი, როგორც წესი, იყენებს შერჩევას.

შერჩევა არის ობიექტების გარკვეული ჯგუფი (ნაწილი) ამორჩეული პოპულაციიდან.

თუ შერჩევის კითხვები (ობიექტები) სწორადაა შერჩეული, მაშინ უმეტეს შემთხვევაში ისინი იქნებიან იგივე ან ანალოგიური იმ მახასიათებლების რაც კითხვებს (ობიექტებს) გააჩნიათ პოპულაციაში. შერჩევის სწორად ამორჩევის ტექნიკა მოყვანილი იქნება ქვემოთ.

პოპულაციის სრულ აღწერას ჩვენ შერჩევის გამოყოფას და მის შესწავლას ვამჯობინებთ თუნდაც იმიტომ, რომ ნაწილის დაკვირვება უფრო იაფია, ვიდრე მთელისა, თუმცა ხარჯების ეკონომიის გარდა შერჩევის უპირატესობას სხვა მოტივებიც განაპირობებენ. შერჩევის უმნიშვნელოვანესი სახეობაა *შემთხვევითი შერჩევა*. შემთხვევითი შერჩევა გულისხმობს, რომ პოპულაციის ყოველი ელემენტისათვის განსაზღვრულია შერჩევაში მოხვედრის შანსი, ალბათობა. როდესაც შერჩევის გეგმა ისეთია, რომ პოპულაციის ყოველი ელემენტისათვის შერჩევაში მოხვედრის შანსი ერთნაირია, ამბობენ, რომ გვაქვს *მარტივი შერჩევა*; შერჩევა შემთხვევითი რიცხვების გამოყენებით; ე. წ. *სისტემატური შერჩევა* (ვთქვათ, ყოველი მე-10 ერთეული); *განშრევებული შემთხვევითი შერჩევა* – უზრუნველყოფს შერჩევაში პოპულაციის სხვადასხვა ჯგუფის წარმომადგენლობას; *კლასტერული შერჩევა* – პოპულაციას ვყოფთ თანაუკვეთ ნაწილებად – კლასტერებად, შემთხვევით ვირჩევთ კლასტერს და შემდეგ შერჩეულს სრულად აღვწერთ.

სტატისტიკური დასკვნების თეორიის ნაწილს, რომელიც წარმოადგენს პოპულაციაზე წინადადების შეფასებისას გადაწყვეტილების მიღების პროცესს, დაფუძნებულს შერჩევიდან მიღებულ ინფორმაციაზე, ***ჰიპოთეზათა შემოწმება*** ეწოდება. მაგალითად, მკვლევარს შეიძლება სურვილი ჰქონდეს იცოდეს ამცირებს თუ არა ახალი პრეპარატი გულის შეტევათა რაოდენობას 70 წელზე უფრო დიდი ასაკის მამაკაცებში. ამის გამოსაკვლევად, არჩეული იქნება 70 წელს გადაცილებულ მამაკაცთა ორი შერჩევა (ჯგუფი). ერთ-ერთი შერჩევის მამაკაცებს ეძლევათ აღნიშნული პრეპარატი, ხოლო მეორე შერჩევაში მყოფ მამაკაცებს აძლევენ რაღაც ნივთიერებას, რომელსაც ჯანმრთელობის თვალსაზრისით ადამიანისათვის არც სარგებლობის მოტანა შეუძლია და არც ზიანის. ექსპერიმენტის ბოლოს ითვლიან მამაკაცების თითოეულ ჯგუფში მომხდარ გულის შეტევათა რაოდენობას, შემდეგ ატარებენ სტატისტიკური ჰიპოთეზის შემოწმების პროცესს და ღებულობენ გადაწყვეტილებას პრეპარატის ეფექტურობის შესახებ.

სტატისტიკოსები სტატისტიკურ მონაცემებს იყენებენ აგრეთვე იმის გასარკვევად, თუ რა კავშირია სიდიდეებს შორის? მაგალითად, უკანასკნ-

ლი ათწლეულების განმავლობაში განსაკუთრებული ყურადღება ექცევა კავშირის დადგენას მოწევასა და ფილტვების კიბოს შორის. 1964 წელს აშშ-ის მთავარმა ქირურგმა გამოაქვეყნა ნაშრომი "მოწევა და ჯანმრთელობა", სადაც ნათქვამია, რომ მონაცემების განხილვისა და შეფასების შემდეგ მისმა ჯგუფმა აღმოაჩინა გარკვეული კავშირი მოწევასა და ფილტვების კიბოს შორის. მას არ უთქვამს, რომ სიგარეტის მოწევა არის ფილტვების კიბოს ფაქტიური მიზეზი, მაგრამ ამბობს, რომ მათ შორის არის კავშირი. ეს დასკვნა დაფუძნებული იყო ჰამონდისა და ჰორნის მიერ 1958 წელს ჩატარებულ გამოკვლევაზე. ამ კვლევისას დაკვირვებულ იქნა 187783 მამაკაცი 45 თვის განმავლობაში. აღმოჩნდა, რომ ფილტვების კიბოთი გამოწვეული სიკვდილიანობის კოეფიციენტი მამაკაცების ამ ჯგუფში 10-ჯერ უფრო დიდი იყო მწეველებში არამწეველებთან შედარებით.

და ბოლოს, წარსული და მიმდინარე მონაცემებისა და პირობების შესწავლით სტატისტიკოსები ცდილობენ ამ ინფორმაციის საფუძველზე გააკეთონ პროგნოზი (იწინასწარმეტყველონ რა იქნება მომავალში). მაგალითად, ავტომობილების დილერს (გამყიდველს), რომელიც ნახავს გასული გაყიდვების ჩანაწერებს კონკრეტული თვეების მიხედვით, შეუძლია გადაწყვიტოს თუ რა ტიპის და რამდენი ავტომობილი (ამა თუ იმ ტიპის) დასჭირდება მომავალი წლის კონკრეტულ თვეში.

სტატისტიკური დასკვნების თეორია *მოიცავს განზოგადოებას შერჩევიდან პოპულაციამდე, ჰიპოთეზათა შემოწმების კრიტერიუმების დადგენას, სიდიდეებს შორის კავშირის დადგენას და პროგნოზის გაკეთებას.*

საბოლოოდ, ჩვენ მივდივართ შემდეგ განმარტებამდე.

სტატისტიკა არის მეცნიერება, რომელიც ერთის მხრივ შეისწავლის მონაცემების შეგროვების, ორგანიზების, კლასიფიკაციის, სისტემატიზაციისა და პირველადი დამუშავების მეთოდებს, ხოლო მეორეს მხრივ, შეისწავლის დამუშავებული მონაცემების ანალიზისა და გამოყენების მეთოდებს თეორიული და პრაქტიკული დასკვნების მიღების მიზნით. შესაბამისად, მათემატიკური სტატისტიკა იყოფა ორ ნაწილად: აღწერითი (ე. წ. დესკრიფციული) სტატისტიკა და სტატისტიკური დასკვნების თეორია, თუ არ ჩავთვლით ალბათობის თეორიას, რომელიც წარმოადგენს ხიდს სტატისტიკის ამ ორ ნაწილს შორის. უფრო ზუსტად ალბათობის თეორია წარმოადგენს სტატისტიკური დასკვნების თეორიის მათემატიკურ საფუძველს.

თავი II
სიდიდე, სიხშირე, სიხშირეების განაწილება

სიდიდეები და მონაცემთა ტიპები. როგორც უკვე ავღნიშნეთ, სტატისტიკოსი კონკრეტულ სიტუაციაში ინფორმაციას ღებულობს შემთხვევითი სიდიდის მონაცემების შეგროვებით. ქვემოთ ჩვენ უფრო დაწვრილებით შევისწავლით სიდიდეების ბუნებასა და მონაცემთა სხვადასხვა ტიპებს.

სიდიდეები შეიძლება დაიყოს *თვისებრივი და რაოდენობრივი* ბუნების სიდიდეებად. *თვისებრივია ის სიდიდეები,* რომლებიც გამოსახავენ რაიმე თვისებას, ან მდგომარეობას გარკვეული კატეგორიებით. მაგალითად, თუ ობიექტები კლასიფიცირდება სქესის მიხედვით (მდედრობითი ან მამრობითი), მაშინ სიდიდე "სქესი" არის თვისებრივი სიდიდე. სხვა მაგალითი თვისებრივი სიდიდის შეიძლება იყოს რელიგიური კუთვნილება ან გეოგრაფიული ადგილმდებარეობა.

რაოდენობრივი სიდიდე არის რიცხობრივი თავისი ბუნებით და შესაძლებელია დალაგდეს ან აღინიშნოს რიცხვთა მწკრივის სახით. მაგალითად, სიდიდე "ასაკი" არის რიცხვითი, და ადამიანები შეიძლება აღინიშნოს (დალაგდეს) მათი ასაკის შესაბამისი რიცხვითი მწკრივის სახით. სხვა მაგალითი რაოდენობრივი სიდიდის არის სიმაღლე, წონა ან სხეულის ტემპერატურა.

რაოდენობრივი სიდიდეები შეიძლება დაიყოს ორ კატეგორიად: *დისკრეტული და უწყვეტი სიდიდეები. დისკრეტული სიდიდე* მოიცემა მნიშვნელობებით 0, 1, 2, 3... და ის არის თვლადი (მნიშვნელობების თვლადი რაოდენობა აქვს). დისკრეტული სიდიდის მაგალითებია: ბავშვების რაოდენობა ოჯახში, სტუდენტების რაოდენობა აუდიტორიაში, კომუტატორის ოპერატორის მიერ დღის განმავლობაში მიღებული სატელეფონო ზარების რაოდენობა.

დისკრეტული სიდიდე ღებულობს მნიშვნელობებს, რომელთა რაოდენობა თვლადია.

უწყვეტ სიდიდეს, განსხვავებით დისკრეტულისაგან, შეუძლია მიიღოს ორ კონკრეტულ მნიშვნელობას შორის არსებული ყველა მნიშვნელობა. მაგალითად, ტემპერატურა უწყვეტი სიდიდეა, რადგან მას შეუძლია მიიღოს ნებისმიერი მნიშვნელობა ტემპერატურის ორ მოცემულ მნიშვნელობას შორის.

უწყვეტი სიდიდე ღებულობს ყველა მნიშვნელობას ნებისმიერ ორ კონკრეტულ მნიშვნელობას შორის (მისი მნიშვნელობების რაოდენობა კონტინიუმის სიმძლავრისაა).

დავუბრუნდეთ კითხვას, რომელიც დასმული იყო ზემოთ: როგორ იკვებებიან აშშ-ის მოქალაქეები? მკვლევარებმა, შემთხვევით შერჩეული ტელეფონის ნომრების საშუალებით, ამოარჩიეს აშშ-ის 23699 მოქალაქე და თითოეულ რესპონდენტს დაუსვეს 6 შეკითხვა:

1. რამდენად ხშირად სვამთ თქვენ ისეთ წვენებს როგორიცაა ფორთოხლის, გრეიფრუტის ან ტომატის წვენი?
2. თუ არ ჩავთვლით წვენებს, რამდენად ხშირად ჭამთ თქვენ ხილს?
3. რამდენად ხშირად ჭამთ თქვენ მწვანე სალათს?
4. რამდენად ხშირად ჭამთ თქვენ კარტოფილს (თუ არ ჩავთვლით კარტოფილის "ჩიფსს")?

5. რამდენად ხშირად ჭამთ თქვენ სტაფილოს?
6. თუ არ ჩაუთვლით სტაფილოს, კარტოფილს, ან სალათს რამდენ პორციას ბოსტნეულს მიირთმევთ თქვენ ჩვეულებრივ?

მკვლევარებმა აღმოაჩინეს, რომ ხილისა და ბოსტნეულის პორციებს მამაკაცები ყოველდღიურად მიირთმევენ ოდნავ ნაკლებს (3.3 პორცია), ვიდრე ქალები (3.7 პორცია). მოსახლეობის მხოლოდ 20% მიირთმევს ხილისა და ბოსტნეულის რეკომენდირებულ 5 ან მეტ პორციას. გარდა ამისა, მათ აღმოაჩინეს, რომ ახალგაზრდა და ნაკლებად განათლებული ადამიანები მიირთმევენ საშუალო მოხმარების დონეზე ნაკლებ ხილისა და ბოსტნეულის პორციებს.

ამ კვლევებზე დაყრდნობით მკვლევარები რეკომენდაციას იძლევიან დიდი საგანმანათლებლო ძალისხმევა მიმართული იქნეს იქითკენ, რომ გაიზარდოს მოსახლეობის კვების რაციონში ხილისა და ბოსტნეულის მოხმარება და გაუმჯობესდეს გარემოს დაცვითი ღონისძიებები, რათა უზრუნველყოფილ იქნეს ხილისა და ბოსტნეულის გაზრდილი მოხმარება.

ბევრ ხელმძღვანელს სჭირდება სტატისტიკის გამოყენება, რომ რათა მოიპოვოს საკუთარი მოთხოვნების (ინტერესების) მხარდაჭერა. მაგალითად, დევიდ ჰენსონი (რომელიც იყო აშშ-ს ფედერალური ავიაციის ადმინისტრაციის უფროსი), თავის სტატიაში სათაურით: "აეროპორტის გაფართოება: გააკეთო მეტი მცირეთი," ამბობს რომ საავიაციო მგზავრობის გაზრდა იწვევს მგზავრების დიდი ნაწილის თავმოყრას აშშ-ს მთავარ აეროპორტებში. მან შესთავაზა რომ აეროპორტებმა უნდა გაზარდონ ტევადობა მიუხედავად სამთავრობო ხარჯების შემცირებისა. სტატიაში მან ჩაამატა შემდეგი ცხრილი, რომელიც უჩვენებს საჰაერო გადაზიდვების ზრდას.

საჰაერო გდაზიდვები

აქ მოყვანილია მონაცემები თუ რამდენმა მილიონმა მგზავრმა ისარგებლა აშშ-ს თვითმფრინავებით აშშ-ში და მის გარეთ	1980	287.9
	1981	274.7
	1982	286.1
	1983	308.2
	1984	334.0
	1985	370.1
	1986	404.7
	1987	441.2
	1988	441.2
	1989	443.6
	1990	456.6
	1991	445.7
	1992	463.0
	1993	468.1
	1994	509.0

რადგანაც მოაცემები წარმოდგენილია ცხრილის სახით, მათ არ აქვთ იმდენი ზეგავლენა მკითხველზე რამდენიც ექნებოდა სტატისტიკური გრაფიკის გამოყენების შემთხვევაში.

ქვემოთ ჩვენ შევისწავლით როგორ მოვახდინოთ მონაცემების ორგანიზება და შემდეგ როგორ ავაგოთ შესაბამისი გრაფიკები ისე, რომ მონაცემები წარმოვადგინოთ შეკუმშულ და ადვილად გასაგებ ფორმაში.

როდესაც ატარებს სტატისტიკურ კვლევას, მკვლევარმა უნდა შეაგროვოს მონაცემები გარკვეული ცელადით კვლევის ფარგლებში. მაგალითად, თუ მკვლევარს სურს შეისწავლოს იმ ადამიანების რაოდენობა, რომლებიც იქნენ დაკბენილი შხამიანი გველების მიერ გარკვეულ გეოგრაფიულ ტერიტორიაზე რამდენიმე წლის განმავლობაში, მას შეუძლია შეაგროვოს მონაცემები ექიმებისაგან, საავადმყოფოებში ან ჯანმრთელობის დეპარტამენტში.

იმისთვის რომ აღწეროს სიტუაცია, გააკეთოს დასკვნები ან ჩაერიოს მოვლენების განვითარებაში, მკვლევარმა უნდა მოახდინოს მონაცემების ორგანიზება გარკვეული თვალსაჩინო, შინაარსიანი გზით. მონაცემების ორგანიზების ყველაზე მოხერხებული გზას წარმოადგენს **განაწილების სიხშირის** აგება.

მონაცემების ორგანიზების შემდეგ, მკვლევარმა ისინი (მონაცემები) ისე უნდა წარმოადგინოს, რომ გასაგები იყოს იმათთვის ვისთვისაც განკუთვნილია მათი წაკითხვა. მონაცემების წარმოდგენის ყველაზე სასარგებლო მეთოდი მდგომარეობს სტატისტიკური დიაგრამებისა და გრაფიკების აგებაში. არსებობს ბევრი განსხვავებული ტიპის დიაგრამა და გრაფიკი, და თითოეულს აქვს თავისი სპეციფიური, განსაკუთრებული მიზანი.

ქვემოთ ჩვენ ავხსნით, თუ როგორ უნდა მოვახდინოთ მონაცემების ორგანიზება **განაწილების სიხშირის** აგებით და როგორ წარმოვადგინოთ მონაცემები დიაგრამებისა და გრაფიკების საშუალებით. ჩვენს მიერ მოყვანილი იქნება დიაგრამებისა და გრაფიკების შემდეგი სახეები: **ჰისტოგრამები, სიხშირეთა პოლიგონი, ოგივა, სვეტოვანი (მართკუთხედებიანი) დიაგრამა, წრიული (სექტორული) დიაგრამა, პარეტოს დიაგრამა და დროითი მწკრივების გრაფიკი.** გარდა ამისა, ჩვენ განვიხილავთ დაწყვილებულ მონაცემებს და ავხსნით როგორი ინტერპრეტაცია უნდა გავუკეთოთ წყვილებს შორის დამოკიდებულებას გაბნევის დიაგრამის საშუალებით.

დავუშვათ, რომ მკვლევარს სურს გამოიკვლიოს მილების რაოდენობა, რასაც გადიან დიდი სავაჭრო ცენტრის მომსახურეები სამსახურამდე ყოველდღე. მკვლევარმა თავიდან უნდა შეაგროვოს მონაცემები, რისთვისაც უნდა გამოკითხოს თანამშრომლები -- დაახლოებით რამდენი მილია მათი სახლიდან სავაჭრო ცენტრამდე. როდესაც მონაცემები შეგროვებულია საწყისი ფორმით, მას ეძახიან *დაუმუშავებელ ან ნედლ მონაცემებს*. ამ შემთხვევაში მონაცემებია:

1	2	6	7	12	13	2	6	9	5
18	7	3	15	15	4	17	1	14	5
4	16	4	5	8	6	5	18	5	2
9	11	12	1	9	2	10	11	4	10
9	18	8	8	4	14	7	3	2	6

ვინაიდან, ასე დაუმუშავებელი მონაცემებიდან მხოლოდ მცირე ინფორმაციის მიღებაა შესაძლებელი, მკვლევარი ახდენს მონაცემების ორგანიზებას განაწილების სიხშირის გამოყენებით. **სიხშირე** არის იმ მნიშვნელობების რაოდენობა, რომლებიც მოხვდნენ განაწილების გარკვეულ კლასში. ზემოთ მოყვანილი მონაცემებისათვის განაწილების სიხშირეს აქვს სახე:

კლასების ზღვრები	სიხშირეთა გამოთვლა	სიხშირე
1—3	///// /////	10
4—6	///// ///// ////	14
7—9	///// /////	10
10—12	///// /	6
13—15	/////	5
16—18	/////	5
		ჯამი 50

ახლა შეგვიძლია გავაკეთოთ ზოგადი დაკვირვება მონაცემებზე შევხედავთ რა მათ განაწილების სიხშირის ფორმით. მაგალითად, აქედან ჩვენ ვხედავთ, რომ თანამშრომლების დიდი უმრავლოესობა ცხოვრობს 9 მილის ფარგლებში სავაჭრო ცენტრიდან.

სიხშირული განაწილება წარმოადგენს ნედლი მონაცემების ორგანიზებას ცხრილის ფორმით, ტიპებისა და სიხშირეების გამოყენებით.

ზემოთ მოყვანილ განაწილებაში კლასებია 1-3, 3-6, და ა.შ. ამ მნიშვნელობებს **კლასების საზღვრები** ეწოდება. მონაცემები 1, 2 და 3 ხვდება (ერთიანდება) პირველ კლასში; მონაცემები 4, 5 და 6 – მეორე კლასში და ა.შ.

სიხშირეთა განაწილების ტიპებიდან ყველაზე ხშირად გამოიყენება ორი სახის სიხშირის განაწილება: *კატეგორიული სიხშირის განაწილება* და *დაჯგუფებული სიხშირის განაწილება*. ქვემოთ ჩვენ მოვიყვანთ მათი აგების პროცედურებს.

კატეგორიულ სიხშირეთა განაწილება

კატეგორიულ სიხშირეთა განაწილება გამოიყენება იმ მონაცემებთან, რომლებიც შეიძლება განაწილდნენ სპეციალურ კატეგორიებში, როგორიცაა ნომინალური ან რიგითი მონაცემების დონე. მაგალითად, ისეთი მონაცემები როგორიცაა პოლიტიკური კუთვნილება, რელიგიური აღმსარებლობა ან კვლევის მთავარი სფეროები, რომლებიც გამოიყენებენ კატეგორიულ სიხშირეთა განაწილებას.

მაგალითი 1. 25 ახალწვეულს გაუკეთეს სისხლის ანალიზი, რათა გაერკვიათ მათი სისხლის ჯგუფი (ტიპი). მონაცემების სიმრავლე ასეთია:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	AB	O
A	O	O	O	AB
AB	A	O	B	A

ავაგოთ სიხშირეთა განაწილება ამ მონაცემებისთვის.

ამოხსნა. რადგან მონაცემები არის კატეგორიული, შესაძლებელია გამოყენებულ იქნეს დისკრეტული კლასები. არსებობს 4 ჯგუფის სისხლი: A, B, O, AB. ეს ჯგუფები შეიძლება გამოყენებულ იქნეს როგორც კლასები განაწილებისათვის.

ქვემოთ ჩვენ მოგვყავს პროცედურები როგორ უნდა ავაგოთ სიხშირეთა განაწილება კატეგორიული მონაცემებისთვის:

ნაბიჯი 1. გავაკეთოთ ცხრილი როგორც ქვემოთაა ნაჩვენები:

A კლასი	B სიხშირეთა გამოთვლა	C სიხშირე	D პროცენტი
A			
B			
O			
AB			

ნაბიჯი 2. დავთვალოთ მონაცემები და შედეგები ჩავწეროთ B სვეტში.

ნაბიჯი 3. დავთვალოთ სიხშირეები და შედეგები ჩავწეროთ C სვეტში.

ნაბიჯი 4. ვიპოვოთ პროცენტი თითოეული კლასისათვის შემდეგი ფორმულის გამოყენებით:

$$\% = \frac{f}{n} \cdot 100\%,$$

სადაც f -- კლასების სიხშირეა, ხოლო n – მნიშვნელობების საერთო რაოდენობა.

მაგალითად, A ჯგუფის სისხლის მქონეთა პროცენტი არის:

$$\% = \frac{5}{25} \cdot 100\% = 20\%.$$

პროცენტი, როგორც წესი, არ არის სიხშირეთა განაწილების ნაწილი, მაგრამ ის შეიძლება იქნეს დამატებული, რამდენადაც იგი გამოყენებული იქნება ისეთი გარკვეული გრაფიკული წარმოდგენის დროს, როგორიცაა წრიული დიაგრამა.

ნაბიჯი 5. ვიპოვოთ C და D სვეტების ჯამები. საბოლოოდ, ცხრილს ექნება ასეთი სახე:

A კლასი	B სიხშირეთა გამოთვლა	C სიხშირე	D პროცენტი
A	/////	5	20
B	///// //	7	28
O	///// ////	9	36
AB	////	4	16
		ჯამი 25	ჯამი 100

მაგალითად, ამ ცხრილიდან ჩვენ ვხედავთ, რომ O ჯგუფის სისხლი აქვს უფრო მეტ ადამიანს, ვიდრე ნებისმიერი სხვა ჯგუფის.

დაჯგუფებულ სიხშირეთა განაწილება

როდესაც მონაცემთა მწკრივი დიდია, მონაცემები უნდა დაჯგუფდეს ისეთ კლასებად, რომელთა მოცულობა (სიგრძე) ერთ ერთეულზე უფრო დიდია. მაგალითად, გემის ბატარეების მუშაობის საათების განაწილება ასეთია:

კლასების ზღვრები	კლასების საზღვრები	სიხშირეთა გამოთვლა	სიხშირე	დაგროვებული სიხშირე
24-30	23.5-30.5	///	3	3
31-37	30.5-37.5	/	1	4
38-44	37.5-44.5	/////	5	9
45-51	44.5-51.5	///// ////	9	18
52-58	51.5-58.5	///// /	6	24
59-65	58.5-65.5	/	1	25
			ჯამი 25	

შემდეგ მაგალითში ჩვენ მოვიყვანთ წინამდებარე სიხშირეთა განაწილების აგების პროცედურას. თუმცა აქვე ავღნიშნავთ რამოდენიმე მომენტს. ამ განაწილებაში, მნიშვნელობებს 24 და 30 პირველი კლასიდან ეწოდება *კლასის ზღვრები*. **კლასის ქვედა ზღვარი** არის 24, ის წარმოადგენს იმ უმცირეს მონაცემს, რომელის მნიშვნელობაც შეიძლება მოთავსდეს (ჩაირთოს) ამ კლასში. **კლასის ზედა ზღვარი** არის 30, ის გვიჩვენებს იმ უდიდეს მონაცემს, რომელიც შეიძლება მოთავსდდეს ამ კლასში. რიცხვებს მეორე სვეტიდან ეწოდება **კლასის საზღვრები**. ეს რიცხვები გამოიყენება რათა დაცალკევდეს კლასები, ასე რომ არ არის ხარვეზი სიხშირეთა მოყვანილ განაწილებაში. ხარვეზები არის ზღვრების სვეტში. მაგალითად, ხარვეზია 30-სა და 31-ს შორის: პირველი სვეტიდან არ ჩანს რომელ კლასს მიეკუთვნება მონაცემი, რომელიც მოთავსებულია 30-სა და 31-ს შორის, მაშინ როცა ეს პრობლემა მეორე სვეტში მოხსნილია.

სტუდენტებს ზოგჯერ ექმნებათ სირთულეები კლასის საზღვრების დადგენაში, როდესაც მოცემულია კლასის ზღვრები. მთავარი წესი მდგომარეობს იმაში, რომ კლასის ზღვრებს უნდა ჰქონდეს იმდენივე დეციმეტრული მნიშვნელობა, რამდენიც მონაცემებს, მაგრამ კლასის საზღვრებს აქვთ ერთი დამატებითი მნიშვნელობის ადგილი და ბოლოში 5. მაგალითად, თუ მონაცემების მნიშვნელობები არის მთელი რიცხვები, როგორიცაა 24, 32, 18, ზღვრები კლასისთვის შეიძლება იყოს 31 და 37, ხოლო საზღვრები არის 30.5 და 37.5. იპოვე საზღვრები 31-დან 0.5-ის გამოკლებით (კლასის ქვედა ზღვარი) და 37-სათვის 0.5-ის მიმატებით (კლასის ზედა ზღვარი):

(ქვედა ზღვარი) – 0.5 = 31 - 0.5 = 30.5 = (ქვედა საზღვარი)

(ზედა ზღვარი) + 0.5 = 37 + 0.5 = 37.5 = (ზედა საზღვარი)

თუ მონაცემები არის გამოსახული მეათედებში, როგორიცაა 6.2, 7.8, 12.6, ჰიპოთეტურად კლასის ზღვრები შეიძლება იყოს 7.8–8.8, და საზღვარი კლასისთვის იქნება 7.75–7.85. იპოვე ეს მნიშვნელობები 7.8-დან 0.05-ის გამოკლებით და 8.8-სათვის 0.05-ის მიმატებით.

შემოვიღოთ კიდევ ერთი ცნება. **კლასის სიგანე** სიხშირეთა განაწილების კლასისთვის მიიღება თუ ერთი კლასის ქვედა (ზედა) ზღვარს გამოვაკლებთ შემდეგი (მომდევნო) კლასის ქვედა (ზედა) ზღვარს. მაგალითად, კლასის სიგანე გემის ბატარეების მუშაობის საათების განაწილების კანონის შემთხვევაში არის 7, რადგან 31 – 24 = 7.

კლასის სიგანე შეიძლება აგრეთვე გამოთვლილ იქნეს თუ ერთი კლასის ზედა საზღვარს გამოვაკლებთ იმავე კლასის ქვედა საზღვარს. ამ შემთხვევაში, ისევ 30.5 - 23.5 = 7.

შენიშვნა: ერთი კლასის ზღვრები არ გამოვაკლოთ ერთმანეთს. ეს მოგვცემს არასწორ პასუხს.

მკვლევარმა ყოველ კონკრეტულ სიტუაციაში თვითონ უნდა გადაწყვიტოს რამდენი უნდა იყოს კლასი და რამხელა უნდა იყოს კლასის სიგანე. იმისათვის რომ ავაგოთ სიხშირეთა განაწილება, უნდა ვიხელმძღვანელოთ შემდეგი წესებით:

1. კლასების რაოდენობა იყოს 5-დან და 20-მდე. თუმცა არ არსებობს რამე რთული და სწრაფი წესი იმის შესახებ, თუ რამდენი კლასი უნდა იყოს სიხშირეთა განაწილებაში, უკიდურესად მნიშვნელოვანია გქონდეს კლასების საკმარისი რაოდენობა, რომ ნათლად აღვწეროთ შეგროვებული მონაცემები.

2. კლასის სიგანე სასურველია იყო ლუწი რიცხვი. ეს საშუალებას იძლევა, რომ თითოეული კლასის შუაწერტილს ჰქონდეს ისეთივე მნიშვნელობა რაც მონაცემს. კლასის **შუაწერტილი** X_m მიიღება თუ ქვედა საზღვრისა და ზედა საზღვარის ჯამს გავყოფთ 2-ზე ან კლასის ქვედა და ზედა ზღვრების ჯამს გავყოფთ 2-ზე:

$$X_m = \frac{\text{ქვედა საზღვარი + ზედა საზღვარი}}{2}$$

ან

$$X_m = \frac{\text{ქვედა ზღვარი + ზედა ზღვარი}}{2}.$$

მაგალითად, შუაწერტილი გემის ბატარეების მუშაობის საათების განაწილების კანონის პირველი კლასისათვის არის:

(24+30)/2=27 ან (23.5+30.5)/2=27.

შუაწერტილი არის კლასის ცენტრის რიცხობრივი მდებარეობა. შუაწერტილი აუცილებელია გრაფიკების ასაგებად. თუ კლასის სიგანე არის კენტი რიცხვი, მაშინ კლასის შუაწერტილი გამოისახება მეათედებში. მაგალითად, თუ კლასის სიგანე არის 6, ხოლო საზღვრები 5.5--11.5, მაშინ შუაწერტილი არის:

(5.5+11.5)/2=17/2=8.5.

მე-2 წესი არის მხოლოდ რჩევა და არ არის მკაცრად დასაცავი, განსაკუთრებით მაშინ როცა კომპიუტერი აგებს გრაფიკს.

3. კლასები უნდა იყოს ურთიერთგამომრიცხავი. ურთიერთგამომრიცხავე კლასებს აქვთ არაგადამკვეთი კლასის ზღვრები, შესაბამისად ამ შემთხვევაში მონაცემები არ შეიძლება მოთავსდეს ორ კლასში ერთდროულად. მაგალითად, ისეთი სიხშირეთა განაწილებები როგორიცაა

<u>ასაკი</u>
10-20
20-30
30-40
40-50

ხშირად არის ლიტერატურაში თუ კვლევებში. თუ ადამიანი არის 40 წლის, მაშინ რომელ კლასში უნდა ჩაიწეროს ეს მონაცემი? უკეთესი გზა ავაგოთ სიხშირეთა განაწილება არის შემდეგი კლასების გამოყენებით:

ასაკი
10-20
21-31
32-42
43-53

4. კლასები უნდა იყოს უწყვეტი. მაშინაც კი როცა კლასში არ არის მონაცემი, ის (კლასი) ჩართული უნდა იყოს სიხშირეთა განაწილებაში. არ უნდა არსებობდეს ცარიელი ადგილი სიხშირეთა განაწილებაში. ერთადერთი გამონაკლისია, მაშინ როცა ნულოვანი სიხშირის მქონე კლასი არის პირველი ან ბოლო. ნულოვანი სიხშირის მქონე კლასი ნებისმიერ ბოლოში შეიძლება ამოღებულ (გამოტოვებულ) იქნეს ყოველგვარი გავლენის გარეშე სიხშირეთა განაწილებაზე.

5. კლასები უნდა იყოს ამომწურავი. უნდა იყო საკმარისი რაოდენობის კლასები რათა "დაბინავდეს" ყველა მონაცემი.

6. კლასები ტოლი უნდა იყოს სიგანეში. ეს მოგვაშორებს დამახინჯებულ ხედვას მონაცემებზე.

ერთი გამონაკლისი ხდება როდესაც განაწილება არის **ღია-ბოლოთი.** მას არ აქვს განსაკუთრებული (კონკრეტული) საწყისი ან საბოლოო მნიშვნელობა. მოვიყვანოთ ორი მაგალითი ღია განაწილების:

ასაკი	წუთი
10-20	110-ის ქვევით
21-31	110-114
32-42	115-119
43-53	120-124
54-64	125-129
65 და ზევით	130-134

სიხშირეთა განაწილება ასაკისთვის არის ღია-ბოლოთი ბოლო კლასისათვის, რომელიც ნიშნავს რომ ყველა ვინც 65 წლისაა ან მეტი უფრო მეტის დათვლილი იქნება ბოლო კლასში. განაწილება წუთებისთვის ღია-ბოლოთია პირველი კლასისათვის, რომლიც ნიშნავს რომ წუთების ყველა მნიშვნელობა, რომელიც 110-ის ქვევითაა დათვლილი იქნება პირველ კლასში.

ქვემოთ მოყვანილი მაგალითი გვიჩვენებს დაჯგუფებული სიხშირეთა განაწილების აგების პროცედურებს, ე. ი. როდესაც კლასები შეიცავენ ერთზე მეტს მონაცემებს.

მაგალითი 2. ქვემოთ მოყვანილი მონაცემები წარმოადგენს რეკორდულ ტემპერატურებს აშშ-ს 50 შტატისთვის. ავაგოთ ამ მონაცემებისათვის დაჯგუფებული სიხშირეთა განაწილება 7 კლასის გამოყენებით:

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

ამოხსნა. წარმოდგენილი რიცხობრივი მონაცემებისთვის დაჯგუფებული სიხშირეთა განაწილების აგების პროცედურები შემდეგია:

ნაბიჯი 1. *განვსაზღვროთ კლასები.*

ვიპოვოთ *უდიდესი და უმცირესი მნიშვნელობები.* უდიდესი $H=134$, ხოლო უმცირესი $L=100$.

ვიპოვოთ *გაბნევის დიაპაზონი:* $R=$ უდიდესი მნიშვნელობა – უმცირესი მნიშვნელობა $=H-L$,

$$R=134-100=34.$$

ავირჩიოთ სასურველი კლასების რაოდენობა (ჩვეულებრივ, 5-დან 20-მდე). ამ შემთხვევაში 7 ნებისმიერადაა შერჩეული.

ვიპოვოთ *კლასის სიგანე,* გაბნევის დიაპაზონის კლასების რაოდენობაზე გაყოფით:

$$\text{სიგანე}=\frac{\text{გაბნევის დიაპაზონი}}{\text{კლასების რაოდენობა}}=\frac{R}{7}=\frac{34}{7}=4.9.$$

შედეგი, თუ ის არის ნაშთიანი, დავამრგვალოთ უახლოეს მთელ რიცხვამდე: $4.9\approx5$ (განასხვავებენ ზევით (მეტობით) და ქვევით (ნაკლებობით) დამრგვალებას. მაგალითად, $85\div6=14.167$ და ზევით დამრგვალება იქნება 15, ანალოგიურად, $53\div4=13.25\approx14$).

ავირჩიოთ საწყისი წერტილი ყველაზე დაბალი კლასის ზღვარისათვის. ეს შეიძლება იყოს მონაცემებიდან ყველაზე უმცირესი მნიშვნელობა ან უმცირეს მონაცემზე ნაკლები ნებისმიერი მოხერხებული რიცხვი. ამ შემთხვევაში ეს იქნება 100. იმისათვის, რომ მივიღოთ მომდევნო კლასის ქვედა ზღვარი საწყის წერტილად აღებულ უმცირეს მონაცემს დავუმატოთ სიგანე. გავაგრძელოთ ეს პროცედურა მანამ სანამ არ ამოვწურავთ შვიდივე კლასს. გვექნება წერტილები 100, 105, 110 და ა.შ. 135.

იმისთვის რომ მივიღოთ პირველი კლასის ზედა ზღვარი მეორე კლასის ქვედა ზღვარს გამოვაკლოთ ერთი ერთეული: 105 – 1 = 104. შემდეგ, პირველი კლასის ზედა ზღვარს მივუმატოთ სიგანე და მივიღებთ მეორე კლასის ზედა ზღვარს: 104 + 5 = 109, მიღებულ ზედა ზღვარს დავუმატოთ ისევ სიგანე და მივიღებთ მესამე კლასის ზედა ზღვარს: 109 + 5 = 114 და ა. შ.

პირველი კლასი იქნება 100–104, მეორე კლასი – 105–109 და ა.შ.

ვიპოვოთ *კლასის საზღვრები:* ყოველი კლასის ქვედა ზღვარს გამოვაკლოთ 0,5 და ყოველი კლასის ზედა ზღვარს მივუმატოთ 0,5. მივიღებთ:

99.5–104.5, 104.5–109.5, და ა.შ.

ნაბიჯი 2. *გამოვსახოთ მონაცემები ჩხირების საშუალებით.*

ნაბიჯი 3. *ვიპოვოთ მონაცემების რიცხობრივი სიხშირე.*

ნაბიჯი 4. *ვიპოვოთ დაგროვილი (ჯამური) სიხშირე.*

დაგროვილი სიხშირის სვეტი შეიძლება დაემატოს განაწილებას ისე რომ მოცემული რიგამდე დაჯამდეს ყველა სიხშირის მონაცემი., შემდეგნაირად: 0 + 2 = 2, 2 + 8 = 10, 10 + 18 = 28, 28 + 13 = 41 და ა.შ.

საბოლოოდ სიხშირეთა განაწილება შემდეგნაირია:

კლასის ზღვრები	კლასის საზღვრები	სიხშირის გამოსახვა	სიხშირე	დაგროვილი სიხშირე
100-104	99.5-104.5	//	2	2
105-109	104.5-109.5	///// //	8	10
110-114	109.5-114.5	///// ///// ///// ///	18	28
115-119	114.5-119.5	///// ///// ///	13	41
120-124	119.5-124.5	///// //	7	48
125-129	124.5-129.5	/	1	49
130-134	129.5-134.5	/	1	50

სიხშირეთა განაწილება გვიჩვენებს რომ კლასი 109.5-114.5 შეიცავს ტემპერატურების უდიდეს რაოდენობას (18), ხოლო შემდეგი კლასი 114.5-119.5 შეიცავს 13 ტემპერატურას. საბოლოო ჯამში, ტემპერატურების უდიდესი რაოდენობა (31) თავმოყრილია 109.5-სა და 119.5-ს შორის.

დაგროვილი სიხშირე გვიჩვენებს მონაცემების რამდენი მნიშვნელობაა კონკრეტულ კლასამდე ამ კლასში შემავალი მნიშვნელობების ჩათვლით. ამ მაგალითში, 28 რეკორდულად მაღალი ტემპერატურა არის ნაკლები ან ტოლი ვიდრე 114. 48 რეკორდულად მაღალი ტემპერატურა არის ნაკლები ან ტოლი ვიდრე 124.

როდესაც გაბნევის დიაპაზონი შედარებით მცირეა, სიხშირეთა განაწილება შეიძლება აიგოს ისე, რომ თითეულ კლასში იყოს მონაცემების მხოლოდ ერთი მნიშვნელობა. ამ ტიპის სიხშირეთა განაწილებას უწოდებენ *არადაჯგუფებულ სიხშირეთა განაწილებას* და ის ნაჩვენები იქნება შემდეგ მაგალითში:

მაგალითი 3. ქვემოთ მოყვანილი მონაცემები წარმოადგენს მილების რაოდენობას, რასაც გადის ერთ გალონი ბენზინით 30 სხვადასხვა ტიპის მანქანა ქალაქში მოძრაობისას. ავაგოთ სიხშირეთა განაწილება.

12	17	12	14	16	18
16	18	12	16	17	15
15	16	12	15	16	16
12	14	15	12	15	15
19	13	16	18	16	14

ამოხსნა.

ნაბიჯი 1. *განვსაზღვროთ კლასები.* ვინაიდან გაბნევის დიაპაზონი არის მცირე (19-12=7), ამიტომ გამოიყენება კლასები, რომლებიც მონაცემების მხოლოდ ერთ მნიშვნელობას შეიცავს. ეს კლასებია 12, 13, 14, 15, 16, 17, 18, 19.

შენიშვნა: თუ მონაცემები არის უწყვეტი ტიპის, მაშინ შეიძლება გამოყენებულ იქნეს სასაზღვრო კლასები. კლასის ქვედა საზღვრის დასადგენ-

ად ყოველი კლასის მნიშვნელობას უნდა გამოვაკლოთ 0.5, ხოლო კლასის ზედა საზღვრის დასადგენად კლასის მნიშვნელობას უნდა მივუმატოთ 0.5.

ნაბიჯი 2. *გამოვსახოთ მონაცემები ჩხირების საშუალებით.*

ნაბიჯი 3. *ვიპოვოთ მონაცემების რიცხობრივი სიხშირე.*

ნაბიჯი 4. *ვიპოვოთ დაგროვილი (ჯამური) სიხშირე.*

დასრულებული დაუჯგუფებელი სიხშირეთა განაწილება შემდეგნაირია:

კლასის ზღვრები	კლასის საზღვრები	სიხშირის გამოსახვა	სიხშირე	დაგროვილი სიხშირე
12	11.5-12.5	///// /	6	6
13	12.5-13.5	/	1	7
14	13.5-14.5	///	3	10
15	14.5-15.5	///// /	6	16
16	15.5-16.5	///// ///	8	24
17	16.5-17.5	//	2	26
18	17.5-18.5	///	3	29
19	18.5-19.5	/	1	30

ამ შემთხვევაში თითქმის ნახევარი (14) მანქანებისა ერთი გალონი ბენზინით გადის 15 ან 16 მილს.

დაჯგუფებული სიხშირეთა განაწილების ასაგებად გადასადგმელი ნაბიჯების ერთობლიობას თავი მოვუყაროთ პროცედურების შემდეგ ცხრილში:

დაჯგუფებული სიხშირეთა განაწილების აგება		
ნაბიჯი 1.	კლასების განსაზღვრა.	
	უმცირესი და უდიდესი მნიშვნელობის პოვნა.	
	გაბნევის დიაპაზონის პოვნა.	
	კლასების სასურველი რაოდენობის შერჩევა.	
	კლასის სიგანის პოვნა გაბნევის დიაპაზონის კლასების რიცხვზე გაყოფით და ზევით დამრგვალება.	
	საწყისი წერტილის შერჩევა (ჩვეულებრივ უმცირესი მონაცემი ან უმცირეს მონაცემზე ნაკლები ნებისმიერი მოხერხებული რიცხვი); სიგანის მიმატება კლასის ქვედა ზღვრების მისაღებად.	
	კლასის ზედა ზღვრების პოვნა.	
	საზღვრების პოვნა.	
ნაბიჯი 2.	სიხშირეების გამოსახვა ჩხირებით.	
ნაბიჯი 3.	რიცხობრივი სიხშირეების გამოთვლა.	
ნაბიჯი 4.	დაგროვილი სიხშირეების გამოთვლა.	

როდესაც ჩვენ ვაგებთ სიხშირეთა განაწილებას, ჩვენ მივყვებით ზემოთაღწერილ პროცედურას. მაგრამ, მეორეს მხრივ, შესაძლებელია რამდენიმე განსხვავებული მაგრამ კორექტული სიხშირეთა განაწილების აგება იგივე მონაცემებით, თუ ჩვენ ავიღებთ განსხვავებულ კლასის სიგანეს, კლასების განსხვავებულ რაოდენობას, ან განსხვავებულ საწყის წერტ-

ილს. შესაბამისად, აქ წარმოდგენილი სიხშირეთა განაწილების აგების მეთოდი არ არის ერთადერთი და არსებობს მისი აგების სხვა მეთოდებიც. კომპიუტერულ პაკეტებში სპეციალურად არსებობს მცირედი ცვლილებების შესაძლებლობა. მაგრამ მიუხედავად ამისა, ნებისმიერი მეთოდში, რომელიც იქნება გამოყენებული, კლასები უნდა იყოს ურთიერთგამომრიცხავი, უწყვეტი, ამომწურავი და თანაბარი სიგანის.

ჩვენ ზემოთ მოვიყვანეთ სხვადასხვა ტიპის სიხშირეთა განაწილებები. პირველი ტიპი, მოყვანილი მაგალით 1-ში, გამოიყენება იმ შემთხვევაში, როცა მონაცემები არის კატეგორიული ტიპის, როგორიცაა სისხლის ჯგუფი ან პოლიტიკური კუთვნილება. ამ ტიპის განაწილებებს უწოდებენ კატეგორიულ სიხშირეთა განაწილებებს. მეორე ტიპი განაწილებების გამოიყენება, როდესაც გაბნევის დიიაპაზონი დიდია ან არსებობს რამოდენიმე ერთეულიანი სიგანის კლასების საჭიროება. ამ ტიპის განაწილებებს უწოდებენ დაჯგუფებული სიხშირეების განაწილებებს და ის მოყვანილია მაგალით 2-ში. სხვა ტიპის განაწილებები გამოიყენება, როცა მონაცემები რიცხობრივია და როცა გაბნევის დიაპაზონი მცირე, ისევე როგორც ეს ნაჩვენებია მაგალით 3-ში. ვინაიდან ყოველი კლასი მხოლოდ ერთი ერთეულია, ამ განაწილებას ეწოდება დაუჯგუფებელი სიხშირეთა განაწილება.

მიზეზები იმისა, რომ აგებული იქნეს ყველა დანარჩენი ტიპის განაწილება, რომელიც გამოიყენება სტატისტიკაში და სასარგებლოა მონაცემების ორგანიზებისა და წარმოდგენისათვის, მდგომარეობს შემდეგში:

1. მოხდეს მონაცემების ორგანიზება შინაარსიანი და გასაგები გზით.
2. მისცეს საშუალება მკითხველს განსაზღვროს განაწილების ბუნება და ფორმა.
3. გაამარტივოს გამოსათვლელი პროცედურები განაწილების სხვადასხვა მახასიათებლების მოსაძებნად.
4. მისცეს საშუალება მკვლევარს დახაზოს დიაგრამები და გრაფიკები მონაცემების წარმოსადგენად.
5. მისცეს საშუალება მკითხველს, რათა მან მოახდინოს შედარება მონაცემთა სხვა სიმრავლეს შორის.

ფაქტორები რომლებიც გამოიყენება ჰისტოგრამებისა და სიხშირეთა პოლიგონის ანალიზისას არსებითად იგივეა, რაც გამოიყენებოდა სიხშირეთა განაწილების ანალიზისას, და მოყვანილი იქნება შემდეგ ნაწილში.

თავი III
ჰისტოგრამა, სიხშირეთა პოლიგონი, ოგივა

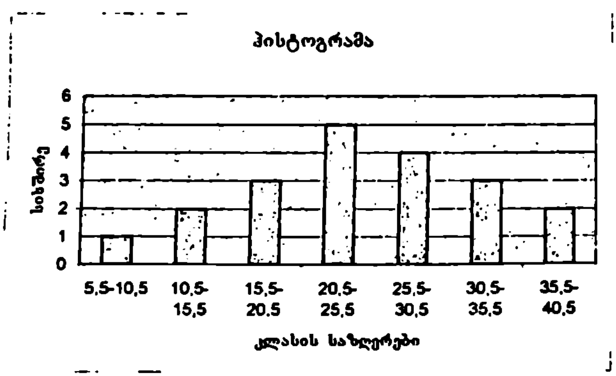
მას შემდეგ რაც მონაცემები ორგანიზებულია სიხშირეთა განაწილების სახით, ისინი შესაძლებელია წარმოდგენილ იქნეს გრაფიკული ფორმით. სტატისტიკაში, გრაფიკებისა და დიაგრამების მიზანია მკითხველსა და მაყურებელს მიაწოდოს მონაცემები თვალსაჩინო ფორმით. უმეტესობა ადამიანებისათვის უფრო ადვილია ჩაწვდეს (გაიაზროს) გრაფიკულად წარმოდგენილი მონაცემების შინაარსს, ვიდრე რიცხობრივი ცხრილების ან სიხშირეთა განაწილებების ფორმით წარმოდგენილ მონაცემებს. ეს განსაკუთრებით არსებითია იმ მომხმარებლისათვის, რომელსაც აქვს მცირე ან საერთოდ არა აქვს სტატისტიკური განათლება.

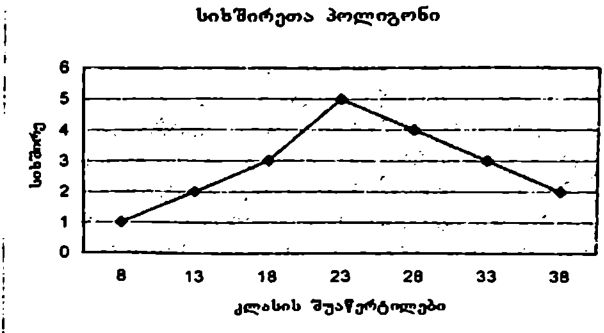
სტატისტიკური გრაფიკები და დიაგრამები შესაძლებელია გამოყენებულ იქნეს როგორც მონაცემთა აღსაწერად, ისე მათი ანალიზისათვის. გრაფიკები და დიაგრამები აგრეთვე სასარგებლოა ადამიანების ყურადღების მისაპყრობად ბეჭდურ ან ზეპირ პრეზენტაციაში. ისინი შესაძლებელია გამოყენებულ იქნეს შედეგის განსახილველად, კრიტიკული აზრის განსამტკიცებლად ან მოხდეს მონაცემთა სიმრავლის განახლება. ისინი აგრეთვე შეიძლება გამოყენებული იქნეს გარკვეული ტენდენციების აღმოსაჩენად ან დროის გარკვეულ პერიოდში სიტუაციის შესაბამისი მოდელის დასადგენად.

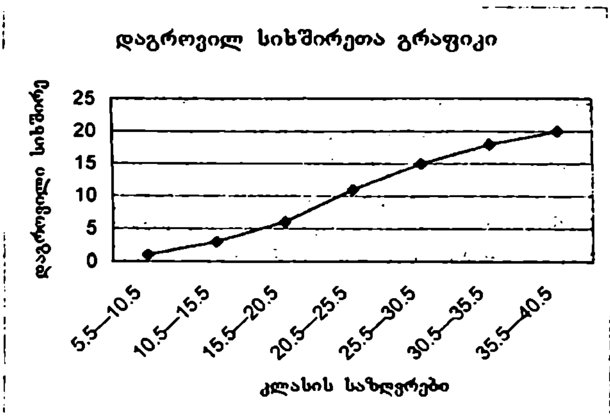
მკვლევარების მიერ სამი ყველაზე ხშირად გამოყენებადი გრაფიკების ფორმებია:

1. *ჰისტოგრამა.*
2. *სიხშირეთა პოლიგონი.*
3. *დაგროვილ სიხშირეთა გრაფიკი ანუ ოგივა.*

ქვემოთ ჩვენ მოგვყავს გრაფიკების სამივე ტიპი, სადაც მონაცემები სამივე გრაფიკში წარმოადგენს კვირის განმავლობაში შემთხვევით შერჩეული 20 მორბენალის მიერ გარბენილი მანძილი გამოსახული მილებში.







ისტორიული შენიშვნა. დიაგრამები პირველად გამოიყენეს შუასაუკუნეების ასტრონომებმა, როცა დაიწყეს ვარსკვლავების მდებარეობის გამოსახვა ცაზე. რომაელი მიწათმზომელები აგრეთვე იყენებდნენ დიაგრამებს, რათა რუკაზე გამოესახათ მიწისზედა ორიენტირები. სტატისტიკური დიაგრამების შემდგომი განვითარება დაკავშირებულია ინჟინერ ვილიამ პლაიფერის სახელთან (1748--1819), რომელიც იყენებდა დიაგრამებს ილუსტრირებული ეკონომიური მონაცემების წარმოსადგენად.

ჰისტოგრამა წარმოადგენს დიაგრამის სახეობას, რომელიც გამოსახავს მონაცემებს სხვადასხვა სიმაღლის მოსაზღვრე ვერტიკალური მართკუთხედების გამოყენებით (გარდა იმ შემთხვევისა, როცა სიხშირე ნულია), რომელთა სიმაღლეები მონაცემების სიხშირეს შეესაბამება.

მაგალითი 1. ავაგოთ ჰისტოგრამა იმ მონაცემების წარმოსადგენად, რომლებიც მოყვანილია ქვემოთ და შეესაბამება ამერიკის 50-ივე შტატის რეკორდულად მაღალ ტემპერატურებს:

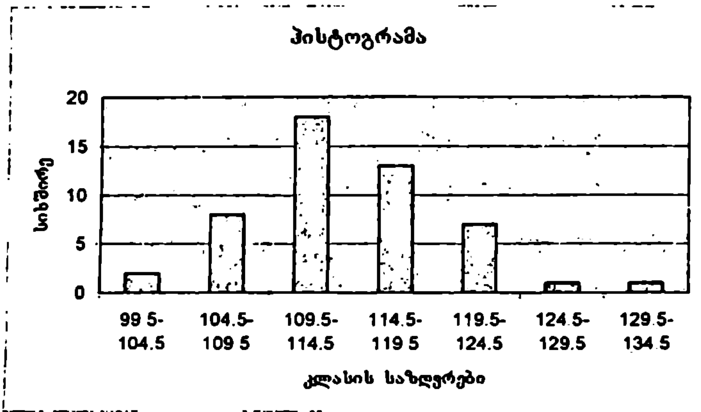
კლასის საზღვრები	სიხშირე
99.5-104.5	2
104.5-109.5	8
109.5-114.5	18
114.5-119.5	13
119.5-124.5	7
124.5-129.5	1
129.5-134.5	1

ამოხსნა.

ნაბიჯი 1. დავხაზოთ საკოორდინატო სიბრტყე და მოვნიშნოთ მასზე მასზე x და y ღერძები. x ღერძი ყოველთვის ჰორიზონტალური ღერძია, ხოლო y ღერძი ყოველთვის ვერტიკალური ღერძია.

ნაბიჯი 2. გადავზომოთ სიხშირეები y ღერძზე, ხოლო კლასის საზღვრები კი – x ღერძზე.

ნაბიჯი 3. გამოვიყენოთ სიხშირეები სიმაღლეებად და ყოველი კლასის ზემოთ დავხაზოთ შესაბამისი მართკუთხედი. მივიღებთ ჰისტოგრამას:



როგორც ეს ჰისტოგრამა გვიჩვენებს, მონაცემთა მნიშვნელობების ყველაზე დიდი რაოდენობის (18) მქონე კლასია 109.5—114.5, შემდეგ მოდის 13 მონაცემით კლასი საზღვრებით 114.5-119.5. ამ დიაგრამას აქვს ერთი პიკი, რომლის ირგვლივაც გროვდება მონაცემები.

იგივე მონაცემთა სიმრავლის წარმოსადგენად შეიძლება გამოყენებულ იქნეს განსხვავებული სახის დიაგრამა (გრაფიკი) – ე. წ. სიხშირეთა პოლიგონი.

სიხშირეთა პოლიგონი წარმოადგენს დიაგრამის (გრაფიკის) სახეობას, რომელიც გამოსახავს მონაცემებს ტეხილის საშუალებით, რომლის წვეროებს წარმოადგენენ კლასების შუაწერტილებისა და სიხშირეების

წყვილი. ამასთანავე, სიხშირეები შეესაბამება შუაწერტილების სიმაღლეებს.

მაგალითი 2. წინა მაგალითში მოყვანილი სიხშირეთა განაწილებისათვის ავაგოთ სიხშირეთა პოლიგონი.

ამოხსნა.

ნაბიჯი 1. ვიპოვოთ თითოეული კლასის შუაწერტილი. გავიხსენოთ, რომ შუაწერტილის საპოვნელად საჭიროა კლასის ქვედა და ზედა საზღვრები შეიკრიბოს და გაიყოს 2-ზე:

$$\frac{99.5+104.5}{2}=102, \quad \frac{104.5+109.5}{2}=107,$$

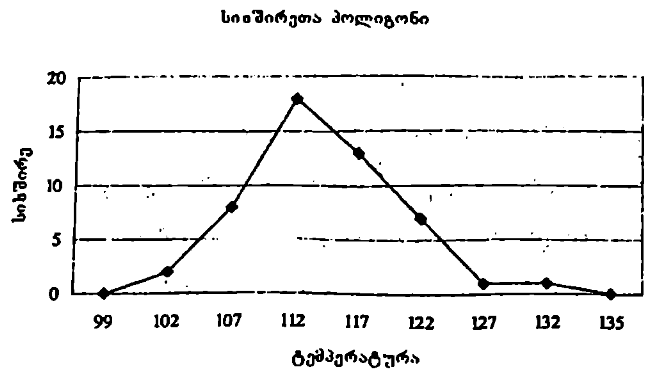
და ა. შ. საბოლოოდ გვექნება შემდეგი ცხრილი:

კლასის საზღვრები	შუაწერტილები	სიხშირე
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1

ნაბიჯი 2. დავხაზოთ x და y ღერძები. x ღერძზე მოვნიშნოთ თითოეული კლასის შუაწერტილი, და შემდეგ შევარჩიოთ მოხერხებული მასშტაბი y ღერძზე სიხშირის გადასაზომად.

ნაბიჯი 3. გამოვიყენოთ შუაწერტილები x-ის მნიშვნელობებად, ხოლო სიხშირეები როგორც y-ის მნიშვნელობები და მოვნიშნოთ შესაბამისი წერტილები საკოორდინატო სიბრტყეზე.

ნაბიჯი 4. შევაერთოთ მეზობელი წერტილები სწორი ხაზებით (მონაკვეთებით). გრაფიკი დავიწყოთ და დავამთავროთ x ღერძზე: პირველი წერტილიდან დავბრუნდეთ უკან x ღერძზე და უკანასკნელი წერტილიდან აგრეთვე მივიდეთ x ღერძამდე იმავე მანძილზე, როგორც ეს ნაჩვენებია ნახაზზე:



სიხშირეთა პოლიგონი და ჰისტოგრამა არის მონაცემთა წარმოადგენის ორი განსხვავებული გზა. ამ ორი გზიდან მკვლევარი თავისი შეხედულებისამებრ ირჩევს რომელი გამოიყენოს. დიაგრამების (გრაფიკების) მესამე ტიპი გამოიყენება კლასების დაგროვილი სიხშირეების წარმოსადგენად. ამ ტიპის დიაგრამას ეწოდება დაგროვილ სიხშირეთა დიაგრამა ან ოგივა. **დაგროვილი სიხშირე** არის ჯამი მოცემული კლასის შესაბამისი სიხშირისა და მანამდე არსებული ყველა სიხშირის.

ოგივა არის დიაგრამის სახეობა რომელიც გამოიყენება სიხშირეთა განაწილების კლასების დაგროვილი სიხშირეების წარმოსადგენად გრაფიკული ფორმით.

მაგალითი 3. მაგალით 1-ში მოყვანილი სიხშირეთა განაწილებისათვის ავაგოთ ოგივა.

ამოხსნა.

ნაბიჯი 1. ვიპოვოთ დაგროვილი სიხშირე თითოეული კლასისათვის:

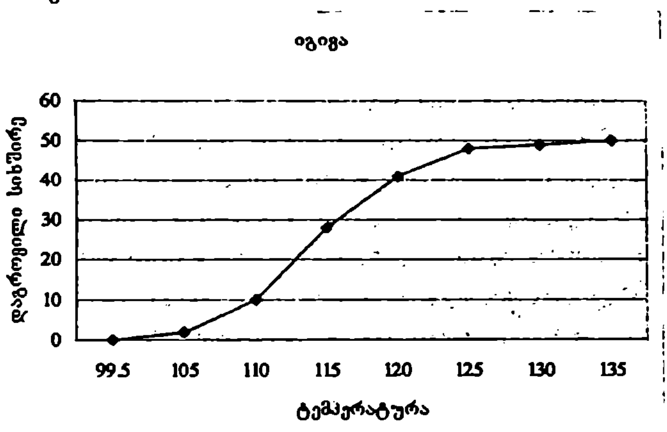
კლასის საზღვრები	დაგროვილი სიხშირე
99.5-104.5	2
104.5-109.5	10
109.5-114.5	28
114.5-119.5	41
119.5-124.5	48
124.5-129.5	49
129.5-134.5	50

ნაბიჯი 2. დავხაზოთ *x* და *y* ღერძები. *x* ღერძზე მოვნიშნოთ თითოეული კლასის საზღვრები, და შემდეგ შევარჩიოთ მოხერხებული მასშტაბი *y* ღერძზე დაგროვილი სიხშირის გადასაზომად. (იმის მიხედვით თუ რა რიცხვები წერია დაგროვილი სიხშირეების სვეტში, შეიძლება გამოყენებული

იქნეს ისეთი გრადუირება (შკალირება) როგორიცაა 0, 1, 2, 3, , ან 0, 5, 10, 15, 20, . . . , ან 0, 1000, 2000, 3000, *y* ღერძზე არ უნდა მოინიშნოს რიცხვები დაგროვილი სიხშირეების სვეტიდან). ამ მაგალითში გამოყენებული იქნება სკალა 0, 5, 10, 15,

ნაბიჯი 3. მოვნიშნოთ დაგროვილი სიხშირეები ყოველი კლასის ზედა საზღვრის თავზე: მოვნიშნოთ საკოორდინატო სიბრტყეზე წერტილები, რომელთა კოორდინატებია კლასების ზედა საზღვრები (აბსცისა) და დაგროვილი სიხშირეები (ორდინატა). ზედა საზღვრები იმიტომ აიღება, რომ დაგროვილი სიხშირეები წარმოადგენენ კლასის ზედა საზღვრამდე მოთავსებული მონაცემთა მნიშვნელობების შესაბამისი სიხშირეების ჯამს.

ნაბიჯი 4. დავიწყოთ პირველი კლასის ზედა საზღვრიდან, 104.5-დან და შევაერთოთ მეზობელი წერტილები სწორი ხაზებით, როგორც ეს ნახაზზეა ნაჩვენები. ბოლოს გრაფიკი გავაგრძელოთ პირველი კლასის ქვედა საზღვრამდე, 99.5-მდე *x* ღერძზე:



დაგროვილ სიხშირეთა დიაგრამა გამოიყენება იმის საილუსტრაციოდ, თუ მონაცემთა რამდენი მნიშვნელობა არ აღემატება გარკვეული კლასის ზედა საზღვარს. მაგალითად, იმის გასაგებად, თუ რამდენი რეკორდულად მაღალი ტემპერატურა არის ნაკლები ვიდრე 114.5, მოვძებნოთ *x* ღერძზე 114.5 და ამ წერტილზე აღვმართოთ ვერტიკალური ხაზი გრაფიკის გადაკვეთამდე. გადაკვეთის წერტილში გავავლოთ ჰორიზონტალური ხაზი *y* ღერძის გადაკვეთამდე. *y* ღერძზე ვღებულობთ მნიშვნელობას 28, როგორც ეს ნახაზზეა ნაჩვენები.

ქვემოთმოყვანილ პროცედურულ ცხრილში მოყვანილია ზემოთაღნიშნული სამი ტიპის დიაგრამის აგების ეტაპები:

სტატისტიკური დიაგრამების აგება
ნაბიჯი 1. საკოორდინატო სიბრტყის დახაზვა და x და y ღერძები მონიშვნა.
ნაბიჯი 2. მოხერხებული მასშტაბის შერჩევა სიხშირეებისათვის ან დაგროვილი სიხშირებისათვის და მათი მონიშვნა ორდინატთა ღერძზე.
ნაბიჯი 3. წარმოვადგინოთ აბსცისთა ღერძზე კლასების საზღვრები ჰისტოგრამისა და ოგივის შემთხვევაში, ან შუაწერტილები სიხშირეთა პოლიგონის შემთხვევაში
ნაბიჯი 4. მოვნიშნოთ წერტილები და დავხაზოთ მართკუთხედები ან ტეხილები

როგორც ზემოთ ვნახეთ, ჰისტოგრამა, სიხშირეთა პოლიგონი, და ოგივა იგება ნედლი (დაუმუშავებელი) მონაცემების სიხშირეების გამოყენებით. ეს განაწილებები შესაძლებელია გადავიყვანოთ განაწილებებში, რომლებიც გამოიყენებენ პროპორციებს (შეფარდებებს) ნაცვლად ნედლი მონაცემების სიხშირეებისა. ამ ტიპის დიაგრამებს უწოდებენ **ფარდობით სიხშირეთა დიაგრამებს.**

დიაგრამები, რომლებიც იყენებენ ფარდობით სიხშირეებს, ნაცვლად სიხშირეებისა, გამოიყენება მაშინ როდესაც პროპორცია მონაცემთა იმ მნიშვნელობების, რომლებიც შედიან მოცემულ კლასში უფრო მნიშვნელოვანია, ვიდრე ფაქტიური რაოდენობა მონაცემთა იმ მნიშვნელობების, რომლებიც შედიან იმავე კლასში. მაგალითად, თუ ჩვენ გვსურს შევადაროთ მოზრდილ ადამიანთა ასაკის განაწილება პენსილვანიის ქალაქ ფილადელფიაში მოზრდილ ადამიანთა ასაკის განაწილებას პენსილვანიის ქალაქ ირიში, მაშინ ჩვენ უნდა გამოვიყენოთ ფარდობით სიხშირეთა განაწილებები. მართლაც, ვინაიდან ფილადელფიის მოსახლეობა შეადგენს 1 478 002 ადამიანს, ხოლო ირის მოსახლეობაა მხოლოდ 105. 270 ადამიანია, ამიტომ თუ ჩვენ ვისარგებლებთ სიხშირეთა განაწილებით, მაშინ ფილადელფიის მონაცემთა მნიშვნელობების შესაბამისი მართკუთხედები გაცილებით დაბალი იქნება, ვიდრე შესაბამისი კლასის მართკუთხედები ირის შემთხვევაში.

იმისათვის რომ სიხშირეები გადავიყვანოთ პროპორციებში ან ფარდობით სიხშირეებში, თითოეული კლასის სიხშირე უნდა გავყოთ ჯამურ სიხშირეზე. ფარდობით სიხშირეთა ჯამი ყოველთვის ერთის ტოლი იქნება. ეს დიაგრამები იქნება ანალოგიური (მსგავსი) სიხშირეების შესაბამისი დიაგრამების, იმ განსხვავებით, რომ ორდინატთა ღერძზე აქ გადაიზომება ფარობითი სიხშირეები ნაცვლად სიხშირეებისა. მოვიყვანოთ ამ დიაგრამების შესაბამისი მაგალითი.

მაგალითი 4. ავაგოთ ჰისტოგრამა, სიხშირეთა პოლიგონი და ოგივა ფარდობითი სიხშირეების გამოყენებით 20 შემთხვევით შერჩეული მორბენალის მიერ კვირის განმავლობაში გარბენილი მანძილების განაწილების მიხედვით:

კლასის საზღვრები	სიხშირე	დაგროვილი სიხშირე
5.5–10.5	1	1
10.5–15.5	2	3
15.5–20.5	3	6
20.5–25.5	5	11
25.5–30.5	4	15
30.5–35.5	3	18
35.5–40.5	2	20
	ჯამი 20	

ამოხსნა.

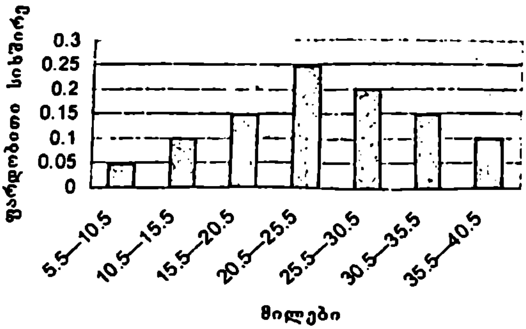
ნაბიჯი 1. გადავიყვანოთ სიხშირეები პროპორციებში ანუ ფარდობით სიხშირეებში თითოეული კლასის სიხშირის გაყოფით დაკვირვებათა საერთო რაოდენობაზე.
5.5—10.5 კლასისათვის ფარდობითი სიხშირე = 1/20=0.05,
10.5—15.5 კლასისათვის ფარდობითი სიხშირე = 2/20=0.10,
15.5—20.5 კლასისათვის ფარდობითი სიხშირე = 3/20=0.15,
და ა. შ.

ნაბიჯი 2. იგივე პროცედურის გამოყენებით ვიპოვოთ ფარდობითი სიხშირეები დაგროვილი სიხშირეების სვეტისათვის. საბოლოოდ, მივიღებთ ფარდობითი სიხშირეების შემდეგ ცხრილს:

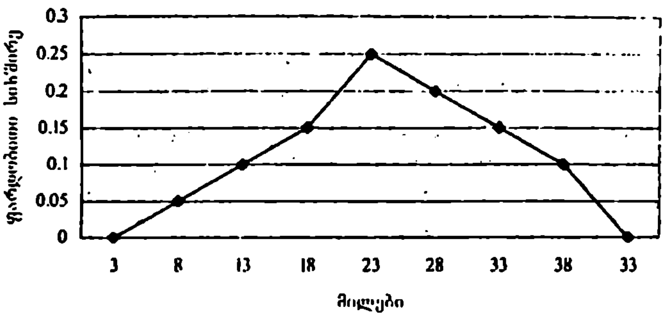
კლასის საზღვრები	შუაწერტილები	ფარდობითი სიხშირე	დაგროვილი ფარდობითი სიხშირე
5.5—10.5	8	0.05	0.05
10.5—15.5	13	0.10	0.15
15.5—20.5	18	0.15	0.30
20.5—25.5	23	0.25	0.55
25.5—30.5	28	0.20	0.75
30.5—35.5	33	0.15	0.90
35.5—40.5	38	0.10	1.00
		ჯამი 1.00	

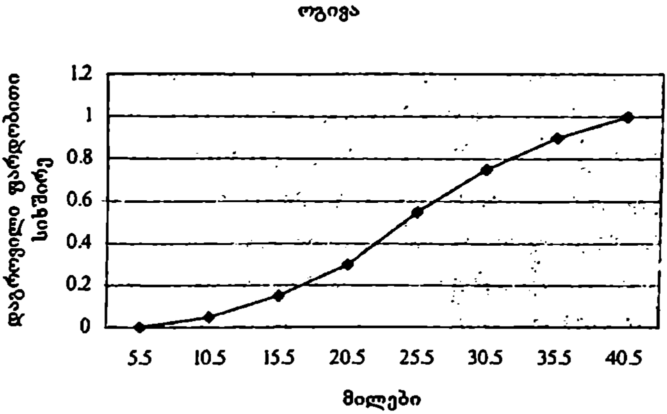
ნაბიჯი 3. დავხაზოთ სამივე ტიპის დიაგრამა. ჰისტოგრამისა და ოგივის შემთხვევებში აბსცისთა ღერძზე გადავზომოთ კლასის საზღვრები, ხოლო სიხშირეთა პოლიგონის შემთხვევაში აბსცისთა ღერძზე გადავზომოთ კლასის შუაწერტილები. ორდინატთა ღერძზე კი გადავზომოთ ფარდობითი სიხშირეები.

ჰისტოგრამა



ფარდობით სიხშირეთა პოლიგონი





ჰისტოგრამებისა და სიხშირეთა პოლიგონის ანალიზისას ყურადღება უნდა მივაქციოთ წირების (გრაფიკების) ფორმას. მაგალითად, აქვს თუ არა მათ ერთი პიკი ან ორი პიკი? არის თუ არა ის შედარებით ბრტყელი? არის თუ არა U-ს ფორმის? არის თუ არა მონაცემთა მნიშვნელობები განფენილი გრაფიკზე, თუ ისინი არიან გარკვეული ცენტრის ირგვლივ თავმოყრილი? აქვს თუ არა ექსტრემუმი ბოლოებში? არის თუ არა ჰისტოგრამა გაწყვეტილი (აკლია მართკუთხედი) ან სიხშირეთა პოლიგონი სადმე ეხება აბსცისთა ღერძს ბოლოების გარდა? არიან თუ არა მონაცემები ერთ ან მეორე ბოლოში თავმოყრილი (გვაქვს თუ არა მარჯვნივ ან მარცხნივ ასიმეტრიული განაწილება)?

მაგალითად, აშშ-ის რეკორდულად მაღალი ტემპერატურების ჰისტოგრამა გვიჩვენებს, რომ საქმე გვაქვს ერთ პიკიან განაწილებასთან, სასაზღვრო კლასით 109.5—114.5, რომელიც შეიცავს ტემპერატურების ყველაზე დიდ რაოდენობას. ამ განაწილებას არა აქვს ცარიელი სასაზღვრო კლასი (არ აკლია მართკუთხედი), და უმაღლეს სასაზღვრო კლასში არის ნაკლები ტემპერატურების რაოდენობა, ვიდრე უმდაბლეს სასაზღვრო კლასში.

გარდა ჰისტოგრამისა, სიხშირეთა პოლიგონისა და ოგივისა, სტატისტიკაში გამოიყენება რამოდენიმე სხვა ტიპის დიაგრამა. ესენია: პარეტოს დიაგრამა, დროითი მწკრივების დიაგრამა და წრიული დიაგრამა.

ზემოთმოყვანილ მაგალითებში მოყვანილი დიაგრამები: ჰისტოგრამა, სიხშირეთა პოლიგონი და ოგივა გვიჩვენებდა თუ როგორ შეიძლება წარმოვადგინოთ მონაცემები, როცა აბსცისთა ღერძზე გამოსახული სიდიდეები რიცხვითია, როგორც სიმაღლე ან წონა. მეორეს მხრივ, როცა აბსცისთა ღერძზე გამოსახული სიდიდეები თვისებრივი ან კატეგორიული ბუნებისაა, გამოიყენება პარეტოს დიაგრამა.

პარეტოს დიაგრამა გამოიყენება კატეგორიული ტიპის მონაცემების სიხშირეთა განაწილების წარმოსადგენად, და ამ შემთხვევაში სიხშირეები

გამოისახება ვერტიკალური მართკუთხედების სიმაღლეებით, რომლებიც დალაგებულია კლების მიხედვით უმაღლესიდან უმცირესისაკენ.

ისტორიული შენიშვნა. ვილფრედ პარეტო (1848--1923) იყო იტალიელი მეცნიერი, რომელმაც განავითარა ეკონომიკის თეორია, სტატისტიკა და სოციალური მეცნიერებები. მისი წვლილი სტატისტიკაში მდგომარეობს იმ მათემატიკური ფუნქციის შემოღებაში, რომელიც გამოიყენება ეკონომიკაში. ამ ფუნქციას გააჩნია მრავალი სტატისტიკური გამოყენება და მას უწოდებენ **პარეტოს განაწილებას**. გარდა ამისა, პარეტომ გამოიკვლია შემოსავლების განაწილება და მისი აღმოჩენა ცნობილია **პარეტოს კანონის** სახელწოდებით.

მაგალითი 5. ქვემოთმოყვანილ ცხრილში მითითებულია აშშ-ის სამართალდამცავ ორგანოთა ოფიცრების მიერ გამოძიებულ სამართალდარღვევათა რაოდენობა, რომლებიც მოხდა 1995 წლის განმავლობაში ეროვნული პარკის ტერიტორიაზე. ავაგოთ ამ მონაცემებისათვის პარეტოს დიაგრამა.

დანაშაულის სახეობა	**რაოდენობა**
მკვლელობა	13
გაუპატიურება	34
ძარცვა	29
თავდასხმა	164

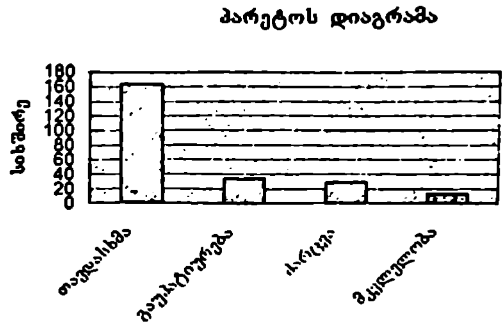
ამოხსნა.

ნაბიჯი 1. დავალაგოთ მონაცემები სიხშირეების მიხედვით უმაღლესიდან უმცირესისაკენ.

დანაშაულის სახეობა	**რაოდენობა**
თავდასხმა	164
გაუპატიურება	34
ძარცვა	29
მკვლელობა	13

ნაბიჯი 2. დავხაზოთ საკოორდინატო სიბრტყე და მოვნიშნოთ x და y ღერძები.

ნაბიჯი 3. დავხაზოთ სიხშირეების შესაბამისი მართკუთხედები როგორც ქვემოთაა ნაჩვენები. როგორც პარეტოს დიაგრამა გვიჩვენებს თავდასხმებს აქვთ უდიდესი სიხშირე, ხოლო მკვლელობებს – უმცირესი.



რჩევები პარეტოს დიაგრამის აგებისას:

1. მართკუთხედები გავაკეთოთ ტოლი სიგანის.
2. სიხშირეების მიხედვით მონაცემები დავალაგოთ უდიდესიდან უმცირესისაკენ.
3. ერთეულები, რომლებიც გამოიყენება სიხშირეებისათვის, ავიღოთ ტოლი ზომის.

როცა ჩვენ ვაანალიზებთ პარეტოს დიაგრამას, შედარებას ვაკეთებთ მართკუთხედების სიმაღლეებზე დაკვირვების მიხედვით.

იმ შემთხვევაში, როცა მონაცემები შეგროვებულია დროის გარკვეულ პერიოდში, ისინი წარმოიდგინება დროითი მწკრივების დიაგრამის სახით.

დროითი მწკრივების დიაგრამა გამოიყენება დროის გარკვეულ პერიოდში მიღებული მონაცემების წარმოსადგენად.

ისტორიული შენიშვნა. დროითი მწკრივების დიაგრამის ისტორია დაახლოებით 1000 წელს ითვლის. პირველად ეს დიაგრამა გამოყენებული იყო პლანეტებისა და მზის მოძრაობის სქემის შესადგენად.

მაგალით 6. პორტის გადაზიდვების მენეჯერის სურვილია გამოიყენოს ქვემოთმოყვანილი მონაცემები იმის საჩვენებლად თუ როგორ იცვლებოდა გადაზიდვები წლების მიხედვით. ავაგოთ დროითი მწკრივების დიაგრამა და გავაკეთოთ დასკვნები.

წლები	**მგზავრები (მილიონებში)**
1990	88.0
1991	85.0
1992	75.7
1993	76.6
1994	75.4

ამოხსნა.

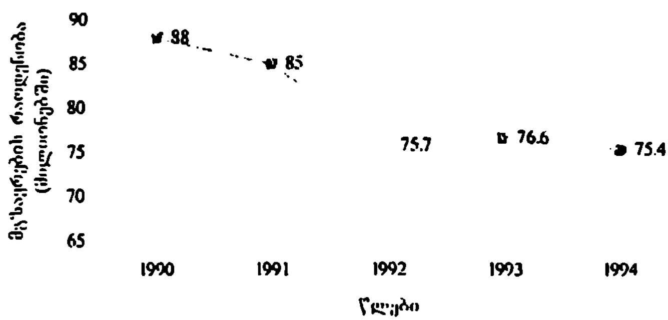
ნაბიჯი 1. დავხაზოთ საკოორდინატო სიბრტყე და მოვნიშნოთ x და y ღერძები.

ნაბიჯი 2. x ღერძზე გადავზომოთ წლები, ხოლო y ღერძზე – კი მგზავრთა რაოდენობა.

ნაბიჯი 3. ცხრილის შესაბამისად ავაგოთ თითოეული წერტილი.

ნაბიჯი 4. შევაერთოთ აგებული წერტილები სწორი ხაზებით ისე როგორც ეს ქვემოთაა ნაჩვენები (არ არის საჭირო მოიძებნოს გლუვი წირი ამ წერტილების შესაერთებლად).

დროითი მწკრივების დიაგრამა

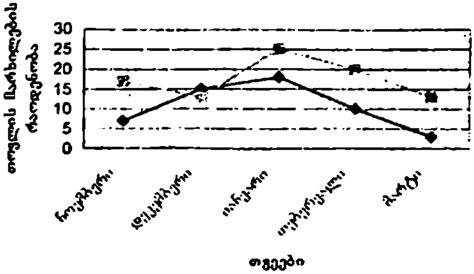


ეს დიაგრამა გვიჩვენებს, რომ მგზავრთა ნაკადი კლებულობდა 1992 წლამდე, ხოლო 1993 და 1994 წლებში იყო მგზავრთა ნაკადის გასწორების (გათანაბრების) ტენდენცია.

დროითი მწკრივების დიაგრამის ანალიზისას ყურადღება უნდა მივაქციოთ აქვს თუ არა მას რაიმე **ტრენდი** (ტენდენცია) დროის გარკვეული პერიოდის განმავლობაში. მაგალითად, აქვს თუ არა აღმავალი ხაზი (რაც მიუთითებს ზრდადობაზე დროის ამ პერიოდში) ან დაღმავალი ხაზი (რაც მიუთითებს კლებადობაზე დროის ამ პერიოდში)? გარდა ამისა, უნდა დავაკვირდეთ აქვს თუ არა ციცაბო დახრა ზევით ან ქვევით. გარკვეულ პერიოდში ციცაბო დახრა ზევით ან ქვევით შესაბამისად გვიჩვენებს, რომ დროის ამ პერიოდში სწრაფ ზრდას ან კლებას.

ერთი და იგივე საკოორდინატო სიბრტყეზე შესაძლებელია შევადაროთ მონაცემთა ორი სიმრავლე ორი დიაგრამის აგებით, რომელსაც უწოდებენ **შედგენილი (რთული) დროითი მწკრივების დიაგრამა**. ქვემოთ მოყვანილი დიაგრამა გვიჩვენებს მაღაზიის მიერ გაყიდული თოვლის მარხილების რაოდენობას ორი სეზონის განმავლობაში

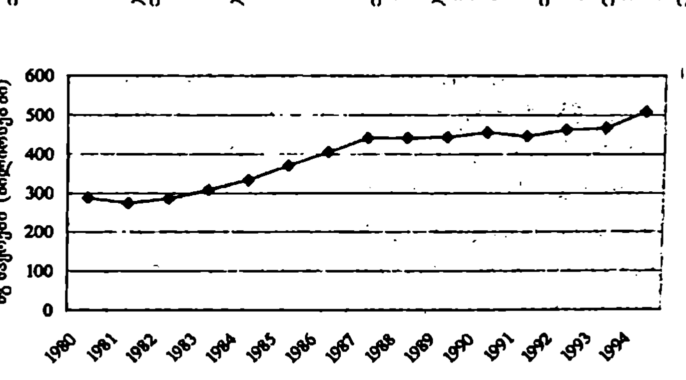
ორი დროითი მწკრივების დიაგრამა



სადაც კვადრატები შეესაბამება 1995 წელს, ხოლო რომბები – 1996 წელს.

აქედან ჩვენ ვხედავთ, მაგალითად, რომ 1995 წლის დეკემბერ – იანვარში იყო თოვლის მარხილების გაყიდვების სწრაფი ზრდის ტენდენცია, რასაც ვერ ვიტყვით 1996 წელზე (მიუხედავად იმისა, რომ ზრდის ტენდენცია აქაც შენარჩუნებული იყო).

დავუბრუნდეთ ახლა ცხრილს, რომელიც უჩვენებდა აშშ-ში საჰაერო გადაზიდვების რაოდენობის მატებას, მაგრამ მას არა აქვს მკითხველზე იმდენი ზეგავლენა რამდენიც ექნებოდა სტატისტიკური დიაგრამების გამოყენების შემთხვევაში. იმის საჩვენებლად, რომ აშშ-ში იმ მგზავრების რაოდენობა, რომლებიც დაფრინავენ თვითმფრინავებით, იზრდება -- გამოვიყენოთ დროითი მწკრივების დიაგრამა. შესაბამისი დიაგრამა მოყვანილია ქვემოთ. ამ გრაფიკის ზრდადობა იმის მაჩვენებელია, რომ აშშ-ში საჰაერო მგზავრების რაოდენობა განსახილველი პერიოდის განმავლობაში ძირითადად იზრდებოდა. თუ შევადარებთ საწყისი წერტილის (დროის საწყის მომენტში – 1980 წელს) სიმაღლეს (მგზავრთა რაოდენობას) ბოლო წერტილის (დროის ბოლო მომენტში – 1994 წელს) სიმაღლეს (მგზავრთა რაოდენობას) დავინახავთ, რომ მგზავრთა რაოდენობა დროის ამ პერიოდში თითქმის გაორმაგდა.



თავი IV
თვისებრივი მონაცემების სიხშირეთა განაწილება

თვისებრივი ნიშნის განაწილება. შეუღლების ცხრილი.

პოპულაციის ელემენტარული ერთეულები ეწოდება ობიექტებს, რომელთა მახასიათებლებიც პოპულაციას შეადგენენ. ერთი და იგივე ელემენტარული ერთეულების მეშვეობით, ბუნებრივია, შეიძლება შეიქმნას ბევრი სხვადასხვა პოპულაცია. მაგალითად, რომელიმე ფირმის მიერ დაქირავებულ მუშაკთა ერთობლიობა მოგვცემს მათი შემოსავლების პოპულაციას, მათი წონების პოპულაციას, მათი პოლიტიკური ორიენტაციების პოპულაციას და ა.შ.

ელემენტარულ ერთეულთა მახასიათებლის მიხედვით განირჩევა თვისებრივი და რაოდენობრივი პოპულაციები. ისინი სხვადასხვა მეთოდებით შეისწავლება. *თვისებრივი ანუ ატრიბუტული მახასიათებელი* ეწოდება ელემენტარული ერთეულის რაიმე თვისებას ან მდგომარეობას (მაგალითად, სქესი, პროფესია, ფერი და სხვა). თვისებრივ მახასიათებლის კონკრეტულ დონეს *ატრიბუტი* ეწოდება. ატრიბუტებია, მაგალითად, პირის ქორწინებითი მდგომარეობის შემდეგი დონეები: მარტოხელა, დაქორწინებული, გაყრილი, ქვრივი. რაოდენობრივია ისეთი მახასიათებელი, რომლის ცალკეული მნიშვნელობები დაკვირვების, გაზომვის ან ათვლის შედეგად მიიღება და რიცხვებით გამოიხატება (სიმაღლე, მასა, ხელფასი, წლიური მოგება და სხვა).

წინასწარ განვიხილოთ შემთხვევა როცა ელემენტარული ერთეული ერთი თვისებრივი ნიშნით ხასიათდება, რომელსაც ორი დონე აქვს -- A და მისი უარყოფა (საწინააღმდეგო) $\bar{A}$. თუ ერთობლიობას გადავნომრავთ 1, 2, . . . , n რიცხვებით, მაშინ ცხადია ნედლი მონაცემები შეიძლება ჩაიწეროს შემდეგი ცხრილის სახით:

ელემენტარული ერთეული	1	2	...	n
ვარიანტი	A	$\bar{A}$	...	A

თუ ამ ცხრილის მონაცემებს დავაჯგუფებთ, გავაერთიანებთ რა პირველ ჯგუფში იმ ერთეულებს, რომელთათვისაც ადგილი აქვს A-ს და აღვნიშნავთ რა მათ რაოდენობას n_A სიმბოლოთი, ხოლო მეორე ჯგუფში გავაერთიანებთ იმ ერთეულებს, რომელთათვისაც A-ს ადგილი არა აქვს (ანუ ადგილი აქვს $\bar{A}$-ს) და აღვნიშნავთ მათ რაოდენობას $n_{\bar{A}}$ სიმბოლოთი (ცხადია, რომ $n_{\bar{A}} = n - n_A$), მაშინ მივიღებთ შემდეგ ცხრილს:

ვარიანტი	A	$\bar{A}$	ჯამი
სიხშირე	n_A	$n_{\bar{A}}$	n

ამგვარ დაჯგუფებას *ორთანრიგიანი დაჯგუფება* ეწოდება, n_A და $n_{\bar{A}}$ *თანრიგების სიხშირეებია*, შესაბამისად. ზემოთ მოყვანილი ცხრილით მივიღეთ *ორდონიანი ნიშნის სიხშირეთა განაწილება*.

თუ ახლა n_A და $n_{\bar{A}}$ სიხშირეებს გავყოფთ n-ზე, მივიღებთ *ფარდობით სიხშირეებს* $h_A = n_A / n$ და $h_{\bar{A}} = n_{\bar{A}} / n$ (ცხადია, რომ $h_A + h_{\bar{A}} = 1$). თუ ზემოთ მოყვანილი ცხრილის მეორე სტრიქონში სიხშირეებს შევცვლით ფარდობითი სიხშირეებით (შესაბამისად, n-ის ნაცვლად ბოლო სვეტში ჩავწერთ 1-ს), მაშინ მივიღებთ *ორდონიანი ნიშნის ფარდობით სიხშირეთა განაწილების* ცხრილს:

ვარიანტი	A	$\bar{A}$	ჯამი
ფარდობითი სიხშირე	n_A / n	$n_{\bar{A}} / n$	1

თუ ნიშანს აქვს ორზე მეტი დონე, $A_1, \dots, A_s, s > 2$, მაშინ შესაბამისი სიხშირეების $n_1, \dots, n_s$ სიმბოლოებით აღნიშვნის შემდეგ ($n_1 + \cdots + n_s = n$), მივიღებთ შემდეგ ცხრილს:

ვარიანტი	A_1	A_2		A_s	ჯამი
სიხშირე	n_1	n_2		n_s	n

უკანასკნელი ცხრილი ასახავს *მრავალთანრიგიან (მრავლობით) დაჯგუფებას*. ისევე როგორც ორთანრიგიანი დაჯგუფების შემთხვევაში, აქაც შესაძლებელია $h_1 = n_1 / n, \qquad , h_s = n_s / n$ ფარდობით სიხშირეებზე გადასვლა, $h_1 + \cdots + h_s = 1$.

ერთი ნიშნის მიხედვით დაჯგუფებას *პირველი რიგის დაჯგუფებას* უწოდებენ. თუ პირველი რიგის დაჯგუფების ყოველი ჯგუფი მეორე ნიშნის მიხედვით დაჯგუფდა, მივიღებთ *მეორე რიგის დაჯგუფებას*. მეორე რიგის დაჯგუფების შესაბამისი ცხრილი იქნება:

II ნიშანი / I ნიშანი	B	$\bar{B}$	ჯამი
A	n_{AB}	$n_{A\bar{B}}$	n_A
$\bar{A}$	$n_{\bar{A}B}$	$n_{\bar{A}\bar{B}}$	$n_{\bar{A}}$
ჯამი	n_B	$n_{\bar{B}}$	n

აქ n_A სიხშირის ჯგუფი დაიყო ორ ნაწილად II ნიშნის მიხედვით – ის ერთეულები, რომელთათვისაც A-სთან ერთად B-ც გვხვდება და მათი რაოდენობაა n_{AB}, და ის ერთეულები, რომელთათვისაც A-სთან ერთად

B უკვე არ გვხვდება (არამედ, გვხვდება $\bar{B}$) და მათი რაოდენობაა $n_{A\bar{B}}$. ანალოგიურად, ორ ნაწილად დაიყოფა $n_{\bar{A}}$ სიხშირის ჯგუფი იმ ერთეულების, რომელთათვისაც ადგილი აქვს $\bar{A}$-ს. ცხადია, რომ სამართლიანია შემდეგი ტოლობები:

$$n_A = n_{AB} + n_{A\bar{B}}, \qquad n_{\bar{A}} = n_{\bar{A}B} + n_{\bar{A}\bar{B}}, \qquad n = n_A + n_{\bar{A}},$$
$$n_B = n_{AB} + n_{\bar{A}B}, \qquad n_{\bar{B}} = n_{A\bar{B}} + n_{\bar{A}\bar{B}}, \qquad n = n_B + n_{\bar{B}}.$$

მაგალითი 1. ქოლერის საწინააღმდეგო აცრებსა და ქოლერის დაავადებას შორის კავშირი, იულისა და გრინვუდის 1915 წლის გამოკვლევის მიხედვით მოყვანილია შემდეგ ცხრილში:

	არ დაავადებულა	დაავადდა	ჯამი
აიცრა	276	3	279
არ აცრილა	473	66	539
ჯამი	749	69	818

ეს ცხრილი აშკარად მეტყველებს აცრების სასარგებლოდ.

ძნელი წარმოსადგენი არ არის, რა შეიცვლება, თუ განვიხილავთ მესამე რიგის დაჯგუფებას, ანუ უმარტივეს შემთხვევაში ორ ორდონიან ნიშანს დაემატება კიდევ ერთი ორდონიანი ნიშანი.

თუ მეორე რიგის დაჯგუფებაში ორდონიანი ნიშნის ნაცვლად განვიხილავთ $A_1,\dots,A_s$ დონეების მქონე პირველ ნიშანსა და $B_1,\dots,B_t$ დონეების მქონე მეორე ნიშანს, და $n_{i,j}$ სიმბოლოებით აღვნიშნავთ იმ ერთეულების რაოდენობას (სიხშირეს), რომელთათვისაც შეთავსებულია (ერთდროულად გვხვდება) ნიშნების A_i და B_j დონეები, მივიღებთ ე. წ. *ორი ნიშნის შეუღლების ცხრილს:*

II (B_j) / I (A_i)	B_1	B_2	B_j	B_t	ჯამი
A_1	$n_{1.1}$	$n_{1.2}$	$n_{1.j}$	$n_{1.t}$	$n_{1\bullet}$
A_2	$n_{2.1}$	$n_{2.2}$	$n_{2.j}$	$n_{2.t}$	$n_{2\bullet}$
A_i	$n_{i.1}$	$n_{i.2}$	$n_{i.j}$	$n_{i.t}$	$n_{i\bullet}$
A_s	$n_{s.1}$	$n_{s.2}$	$n_{s.j}$	$n_{s.t}$	$n_{s\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet j}$	$n_{\bullet t}$	n

აქ შემოღებულია შემდეგი აღნიშვნები: $n_{1\bullet}$, $n_{s\bullet}$ არის პირველი ნიშნის $A_1,\dots,A_s$ დონეების შესაბამისი სიხშირეები, ხოლო $n_{\bullet 1}$, $n_{\bullet t}$ —

მეორე ნიშნის $B_1,\dots,B_l$ დონეების შესაბამისი სიხშირეები. სამართლიანია შემდეგი ტოლობები (ვარსკვლავი ინდექსად ნიშნავსშესაბამისი ინდექსით შეჯამებას):

$$n_{1\bullet}=n_{1,1}+\cdots+n_{1,l}, \qquad n_{s\bullet}=n_{s,1}+\cdots+n_{s,l},$$
$$n_{\bullet 1}=n_{1,1}+\cdots+n_{s,1}, \qquad n_{\bullet l}=n_{1,l}+\cdots+n_{s,l},$$
$$n_{\bullet\bullet}=n=n_{1\bullet}+\cdots+n_{s\bullet}=n_{\bullet 1}+\cdots+n_{\bullet l}.$$

მაგალითი 2. დავაკვირდეთ, როგორაა შეუღლებული ორი ნიშანი – ქალაქში თუ სოფელში ცხოვრება (შესაბამისად, A და $\bar{A}$) და ოჯახის სულადობა სამცხე-ჯავახეთში 1998 წლის მონაცემების მაგალითზე:

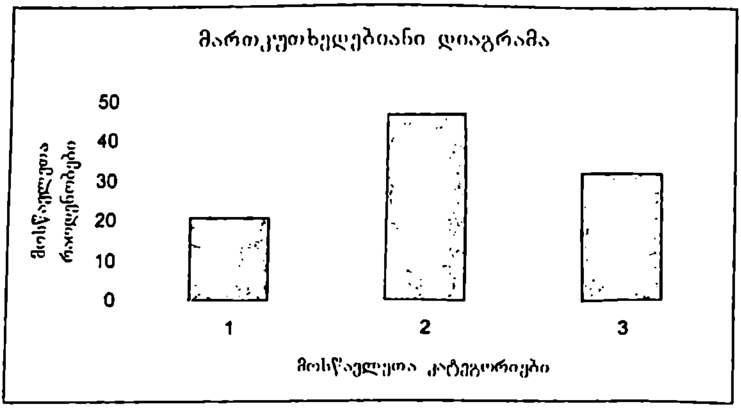
ქალაქი ან სოფელი	ოჯახის სულადობა										ჯამი
	1	2	3	4	5	6	7	8	9	10	
ქალაქი	5	6	6	8	2	9	4	0	1	1	42
სოფელი	10	20	13	11	24	11	0	0	0	0	89
ჯამი	15	26	19	19	26	20	4	0	1	1	131

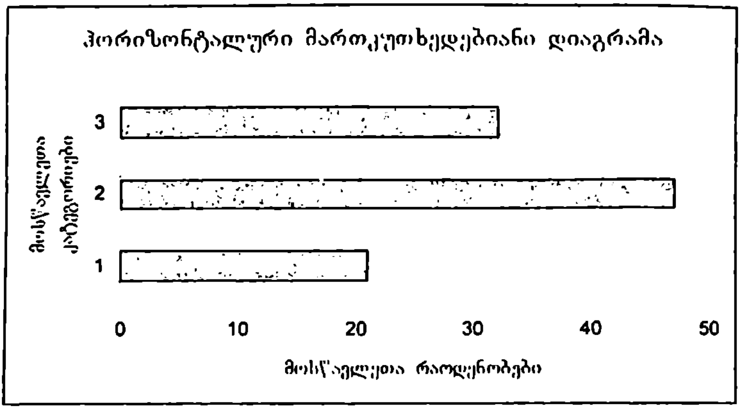
ცხრილიდან შეიძლება დავასკვნათ, რომ ოჯახის სულადობის მხრივ სოფლად და ქალაქად განსხვავებული მდგომარეობაა.

თვისებრივ მონაცემთა განაწილების გრაფიკული წარმოდგენისათვის ორი ძირითადი საშუალება არსებობს: მართკუთხედებიანი (სვეტოვანი) დიაგრამები და წრიული (სექტორებიანი) დიაგრამები.

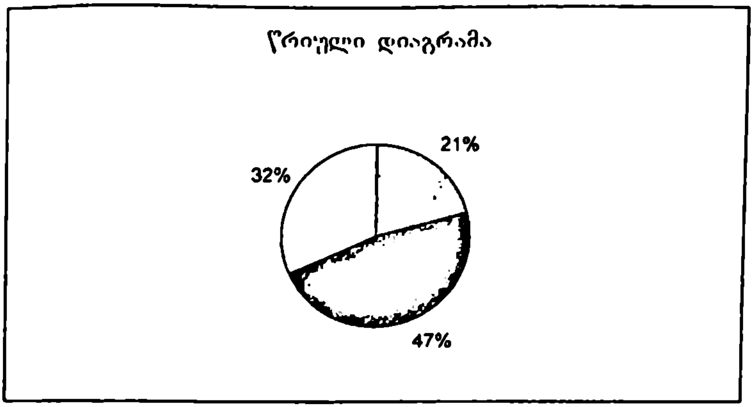
მაგალითი: სკოლის დამამთავრებელი კლასების მოსწავლეთა აკადემიური მოსწრება ნიშნების მიხედვით: I კატეგორია: “9 — 10 ბალიანი” – 21, II კატეგორია: “7 – 8 ბალიანი” – 47, III კატეგორია: “5 – 6 ბალიანი” – 32.

ავიღოთ აბსცისთა ღერძზე მდებარე ტოლი ფუძეების მქონე სამი მართკუთხედი, რომელთა სიმაღლეები პროპორციულია მოსწავლეთა რაოდენობების და ისინი ერთმანეთისაგან ტოლი დაშორებით განვალაგოთ. ორდინატთა ღერძზე გადაიზომება სიხშირეები. ასე მიღებულ ნახატს მართკუთხედებიანი დიაგრამა ეწოდება. ცხადია, რომ ვერტიკალურ ღერძზე შეიძლებოდა გადაგვეზომა ფარდობითი სიხშირე.





წრიული დიაგრამა განკუთვნილია დროის მოცემულ მომენტში ან მოცემულ შუალედში ობიექტთა მოცემული რაოდენობის კომპონენტებად დაყოფის ანუ ამ ობიექტთა კლასიფიკაციის აღსანიშნავად, იგი უფრო მოხერხებულია პროცენტებში გამოთქმული ნაწილების ან ფარდობითი სიხშირეების გამოსახატად. ჩვენს მაგალითში ვაგებთ სექტორებს, რომლებიც შეადგენენ სრული კუთხის – 360-ის შესაბამისად 21/100-ს, ანუ 21%. 47/100-ს ანუ 47%, 32/100-ს ანუ 32%.



სტატისტიკაში ინტენსიურად გამოიყენება წრიული დიაგრამა. წრიული დიაგრამის გამოყენების მიზანია ვაჩვენოთ მთელის ნაწილებს შორის ურთიერთმიმართება სექტორების ზომების ვიზუალური შედარების გზით. გამოიყენება როგორც პროცენტები, ისე პროპორციები. სიდიდეები შეიძლება იყოს როგორც რიცხობრივი, ისე კატეგორიული ბუნების.

წრიული დიაგრამა ეს არის სექტორებად დაყოფილი წრე, სადაც თითოეული სექტორი შეესაბამება ნებისმიერი კატეგორიის განაწილებაში მიღებული სიხშირეების პროცენტულ გამოსახულებებს ან ფარდობით სიხშირეებს.

მაგალითი 1. ქვემოთ მოყვანილია 1998 წლის განმავლობაში მსუბუქ საუზმეზე დახარჯული ფუნტების რაოდენობის განაწილება საკვების სახეობების მიხედვით. ავაგოთ წრიული დიაგრამა.

მსუბუქი საუზმე	ფუნტები (სიხშირე)
კარტოფილის "ჩიფსი"	11.2 მილიონი
სიმინდის "ჩიფსი"	8.2 მილიონი
ფუნთუშა	4.3 მილიონი
სიმინდის ბურბუშელა	3.8 მილიონი
კექსი	2.5 მილიონი
	ჯამი $n = 30.0$ მილიონი

ამოხსნა.

ნაბიჯი 1. ვინაიდან წრე შედგება 360 გრადუსისაგან, თითოეული კლასის სიხშირე უნდა გადავიყვანოთ წრის პროპორციულ ნაწილში. ეს გადაყვანა უნდა მოხდეს შემდეგი ფორმულის გამოყენებით

$$\text{გრადუსი} = \frac{f}{n} \cdot 360^0,$$

სადაც f – არის თითოეული კლასის შესაბამისი სიხშირე,

ხოლო n -- არის სიხშირეთა ჯამი. ამიტომ გვექნება შემდეგი თანაფარდობები (ცხადია, რომ ყველა გრადუსის ჯამი უნდა იყოს 360^0):

კარტოფილის "ჩიფსი" — $\frac{11.2}{30}\cdot 360^0 = 134^0$

სიმინდის "ჩიფსი" — $\frac{8.2}{30}\cdot 360^0 = 98^0$

ფუნთუშა — $\frac{4.3}{30}\cdot 360^0 = 52^0$

სიმინდის ბურბუშელა — $\frac{3.8}{30}\cdot 360^0 = 46^0$

ქექსი — $\frac{2.5}{30}\cdot 360^0 = 30^0$

ჯამი ——— $= 360^0$

ნაბიჯი 2. თითოეული სიხშირე უნდა გადავიყვანოთ აგრეთვე პროცენტებში, რისთვისაც უნდა ვისარგებლოთ ფორმულით

$$\% = \frac{f}{n}\cdot 100\% .$$

შესაბამისად, ჩვენ მივიღებთ შემდეგ პროცენტებს (ცხადია მათი ჯამი უნდა იყოს 100%):

კარტოფილის "ჩიფსი" — $\frac{11.2}{30}\cdot 100\% = 37.3\%$

სიმინდის "ჩიფსი" — $\frac{8.2}{30}\cdot 100\% = 27.3\%$

ფუნთუშა — $\frac{4.3}{30}\cdot 100\% = 14.3\%$

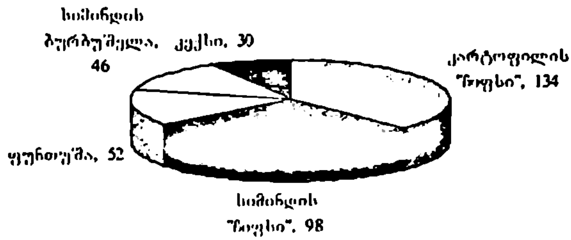
სიმინდის ბურბუშელა — $\frac{3.8}{30}\cdot 360^0 = 46^0$

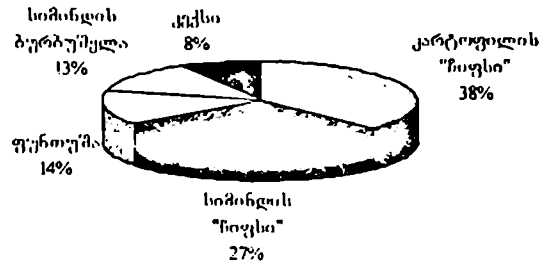
ქექსი — $\frac{2.5}{30}\cdot 100\% = 8.3\%$

ჯამი ——— = 100%.

შენიშვნა: დამრგვალების გამო გრადუსების ჯამი შეიძლება ყოველთვის არ იყოს 360 გრადუსი, ისევე როგორც პროცენტების ჯამი – 100%.

ნაბიჯი 3. ვისარგებლოთ ტრანსპორტირით, დავხაზოთ წრიული დიაგრამა და მასზე მივუთითოთ თითოეული სექტორის დასახელება და შესაბამისი ერთეული, როგორც ეს ნახაზზეა ნაჩვენები.





მაგალითი 2. ააგეთ წრიული დიაგრამა არმიის ახალწვეულთა სისხლის ჯგუფების სიხშირეთა განაწილების ქვემოთ მოყვანილი ცხრილის მიხედვით:

კლასი	სიხშირე
A	5
B	7
O	9
AB	4
	ჯამი 25

ამოხსნა.

ნაბიჯი 1. ვიპოვოთ თითოეული კლასის შესაბამისი გრადუსული ზომა შემდეგი ფორმულის გამოყენებით:

$$გრადუსი = \frac{f}{n} \cdot 360^0$$

მაშინ თითოეული კლასისათვის მივიღებთ:

$$A \quad — \quad \frac{5}{25} \cdot 360^0 = 72^0$$

$$B \quad — \quad \frac{7}{25} \cdot 360^0 = 100.8^0$$

$$O \quad — \quad \frac{9}{25} \cdot 360^0 = 129.6^0$$

$$AB \quad — \quad \frac{4}{25} \cdot 360^0 = 57.6^0.$$

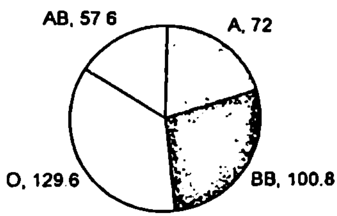
ნაბიჯი 2. ვიპოვოთ თითოეული კლასის შესაბამისი პროცენტი, შემდეგი ფორმულის მიხედვით:

$$\% = \frac{f}{n} \cdot 100\%.$$

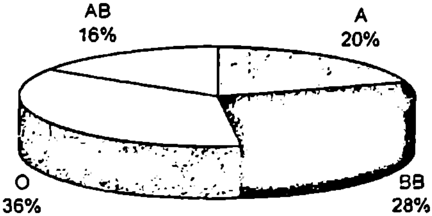
მივიღებთ შესაბამისად პროცენტების შემდეგ რიცხვებს: A – 20%, B – 28%, O – 36% , AB – 16%.

ნაბიჯი 3. ვისარგებლოთ ტრანსპორტირით, ავაგოთ წრიული დიაგრამა და მასზე მოვნიშნოთ თითოეული სექტორის სახელი. გვექნება:

წრიული დიაგრამა



წრიული დიაგრამა



ამ დიაგრამიდან ჩანს, რომ ყველაზე მეტად გავრცელებული სისხლის ჯგუფია O (36%). იმისათვის, რომ გავაანალიზოთ წრიულ დიაგრამაზე წარმოდგენილი მონაცემები, უნდა შევადაროთ სექტორები. მაგალითად, არის თუ არა რომელიმე სექტორი შედარებით დიდი ყველა დანარჩენთან შედარებით? ეს დიაგრამა გვიჩვენებს, რომ ყველა დანარჩენ სისხლის ტიპს შორის ყველაზე მეტად გავრცელებულია O ტიპის სისხლი. ადამიანები, რომელთაც აქვთ AB ტიპის სისხლი, არიან უმცირესობაში. ადამიანების ორჯერ უფრო მეტ რაოდენობას O ტიპის სისხლი, AB ტიპის სისხლთან შედარებით.

თავი V
რაოდენობრივი მონაცემების სიხშირეთა განაწილება

რაოდენობრივია ისეთი მახასიათებელი, რომლის ცალკეული მნიშვნელობები დაკვირვების, გაზომვის ან თვლის შედეგად მიიღება და რიცხვით გამოიხატება (სიმაღლე, მასა, ხელფასი, წლიური მოგება და სხვა). ვთქვათ, ელემენტარული ერთეული ხასიათდება ერთი რაოდენობრივი x ნიშნით და ნედლი მონაცემები შედგება შემდეგი n დაკვირვებისგან: $x_1,...,x_n$. *ვარიაციული მწკრივი* ეწოდება ნედლი მონაცემების დალაგებას არაკლებადი მიმდევრობის სახით $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. თუ ახლა $x^{(1)}$-ით აღვნიშნავთ მინიმალურ დაკვირვებულ მნიშვნელობას, $x^{(2)}$-ით – სიდიდით მეორე ნიშვნელობას ვარიაციულ მწკრივში და ა. შ. $x^{(s)}$-ით კი – მაქსიმალურს, მაშინ მივიღებთ ზრდად ვარიაციულ მწკრივს: $x^{(1)} < x^{(2)} <...< x^{(s)}$. თუ n_1-ით აღვნიშნავთ ვარიაციულ მწკრივში $x^{(1)}$-ის სიხშირეს, და ა. შ. n_s-ით აღვნიშნავთ ვარიაციულ მწკრივში $x^{(s)}$-ის სიხშირეს, მივიღებთ ზრდადი ვარიაციული მწკრივის შესაბამის სიხშირეებს, ანუ სიხშირეთა განაწილებას:

$x^{(i)}$	$x^{(1)}$	$x^{(2)}$		$x^{(s)}$	ჯამი
n_i	n_1	n_2		n_s	n

თუ გადავალთ ფარდობით სიხშირეებზე: $h_1 = n_1 / n ,...,$ $h_s = n_s / n$ (გასაგებია, რომ მაშინ $h_1 + \cdots + h_s = 1$), გვექნება ფარდობით სიხშირეთა განაწილება:

$x^{(i)}$	$x^{(1)}$	$x^{(2)}$		$x^{(s)}$	ჯამი
h_i	h_1	h_2		h_s	1

ეს ცხრილები შეიძლება წერტილოვანი დიაგრამის ან მესერული დიაგრამის სახით გამოისახოს.

მაგალითი 1. მსხვილი კომპანიის კადრების განყოფილების უფროსს დაევალა ჩაეტარებინა თანამშრომელთა სამსახურში არ გამოცხადების ანალიზი. ქვევით მოცემულია უკანასკნელი 21 სამუშაო დღის განმავლობაში სამსახურში არ გამოცხადებულ თანამშრომელთა რაოდენობები.

1, 1, 2, 2, 1, 0, 2, 3, 2, 2, 1, 1, 4, 3, 1, 0, 2, 1, 4, 1, 3.

ასეთი მონაცემები მიეკუთვნებიან *დისკრეტული ტიპის მონაცემებს* (ის ღებულობს მხოლოდ მთელ მნიშვნელობებს). მოცემულ რიცხვთა ერთობლიობა წარმოადგენს $n = 21$ *მოცულობის შერჩევას* (შერჩევის მოცულობა დაკვირვებათა რაოდენობის ტოლია). ასეთი სახით ჩაწერილი მონაცემების გამოყენება შემდგომი ანალიზისათვის მეტად ძნელია (მათ *ნედლი მონაცემები* ეწოდება). ამიტომ, პირველ რიგში მონაცემები დავალაგოთ

ზრდადობის მიხედვით (გავაკეთოთ ვარიაციული მწკრივი). მივიღებთ შემდეგ მწკრივს:

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2,2, 2, 2, 2, 2, 3, 3, 3, 4, 4.

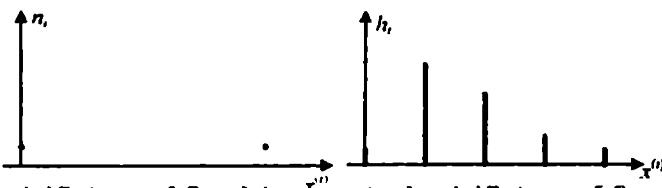
ვარიაციული მწკრივიდან ჩვენ ადვილად შეგვიძლია ამოვწეროთ მასში შემავალი განსხვავებული რიცხვითი მნიშვნელობები, ანუ *ვარიანტები* და ამ მნიშვნელობათა *სიხშირეები*, ანუ ვარიანტების გამეორებათა რაოდენობები. მოცემულ ვარიაციულ მწკრივში 5 განსხვავებული მნიშვნელობაა: 0, 1, 2, 3, 4. ამასთანავე, უკანასკნელი 21 სამუშაო დღის განმავლობაში მხოლოდ ორჯერ იყო სრული გამოცხადება, რვაჯერ არ გამოცხადდა მარტო ერთი თანამშრომელი და ა. შ. ყველა ამ ინფორმაციას შეგვიძლია თავი მოვუყაროთ ერთ ცხრილში, რომელიც გვაძლევს სამსახურში არ გამოცხადებულ თანამშრომელთა სიხშირული განაწილების სურათს:

არ გამოცხადებულთა რაოდენობა (ვარიანტი)	0	1	2	3	4	დაკვირვებული დღეების რაოდენობა
გაცდენილი დღეების რაოდენობა (სიხშირე)	2	8	6	3	2	21

თუ გადავალთ ფარდობით სიხშირეებზე (რომელიც მიიღება სიხშირეების გაყოფით შერჩევის მოცულობაზე), მაშინ მივიღებთ ფარდობით სიხშირეთა განაწილების ცხრილს:

ვარიანტები	0	1	2	3	4	ჯამი
ფარდობითი სიხშირე	$\frac{2}{21}$	$\frac{8}{21}$	$\frac{6}{21}$	$\frac{3}{21}$	$\frac{2}{21}$	1

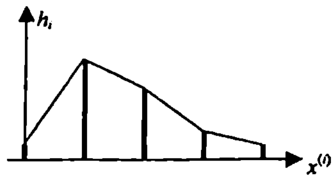
ეს ცხრილები შეიძლება *წერტილოვანი ან მესერული დიაგრამების* საშუალებით გამოისახოს:



სიხშირეთა განაწილების წერტილოვანი დიაგრამა

ფარდობით სიხშირეთა განაწილების მესერული დიაგრამა

გარდა ამისა, მონაცემთა წარმოსადგენად გამოიყენება *სიხშირეთა პოლიგონი*. იგი მიიღება წერტილოვანი დიაგრამის წერტილების მიმდევრობითი შეერთებით. მას შემდეგი სახე აქვს:



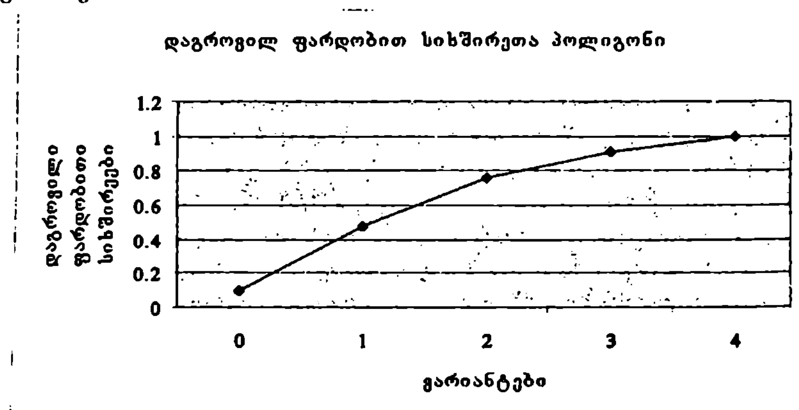
გადავიდეთ *დაგროვილ ფარდობით სიხშირეებზე*. ვარიანტის მოცემული მნიშვნელობისათვის დაგროვილი ფარდობითი სიხშირე განისაზღვრება როგორც იმ ვარიანტების ფარდობითი სიხშირეების ჯამი, რომლებიც არ აღემატება ვარიანტის მოცემულ მნიშვნელობას. მოცემულ მაგალითში დაგროვილი ფარდობითი სიხშირეების ცხრილს ექნება შემდეგი სახე:

ვარიანტები	0	1	2	3	4
დაგროვილი ფარდობითი სიხშირე	$\frac{2}{21}$	$\frac{2}{21}+\frac{8}{21}$	$\frac{2}{21}+\frac{8}{21}+\frac{6}{21}$	$\frac{2}{21}+\frac{8}{21}+\frac{6}{21}+\frac{3}{21}$	1

საბოლოოდ ეს ცხრილი მიიღებს შემდეგ სახეს:

ვარიანტები	0	1	2	3	4
დაგროვილი ფარდობითი სიხშირე	$\frac{2}{21}$	$\frac{10}{21}$	$\frac{16}{21}$	$\frac{19}{21}$	1

ამასთანავე, *დაგროვილ ფარდობით სიხშირეთა პოლიგონს* ექნება შემდეგი სახე:



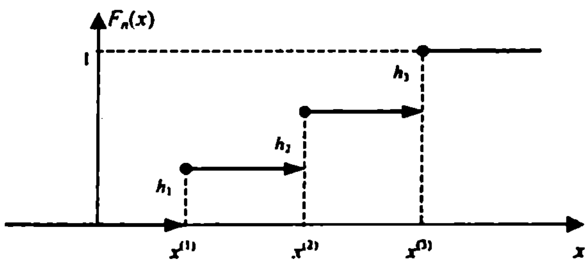
საზოგადოდ, მოცემული ნამდვილი x რიცხვისათვის *დაგროვილი ფარდობითი სიხშირე* $F_n(x)$ განისაზღვრება როგორც იმ $x^{(i)}$ მნიშვნელობა-

თა ფარდობითი სიხშირეების ჯამი, რომლებიც x-ს არ აღემატება (ანუ x-ზე ნაკლებია ან მისი ტოლია). შესაბამისად, გვაქვს:

$$F_n(x) = \begin{cases} 0, & x < x^{(1)}, \\ \dfrac{n_1}{n}, & x^{(1)} \le x < x^{(2)}, \\ \dfrac{n_1}{n} + \cdots + \dfrac{n_k}{n}, & x^{(k)} \le x < x^{(k+1)}, \\ 1, & x^{(s)} \le x. \end{cases}$$

ცხადია, რომ $F_n(x)$ ფუნქცია უბან-უბან მუდმივია და მისი ზრდა ხდება $x^{(1)}$, $x^{(2)}$, ,$x^{(s)}$ წერტილებში. $x^{(1)}$-მდე $F_n(x)$ ნულია, $x^{(s)}$-ის მარჯვნივ (მისი ჩათვლით) კი – 1. ყოველ $x^{(i)}$ წერტილში $F_n(x)$ ფუნქცია მატულობს (იზრდება) $h_i = n_i / n$ ფარდობითი სიხშირის ტოლი სიდიდით, სწორედ ეს უკანასკნელი გარემოება ნიშნავს, რომ $F_n(x)$ მოიცავს ფარდობით სიხშირეთა განაწილებით მოცემულ მთელ ინფორმაციას.

დაგროვილ *(კუმულაციურ, კუმულატიურ)* სიხშირეთა, ან ფარდობითი სიხშირეთა ფუნქციას ზოგჯერ *ემპირიულ განაწილების ფუნქციას* ან *კუმულატას* უწოდებენ. მის გრაფიკს ($s = 3$-ის შემთხვევაში) ექნება შემდეგი სახე:



უკანასკნელ მაგალითში მოყვანილი მონაცემებისათვის დაგროვილ ფარდობით სიხშირეთა ფუნქცია იქნება:

$$F_{21}(x) = \begin{cases} 0, & x < 0, \\ 2/21, & 0 \le x < 1, \\ 10/21, & 1 \le x < 2, \\ 16/21, & 2 \le x < 3, \\ 19/21, & 3 \le x < 4, \\ 1, & x \ge 4. \end{cases}$$

თავი VI

მონაცემთა დაჯგუფება. სიხშირეთა ინტერვალური განაწილება

$x^{(1)}<x^{(2)}<...<x^{(l)}$ ვარიანტების დიდი რაოდენობის შემთხვევაში კუმულატა ძნელად აღსაქმელია და მისი გამოყენება შემდგომი ანალიზისათვის შეზღუდულია. ამ შემთხვევაში, განსაკუთრებით კი უწყვეტ მონაცემთა აღწერისას მიმართავენ მონაცემთა დაჯგუფებას და სიხშირეებს და კუმულაციურ სიხშირეებს დაჯგუფებული მონაცემებისათვის გამოითვლიან. როგორც ცნობილია, მონაცემებს უწოდებენ *უწყვეტს*, თუ ისინი ღებულობენ ნებისმიერ მნიშვნელობებს რაიმე საზღვრებში.

დაჯგუფების მიზნით ინტერვალი, რომელშიც მოთავსებულია დაკვირვებული მნიშვნელობები x_{min} მინიმალური მნიშვნელობიდან x_{max} მაქსიმალურ მნიშვნელობამდე დაიყოფა ტოლი ან არატოლი სიგრძის რამდენიმე ინტერვალად:

$x_{min}=a_0 \le x < a_1, a_1 \le x < a_2, ..., a_{l-1} \le x \le a_l = x_{max}$

შემდეგ გამოიყოფა ამ ინტერვალებში მოთავსებულ დაკვირვებათა ჯგუფები, რომელთა სიხშირეები n_1, ,n_l სიმბოლოებით აღინიშნება. შესაბამისად, ჩვენ შეგვიძლია შევადგინოთ სიხშირეთა და ფარდობით სიხშირეთა ინტერვალური განაწილებების ცხრილი:

ინტერვალი	$[a_0,a_1)$	$[a_1,a_2)$		$[a_{l-1},a_l]$	ჯამი
სიხშირე	n_1	n_2		n_l	n
ფარდობითი სიხშირე	n_1/n	n_2/n		n_l/n	1

ზოგჯერ სიხშირეთა (ფარდობით სიხშირეთა) ინტერვალური განაწილების ნაცვლად იხილავენ მის დისკრეტიზაციას ანუ განაწილებას, რომელიც ყოველი ინტერვალის შუაწერტილს მიაწერს ინტერვალის შესაბამის სიხშირეს (ფარდობით სიხშირეს). გარკვეული აზრით ეს უფრო მოხერხებულია (თუნდაც მისი სიმარტივის გამო). შესაბამისად, იხილავენ სიხშირეთა (ფარდობით სიხშირეთა) განა-წილების ცხრილის ანალოგს:

ინტერვალის ცენტრი	$(a_0+a_1)/2$	$(a_1+a_2)/2$	. .	$(a_{l-1}+a_l)/2$	ჯამი
სიხშირე	n_1	n_2	.	n_l	n
ფარდობითი სიხშირე	n_1/n	n_2/n	.	n_l/n	1

მონაცემების დაჯგუფებისას როგორც წესი, ტოლი სიგრძის ინტერვალებს ამჯობინებენ, თუმცა ზოგჯერ გამართლებულია განსხვავებული სიგრძის ინტერვალების განხილვაც. ინტერვალების არაერთგვაროვნების მაჩვენებელია შემდეგი მაგალითი. დავუშვათ, ერთი ინტერვალის სიგრძეა 10 და იგი 30 დაკვირვებას შეიცავს, მეორის სიგრძეა 4 და მასში 20 დაკვირვებაა. მაშინ პირველ ინტერვალში სიგრძის ერთეულზე მოდის საშუალოდ

3 დაკვირვება, ხოლო მეორეში – 5. ამიტომ გამართლებულია (მიზანშეწონილია) ე. წ. *სიხშირეთა სიმკვრივის* და *ფარდობით სიხშირეთა სიმკვრივის* ცნებების შემოღება.

სიხშირეთა სიმკვრივეს უწოდებენ შეფარდებას $\xi_k = n_k / \Delta a_k$, $k = 1,2,\dots,l$. სადაც $\Delta a_k = a_k - a_{k-1}$ წარმოადგენს ინტერვალის სიგრძეს. ანალოგიურად, ფარდობით სიხშირეთა სიმკვრივეს უწოდებენ შეფარდებას

$$\eta_k = h_k / \Delta a_k = n_k / n\Delta a_k, \quad k = 1,2,\dots,l.$$

თუ ახლა ინტერვალების სიხშირეთა სიმკვრივეებს შესაბამის ინტერვალებში აბსცისთა ღერძის პარალელური მონაკვეთებით გამოვსახავთ ღერძიდან ξ_k სიმაღლეზე, მივიღებთ $\xi(x)$ *სიხშირეთა ჰისტოგრამას*, ხოლო ფარდობით სიმკვრივეთა ანალოგიური გამოსახვით აბსცისთა ღერძის პარალელური მონაკვეთებით ღერძიდან η_k სიმაღლეზე – $\eta(x)$ *ფარდობით სიხშირეთა ჰისტოგრამას*. ყოველ $[a_{k-1}, a_k)$ ინტერვალზე, როგორც ფუძეზე, აგებულია პირველ შემთხვევაში n_k ფართობის მართკუთხედი, ხოლო მეორეში – $\xi_k = n_k / n$ ფართობისა. $\xi(x)$ სიხშირეთა ჰისტოგრამისათვის ამ ფართობების ჯამია n, ხოლო $\eta(x)$ ფარდობით სიხშირეთა ჰისტოგრამისათვის – 1.

ჰისტოგრამის აგების მოყვანილი წესი მოკლედ შეიძლება შემდეგი ცხრილის სახით გამოისახოს (სადაც $\xi(x)$ – აღნიშნავს სიხშირეთა ჰისტოგრამის ორდინატას, ხოლო $\eta(x)$ – აღნიშნავს ფარდობით სიხშირეთა ჰისტოგრამის ორდინატას):

ინტერვალი	$x < a_0$	$[a_0; a_1)$	$[a_1, a_2)$		$[a_{l-1}, a_l]$	$x \geq a_l$
$\xi(x)$	0	$\xi_1 = n_1 / \Delta a_1$	$\xi_2 = n_2 / \Delta a_2$		$\xi_l = n_l / \Delta a_l$	0
$\eta(x)$	0	$\eta_1 = h_1 / \Delta a_1$	$\eta_2 = h_2 / \Delta a_2$		$\eta_l = h_l / \Delta a_l$	0

სიხშირეთა ინტერვალური განაწილების აგებისას ინტერვალების რაოდენობისა და სიგანის განსაზღვრისათვის საყოველთაოდ მიღებული წესი არ არსებობს. სტატისტიკურ ლიტერატურაში რეკომენდებულია 5-დან 20-მდე ტოლი სიგრძის ინტერვალის აგება, მაგრამ საკითხი ყოველ კერძო შემთხვევაში ცალკეა გადასაწყვეტი. იმ შემთხვევაში, როცა ტოლი Δ სიგრძის ინტერვალებს ვაგებთ, ცხადია, რომ ინტერვალების რაოდენობა l იქნება:

$$l = \frac{x_{max} - x_{min}}{\Delta}.$$

მაგალითი 1. ქვემოთმოყვანილ ცხრილში მოცემულია დიდი ზომის ავტომანქანების საწვავის დანახარჯის დასადგენად ჩატარებული 100 დაკვ-

ირვების მონაცემები. ცხრილი გვიჩვენებს 1 გალონი (=3.78 ლიტრი) საწვავით ავტომანქანის გავლილ მანძილს მილებში (1 მილი = 1.609 კმ):

მილი/გალონზე									
19.0	20.8	22.0	22.7	20.0	18.9	16.6	16.8	20.8	14.7
15.1	21.8	21.1	21.5	21.1	15.5	19.3	15.1	20.6	16.8
18.2	20.5	15.3	16.2	16.3	22.8	22.7	21.9	22.5	17.1
19.1	21.6	19.0	18.3	18.6	22.1	17.5	22.9	21.7	18.7
21.9	20.2	14.5	14.1	22.9	20.2	17.3	22.6	19.3	21.7
21.5	22.6	18.7	19.2	22.8	21.6	21.7	20.5	22.7	20.4
18.8	15.1	16.5	20.5	19.1	17.4	19.7	19.2	16.4	21.9
14.3	19.2	19.7	17.1	21.4	21.9	21.7	19.2	23.9	19.6
20.9	18.5	20.2	18.2	20.2	22.4	20.4	21.6	21.3	22.4
20.5	18.1	20.7	21.3	16.9	20.3	23.9	18.8	21.1	21.9

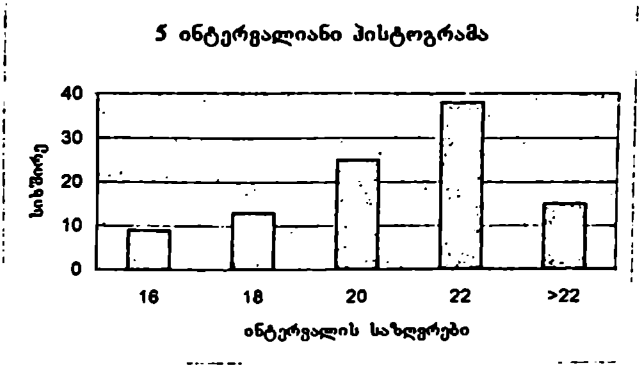
ამ მონაცემების დამუშავება მდგომარეობს შემდეგში:

I. დაკვირვებულ მონაცემებს შორის უნდა მოიძებნოს მინიმალური და მაქსიმალური მნიშვნელობები. წარმოდგენილ მონაცემებში უმცირესი რიცხვითი მნიშვნელობაა x_{min}=14.1, ხოლო უდიდესი – x_{max}=23.9. ეს ორი რიცხვი განსაზღვრავს მონაცემთა გაფანტულობის უმარტივეს მახასიათებელს – *გაბნევის დიაპაზონს*, რომელიც არის სხვაობა მონაცემთა მაქსიმალურ და მინიმალურ რიცხვით მნიშვნელობებს შორის (გაბნევის დიაპაზონი გამოთვლისა და ინტერპრეტაციის სიმარტივის გამო ფართოდ გამოიყენება ისეთი ამოცანების გადაწყვეტისას, სადაც გამოთვლის სისწრაფესა და მარტივ შინაარსს გადამწყვეტი როლი ენიჭება, მაგალითად, ხარისხის კონტროლის სფეროში). შესაბამისად, მონაცემთა გაბნევის დიაპაზონი ტოლია 23.9-14.1=9.8.

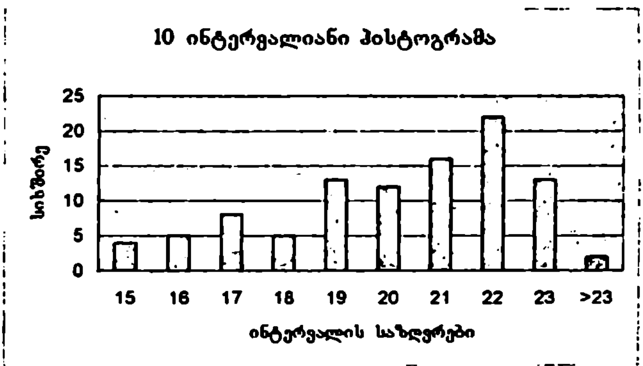
II. შემდეგი ნაბიჯია – მონაცემთა დაჯგუფების ინტერვალების გამოყოფა. ამისათვის უმარტივეს შემთხვევაში გაბნევის დიაპაზონს ყოფენ ერთი და იგივე სიგრძის დაახლოებით 5 – 20 ქვეინტერვალად. მოცემულ შემთხვევაში ავირჩიოთ 5 ტოლი სიგრძის ინტერვალი. სიმარტივისათვის მინიმალური და მაქსიმალური მნიშვნელობები დავამრგვალოთ უახლოეს მთელ რიცხვებამდე: კერძოდ, მინიმუმად ავიღოთ 14, ხოლო მაქსიმუმად – 24. მაშინ თითოეული ქვეინტერვალის სიგრძედ უნდა ავიღოთ Δ=(24-14)/5=2. შესაბამისად, დაჯგუფების ინტერვალები იქნება: (–∞,16], (16,18], (18,20], (20,22], (22,+∞). გამოვთვალოთ ამ ქვეინტერვალებში მოხვედრილი მონაცემების სიხშირეები და ჩავწეროთ ისინი ცხრილის სახით:

ინტერვალი	სიხშირეთა გამოთვლა	სიხშირე
(–∞,16]	///// ////	9
(16,18]	///// ///// ///	13
(18,20]	///// ///// ///// ///// /////	25
(20,22]	///// ///// ///// ///// ///// ///// ///// ///	38
(22,+∞)	///// ///// /////	15
	სულ	100

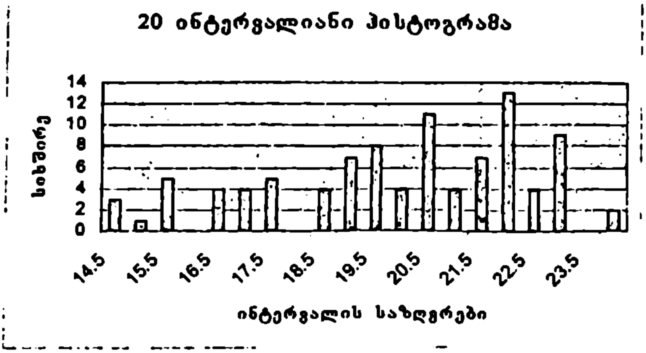
ამ ცხრილის მიხედვით აგებულ ჰისტოგრამას ექნება შემდეგი სახე:



თუ იგივე მონაცემებს დავაჯგუფებთ 10 ქვეინტერვალში, მაშინ ჰისტოგრამას ექნება შემდეგი სახე:



და ბოლოს, თუ იგივე მონაცემებს დავაჯგუფებთ 20 ქვეინტერვალ-ში, მაშინ უკვე ჰისტოგრამას ექნება შემდეგი სახე:



როგორც ვხედავთ, მონაცემთა დაჯგუფების ქვეინტერვალების რაოდენობის ზრდასთან ერთად ჰისტოგრამა სულ უფრო და უფრო დაკბილული ხდება. ამასთანავე, ამ შემთხვევაში დაჯგუფების 5 და 10 ინტერვალი იძლევა უფრო მკაფიოდ გამოხატულ კანონზომიერებას, მაშინ როდესაც 20 ინტერვალის შემთხვევაში გაჩნდა ბევრი, მოვლენის შინაარსიდან გამომდინარე, აზრს მოკლებული "რხევა", რაც გამოწვეულია მონაცემთა მოცემული რაოდენობისათვის დაჯგუფების ინტერვალთა დიდი რაოდენობით.

დავალება. აიღეთ დაჯგუფების 10 (20) ინტერვალი, შეადგინეთ სიხშირეთა ინტერვალური განაწილების ცხრილი ორივე შემთხვევაში, ააგეთ შესაბამისი ჰისტოგრამები და შეადარეთ ზემოთ დახაზულ ჰისტოგრამებს.

სიხშირეთა განაწილება გრაფიკულად შეიძლება გამოისახოს აგრეთვე ე. წ. *ფოთლებიანი ღეროების მსგავსი დიაგრამით*, რაც გულისხმობს მოცემული რიცხვებიდან ათობითი ნიშნის გამოყოფას (ფოთლები) და დანარჩენი ნიშნების შესაბამისი რიცხვების ზრდის მიხედვით ჩამოწერას თითოჯერ ზევიდან ქვემოთ ვერტიკალურად (ვერტიკალური ღერო). შემდეგ, ვერტიკალურ ღეროზე დატანილი ფიქსირებული პირველი რამოდენიმე საერთო ათობითი ნიშნის გვერდით, ამ რიცხვების ბოლო ათობითი ნიშნების (ფოთლების) ამოწერას რიგრიგობით ჰორიზონტალურად (ჰორიზონტალური ღერო).

მაგალითი 1-სათვის "ვერტიკალურ ღეროს" შეადგენენ ორნიშნა რიცხვები (ერთეულები და ათეულები), "ფოთლებს" – მეათედები, ხოლო მოცემული ორნიშნა რიცხვის შესაბამისი "ფოთლები" – "ჰორიზონტალურ ღეროს".

ვერტიკალური ხაზის მარჯვნივ მიწერილია მეათედები. მთლიანად სურათი არსებითად ჰისტოგრამის სახეს იღებს (თუ სურათს 90 გრადუსით მოვაბრუნებთ საათის ისრის საწინააღმდეგო მიმართულებით). ცხადია, რომ ათობითი ნიშნების დიდი რაოდენობისათვის ფოთლებიანი ღეროების მსგავსი დიაგრამა შეიძლება რთული ასაგები შეიქმნეს. ამგვარი დიაგრამა თვალსაჩინოა და ზოგ სიტუაციაში აადვილებს გამოთვლებს. მაშინ როცა ჰისტოგრამის აგებისას მონაცემებში არსებული ინფორმაცია სრულად არ ინახება, ფოთლებიანი ღეროების მსგავსი დიაგრამა არ კარგავს არც ერთ

მონაცემს და ამას გარდა ჰისტოგრამის ფორმასაც იძენს. ეს დიაგრამა ე. წ. მონაცემთა დაზვერვითი ანალიზის ერთ-ერთი იარაღია. ამ დარგის შემქმნელია ცნობილი ამერიკელი სტატისტიკოსი ჯ. ტიუკი. ეს თეორია გადმოცემულია ჯ. ტიუკის წიგნში: “Exploratory Data Analysis”. Addison-Wesley, 1977.

ფ ო თ ლ ე ბ ი (ბ ო ლ ო ა თ ო ბ ი თ ი ნიშანი)

ვ 14 3 5 1 7

ე 15 1 1 3 5 1

რ 16 5 2 3 9 6 8 4 8

ტ 17 1 4 5 3 1

ი 18 2 8 5 1 7 3 2 6 9 8 7

კ 19 0 19 2 0 7 2 1 3 7 2 2 3 6

ა 20 9 5 8 5 2 2 7 5 0 2 2 3 4 5 8 6 4

ლ 21 9 5 8 6 1 5 3 1 4 6 9 7 7 9 6 7 3 1 7 9 9ჰორიზონტალური ღეროები

უ 22 6 0 7 9 8 8 1 4 7 9 6 5 7 4

რ 23 9 9

ი

ღ

ე

რ

ო

ფოთლებიანი ღეროების მსგავსი დიაგრამა არის მონაცემთა ორგანიზების მეთოდი, რომელიც შედგება მონაცემთა დახარისხების და გრაფიკულად გამოსახვის კომბინაციისაგან. დაგროვილ სიხშირეთა განაწილებასთან შედარებით მისი უპირატესობა იმაშია, რომ იგი უკეთ ინახავს ფაქტიურ მონაცემებს გრაფიკულ წარმოდგენასთან შედარებით.

ფოთლებიანი ღეროების მსგავსი დაგრამა წარმოადგენს მონაცემების განლაგების სქემას (გეგმას), რომელშიც დაკვირვებული თითოეული მონაცემის ერთი ნაწილი გამოიყენება როგორც ვერტიკალური ღერო, ხოლო მეორე ნაწილი კი როგორც ფოთოლი, რათა მოხდეს ჯგუფებისა და კლასების ფორმირება.

მაგალითი 1. ქვემოთ მოყვანილია იმ პაციენტების რაოდენობა, რომლებმაც ამბულატორიული მკურნალობის ცენტრს მიმართეს კარდიოგრამის გადასაღებად 20 დღიანი შერჩევის განმავლობაში. ავაგოთ ამ მონაცემებისათვის ფოთლებიანი ღეროების მსგავსი დიაგრამა.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

ამოხსნა.

ნაბიჯი 1. დავალაგოთ მონაცემები ზრდადობის მიხედვით:
02, 13, 14, 20, 23, 25, 31, 32, 32, 32,
32, 33, 36, 43, 44, 44, 45, 51, 52, 57.

შენიშვნა: მონაცემების ზრდის მიხედვით დალაგება არ არის აუცილებელი, მაგრამ ის მოხერხებულია ფოთლებიანი ღეროების მსგავსი დიაგრამის ასაგებად.

ნაბიჯი 2. განვაცალკევოთ (დავაჯგუფოთ) მონაცემები პირველი ციფრის მიხედვით, ისე როგორც ქვემოთაა ნაჩვენები:
02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57.

ნაბიჯი 3. ავაგოთ დიაგრამა, რისთვისაც მონაცემის პირველი ციფრი გამოვიყენოთ ვერტიკალურ ღეროდ, ხოლო მომდევნო ციფრი კი – ფოთლად. მაგალითად, მონაცემისათვის 32, პირველი ციფრი 3, არის ღერო, ხოლო მეორე ციფრი 2, არის ფოთოლი. მონაცემისათვის 14, 1 არის ღერო, ხოლო 4 კი ფოთოლი და ა. შ. საბოლოოდ, მივიღებთ ქვემოთ მოყვანილ დიაგრამას:

0	2					
1	3	4				
2	0	3	5			
3	1	2	2	2	2	3	6
4	3	4	4	5		
5	1	2	7			

აგებული ფოთლებიანი ღეროების მსგავსი დიაგრამა გვიჩვენებს, რომ მონაცემების ამ განაწილებას პიკი აქვს ცენტრში და მონაცემებში არ არის წყვეტა. 20 დღიდან 7 დღის განმავლობაში იმ პაციენტების რაოდენობა, რომლებმაც გადაიღეს კარდიოგრამა მოთავსებულია 31-დან 36-მდე. დიაგრამა აგრეთვე გვიჩვენებს, რომ ამბულატორიულ ცენტრში პაციენტების რაოდენობა დღის განმაელობაში მერყეობდა მინიმუმ 2 პაციენტიდან მაქსიმუმ 57 პაციენტამდე.

დავალება. სადაზღვევო კომპანიის ხელმძღვანელობას აინტერესებს გამოიკვლიოს გასული ზაფხულის 30 დღის განმავლობაში დიდ ქალაქში მოპარულ ავტომანქანათა განაწილება. დაკვირვებული ნედლი მონაცემები მოყვანილია ქვემოთ:

52	62	51	50	69
58	77	66	53	57
75	56	55	67	73
79	59	68	65	72
57	51	63	69	75
65	53	78	66	55

ავაგოთ ფოთლებიანი ღეროების მსგავსი დიაგრამა შემდეგი კლასების გამოყენებით: 50—54, 55—59, 60—64, 65—69, 70—74, 75—79 და გავაკეთოთ შესაბამისი დასკვნები.

თავი VII

ცენტრალური ტენდენციის საზომები (საშუალო, მედიანა, მოდა)

ჩვენ უკვე განვიხილეთ თუ როგორ შეიძლება ნედლი მონაცემებიდან სასარგებლო ინფორმაციის მოპოვება (მიღება) პირველ ეტაპზე მათი სიხშირული განაწილების სახით ორგანიზებით, ხოლო შემდგომ კი მონაცემების წარმოდგენით სახადასხვა დიაგრამების გამოყენებით. ახლა ჩვენ ვიწყებთ იმ სტატისტიკური მეთოდების შევისწავლას, რომლებიც მონაცემების დაჯამებისა და შესაბამისი დასკვნების გაკეთების საშუალებას მოგვცემენ. ამ მეთოდებს შორის ყველაზე ცნობილია *საშუალოს* მოძებნის მეთოდი.

მაგალითად, თქვენ შეიძლება წაიკითხოთ ასეთი ინფორმაცია (Harper's magazine, June 1995, p.11): საშუალო სიჩქარე, რომლითაც დღის განმავლობაში ავტომობილები გადაკვეთავენ მანჰეტენის ცენტრს არის 53 მილი/საათში ან წუთების საშუალო რიცხვი, რომელსაც ამერიკელი მამები ყოველდღიურად მარტო ატარებენ თავიანთ ოთხი წლის ბავშვებთან არის 42 წუთი.

ამ მაგალითებში, სიტყვა *საშუალო* გაურკვეველია (არაცალსახაა), ვინაიდან საშუალოს მისაღებად შესაძლებელია გამოყენებულ იქნას რამოდენიმე განსხვავებული მეთოდი. პოპულარულ ენაზე, საშუალო ნიშნავს განაწილების ცენტრს ან უფრო ტიპიურ შემთხვევას. საშუალოს საზომებს აგრეთვე უწოდებენ *ცენტრალური ტენდენციის საზომებს* და მათ მიეკუთვნება *საშუალო, მედიანა,* და *მოდა.*

მონაცემთა სიმრავლის საშუალოს ცოდნა არ არის საკმარისი მონაცემთა სიმრავლის სრულად აღსაწერად. მაგალითად, თუ ფეხსაცმლის მაღაზიის მეპატრონემ იცის, რომ მამაკაცის ფეხსაცმლის საშუალო ზომა არის ზომა 40, ის ვერ იქნება ბიზნესში დიდხანს (ვერ იქნება წარმატებული ბიზნესმენი), თუ კი ის შეუკვეთავს მხოლოდ 40 ზომის ფეხსაცმელებს. როგორც ეს მაგალითი გვიჩვენებს, დამატებით, საშუალოს ცოდნასთან ერთად, საჭიროა ვიცოდეთ თუ როგორაა მონაცემთა მნიშვნელობები გაბნეული. უფრო ზუსტად, არის თუ არა მონაცემთა მნიშვნელობები დაჯგუფებული (თავმოყრილი) საშუალოს ირგვლივ, თუ ისინი უფრო თანაბრად არიან გაბნეული მთელი განაწილების გასწვრივ? საზომებს, რომლებიც განსაზღვრავენ მონაცემთა მნიშვნელობების გაფანტულობას (გაბნევას) უწოდებენ *გადახრის საზომებს* ან *გაფანტულობის (გაბნევის) საზომებს.* ამ ტიპის საზომებია: *გაბნევის დიაპაზონი, სტანდარტული გადახრა,* და *საშუალო კვადრატული გადახრა.*

გარდა ამისა, მონაცემთა აღსაწერად კიდევ აუცილებელია სხვა ტიპის საზომები. ამ საზომებს უწოდებენ *მონაცემთა პოზიციის საზომებს.* ისინი გვიჩვენებენ სპეციფიური მნიშვნელობების მდებარეობას მომაცემთა სიმრავლეში ან მათს ფარდობით მდებარეობას მონაცემთა სხვა მნიშვნელობებთან მიმართებაში. ყველაზე გავრცელებული პოზიციის მახასიათებლებია: *პროცენტილები, დეცილები,* და *კვარტილები.* ეს მახასიათებლები ინტენსიურად გამოიყენება ფსიქოლოგიაში და განათლებაში.

სტატისტიკის ნაწილს, რომელიც სწავლობს ცენტრალური ტენდენციის საზომებს, გადახრას, და პოზიციის საზომებს *ტრადიციულ სტატისტი-*

კას უწოდებენ. ქვემოთ ჩვენ აგრეთვე შევისწავლით მეთოდს (ტექნიკას), რომელსაც *მონაცემთა დაზვერვითი ანალიზი* ეწოდება. ეს მეთოდი მოიცავს *ფოთლებიანი ღეროების მსგავს დიაგრამას და ე. წ. ბოქსპლოტს*. აქ სტატისტიკოსი ცდილობს გამოიკვლიოს (დაადგინოს) რას უჩვენებს მონაცემი, განსხვავებით ტრადიციული ტექნიკისაგან, რომელიც გამოიყენება მონაცემების შესახებ სხვადასხვა ჰიპოთეზების (დაშვებების წინადადებების) განსამტკიცებლად (შესამოწმებლად).

იმ შემთხვევაში, როცა პოპულაცია მცირეა, არ არის აუცილებელი გამოვიყენოთ შერჩევა და გამოიყენება სრული პოპულაცია ინფორმაციის მისაღებად. მაგალითად, დავუშვათ, რომ სადაზღვევო კომპანიის მენეჯერს სურს იცოდეს კომპანიის პროდუქტების საშუალო კვირეული გაყიდვები. თუ სადაზღვევო კომპანიას ყავს გამყიდველების (აგენტების) დიდი რიცხვი ქვეყნის მასშტაბით, მას შეუძლია გამოიყენოს შერჩევა და იქიდან გააკეთოს დასკვნები გაყიდვების მთლიან მოცულობაზე. მაგრამ, თუ კომპანიას ყავს მხოლოდ რამოდენიმე გამყიდველი, ვთქვათ, მხოლოდ 87 აგენტი, მან უნდა გამოიყენოს გაყიდული პროდუქტების მთელი რაოდენობა შემთხვევით შერჩეული კვირისათვის და ამდენად გამოიყენოს მთელი პოპულაცია.

ისტორიული შენიშვნა. 1796 წელს ადოლფ ქვითლეტი იკვლევდა ფრანგი ახალწვეულების მახასიათებლებს (სიმაღლე, წონა და სხვა) რათა განესაზღვრა "საშუალო კაცი". ფლორენც ნაითინგეილი იმდენად მოექცა ქვითლეტის შრომების გავლენის ქვეშ, რომ მან დაიწყო ყირიმის ომის მიმდინარეობისას სამხედრო ჰოსპიტალის სამედიცინო ჩანაწერების (დოკუმენტების) შეგროვება და ანალიზი. მისი შრომის საფუძველზე ჰოსპიტალმა დაიწყო სწორი სამედიცინო ჩანაწერების გაკეთება თავიანთი პაციენტების ავადმყოფობის ისტორიაში, რითაც ექიმებს პაციენტების მოვლის და განკურნების უკეთესი საშუალება მიეცა.

საზომებს (მახასიათებლებს), რომელთა საპოვნელად გამოიყენება პოპულაციის ყველა მონაცემის მნიშვნელობები, ეწოდება *პარამეტრები*. ხოლო საზომებს (მახასიათებლებს), რომელთა მისაღებად გამოიყენება შერჩევის მონაცემების მნიშვნელობები, ეწოდება *სტატისტიკები*. შესაბამისად, გაყიდვების საშუალოს, რომელიც მიიღება გაყიდული პროდუქტების შერჩევიდან, ეწოდება *სტატისტიკა*, და გაყიდვების საშუალოს, რომელიც მიიღება მთლიანი პოპულაციიდან, ეწოდება *პარამეტრი*.

სტატისტიკა წარმოადგენს მახასიათებელს ან საზომს, რომელიც მიიღება შერჩევის მონაცემთა მნიშვნელობების გამოყენებით.

პარამეტრი წარმოადგენს მახასიათებელს ან საზომს, რომელიც მიიღება გარკვეული პოპულაციის ყველა მონაცემთა მნიშვნელობების გამოყენებით.

სანამ გადავალთ უშუალოდ მახასიათებლებზე, მოვიყვანოთ **დამრგვალების ზოგადი წესი**: სტატისტიკაში დამრგვალების ძირითადი წესი მდგომარეობს იმაში, რომ სანამ გამოთვლები გრძელდება, დამრგვალება არ უნდა მოხდეს, ვიდრე არ იქნება მიღებული საბოლოო პასუხი. როდესაც დამრგვალება ხდება შუალედურ ეტაპზე, მაშინ იზრდება განსხვავება პასუხსა და ზუსტ შედეგს შორის. მაგრამ სახელმძღვანელოებში და ამოხსნის ნიმუშებში, არ არის პრაქტიკული (მოხერხებული) საშუალედო გამო-

თელებში დავტოვოთ ბევრი ათწილადი ნიშანი. შესაბამისად, თუ მაგალითის მნიშვნელობები შეიცავს ათწილადების საკმაო რაოდენობას (ჩვეულებრივ, მძიმის შემდეგ სამი ან ოთხი ნიშანი), მაშინ იმისათვის რომ მივიღოთ ადექვატური პასუხი, დამრგვალება უნდა მოხდეს გამოთვლების ბოლო ეტაპზე.

საშუალო, რომელიც აგრეთვე ცნობილია როგორც არითმეტიკული საშუალო, მიიღება თუ შევკრიბავთ მონაცემების მნიშვნელობებს და გავყოფთ მნიშვნელობების საერთო რაოდენობაზე. მაგალითად, 3, 2, 6, 5 და 4-ის საშუალო მიიღება შეკრებით 3+2+6+5+4 = 20 და 5-ზე გაყოფით. შესაბამისად, ამ მონაცემების საშუალოა 20/5=4. მონაცემების მნიშვნელობები გამოისახება x-ებით. მონაცემთა ამ სიმრავლეში $x_1=3$, $x_2=2$, $x_3=6$, $x_4=5$ და $x_5=4$. x-ის ყველა მნიშვნელობათა ჯამის აღსანიშნავად გამოიყენება $\sum$ სიმბოლო (ბერძნული დიდი სიგმა ასო), და $\sum x$ აღნიშნავს მონაცემთა სიმრავლის x-ის ყველა მნიშვნელობათა ჯამს.

საშუალო (mean) არის მნიშვნელობათა ჯამი გაყოფილი მნიშვნელობათა საერთო რაოდენობაზე. შერჩევის საშუალოს აღნიშნავენ $\bar{x}$ სიმბოლოთი:

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i\,,$$

სადაც n გვიჩვენებს მნიშვნელობათა საერთო რაოდენობას შერჩევაში.

პოპულაციის შემთხვევაში, საშუალოს აღსანიშნავად, ჩვენ გამოვიყენებთ ბერძნულ ასოს μ (მიუ):

$$\mu=\frac{x_1+x_2+\cdots+x_N}{N}=\frac{1}{N}\sum_{i=1}^{N}x_i\,,$$

სადაც N გვიჩვენებს მნიშვნელობათა საერთო რაოდენობას პოპულაციაში.

მაგალითი 1. ფირმის წლიური საპროცენტო შემოსავალი ბოლო ათი წლის განმავლობაში იყო: 8.1, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.5. როგორია ფირმის საშუალო წლიური შემოსავალი ამ ათი წლის განმავლობაში?

ცხადია, რომ

$$\bar{x}=(8.1+(-6.2)+20.9+(-2.7)+33.6+42.9+24.4+5.2+3.1+30.5)/10=16\,.$$

ე. ი. საშუალო წლიური შემოსავალი არის 16%.

არითმეტიკულ საშუალოს შეიძლება მიეცეს შემდეგი მარტივი ინტერპრეტაცია: არითმეტიკული საშუალო არის შერჩევის ერთ ინდივიდზე ან ობიექტზე მოსული ჯამურ დაკვირვებათა წილი (საშუალო რაოდენობა).

მაგალითი 2. 10 ოჯახში ბავშვთა რაოდენობებია:

3, 2, 1, 1, 2, 3, 3, 1, 2, 2.

მაშინ ერთ ოჯახზე მოსული ბავშვების საშუალო რაოდენობა იქნება

$$(3+2+1+1+2+3+3+1+2+2)/10=2\,.$$

ამრიგად, თუ 10 ოჯახს ჯამში ყავს 20 ბავშვი, მაშინ არითმეტიკული საშუალო უდრის ბავშვთა იმ რაოდენობას, რომელიც მოვიდოდა ერთ ოჯახზე იმ შემთხვევაში რომ ბავშვები თანაბრად გაენაწილებინათ.

დავუშვათ, n მუშის ჯამური ხელფასია P ლარი. მაშინ არითმეტიკული საშუალო ტოლია შეფარდების P/n და იგი არის ერთ მუშაზე მოსული ხელფასი იმ შემთხვევაში, თუ ეს ხელფასი თანაბრად განაწილდება ყოველ მათგანზე. მეორეს მხრივ, თუ ცნობილია, რომ n მუშის საშუალო ხელფასია $\bar{x}$, მაშინ მათი ჯამური ხელფასი იქნება $P=n\bar{x}$ (ანუ $n\bar{x}=\sum_{i=1}^{n} x_i$).

საშუალოს დამრგვალების წესი: საშუალო უნდა დამრგვალდეს ერთი ათობითი ნიშნით მეტზე ვიდრე ნედლი მონაცემებია წარმოდგენილი. მაგალითად, თუ ნედლი მონაცემები მოცემულია მთელი რიცხვებით, მაშინ საშუალო უნდა დამრგვალდეს მეათედებამდე სიზუსტით. თუ ნედლი მონაცემები გამოსახულია მეათედებში, მაშინ საშუალო უნდა დამრგვალდეს მეასედებამდე სიზუსტით. ამ წესის ახსნა იმაში მდგომარეობს, რომ განასხვაოს საშუალო ნედლი მონაცემებისაგან.

დაჯგუფებული მონაცემების შემთხვევაში საშუალოს პოვნის პროცედურა იყენებს კლასების შუაწერტილებს. ამ პროცედურას ჩვენ ქვემოთ მოვიყვანთ.

მაგალითი 3. მოცემულია კვირის განმავლობაში შემთხვევით შერჩეული 20 მორბენალის მიერ გარბენილი მანძილების სიხშირული განაწილების ცხრილი

კლასის საზღვრები	სიხშირე
5.5—10.5	1
10.5—15.5	2
15.5—20.5	3
20.5—25.5	5
25.5—30.5	4
30.5—35.5	3
35.5—40.5	2
	$n=20$

ვიპოვოთ საშუალო.

ამოხსნა.

ნაბიჯი 1. გავაკეთოთ შემდეგი სახის ცხრილი

A კლასები	*B* სიხშირე (f)	*C* შუაწერტილი (x_m)	*D* $f \cdot x_m$
5.5—10.5	1		
10.5—15.5	2		
15.5—20.5	3		
20.5—25.5	5		
25.5—30.5	4		
30.5—35.5	3		
35.5—40.5	2		
	$n=20$		

ნაბიჯი 2. ვიპოვოთ თითოეული კლასის შუაწერტილი და შევიტანოთ C სვეტში:

$$x_m = \frac{5.5+10.5}{2} = 8, \quad \frac{10.5+15.5}{2} = 13, \text{ და ა. შ.}$$

ნაბიჯი 3. ყოველი კლასის შესაბამისი სიხშირე გავამრავლოთ კლასის შუაწერტილზე, როგორც ქვემოთაა ნაჩვენები და შედეგები შევიტანოთ D სვეტში:

$$1 \cdot 8 = 8, \quad 2 \cdot 13 = 26, \text{ და ა. შ.}$$

დასრულებულ ცხრილს ექნება შემდეგი სახე:

A კლასები	B სიხშირე (f)	C შუაწერტილი (x_m)	D $f \cdot x_m$
5.5—10.5	1	8	8
10.5—15.5	2	13	26
15.5—20.5	3	18	54
20.5—25.5	5	23	115
25.5—30.5	4	28	112
30.5—35.5	3	33	99
35.5—40.5	2	38	76
	$n = 20$		$\sum f \cdot x_m = 490$

ნაბიჯი 4. ვიპოვოთ D სვეტის მნიშვნელობათა ჯამი.

ნაბიჯი 5. საშუალოს მისაღებად მიღებული ჯამი გავყოთ n-ზე:

$$\bar{x} = \frac{\sum f \cdot x_m}{n} = \frac{490}{20} = 24.5 \text{ მილი.}$$

მოვიყვანოთ საშუალოს თვისებები:

1. მონაცემთა დაკვირვებული მნიშვნელობების საშუალოდან გადახრების ჯამი ნულის ტოლია:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

მართლაც, ჯამის თვისებებიდან გამომდინარე გვაქვს:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n(\frac{1}{n}\sum_{i=1}^{n} x_i) - n\bar{x} = n\bar{x} - n\bar{x} = 0. \quad \blacksquare$$

ეს თვისება გვეხმარება გავარკვიოთ საშუალოს ფიზიკური შინაარსი. წარმოვიდგინოთ უწონო ერთგვაროვანი ღერძი, რომელზედაც წერტილებში კოორდინატებით $x_1, \quad , x_n$ მოთავსებულია ერთეულოვანი მასები. ამ შემთხვევაში, მოყვანილი თვისებიდან გამომდინარე, $\bar{x}$ წარმოადგენს იმ საყრდენი წერტილის კოორდინატს, რომლის მიმართაც ღერძი იმყოფება წონასწორობაში ანუ $\bar{x}$ წარმოადგენს სიმძიმის ცენტრს. (ამაში ადვილად

დავრწმუნდებით, თუ გავიხსენებთ სიმძიმის ცენტრის განმარტებას: თუ $x_1, \ldots, x_n$ წერტილებში მოთავსებულია $m_1, \ldots, m_n$ მასები, მაშინ *სიმძიმის ცენტრი* არის იმ a კოორდინატის მქონე წერტილი, რომლისთვისაც სრულდება ტოლობა: $\sum_{i=1}^{n}(x_i - a)m_i = 0$. ჩვენს შემთხვევაში: $m_i = 1, i = 1, \ldots, n$ და მაშასადამე, $a = \bar{x}$).

2. მონაცემთა დაკვირვებული მნიშვნელობების საშუალო ის რიცხვია, რომლიდანაც ცალკეული დაკვირვებების გადახრების კვადრატების ჯამი მინიმალურია, ნებისმიერი a რიცხვისათვის:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - a)^2$$

მართლაც, ჯამის თვისებებისა და საშუალოს წინა თვისების თანახმად, გვაქვს:

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - a)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - a)^2 =$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2 \ge \sum_{i=1}^{n}(x_i - \bar{x})^2 . \qquad \blacksquare$$

3. თუ ძველ მონაცემებს შევუცვლით ათვლის წერტილსა და მასშტაბს, მაშინ ახალი მონაცემების საშუალო მიიღება ძველისაგან ათვლის წერტილისა და მასშტაბის ისეთივე შეცვლით.

თუ არსებული $x_1, \ldots, x_n$ მონაცემებიდან გადავალთ ახალ $y_1 = ax_1 + b, \ldots, y_n = ax_n + b$ მონაცემებზე (სადაც a და b ნამდვილ რიცხვთა ნებისმიერი წყვილია), მაშინ ადვილი შესამოწმებელია, რომ

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = a\bar{x} + b .$$

4. ტოლმოცულობიან მონაცემთა წრფივი კომბინაციის საშუალო ტოლია საშუალოთა წრფივი კომბინაციის.

თუ $x_1, \ldots, x_n$ და $y_1, \ldots, y_n$ ტოლი მოცულობის ორი შერჩევაა და გადავალთ ახალ შერჩევაზე $z_1 = ax_1 \pm by_1, \ldots, z_n = ax_n \pm by_n$, მაშინ

$$\bar{z} = a\bar{x} \pm b\bar{y} .$$

5. ორი შერჩევის გაერთიანებით მიღებული შერჩევის საშუალო ცალკეულ საშუალოთა შეწონილი ჯამის ტოლია.

თუ $x_1, \ldots, x_n$ და $y_1, \ldots, y_m$ ორი შერჩევაა, მაშინ $x_1, \ldots, x_n, y_1, \ldots, y_m$ გაერთიანებული შერჩევის საშუალო მოიცემა ფორმულით:

$$\bar{z} = \frac{n}{n+m}\bar{x} + \frac{m}{n+m}\bar{y} .$$

დავალება. შეამოწმეთ საშუალოს მე-3, მე-4 და მე-5 თვისებები.

ეს თვისებები ძალიან მნიშვნელოვანია პრაქტიკული გამოყენების თვალსაზრისით. ვინაიდან მონაცემთა ახალი მასივის გაჩენა არ საჭირო-

ებს საშუალოს თავიდან დათვლას, არამედ ის გამოითვლება უკვე დათვლილი საშუალოებიდან შესაბამისი წესის მიხედვით. მაგალითად, მე-4 თეისებიდან გამომდინარე მონაცემთა ჯამის საშუალოს გამოსათვლელად საკმარისია მონაცემთა თითოეული ჯგუფის საშუალოების შეკრება (მე-4 თეისებაში უნდა ავიღოთ $a = b = 1$).

მიუხედავად იმისა, რომ საშუალო ადვილი გამოსათვლელია, გააჩნია მარტივი ინტერპრეტაცია, მონაცემთა ახალი მასივის გაჩენისას ინარჩუნებს არითმეტიკულ ოპერაციებს და წარმოადგენს მომაცემთა ცენტრის ერთ-ერთ ყველაზე გავრცელებულ რიცხვით მახასიათებელს, აღსანიშნავია, რომ საშუალოს გააჩნია ერთი სერიოზული ნაკლი, იგი ზედმეტად რეაგირებს ექსტრემალური დაკვირვებების უმნიშვნელო რაოდენობაზეც კი. ასეთი დაკვირვებები შეიძლება იყოს ორი ტიპის. პირველ ტიპს მიეკუთვნება ე. წ. "ამოვარდნილი" დაკვირვებები ანუ ის დაკვირვებები (მონაცემები), რომლებიც მკვეთრადაა განსხვავებული მონაცემთა ძირითადი მასივისაგან. ამ შემთხვევაში საშუალო მკვეთრად იქნება წანაცვლებული "ამოვარდნილი" დაკვირვებების მიმართულებით. ექსტრემალურ დაკვირვებებთან საქმე გვაქვს იმ შემთხვევაშიც, როცა მონაცემებით აგებული (გაგლუვებული) ჰისტოგრამა ძლიერ ასიმეტრიულია რომელიმე მიმართულებით, ანუ მას გააჩნია რომელიმე მიმართულებით გაწელილი გრძელი "კუდი", ხოლო საწინააღმდეგო მიმართულებით კი – "მოკლე" კუდი. ამ შემთხვევაშიც საშუალო წანაცვლებულია გრძელი "კუდის" მიმართულებით და მონაცემთა ძირითადი მასა მის ერთ მხარეს აღმოჩნდება. ასეთი მონაცემებისათვის საშუალო არ წარმოადგენს ცენტრალური ტენდენციის საიმედო საზომს.

მაგალითი 4. ვთქვათ, მუსიკალურ ნაწარმოებთა კატალოგიდან ამოვარჩიეთ ხუთი ნაწარმოები, რომელთა ხანგრძლივობებია (წუთებში): 37, 46, 40, 57 და 50. მაშინ $\bar{x} = 46$. თუ ახლა მეხუთე ნაწარმოებს, რომლის ხანგრძლივობა იყო 50 წუთი, შევცვლით სხვა ნაწარმოებით, რომლის ხანგრძლივობაა 200 წუთი, მაშინ საშუალო იქნება $\bar{x} = 76$. ანუ საშუალომ გადაინაცვლა ექსტრემალური მნიშვნელობისაკენ, ყველა დანარჩენი მონაცემი საგრძნობლად ნაკლებია მასზე. ამ შემთხვევაში $\bar{x}$ არ არის განლაგების (ლოკალიზაციის) მდგრადი მახასიათებელი. თუ მეხუთე მონაცემს შევცვლით 400-ით, მაშინ $\bar{x} = 116$.

მკვეთრად ასიმეტრიული (მარჯვენივ) ფორმა აქვს მოსახლეობის შემოსავლების სიხშირულ განაწილებებს, ამიტომ $\bar{x}$ მკვეთრად წანაცვლებული აღმოჩნდება მარჯვენივ (ექსტრემალურად დიდი შემოსავლების მიმართულებით) და გასაშუალოებული შემოსავლების გამოყენება ცხოვრების დონის აღსაწერად აზრს მოკლებულია.

მაგალითი 5. ქვევით მოყვანილია ავსტრალიის ერთ-ერთი უნივერსიტეტის თანამშრომელთა წლიური შემოსავლები (1000 დოლარებში): 28, 109, 26, 32, 30, 26, 29. მაშინ საშუალო იქნება $\bar{x} = 40$. როგორც ვხედავთ, ყველა მონაცემი გარდა ერთისა (109) 40-ზე ნაკლებია და ამდენად, საშუალო არ აღწერს ტიპობრივ შემოსავალს.

მედიანა (median) განეკუთვნება მონაცემთა მდებარეობის ცენტრალური ტენდენციის მდგრადი საზომების ჯგუფს (განსხვავებით საშუალოსა-

გან). მონაცემთა რაიმე რიცხვითი მახასიათებლის *მდგრადობა* ნიშნავს, რომ დაკვირვებათა მცირე რაოდენობის ცვლილებას შეზღუდული გავლენა აქვს აღნიშნულ მახასიათებელზე, მიუხედავად იმისა, თუ როგორია ამ ცვლილების სიდიდე. სწორედ ამ თვისების გამო მედიანა, საშუალოს შემდეგ, წარმოადგენს ცენტრალური ტენდენციის ყველაზე გავრცელებულ საზომს.

როგორც ცნობილია, გეომეტრიაში მედიანა არის სამკუთხედის წვეროს მოპირდაპირე გვერდის შუაწერტილთან შემაერთებელი მონაკვეთი, ხოლო ფართე გაგებით მედიანა არის "შუა"-ს სინონიმი. სტატისტიკაშიც, მედიანა განიმარტება, როგორც მონაცემთა ვარიაციული მწკრივის შუაში მდგომი მნიშვნელობა იმ აზრით, რომ მასზე ნაკლები რიცხვითი მნიშვნელობის მონაცემთა რაოდენობა მასზე დიდი რიცხვითი მნიშვნელობის მონაცემთა რაოდენობის ტოლია. გასაგებია, რომ თუ დაკვირვებულ მონაცემთა რაოდენობა კენტია, მაშინ მედიანა მართლაც არის ვარიაციული მწკრივის შუა ელემენტი. თუ დაკვირვებათა რაოდენობა ლუწია $n=2k$, მაშინ მედიანის განმარტებას დააკმაყოფილებს ნებისმიერი რიცხვი (x_k, x_{k+1}) რიცხვითი შუალედიდან. ე. ი. ამ შემთხვევაში მედიანა ცალსახად არ განიმარტება. შესაბამისად, ამ გაუგებრობის თავიდან აცილების მიზნით მიღებულია შემდეგი შეთანხმება: როცა $n=2k$, მაშინ მედიანას უწოდებენ ვარიაციული მწკრივის შუაში მდგომი ორი ელემენტის საშუალო არითმეტიკულს.

დავუშვათ, რომ x_1, $\ldots$ $,x_n$ შერჩევის ელემენტებია, ხოლო $x_{(1)}$, $\ldots$ $,x_{(n)}$ კი – შესაბამისი ვარიაციული მწკრივია. მედიანა აღინიშნება $\tilde{x}$ სიმბოლოთი და იგი აიგება შემდეგი წესით:

$$\tilde{x} = x_{((n+1)/2)}, \text{ როცა } n \text{ კენტია, და}$$

$$\tilde{x} = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}, \text{ როცა } n \text{ ლუწია.}$$

ამრიგად, მედიანის მოსაძებნად საჭიროა შემდეგი ნაბიჯები: I. დავალაგოთ ნედლი მონაცემები ზრდადობის მიხედვით (ავაგოთ ვარიაციული მწკრივი) და II. შევარჩიოთ (ან ავაგოთ) შუაწერტილი. როგორც ვხედავთ, მედიანის მოსაძებნად საჭიროა ვარიაციული მწკრივის აგება (განსხვავებით საშუალოს მოძებნის შემთხვევისაგან), რაც დაკვირვებათა დიდი რაოდენობის დროს გარკვეულ სიძნელეებთანაა დაკავშირებული.

მაგალითი 1. ქვემოთ მოყვანილია შვიდი სამხედრო ახალწვეულის წონა (ფუნტებში, 1 ფუნტი = 453,6 გრამი):

180, 201, 220, 191, 219, 209, 186.

იპოვეთ მედიანა.

ამოხსნა.

ნაბიჯი 1. დავალაგოთ მონაცემები ზრდადობის მიხედვით:

180, 186, 191, 201, 209, 219, 220.

ნაბიჯი 2. შევარჩიოთ შუაწერტილი. ვინაიდან $n=7$, ამიტომ მედიანა იქნება $\tilde{x} = x_{((7+1)/2)} = x_{(4)} = 201$.

მაგალითი 2. აშშ-ში რვა წლის განმავლობაში მომხდარი ტორნადოების რაოდენობებია:

$$684,\ 764,\ 656,\ 702,\ 856,\ 1133,\ 1132,\ 1303.$$

იპოვეთ მედიანა.

ამოხსნა. ვარიაციული მწკრივი იქნება:

$$656,\ 684,\ 702,\ 764,\ 856,\ 1132,\ 1133,\ 1303.$$

ვინაიდან შუაწერტილი მდებარეობს 764-სა და 856-ს შორის, ამიტომ მედიანას ვპოულობთ ამ ორი მნიშვნელობის შეკრებითა და 2-ზე გაყოფით:

$$\tilde{x}=\frac{x_{(8/2)}+x_{((8/2)+1)}}{2}=\frac{x_{(4)}+x_{(5)}}{2}=\frac{764+856}{2}=810,$$

ე. ი. ტორნადოების რაოდენობის მედიანაა 810.

მოვიყვანოთ მედიანის თვისებები:

I. შერჩევის მედიანა წარმოადგენს შემდეგი განტოლების ამოხსნას:

$$\sum_{i=1}^{n} sign(x_i-\tilde{x})=0,$$

სადაც $sign x=1$, თუ $x>0$; $sign 0=0$ და $sign x=-1$, თუ $x<0$.

ე. ი. დაკვირვებების მედიანისაგან გადახრის ნიშნების ჯამი ნულის ტოლია (გავიხსენოთ, რომ საშუალოს შემთხვევაში თვითონ გადახრების ჯამი იყო ნულის ტოლი).

II. მედიანა წარმოადგენს ისეთ სიდიდეს, რომლიდანაც ცალკეული დაკვირვებების გადახრების აბსოლუტური სიდიდეების ჯამი მინიმალურია, ანუ ნებისმიერი a რიცხვისათვის ადგილი აქვს უტოლობას:

$$\sum_{i=1}^{n}|x_i-\tilde{x}|\le\sum_{i=1}^{n}|x_i-a|$$

(შეადარეთ საშუალოს მე-2 თვისებას).

III. თუ ძველ მონაცემებს შევუცვლით ათვლის წერტილსა და მასშტაბს, მაშინ ახალი მონაცემების მედიანა მიიღება ძველისაგან ათვლის წერტილისა და მასშტაბის ისეთივე შეცვლით ანუ თუ არსებული x_1, $,x_n$ მონაცემებიდან გადავალთ ახალ $z_1=ax_1+b$, $,z_n=ax_n+b$ მონაცემებზე (სადაც a და b ნამდვილ რიცხვთა ნებისმიერი წყვილია), მაშინ:

$$\tilde{z}=a\tilde{x}+b.$$

აღსანიშნავია, რომ მედიანისათვის ძალაში აღარაა საშუალოს მე-4 და მე-5 თვისებების ანალოგიური თვისება.

დავალება. შეამოწმეთ მედიანის I—III თვისებები.

მაგალითი 3. დავუბრუნდეთ საშუალოზე მე-5 მაგალითის მონაცემებს და წარმოვადგინოთ ისინი ვარიაციული მწკრივის სახით:

$$x_{(1)}=26,\ x_{(2)}=26,\ x_{(3)}=28,\ x_{(4)}=29,\ x_{(5)}=30,\ x_{(6)}=32,\ x_{(7)}=109.$$

მაშინ, ვინაიდან $n=8$ კენტია, მედიანა იქნება $\tilde{x}=x_{((7+1)/2)}=x_{(4)}=29$. გავიხსენოთ, რომ საშუალო იყო $\bar{x}=40$. საიდანაც ჩანს, რომ მედიანა ამ მაგალითში უკეთ აღწერს ტიპობრივ შემოსავალს, ვიდრე საშუალო.

მოვახდინოთ ახლა დაკვირვებათა ცვლილება და შევხედოთ როგორ იქცევა საშუალო და მედიანა. ვთქვათ, დაკვირვება $x_{(7)}=109$ შევცვალეთ

200-ით. მაშინ შეცვლილი მონაცემებისათვის მედიანა დარჩება იგივე $\bar{x}=29$, ხოლო საშუალო წაინაცვლებს დიდი დაკვირვების (200-ის) მიმართულებით $\bar{x}=53$. ამდენად, მედიანა მდებარეობის უფრო მდგრადი მახასიათებელია, ვიდრე საშუალო.

ვთქვათ, ახლა დაკვირვება $x_{(2)}=26$ შევცვალეთ 300-ით. მაშინ ახალი მონაცემების ვარიაციული მწკრივი იქნება:

$$x_{(1)}=26,\ x_{(2)}=28,\ x_{(3)}=29,\ x_{(4)}=30,\ x_{(5)}=32,\ x_{(6)}=109,\ x_{(7)}=300.$$

ამ შემთხვევაში მედიანა იცვლება მხოლოდ ერთი ერთეულით $\bar{x}=x_{(4)}=30$, მაშინ როდესაც საშუალო კვლავ მკვეთრად იზრდება $\bar{x}=78.85$. რაც ისევ მეტყველებს მედიანის, როგორც ცენტრის მახასიათებლის მდგრადობაზე.

მედიანას გააჩნია კიდევ ერთი სასარგებლო თვისება. იმ შემთხვევაში, როცა შერჩევის ინდივიდები დალაგებულია შესასწავლი ნიშნის (მაჩვენებლის) სიდიდის მიხედვით (მაგალითად, ადამიანები – სიმაღლის მიხედვით) მედიანის გამოსათვლელად საჭიროა მხოლოდ შუაში მოქცეული ინდივიდის (თუ n კენტია) ან ორი შუა ინდივიდის (თუ n ლუწია) გაზომვა.

მედიანა განსაკუთრებით მნიშვნელოვან ინფორმაციას იძლევა მაშინ, როცა მოცემულ შერჩევაში დაკვირვებების შედარებით მცირე რაოდენობა მნიშვნელოვნად განსხვავდება დანარჩენი სიმრავლისაგან. მაგალითად, მოსახლეობის გარკვეული ჯგუფის შემოსავლების განაწილების მონაცემებში მედიანა მოსახლეობის ამ ჯგუფის ცხოვრების დონის უფრო აზრიან სურათს იძლევა, ვიდრე საშუალო (რომელზედაც დიდ გავლენას ახდენს ერთი ინდივიდის შემოსავალი).

თუმცა ეს არ უნდა გავიგოთ ისე, რომ მედიანას ყოველთვის უნდა მიენიჭოს უპირატესობა საშუალოსთან შედარებით. საშუალო გამოითვლება ყველა მონაცემის მეშვეობით (ყველა მონაცემის სიდიდის გათვალისწინებით) და შეიცავს მეტ ინფორმაციას, ვიდრე მედიანა, მისი მგრძნობიარობა კი ხშირ შემთხვევაში მას მოცემული შერჩევის კარგ მახასიათებლად აქცევს. საშუალო შეიძლება გამოყენებულ იქნას როგორც საზომი, რომელიც ასახავს, თუ საშუალოდ რა სიდიდისაა შერჩევის ელემენტები. ორი სხვადასხვა ჯგუფის ელემენტების შედარების მიზნით ჩვეულებრივ, პირველ რიგში, ადარებენ მათ არითმეტიკულ საშუალოებს.

მაგალითი 4. მოცემულია ორი A და B ფირმის შემოსავლიანობა (პროცენტებში) ბოლო 10 წლის განმავლობაში:

ფირმა A: 8.3, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.5;

ფირმა B: 12.1, 6.4, 12.2, 27.8, 25.7, 18.2, 10.7, -1.3, 11.4, -2.8.

თუ გამოვთვლით საშუალოებს A და B ფირმების შემოსავლიანობებისათვის, მივიღებთ:

$$\bar{x}_A=16\% \text{ და } \bar{x}_B=12\%.$$

ამ ორი საშუალოს შედარება გვაძლევს საფუძველს დავასკვნათ, რომ A ფირმის შემოსავლიანობა საშუალოდ უფრო მაღალია, ვიდრე B ფირმისა (შემდგომი ანალიზი კი წარმოებს სხვა მახასიათებლების საშუალებით).

ცენტრალური ტენდენციის მესამე საზომს წარმოადგენს **მოდა** (**mode**).

მოდა წარმოადგენს იმ მნიშვნელობას შერჩევიდან, რომელიც ყველაზე ხშირად გვხვდება. ზოგჯერ ამბობენ, რომ მოდა არის ყველაზე ტიპიური შემთხვევა. მონაცემთა სიმრავლეს შეიძლება ჰქონდეს ერთზე მეტი მოდა ან საერთოდ არ ჰქონდეს მოდა. ვნახოთ მაგალითების მიხედვით.

მაგალითი 1. ქვემოთ მოყვანილია აშშ-ის მრავალჯერადი გამოყენების კოსმოსური ხომალდის ფრენის ხანგრძლივობა (დღეებში) 1992—94 წლებში:

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11.

ვიპოვოთ მოდა.

ამოხსნა. მოდის მოსაძებნად სასარგებლოა (მოხერხებულია) მონაცემების დალაგება ზრდის მიხედვით, თუმცა ეს არ არის აუცილებელი:

6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14.

ვინაიდან 8-დღიანი ფრენა მოხდა 5-ჯერ (8-ის სიხშირე უფრო დიდია ვიდრე ნებისმიერი სხვა რიცხვის), ამიტომ ამ მონაცემების მოდა არის 8.

მაგალითი 2. ვიპოვოთ სამხრეთ-დასავლეთ პენსილვანიის 10 შერჩეულ საგრაფოში ქვანახშირის მოპოვებაში დასაქმებული მუშათა რიცხვის მოდა. ეს მონაცემებია:

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752.

ამოხსნა. ვინაიდან ყველა მნიშვნელობა გვხვდება მხოლოდ ერთხელ, აქ მოდა არ არსებობს.

შენიშვნა: არ უნდა ვთქვათ, რომ მოდა არის ნული. ეს იქნება არაკორექტული, ვინაიდან მონაცემთა ზოგიერთი სიმრავლისათვის (მაგალითად, როგორიცაა ტემპერატურა) ნული შეიძლება აღმოჩნდეს ფაქტიური (რეალური) მნიშვნელობა.

მაგალითი 3. შემოწმებულ (გაზომილ) იქნა 11 სხვადასხვა ავტომობილის სამუხრუჭე მანძილის სიდიდე, როცა ისინი მოძრაობდნენ 15 მილი/საათში სიჩქარით. ვიპოვოთ მოდა, თუ მიღებული იყო შემდეგი მონაცემები:

15, 18, 18, 18, 20, 22, 24, 24, 24, 26, 26.

ამოხსნა. ვინაიდან 18 და 24 ორივე გვხვდება მაქსიმალურ რიცხვჯერ – 3-ჯერ, ამიტომ მოდა არის 18 და 24. ამ შემთხვევაში ამბობენ, რომ მონაცემთა სიმრავლე არის *ბიმოდალური.*

შეიძლება ისე მოხდეს, რომ შერჩევაში ისეთი მნიშვნელობა, რომელიც ყველაზე ხშირად გვხვდება არ აღმოჩნდეს. ამ შემთხვევაში შეიძლება ვილაპარაკოთ ე. წ. *მოდალურ ინტერვალზე,* რომელიც ასე განიმარტება: თუ აგებული გვაქვს ასე თუ ისე მისაღები ჰისტოგრამა (ანუ ჰისტოგრამა, რომელიც ადეკვატურად აღწერს მონაცემთა სიხშირულ განაწილებას), მაშინ **მოდალური ინტერვალი** ეწოდება იმ ინტერვალს, რომელსაც შეესაბამება ჰისტოგრამის მაქსიმალური მნიშვნელობა, ანუ მოდალური ინტერვალი არის ის ინტერვალი, რომელშიც მოხვდა მონაცემთა ყველაზე დიდი რაოდენობა. გარკვეულობისათვის, მოდას უწოდებენ მოდალური ინტერვალის შუაწერტილს. თუ შერჩევის მოცულობა დიდია, მაშინ სწორედ მონაცემთა დაჯგუფებას და მოდალური ინტერვალის მოძებნას ამჯობინებენ.

გასაგებია, რომ მოდის (მოდალური ინტერვალის) მოძებნა საკმაოდ რთულია. იმ ინტერვალის მითითება, სადაც მოხვდა მონაცემთა მაქსიმალური რაოდენობა, დამოკიდებულია იმაზე, თუ როგორ დავაჯგუფებთ მონაცემებს, კერძოდ, იმაზე თუ რა სიგრძის ინტერვალებად დავყოფთ ინტერვალს (x_{min}, x_{max}). შესაბამისად, მოდალური ინტერვალის დადგენას თან ახლავს ყველა ის სიძნელე, რაც დაკავშირებულია ჰისტოგრამის აგებასთან.

მაგალითი 4. ვიპოვოთ მოდალური კლასი კვირის განმაელობაში შემთხვევით შერჩეული 20 მორბენალის მიერ გარბენილი მანძილების სიხშირული განაწილების ცხრილის მიხედვით:

კლასის საზღვრები	სიხშირე
5.5—10.5	1
10.5—15.5	2
15.5—20.5	3
20.5—25.5	5
25.5—30.5	4
30.5—35.5	3
35.5—40.5	2
	$n = 20$

ამოხსნა. მოდალური კლასი იქნება კლასი 20.5—25.5, ვინაიდან მას აქვს ყველაზე დიდი სიხშირე. როგორც ზემოთ იყო აღნიშნული, ზოგჯერ მოდას უწოდებენ მოდალური კლასის შუაწერტილს. შესაბამისად, მოდად შეიძლება მიჩნეულ იქნეს (20.5+25.5)/2=23 მილის გარბენა კვირაში.

შერჩევის მოდის ინტერპრეტაცია წააგავს ყოველდღიურ ცხოვრებაში სიტყვა "მოდის" მნიშვნელობას. მაგრამ, როგორც ის, რაც "მოდაშია" (ანუ ყველაზე ხშირია, გავრცელებულია) შეიძლება სულაც არ აღწერდეს საერთო სტილს, ასევე შერჩევითი მოდაც ახასიათებს დაკვირვებათა მხოლოდ მცირე ნაწილს. შერჩევის მოდა იძლევა მნიშვნელოვან ინფორმაციას როგორც სხვადასხვა საქონლის გამომშვები ფირმისათვის, ისე ამ საქონლის მომხმარებლებისათვის -- მაღაზიის მფლობელებისათვის; კონსტრუქტორებისათვის, რომლებიც აწოდებენ საქონელს სპეციფიკურ ბაზრებს. მაგალითად, ტანსაცმლის გამომშვებმა ფირმამ (შესაბამისად, ტანსაცმლის მაღაზიის მფლობელმა) უნდა იცოდეს კოსტუმების ყველაზე გავრცელებული ზომები, რათა გაითვალისწინოს ეს ფაქტი საქონლის გამოშვებისას (შესაბამისად, მომარაგებისას); საათების მწარმოებელმა უნდა იცოდეს, თუ რა კლასის საათები იყიდება ყველაზე ხშირად, იმ მიზნით, რომ გაზარდოს ამ კლასის პროდუქციის წარმოება.

მოდა არის ერთადერთი ცენტრალური ტენდენციის საზომებიდან, რომლის გამოყენებაც შეიძლება სახელდებითი ან კატეგორიული მონაცემებისათვის, ვინაიდან, ცხადია რომ კატეგორიზებული მონაცემებისათვის საშუალოსა და მედიანის გამოყენება უაზრობაა, ხოლო მოდა კი იძლევა სასარგებლო ინფორმაციას ყველაზე ტიპიური შემთხვევის გამოსავლენად.

მაგალითი 5. სტუდენტებში ჩატარებულ იქნა გამოკითხვა იმის შესახებ თუ რა სფეროში აპირებდნენ ისინი მოღვაწეობას სწავლის დასრულების შემდეგ. მიღებულ იქნა სიხშირეთა განაწილების შემდეგი ცხრილი:

მოღვაწეობის სფერო	სიხშირე
ბიზნესი	1425
ჰუმანიტარული მეცნიერებები	878
კომპიუტერული მეცნიერებები	632
განათლება	471
სახელმწიფო სტრუქტურები	95

ამოხსნა. ვინაიდან ყველაზე მაღალი სიხშირის მქონე კატეგორიაა ბიზნესი, ამიტომ ყველაზე ტიპიური დასაქმების სფეროა ბიზნესში დასაქმება.

მონაცემთა განლაგების ცენტრის სამივე მახასიათებელი: საშუალო, მედიანა და მოდა წარმოადგენს პოპულაციის შესაბამისი მახასიათებლების: საშუალოს, მედიანისა და მოდის შერჩევით ანალოგებს. ალბათობის თეორიაში მტკიცდება, რომ შერჩევის მოცულობის ზრდისას შერჩევის რიცხვითი მახასიათებლები იკრიბება (უახლოვდება) პოპულაციის ანალოგიურ მახასიათებლებს. აქედან გამომდინარე, შერჩევითი მახასიათებლები შეიძლება ჩაითვალოს პოპულაციის მახასიათებლების შეფასებად და გამოყენებულ იქნას ამ უკანასკნელის შესახებ სტატისტიკური დასკვნების მასალებად.

თავი VIII
მონაცემთა გაფანტულობის საზომები

სტატისტიკაში, იმისათვის რომ ზუსტად (ადექვატურად) აღვწეროთ მონაცემთა სიმრავლე, სტატისტიკოსმა უნდა იცოდეს უფრო მეტი, ვიდრე ცენტრალური ტენდენციის საზომებია. ჩვენ შევისწავლეთ მონაცემთა ცენტრალური ტენდენციის (ანუ მონაცემთა ცენტრის მდებარეობის) სამი რიცხვითი მახასიათებელი: შერჩევითი საშუალო, შერჩევითი მედიანა და შერჩევითი მოდა. გასაგებია, რომ ეს სამი მახასიათებელი სრულად ვერ წარმოადგენს მონაცემთა სიმრავლის თვისებებს. მეორე ეტაპზე, ბუნებრივად იბადება კითხვა რამდენად ტიპიურია მონაცემთა ცენტრის მახასიათებელი კერძოდ, საშუალო, მონაცემთა მთელი სიმრავლისათვის? როგორია მონაცემთა გაფანტულობა საერთოდ და კონკრეტულად კი საშუალოს მიმართ? როგორია მონაცემთა ცვალებადობის ხარისხი და რა სიდიდეებით შეიძლება მისი დახასიათება?

მაგალითი 1. დავუშვათ, რომ მოცემულია მონაცემთა ორი მწკრივი:

A	8	9	10	10	13
B	1	5	10	16	18.

გამოვთავალოთ თითოეულის საშუალო. გვაქვს:

$$\bar{x}_A=(8+9+10+10+13)/5=10 \text{ და } \bar{x}_B=(1+5+10+16+18)/5=10.$$

აღმოჩნდა, რომ ორივე მონაცემს აქვს ერთი და იგივე საშუალო, მაგრამ ცხადია, რომ B მწკრივის ელემენტთა ცვალებადობა უფრო ძლიერია, ვიდრე A მწკრივის ელემენტების. A მწკრივის ელემენტები უფრო მჭიდროდ არიან თავმოყრილი (კონცენტრირებული) საშუალოს ირგვლივ, ვიდრე B მწკრივის ელემენტები.

ხშირ შემთხვევაში, თვალით შემჩნევა იმისა, თუ რომელი მწკრივის ცვალებადობა უფრო დიდია, საკმაოდ ძნელია (მაგალითად, მონაცემთა დიდი რაოდენობის დროს). ამიტომ აუცილებელია ცვალებადობის ანუ გაფანტულობის საზომების შემოღება, ე. ი. ისეთი რიცხვითი მახასიათებლების შემოღება, რომელიც საშუალებას მოგვცემს შევაფასოთ მონაცემთა გაბნევის ხარისხი.

მონაცემთა გაფანტულობის ერთ-ერთი უმარტივესი რიცხვითი მახასიათებელია **გაბნევის დიაპაზონი (range)**, რომელიც წარმოადგენს შერჩევის მაქსიმალური (უდიდესი) და მინიმალური (უმცირესი) მნიშვნელობების სხვაობას. სხვა სიტყვებით – გაბნევის დიაპაზონი წარმოადგენს ვარიაციული მწკრივის უკანასკნელი და პირველი წევრების სხვაობას, თუმცა გაბნევის დიაპაზონის დასადგენად არ არის აუცილებელი (უბრალოდ მოხერხებულია) ვარიაციული მწკრივის შექმნა (მონაცემების დალაგება ზრდადობის მიხედვით). გაბნევის დიაპაზონი აღინიშნება R ასოთი. შესაბამისად, თუ $x_1,\dots,x_n$ მოცემული n მოცულობის მქონე შერჩევაა (ხოლო $x_{(1)},\dots,x_{(n)}$ კი – შესაბამისი ვარიაციული მწკრივი), მაშინ:

$$R=\max_{1\le i\le n} x_i-\min_{1\le i\le n} x_i=x_{(n)}-x_{(1)},$$

სადაც $x_{(n)} = \max_{1\le i\le n} x_i$ $(x_{(1)} = \min_{1\le i\le n} x_i)$ – მაქსიმალური (შესაბამისად, მინიმალური) სიდიდის მონაცემია.

ზემოთ მოყვანილ მაგალითში მონაცემთა A მწკრივის გაბნევის დიაპაზონი არის $R_A = 13 - 5 = 8$, ხოლო B მწკრივის კი – $R_B = 18 - 1 = 17$.

მონაცემთა გაბნევის დიაპაზონი გამოთვლისა და ინტერპრეტაციის სიმარტივის გამო ფართოდ გამოიყენება ისეთი ამოცანების გადაწყვეტისას, სადაც გამოთვლის სისწრაფესა და მარტივ შინაარს გადამწყვეტი მნიშვნელობა ენიჭება. მაგალითად, ხარისხის კონტროლის სფეროში. სხვა შემთხვევებში კი ამ მახასიათებლის მიმართ წარმოიშობა სერიოზული პრობლემები. გაბნევის დიაპაზონზე დიდ გავლენას ახდენს იშვიათი ან ე. წ. ამოვარდნილი დაკვირვებები. მაგალითად, დავუშვათ, რომ ჩვენ რომელიმე რეგიონში ვაწარმოებთ შერჩევას მამაკაცთა სიმაღლის განაწილების შერჩევის მიზნით. ვთქვათ, ამ რეგიონში ცხოვრობს მამაკაცი, რომლის სიმაღლეა 210 სმ. ცხადია, რომ თუ შერჩევაში შემთხვევით აღმოჩნდება ეს მამაკაცი (რაც იშვიათი ხდომილებაა), გაბნევის დიაპაზონი უფრო დიდი იქნება, ვიდრე იმ შემთხვევაში, როცა მის ნაცვლად შერჩევაში მოხვდება სხვა ტიპიური სიმაღლის მქონე მამაკაცი. ანალოგიური სურათი შეიძლება გვქონდეს მოსახლეობის შემოსავლების განაწილების შესწავლისას, როდესაც განსაკუთრებით დიდი შემოსავლების მქონე ადამიანი აღმოჩნდება (იშვიათი ხდომილება) ჩვენს შერჩევაში.

შემდეგი მაგალითი გვიჩვენებს, რომ ერთ ექსტრემალურად მაღალ (სხვა შემთხვევაში შეიძლება ექსტრემალურად დაბალ) მონაცემის მნიშვნელობას შეუძლია მნიშვნელოვანი გავლენა მოახდინოს გაბნევის დიაპაზონზე.

მაგალითი 2. ქვემოთ მოყვანილია ერთ-ერთი წარმოების მუშაკთა წლიური შემოსავალი დოლარებში.

თანამდებობა	შემოსავალი
მეპატრონე	100 000
მენეჯერი	40 000
გაყიდვების აგენტი	30 000
I თანრიგის მუშა	25 000
II თანრიგის მუშა	18 000
III თანრიგის მუშა	15 000

იპოვეთ გაბნევის დიაპაზონი.

ამოხსნა. გაბნევის დიაპაზონი $R = 100000 - 15000 = 85000$.

შერჩევის გაბნევის დიაპაზონის, როგორც მონაცემთა გაფანტულობის საზომის მთავარი ნაკლი ისაა, რომ ის არ შეიცავს ინფორმაციას იმის შესახებ, თუ როგორაა გაბნეული დანარჩენი (საშუალედო) მონაცემები მაქსიმალურ და მინიმალურ მნიშვნელობებს შორის. შესაძლოა ეს მონაცემები გროვდებოდნენ საშუალო მნიშვნელობასთან, ან ექსტრემალური მნიშვნელობების მახლობლობაში, ან თანაბრად იყვნენ განაწილებული მათ შორის და სხვა. საშუალედო მონაცემების განლაგებისა და შესაბამისად, მა-

თი გაფანტულობის შესახებ ინფორმაციას იძლევა ე. წ. **პროცენტილები (percentiles).**

პროცენტილები ერთდროულად წარმოადგენს მონაცემთა განლაგებისა და მათი გაფანტულობის საზომს. ვთქვათ, P რაიმე რიცხვია, მოთავსებული 0-სა და 100-ს შორის $0 < p < 100$.

მონაცემთა სიმრაველის P *რიგის პროცენტილი* (უბრალოდ P*-პროცენტილი*) ეწოდება ისეთ $\tilde{x}_P$ მნიშვნელობას (სიდიდეს), რომელსაც გააჩნია შემდეგი თვისება: მონაცემთა არაუმეტეს P%-ისა ნაკლებია $\tilde{x}_P$-ზე და არაუმეტეს (100-P)%-ისა მეტია $\tilde{x}_P$-ზე.

მიღებულია P რიცხვის წარმოდგენა $P = 100 \cdot \alpha$ სახით, სადაც $0 < \alpha < 1$. მაგალითად, თუ P=10-ს, მაშინ $\alpha = 0.1$, თუ P=25-ს, მაშინ $\alpha = 0.25$. ამ შემთხვევაში P*-პროცენტილის* განსაზღვრება ასე გამოითქმის: $\tilde{x}_P$ ისეთი რიცხვია, რომ ერთდროულად უნდა სრულდებოდეს შემდეგი ორი უტოლობა:

$$N\{x_i : x_i < \tilde{x}_P\} \le \alpha n,$$

$$N\{x_i : x_i > \tilde{x}_P\} \le (1-\alpha)n,$$

სადაც $N\{x_i : \cdots\}$ სიმბოლოთი აღნიშნულია იმ დაკვირვებების რაოდენობა, რომელთათვისაც სრულდება ორი წერტილის შემდეგ მოთავსებული თვისება. $P = 100 \cdot \alpha$ პროცენტილი არის α-კვანტილის (F განაწილების α-კვანტილი x_α განიმარტება შემდეგნაირად: $x_\alpha = \min\{x : F(x) \ge \alpha\}$) შერჩევითი ანალოგი.

სიმარტივისათვის, ქვემოთ ჩვენ ვიგულისხმებთ, რომ ვარიაციული მწკრივი მკაცრად ზრდადი მიმდევრობაა: $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$.

ადვილი მისახვედრია, რომ მედიანა წარმოადგენს 50-პროცენტილს. მართლაც, გავიხსენოთ მედიანის განმარტება:

$\tilde{x} = x_{((n+1)/2)}$, როცა n კენტია, და

$\tilde{x} = \dfrac{x_{(n/2)} + x_{((n/2)+1)}}{2}$, როცა n ლუწია.

შესაბამისად, თუ n კენტია, მაშინ $N\{x_i : x_i < \tilde{x}\} = N\{x_i : x_i > \tilde{x}\} = \dfrac{n-1}{2} < 0.5n$ (გავიხსენოთ, რომ $\alpha = 0.5$, თუ $P = 50$), ხოლო თუ n ლუწია, მაშინ $N\{x_i : x_i < \tilde{x}\} = N\{x_i : x_i > \tilde{x}\} = 0.5n$.

თუ $P = 100 \cdot \alpha$, მაშინ მედიანის ანალოგიურად P-პროცენტილის ($\tilde{x}_P$-ს) გამოსათვლელი გამოსახულება იქნება:

$\tilde{x}_P = x_{([n\alpha]+1)}$, როცა $n\alpha$ არ არის მთელი რიცხვი, და

$\tilde{x}_P = \dfrac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2}$, როცა $n\alpha$ მთელი რიცხვია.

მაგალითი 3. თუ მოცემულია $n = 15$ მოცულობის მქონე შერჩევა, და $P = 10$, მაშინ $\alpha = 0.1$. შესაბამისად, $n\alpha = 1.5$ და ვინაიდან $n\alpha$ არ არის

მთელი რიცხვი, ხოლო მისი მთელი ნაწილია [1.5] = 1, ამიტომ ზემოთ მოყვანილი ფორმულის თანახმად: $\tilde{x}_{10} = x_{(2)}$, ანუ 10-პროცენტილი ვარიაციული მწკრივის მეორე წევრის ტოლია. მართლაც,

$$N\{x_i : x_i < x_{(2)}\} = 1 < n\alpha = 1.5 \quad \text{და} \quad N\{x_i : x_i > x_{(2)}\} = 13 < n(1-\alpha) = 13.5.$$

მაგალითი 4. თუ კი $n = 20$, $P = 10$, მაშინ $\alpha = 0.1$, $n\alpha = 2$ – მთელი რიცხვია და ისევ ზემოთ მოყვანილი ფორმულის თანახმად:

$$\tilde{x}_{10} = \frac{x_{(2)} + x_{(3)}}{2},$$

ანუ 10-პროცენტილი არის $[x_{(2)}, x_{(3)}]$ ინტერვალის შუა წერტილი. ამ შემთხვევაშიც

$$N\{x_i : x_i < \tilde{x}_{10}\} = 2 = n\alpha \quad \text{და} \quad N\{x_i : x_i > \tilde{x}_{10}\} = 18 = n(1-\alpha).$$

25-პროცენტილს ეწოდება **პირველი კვარტილი (quartile)** და აღინიშნება Q_1-ით, ხოლო 75-პროცენტილს ეწოდება **მესამე კვარტილი** და აღინიშნება Q_3-ით. 50-პროცენტილს (რომელიც წარმოადგენს მედიანას), აგრეთვე უწოდებენ **მეორე კვარტილს**. სხვა სიტყვებით, რომ ვთქვათ, $(Q_1, \tilde{x}, Q_2)$ კვარტილები მონაცემთა სიმრავლეს პირობითად ოთხ ტოლ ნაწილად ყოფს.

კვარტილების მოსაძებნად, ჩვენ შეგვიძლია ვისარგებლოთ როგორც პროცენტილების გამოთვლის ფორმულებით, ისე შემდეგი წესითაც: ჯერ მოვძებნოთ მედიანა, შემდეგ ვარიაციულ მწკრივში მის მარცხნივ განლაგებული მონაცემების (ანუ მონაცემთა პირველი ნახევრის) მედიანა, რომელიც მოგვცემს პირველ კვარტილს, Q_1-ს, და ასევე, ვარიაციულ მწკრივში მის მარჯვენივ განლაგებული მონაცემების (ანუ მონაცემთა მეორე ნახევრის) მედიანა, რომელიც მოგვცემს მესამე კვარტილს, Q_3-ს.

დავალება. შეამოწმეთ რომ უკანასკნელი წესით გამოთვლილი კვარტილები დაემთხვევა პროცენტილების ფორმულით გამოთვლილ შესაბამის კვარტილებს.

მაგალითი 5. ბოსტნეულის მაღაზიის 50 მყიდველის მიერ დახარჯული თანხა (ვარიაციულ მწკრივად დალაგებული და დამრგვალებული) შემდეგია:

2 6 6 8 9 10 11 11 12 12 13 13 <u>14</u> 14 14 15 15 16
17 18 18 18 19 19 20 || 20 20 21 23 26 27 28 29 30 31 32
33 <u>33</u> 34 36 37 39 40 43 45 52 61 63 64 69.

ვიპოვოთ მედიანა და სხვა კვარტილები.

ამოხსნა. რადგან შერჩევის მოცულობა $n = 50$ ლუწია, ამიტომ

$$\tilde{x} = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2} = \frac{x_{(25)} + x_{(26)}}{2} = 20.$$

შევნიშნოთ, რომ თუმცა $\tilde{x} = 20$, ის არ ემთხვევა ვარიაციული მწკრივის არცერთ წევრს: მისი პოზიცია არის ვარიაციული მწკრივის 25-ე და 26-ე პოზიციების საშუალო არითმეტიკული, ეს პოზიცია აღნიშნულია ‖ სიმბო-

ლოთი. ამ პოზიციის მარცხნივ მოთავსებულია 25 წევრი, რომელთა მედიანა ანუ სრულ მონაცემთა პირველი კვარტილია

$$Q_1 = x_{((25+1)/2)} = x_{(13)} = 14 .$$

მედიანის მარჯვნივ აგრეთვე 25 წევრია მოთავსებული, რომელთა მედიანა ანუ სრულ მონაცემთა მესამე კვარტილია მედიანის პოზიციიდან მარჯვენივ მე-13 პოზიციაზე მყოფი წევრი:

$$Q_3 = 33 .$$

ამრიგად, მოყვანილი მონაცემებისათვის $\tilde{x} = 20$, $Q_1 = 14$, $Q_3 = 33$. დავრწმუნდეთ, რომ პირველი და მესამე კვარტილებისათვის პროცენტილის გამოსათვლელი ფორმულებით მივიღებთ იგივე მნიშვნელობებს.

მართლაც, რადგან პირველი კვარტილი წარმოადგენს 25-პროცენტილს, $\alpha = 0.25$, $n = 50$, ამიტომ $n\alpha = 12.5$ მთელი არაა და [12.5] = 12, შესაბამისად გვექნება:

$$Q_1 = x_{([n\alpha]+1)} = x_{(13)} = 14 .$$

ანალოგიურად, მესამე კვარტილის გამოთვლისას $\alpha = 0.75$, $n\alpha = 37.5$, რომელიც არაა მთელი რიცხვი, [37.5] = 37 და ამიტომ, ისევ

$$Q_3 = x_{([n\alpha]+1)} = x_{(38)} = 33 .$$

კვარტილები მედიანასთან ერთად, ე. ი. სამეული $(Q_1, \tilde{x}, Q_2)$ გვაძლევს გარკვეულ წარმოდგენას მონაცემთა ცენტრის, გაბნევისა და განაწილების (გაგლუვებული ჰისტოგრამისა და პოლიგონის) ფორმის შესახებ. ამ შემთხვევაში $13 = Q_3 - \tilde{x} > \tilde{x} - Q_1 = 6$ (მესამე კვარტილი უფრო შორსაა მედიანისაგან, ვიდრე პირველი კვარტილი), რაც მიუთითებს იმაზე, რომ განაწილება მარჯვნივ ასიმეტრიულია (მისი მარჯვენა ბოლო უფრო ასიმეტრიულია, ვიდრე მარცხენა).

მანძილი კვარტილებს შორის წარმოადგენს გაფანტულობის მარტივ საზომს, რომელიც ახასიათებს ვარიაციულ მწკრივში კვარტილებს შორის მოთავსებულ მონაცემთა გაფანტულობის ხარისხს (შუა 50%-იანი ზონის გაბნევას). მას ეწოდება **კვარტილთშორისი გაბნევის დიაპაზონი (Interquartial range – IQR)**:

$$IQR = Q_3 - Q_1 .$$

ზემოთ მოყვანილ მაგალითში $IQR = 33 - 14 = 19$.

პირველი და მესამე კვარტილები და კვარტილთშორისი გაბნევის დიაპაზონი შერჩევის მდგრადი მახასიათებლებია. არცერთი მათგანი არ რეაგირებს $[Q_1, Q_3]$ ინტერვალის გარეთ მოთავსებულ ნებისმიერ ისეთ ცვლილებაზე, რომელიც ამ დაკვირვებებს კვლავ $[Q_1, Q_3]$ ინტერვალის გარეთ ტოვებს, თანაც ინტერვალის იმავე მხარეს, სადაც ადრე იმყოფებოდა. უკანასკნელ მაგალითში, პირველი 12 წევრის ნებისმიერი ისეთი ცვლილება, როცა არცერთი არ ღებულობს 14-ზე მეტ მნიშვნელობას და

ბოლო 12 წევრის ისეთი ცვლილება, როცა არცერთი არ ღებულობს 33-ზე ნაკლებ მნიშვნელობას, არ შეცვლის არც პირველ კვარტილს, არც მესამე კვარტილს და არც კვარტილთშორისი გაბნევის დიაპაზონს.

ვინაიდან მედიანა და პირველი და მესამე კვარტილები არ შეიცავენ ინფორმაციას განაწილების (გაგლუვებული ჰისტოგრამის) კუდების შესახებ, ამიტომ განაწილების ფორმის შესახებ უფრო სრულ ინფორმაციას იძლევა შემდეგი ხუთეული:

$$x_{min},\quad Q_1,\quad \tilde{x},\quad Q_3,\quad x_{max}.$$

ბოსტნეული მაღაზიის მყიდველთა დანახარჯების მონაცემებისათვის აღნიშნული ხუთეულია:

$$x_{min}=2,\quad Q_1=14,\quad \tilde{x}=20,\quad Q_3=33,\quad x_{max}=69.$$

10-პროცენტილს **დეცილი (decile)** ეწოდება. მონაცემთა განაწილების დასახასიათებლად იყენებენ აგრეთვე შემდეგ ხუთეულსაც:

პირველი დეცილი	Q_1	$\tilde{x}$	Q_3	მე-9 დეცილი
↓	↓	↓	↓	↓
10	25	50	75	9

პროცენტილები

მაგალითი 6. ავაგოთ უკანასკნელი ხუთეული მონაცემთა შემდეგი სიმრავლისათვის:

7 18 12 17 29 18 4 27 30 2 4 10 21 5 8.

შევქმნათ ვარიაციული მწკრივი (დავალაგოთ ეს მონაცემები ზრდადობის მიხედვით):

2 4 4 5 7 8 10 12 17 18 18 21 27 29 30.

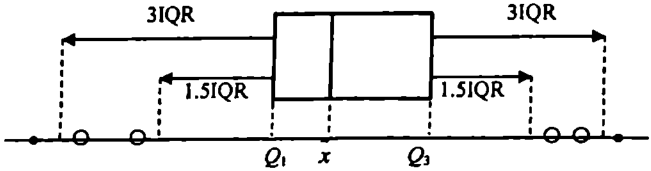
გვაქვს: $x_{min}=2$, $x_{max}=30$, $n=15$. შესაბამისად, მედიანა – $\tilde{x}=x_{(8)}=12$; პირველი დეცილი – $\tilde{x}_{(10)}=x_{([1.5]+1)}=x_{(2)}=4$; პირველი კვარტილი (ანუ პირველი 7 მონაცემის მედიანა) – $Q_1=\tilde{x}_{(25)}=x_{([3.75]+1)}=x_{(4)}=5$; მესამე კვარტილი (ანუ ბოლო 7 მონაცემის მედიანა) – $Q_3=\tilde{x}_{(75)}=x_{([0.75\cdot 15]+1)}=x_{(12)}=21$; მე-9 დეცილი – $\tilde{x}_{(90)}=x_{([0.9\cdot 15]+1)}=x_{(14)}=29$. ი. ი. საძიებელი ხუთი მახასიათებელია:

<u>*I დეცილი*</u> = 4, $Q_1=5$, $\tilde{x}=12$, $Q_3=21$, <u>*მე-9 დეცილი*</u> = 29.

გარდა ამისა, კვარტილთშორისი გაბნევის დიაპაზონი – $IQR=Q_3-Q_1=16$.

უკანასკნელ პერიოდში განსაკუთრებული პოპულარობით სარგებლობს მონაცემთა განაწილების ვიზუალური წარმოჩენის ახალი საშუალება – ე. წ. **ბოქსპლოტი (boxplot)**. როგორც ცნობილია, ფოთლებიანი ღეროების მსგავსი დიაგრამა და ჰისტოგრამა გვაძლევს მონაცემთა სიმრავლის ზოგად დახასიათებას, ბოქსპლოტი კი იძლევა მონაცემთა სიმრავლის გრაფიკულ დახასიათებას. ბოქსპლოტი მდგრადი საზომების – (Q_1, $\tilde{x}$, Q_3, IQR) მეშვეობით საშუალებას იძლევა გამოიკვეთოს მონაცემთა სიმრავლის ზო-

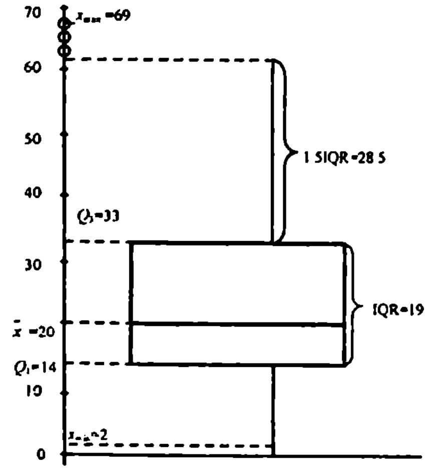
გიერთი თავისებურება: ა). განლაგების ცენტრი; ბ). გაფანტულობა; გ). გავრცობა და სიმეტრიულობიდან გადახრის ბუნება და ბოლოს, მოხდეს დ). **"ამოვარდნილი (outlier)"** დაკვირვებების იდენტიფიკაცია. აღვწეროთ, თუ როგორ ხდება ბოქსპლოტის აგება:



ბოქსპლოტი

1. გავავლოთ ჰორიზონტალური ღერძი, რომელზედაც მოვნიშნოთ Q_1, $\bar{x}$ და Q_3;
2. ავაგოთ მართკუთხედი, რომლის გვერდები მართობულია ჰორიზონტალური ღერძის და რომლის მარცხენა გვერდი Q_1-ის, ხოლო მარჯვენა კი Q_3-ის ზემოთაა მოთავსებული;
3. გავავლოთ ვერტიკალური წრფე მართკუთხედში მედიანის ზევით;
4. მართკუთხედის თითოეული გვერდიდან გავავლოთ ჰორიზონტალური ღერძის პარალელური ორი წრფე, პირველზე გადავზომოთ $1.5IQR$, ხოლო მეორეზე – $3IQR$ სიგრძის მონაკვეთები;
5. ყოველი მონაცემის შესაბამის წერტილზე, რომელიც მართკუთხედის გვერდიდან გადაზომილ $1.5IQR$-სა და $3IQR$-ს შორისაა მოთავსებული შემოვხაზოთ წრეწირი. ამ მონაცემებს ეწოდება ზომიერი "ამოვარდნები", ხოლო იმ მონაცემების შესაბამის წერტილებზე, რომლებიც მოთავსებულია მართკუთხედის გვერდიდან გადაზომილი $3IQR$-ის გარეთ შემოვხაზოთ გამუქებული წრე. ასეთი წრეებით შემოხაზულ მონაცემებს ექსტრემალური "ამოვარდნები" ეწოდება.

ხშირად გრაფიკს აბრუნებენ 90^0-იანი კუთხით და გამოსახავენ ოდნავ მოდიფიცირებული სახით. კერძოდ, ვერტიკალურ ღერძზე გადაიზომება x_{min}, Q_1, $\bar{x}$, Q_3, x_{max} და მხოლოდ $1.5IQR$ მანძილი მართკუთხედის გვერდებიდან და ისიც მხოლოდ იმ შემთხვევაში, როდესაც x_{min} ან x_{max} აღმოჩნდება მის გარეთ. ზოგჯერ ამ გრაფიკზე პირველ და მეცხრე დეცილსაც მიუთითებენ. მაგალითად, ბოსტნეულის მაღაზიის მყიდველთა დანახარჯების მონაცემებისათვის ბოქსპლოტს ექნება შემდეგი სახე:



გარდა ზომიერი და ექსტრემალური „ამოვარდნების" საიდენტიფიკაციო წესისა, გამოიყენება ე. წ. უხეში წესი იმ მონაცემების აღმოსაჩენად, რომელთა შესახებ შეიძლება გაჩნდეს ეჭვი, რომ ისინი არ არიან ტიპიური მონაცემთა მოცემული სიმრავლისათვის. ეს წესი მდგომარეობს შემდეგში: გამოვყოთ ის მონაცემები, რომლებიც აღემატებიან ($Q_3+1.5\,IQR$)-ს, და ის მონაცემები, რომლებიც ნაკლებია ($Q_1-1.5\,IQR$)-ზე. უკანასკნელ ნახაზზე ეს მონაცემებია 63, 64, 69 და ისინი აღნიშნულია წრეებით. ეს ის მონაცემებია, რომლებიც შესაძლოა ამოვარდნებს წარმოადგენს. იმის დასადგენად, არიან თუ არა ეს მონაცემები „ამოვარდნილები" (ანუ მკვეთრად მოშორებულნი არიან თუ არა მონაცემთა ძირითად მასივს), უნდა აიგოს ფოთლებიანი ღეროების მსგავსი დიაგრამა მაგალითი 5-ის მონაცემებისათვის.

დავალება. ააგეთ ფოთლებიანი ღეროების მსგავსი დიაგრამა მაგალითი 5-ის მონაცემებისათვის, დარწმუნდით, რომ სიხშირეთა განაწილება მკაცრად ასიმეტრიულია და მონაცემები 63, 64 და 69 არ წარმოადგენს „ამოვარდნილ" მონაცემებს.

რანგები, პროცენტული რანგები და მათი კავშირი პროცენტილებთან

დავუშვათ, რომ მოცემულია n მოცულობის განსხვავებული მნიშვნელობების მქონე შერჩევა $x_1,\dots,x_n$, სადაც $x_i \neq x_j$, თუ $i \neq j$, ე. ი. არც ერთი მნიშვნელობა არ მეორდება. შევადგინოთ ვარიაციული მწკრივი (დავალაგოთ მონაცემები ზრდის მიხედვით):

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}.$$

x_i დაკვირვების **რანგი (rank)** ეწოდება ამ დაკვირვების ნომერს ვარიაციულ მწკრივში (იგი წარმოადგენს მონაცემთა პოზიციის ერთ-ერთ მახ-

ასიათებელს). სხვა სიტყვებით რომ ვთქვათ, x_i-ს რანგი ეწოდება ისეთ მთელ რიცხვს r_i, რომლისთვისაც სრულდება ტოლობა:

$$x_{(r_i)} = x_i .$$

სინამდვილეში x_i დაკვირვების რანგი r_i არის იმ დაკვირვებების რაოდენობა, რომლებიც არ აღემატებიან x_i დაკვირვებას:

$$r_i = N\{x_j : x_j \le x_i\},$$

ამასთანავე გასაგებია, რომ მაქსიმალური დაკვირვების რანგი დაემთხვევა შერჩევის მოცულობას.

მაგალითი 1. განვიხილოთ შემდეგი მონაცემები: 5, 6, 2, 7, 9, 8, 10 და თითოეულ მათგანს მივუწეროთ თავისი რანგი.

ამოხსნა. ვარიაციული მწკრივი არის: 2, 5, 6, 7, 8, 9, 10. ამიტომ $x_1 = 5$-ის რანგი იქნება $r_1 = 2$ (დაკვირვებას 5 არ აღემატება ორი დაკვირვება: 2 და 5); $x_2 = 6$-ის რანგია $r_2 = 3$ (დაკვირვებას 6 არ აღემატება სამი დაკვირვება: 2, 5 და 6) და ა. შ. საბოლოოდ მონაცემთა მოცემული სიმრავლე ქვეშ მიწერილი რანგებით ასე გამოიყურება:

5	6	2	7	9	8	10
$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
2	3	1	4	6	5	7

მაგალითი 2. მონაცემებია: 18, 13, 15, 12, 16, 20. მივუწეროთ რანგები. გვაქვს:

18	13	15	12	16	20
5	2	3	1	4	6

აღსანიშნავია, რომ რანგები ინვარიანტულია (არ იცვლება) მონაცემთა ნებისმიერი გადალაგების მიმართ. თუ ამ მაგალითის მონაცემებს დავალაგებთ სხვა თანმიმდევრობით, მაგალითად, შემდეგნაირად: 15, 20, 12, 18, 13, 16, მაშინ რანგები იქნება

15	20	12	18	13	16
3	6	1	4	2	5

ანუ ყოველ წევრს იგივე რანგი მიეწერა რაც ადრე.

ბუნებრივად იბადება კითხვა: თუ მოცემულია შერჩევა $x_1, \dots, x_n$, ისეთი, რომ პირობა $x_i \ne x_j$, თუ $i \ne j$ არ სრულდება, ანუ ზოგიერთი წევრი შერჩევაში მეორდება, მაშინ როგორი წესით მოხდება რანგების მიწერა?

ვთქვათ, ჩვენი მონაცემებია: 18, 28, 23, 29, 32, 18, 21, 14, 18, 14. რანგის განმარტების თანახმად ძველი წესის თანახმად ვარიაციული მწკრივის სახით გადაწერილ ამ მონაცემებს რანგები ასე უნდა მიეწეროს:

14	14	18	18	18	21	23	28	29	32
1	2	3	4	5	6	7	8	9	10

როგორც ვხედავთ, 14-ის ტოლი დაკვირვება შეგვხვდა ორჯერ და ძველი წესით პირველს მივაწერეთ რანგი 1, ხოლო მეორეს კი – რანგი 2 (ანალოგიურად, 18 შეგვხვდა სამჯერ და შესაბამისად, მას დაკავებული პოზიციის მიხედვით მიეწერა რანგები: 3, 4 და 5). გამოდის, რომ ერთ-ერთ მათგანს მივანიჭეთ უპირატესობა. უმჯობესია განმეორებადი დაკვირვებებ-

ისათვის მიგვეწერა ერთი და იგივე რანგი შემდეგი წესით: ტოლი დაკვირვებებიდან თითოეულს მივანიჭოთ მათზე მოსული რანგების საშუალო არითმეტიკული. მაგალითად, დაკვირვებებს, რომელთა მნიშვნელობებია 14, უნდა მიენიჭოს რანგი 1.5 (რადგანაც (1+2)/2=1.5), 18-ის ტოლ დაკვირვებებს – რანგი 4 (ვინაიდან (3+4+5)/3=4).

მეორეს მხრივ, იმ შემთხვევაშიც კი, როდესაც შერჩევაში არცერთი ელემენტი არ მეორდება, ელემენტის რანგის მითითება არ იძლევა ამომწურავ ინფორმაციას მის მდებარეობაზე, თუ ამავე დროს ჩვენთვის ცნობილი არ არის შერჩევის მოცულობაც. მაგალითად, თუ ვიცით, რომ x^* ელემენტს n მოცულობის შერჩევიდან და y^* ელემენტს მეორე, m მოცულობის შერჩევიდან გააჩნიათ ერთი და იგივე რანგი, ჩვენ ჯერ კიდევ ვერ დავასკვნით, რომ მათი პოზიციები იდენტურია, თუ არ გვექნება დამატებითი ინფორმაცია იმის შესახებ, რომ $m=n$.

საზოგადოდ, როგორ შევადაროთ ერთმანეთს იმ ორი მონაცემის პოზიციები, რომლებიც განსხვავებული მოცულობის სხვადასხვა შერჩევას ეკუთვნის და რანგებიც განსხვავებულები აქვს? (ამ ტიპის კითხვები წარმოიშობა, მაგალითად, ქართული ეროვნული გამოცდების დროს, როცა შეფასების ცენტრმა უნდა შეადაროს სხვადასხვა აბიტურიენტების პოზიციები (რეიტინგები) სხვადასხვა დისციპლინებში ან მაშინაც კი, როცა ერთი დისციპლინის ფარგლებში მათ შეასრულეს სხვადასხვა საგამოცდო ვარიანტი). ამ კითხვაზე პასუხს იძლევა ე. წ. **პროცენტული რანგის (percentile rank)** ცნება.

თუ n შერჩევის მოცულობაა, ხოლო დაკვირვების რანგია r, მაშინ ამ დაკვირვების პროცენტული რანგი მოიცემა შემდეგი ფორმულით:

$$\frac{r-1}{n}\cdot 100\%+\frac{0.5}{n}\cdot 100\%=\frac{2r-1}{2n}\cdot 100\%.$$

მოცემული დაკვირვების *პროცენტული რანგი* წარმოადგენს ამ დაკვირვების ქვევით მდგომ მონაცემებზე მოსული პროცენტებისა ($\frac{r-1}{n}\cdot 100\%$) და თავად დაკვირვებაზე მოსული პროცენტების ნახევარის ($\frac{0.5}{n}\cdot 100\%$) ჯამს.

ამრიგად, მოცემული დაკვირვების პროცენტული რანგი მიგვითითებს დაკვირვებულ მნიშვნელობათა რა პროცენტს (მინუს ამ დაკვირვებაზე მოსული პროცენტების ნახევარი) აღემატება ეს მონაცემი. ცხადია, რომ რაც უფრო მაღალია დაკვირვების პროცენტული რანგი, მით უფრო უკეთესია მისი პოზიცია (თუ ნიშნები განლაგებულია აღმავლობის მიხედვით).

ფაქტიურად, პროცენტული რანგი გვიჩვენებს, ამა თუ იმ შერჩევის კონკრეტული მონაცემი, მონაცემების რამდენი პროცენტის წინაა. შესაბამისად, თუ ვიგულისხმებთ, რომ მონაცემებს სხვადასხვა შერჩევებში მოხვედრის თანაბარი შანსები აქვთ, მაშინ ბუნებრივი იქნება ერთ შერჩევაში უკეთეს პოზიციაზე მდგომ მონაცემს მიენიჭოს უპირატესობა სხვა შერჩევაში უარეს პოზიციაზე მდგომ მონაცემთან შედარებით. მაგალითად, ეროვნულ გამოცდებზე ზოგადი უნარების გამოცდის პირველ ვარიანტში 37

პროცენტული რანგის მქონე აბიტურიენტს უპირატესობა ენიჭება მესამე ვარიანტში 36 პროცენტული რანგის მქონე აბიტურიენტთან შედარებით, ან ინგლისურ ენაში 43 პროცენტული რანგის მქონე აბიტურიენტს უპირატესობა ენიჭება გერმანულ ენაში 42.9 პროცენტული რანგის მქონე აბიტურიენტთან შედარებით.

მაგალითი 3. ვთქვათ, კლასში არის 10 მოსწავლე და გიორგის ნიშნები უფრო მაღალია, ვიდრე ხუთი მოსწავლის. კოტე სწავლობს 50 მოსწავლისაგან შემდგარ კლასში და მისი ნიშნები უფრო მაღალია, ვიდრე 17 მოსწავლის. რომელს უკავია უკეთესი პოზიცია თავის კლასში, კოტეს თუ გიორგია?

ამოხსნა. უპირველეს ყოვლისა შევნიშნოთ, რომ მოსწავლეები დალაგებულია ნიშნების ზრდის მიხედვით და გიორგის რანგია 6, ხოლო კოტეს რანგია 18. ჩავატაროთ თითოეულის პოზიციის პროცენტული ანალიზი. გიორგი უკეთესია 5 მოსწავლეზე, რომლებიც შეადგენენ კლასის 50%-ს, თვითონ გიორგიზე მოდის 10%, ხოლო 40%-ს აქვს გიორგიზე უკეთესი ნიშნები. ამ მონაცემებით გიორგის პროცენტული რანგია:

$$50\%+5\%=55\%.$$

კოტეს ნიშნები უკეთესია 17 მოსწავლის ნიშნებზე, რომელიც შეადგენს კლასის 34%-ს, კოტეზე მოდის 2%, ამიტომ მისი პროცენტული რანგია:

$$34\%+1\%=35\%.$$

ვინაიდან, გიორგის პროცენტული რანგი (55) უფრო მაღალია, ვიდრე კოტესი (35), ამიტომ გიორგის პოზიცია თავის კლასში უკეთესია, ვიდრე კოტესი საკუთარ კლასში.

ეს მაგალითი ამავე დროს გვიჩვენებს, რომ რანგის ცნების შემოსაღებად არ არის აუცილებელი გაგვაჩნდეს რიცხვითი მონაცემები, საკმარისია მხოლოდ ობიექტების ან ინდივიდების სიმრავლის რაიმე თვისების ან ნიშნის (მაგალითად, ნიჭიერება, სიმაღლე, პოპულარობა და ა. შ.) გაძლიერების ან ზრდის მიხედვით დალაგება და ყოველი მათგანისათვის რიგობრივი ნომრის მიწერა.

პროცენტილებსა და პროცენტულ რანგებს შორის შემდეგი კავშირია: თუ შერჩევის ელემენტის პროცენტული რანგია P, მაშინ ეს ელემენტი წარმოადგენს შერჩევის P-პროცენტილს (თუმცა არ არის სავალდებულო, რომ P-ს ნებისმიერი მნიშვნელობისათვის P-პროცენტილი წარმოადგენდეს შერჩევის ელემენტს). მართლაც, ვთქვათ შერჩევის მოცულობაა n და რომელიმე ელემენტის პროცენტულია რანგია P. ვიპოვოთ ამ ელემენტის რანგი r. პროცენტული რანგის განმარტების თანახმად:

$$\frac{2r-1}{2n}=\frac{P}{100}=:\alpha$$

(შეფარდების მნიშვნელობა აღნიშნულია α-თი). აქედან, ცხადია, რომ $r=(2n\alpha+1)/2$. ვაჩვენოთ ახლა, რომ $x_{(r)}$ წარმოადგენს P-პროცენტილს. რადგანაც r მთელი რიცხვია (შესაბამისად, მთელი კენტი რიცხვია $2r-1$), ამიტომ $2n\alpha=2r-1$ კენტია. ე. ი. $n\alpha$ არ არის მთელი რიცხვი და ამიტომ P-პროცენტილის გამოსათვლელი ფორმულის თანახმად გვექნება:

$$\tilde{x}_p = x_{([n\alpha]+1)} = x_{([\frac{2r-1}{2}]+1)} = x_{([r-\frac{1}{2}]+1)} = x_{(r-1+1)} = x_{(r)}.$$

პროცენტილები და პროცენტული რანგები გამოიყენება როგორც შედარების საშუალება განათლების სფეროში, ფსიქოლოგიასა და სხვა დარგებში, რომლებსაც საქმე აქვთ შეფასებებთან (ნიშნებთან) და ბალებთან. მაგალითად, ამერიკის შეერთებული შტატების კოლეჯებში ჯერ შემსვლელთა პროცენტულ რანგებს განიხილავენ ნიშნების მიხედვით და მხოლოდ ამის შემდეგ – თავად ნიშნებს. შეიძლება ვილაპარაკოთ რომელიმე ბანკის პროცენტულ რანგზეც ბანკების გარკვეულ ჯგუფში, ერთი რაიონის წარმოების დონის პროცენტულ რანგზე ქვეყნის წარმოებაში და სხვა.

დისპერსია და სტანდარტული გადახრა

ისევ დავუბრუნდეთ ადრე განხილულ მაგალითს. დავუშვათ, რომ მოცემულია მონაცემთა ორი მწკრივი:

A	8	9	10	10	13
B	1	5	10	16	18.

როგორც ჩვენ ვნახეთ, $\bar{x}_A = \bar{x}_B = 10$. გარდა ამისა, დავასკვენით, რომ B მწკრივის ელემენტების ცვალებადობა უფრო მაღალია, ვიდრე A მწკრივის ელემენტებისა. ჩვენ ამ შემთხვევაში ვიზუალურად დავინახეთ, რომ A მწკრივის წევრები უფრო ნაკლებად არიან გადახრილი საშუალოსაგან, ვიდრე B მწკრივისა. გასაგებია, რომ თუ მონაცემთა სიმრავლე დიდი მოცულობისაა, ასეთი ფაქტების ვიზუალურად აღმოჩენა შეუძლებელია. შესაბამისად, ბუნებრივად ჩნდება საჭიროება ისეთი საზომის შემოღებისა, რომელიც გაგვიადვილებდა მონაცემების თავისი საშუალოს მიმართ გაფანტულობის შესახებ დასკვნის გაკეთებას. სწორედ ასეთ საზომს წარმოადგენს **შერჩევითი დისპერსია (sample variance)**.

გამოვთვალოთ საშუალოდან გადახრები ზემოთ მოყვანილი მწკრივებისათვის (თითოეულ მონაცემს გამოვაკლოთ საშუალო):

A –	$\bar{x}_A - x_i$:	-2	-1	0	0	3
B –	$\bar{x}_B - x_i$:	-9	-5	0	6	8

პირველი რაც შეიძლება აზრად მოგვივიდეს: შეიძლება გვეფიქრა, რომ გაფანტულობის საზომად შეიძლებოდა აგვეღო ამ გადახრების არითმეტიკული საშუალო, მაგრამ ორივე შემთხვევაში ეს იქნებოდა ნული (გაიხსენეთ საშუალოს პირველი თვისება). შესაბამისად, ეს სიდიდე არანაირ ინფორმაციას არ მოგვცემს.

შემდეგი ბუნებრივი ნაბიჯი იქნებოდა, რომ აგვეღო გადახრათა მოდულების არითმეტიკული საშუალო. ასე გამოთვლილ სიდიდეს **საშუალო აბსოლუტური გადახრა (Mean Absolute Deviation –** ***MAD*****)** ეწოდება. გვაქვს:

$$MAD_A = (2+1+0+0+3)/5 = 1.2 \text{ და}$$

$$MAD_B = (9+5+0+6+8)/5 = 7.2,$$

ანუ, $MAD_A < MAD_B$.

ფორმალურად საშუალო აბსოლუტური გადახრის ანალიზური გამოსახულება ასე გამოიყურება:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

როგორც ვხედავთ, გაფანტულობის ამ საზომმაც დაადასტურა ჩვენი ვარაუდი. მაგრამ სტატისტიკური დასკვნების თეორიაში ეს მაჩვენებელი საკმაოდ იშვიათად გამოიყენება. ამ თეორიაში უდიდეს როლს თამაშობს *შერჩევითი დისპერსია*, რომელიც წარმოადგენს საშუალოდან გადახრების კვადრატების საშუალო არითმეტიკულს და აღინიშნება s^2 სიმბოლოთი. ჩვენს შემთხვევაში გვაქვს:

$$s_A^2 = (4+1+0+0+9)/5 = 2.6 \text{ და}$$

$$s_B^2 = (81+25+0+36+64)/5 = 41.2,$$

ე. ი. ისევ, $s_A^2 < s_B^2$.

თუ მოცემულია n მოცულობის $x_1,...,x_n$ შერჩევა, მაშინ შერჩევითი დისპერსიის ანალიზური გამოსახულება შემდეგია:

$$s_n^2 = \frac{1}{n}[(x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

შერჩევითი დისპერსია შეიძლება გადაიწეროს გამოთვლებისათვის მოხერხებული უფრო მარტივი ფორმით:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2.$$

თუ $x_1,...,x_N$ წარმოადგენს პოპულაციას (და არა შერჩევას), მაშინ პოპულაციის საშუალოა

$$\mu = \frac{1}{N}\sum_{i=1}^{N}x_i,$$

ხოლო

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

გამოსახულებას *პოპულაციის დისპერსია* ეწოდება.

შევნიშნავთ, რომ შერჩევითი დისპერსია წარმოადგენს პოპულაციის დისპერსიის ემპირიულ ანალოგს, რომელიც დაკვირვებათა რიცხვის უსასრულოდ ზრდისას მიისწრაფის პოპულაციის დისპერსიისაკენ.

მოსახერხებელია გაფანტულობის ოდნავ მოდიფიცირებული საზომის განხილვა, რომელსაც *შესწორებულ შერჩევით დისპერსიას* უწოდებენ:

$$s_n'^2 = \frac{n}{n-1}s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - n(\bar{x})^2],$$

ან პოპულაციის შემთხვევაში:

$$\sigma'^2 = \frac{N}{N-1}\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$$

შევნიშნოთ, რომ თუ პოპულაციის μ საშუალო ცნობილია, მაშინ შერჩევითი დისპერსია განისაზღვრება ფორმულით

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 ,$$

უცნობი საშუალოს შემთხვევაში კი μ-ს ნაცვლად უნდა ვისარგებლოთ შერჩევითი $\bar{x}$ საშუალოთი, მაგრამ როგორც ცნობილია (საშუალოს მეორე თვისება): $\sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - \mu)^2$, ანუ x_i-ები უფრო ახლოსაა $\bar{x}$-თან, ვიდრე μ-სთან. ამის საკომპენსაციოდ $s_n'^2$-ის ფორმულაში n-ის ნაცვლად გამოიყენება $n-1$.

მოვიყვანოთ შერჩევითი დისპერსიის თვისებები:

I. თუ არსებული $x_1,...,x_n$ მონაცემებიდან შევქმნით ახალ მონაცემებს $y_i = ax_i + b$, $i = 1,...,n$, მაშინ $s_y^2 = a^2 s_x^2$ და $s_y'^2 = a^2 s_x'^2$.

მართლაც, ვინაიდან საშუალოს გადათვლის წესის თანახმად $\bar{y} = a\bar{x} + b$, ამიტომ $y_i - \bar{y} = ax_i + b - a\bar{x} - b = a(x_i - \bar{x})$ და შესაბამისად,

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}a^2(x_i - \bar{x})^2 = a^2 s_x^2 .$$

II. თუ $x_1,...,x_n$ და $y_1,...,y_m$ ორი შერჩევაა, მაშინ გაერთიანებული შერჩევისათვის $x_1,...,x_n, y_1,...,y_m$ შერჩევითი დისპერსია და შესწორებული შერჩევითი დისპერსიები გადაითვლება შემდეგნაირად:

$$s^2 = \frac{n}{m+n}s_x^2 + \frac{m}{m+n}s_y^2 + \frac{mn}{(m+n)^2}(\bar{x} - \bar{y})^2 \text{ და}$$

$$s'^2 = \frac{n-1}{m+n}s_x'^2 + \frac{m-1}{m+n}s_y'^2 + \frac{mn}{(m+n-1)(m+n)}(\bar{x} - \bar{y})^2 .$$

(იმ კერძო შემთხვევაში, როცა $m = n$ და $\bar{x} = \bar{y}$, გაერთიანებული შერჩევის შერჩევითი დისპერსია იქნება თავიდან აღებული შერჩევების შერჩევითი დისპერსიების საშუალო არითმეტიკული: $s^2 = (s_x^2 + s_y^2)/2$).

მართლაც, თუ გავითვალისწინებთ, რომ გაერთიანებული შერჩევის საშუალოა $\frac{n}{m+n}\bar{x} + \frac{m}{m+n}\bar{y}$, გვაქვს:

$$s^2 = \frac{1}{m+n}\left(\sum_{i=1}^{n}(x_i - \frac{n}{m+n}\bar{x} - \frac{m}{m+n}\bar{y})^2 + \sum_{i=1}^{m}(y_i - \frac{n}{m+n}\bar{x} - \frac{m}{m+n}\bar{y})^2\right) =$$

$$= \frac{n}{m+n}s_x^2 + \frac{m}{m+n}s_y^2 + \frac{mn}{(m+n)^2}(\bar{x} - \bar{y})^2 .$$

როგორც ვხედავთ, შერჩევითი დისპერსიის განზომილება უდრის დაკვირვებათა განზომილების კვადრატს. სტატისტიკოსს ხშირად ესაჭიროება გაფანტულობის ისეთი საზომი, რომელიც იმავე ერთეულებში იზომება, რაც საწყისი მონაცემები. ასეთი საზომი მიიღება შერჩევითი დისპერსიიდან კვადრატული ფესვის ამოღებით და მას **სტანდარტული გადახრა (standard deviation)** ან **საშუალო კვადრატული გადახრა (mean square deviation)** ეწოდება.

სტანდარტული გადახრა წარმოადგენს არითმეტიკულ კვადრატულ ფესვს შერჩევითი დისპერსიიდან

$$s=+\sqrt{s^2} \quad (s'=+\sqrt{s'^2}),$$

ანალოგიურად, *პოპულაციის სტანდარტული გადახრა* არის არითმეტიკული კვადრატული ფესვი პოპულაციის დისპერსიიდან

$$\sigma=+\sqrt{\sigma^2}$$

შემდგომში დისპერსიისა და სტანდარტული გადახრის გამოსათვლელად გამოიყენება შედარებით მოხერხებული ფორმულები, რომლებიც მათემატიკურად ექვივალენტურია ზემოთმოყვანილი ფორმულების, მაგრამ არ საჭიროებს საშუალოს წინასწარ გამოთვლას. ამ ფორმულებს უწოდებენ დისპერსიისა და სტანდარტული გადახრის შემოკლებულ ფორმულებს და მათ აქვთ შემდეგი სახე:

$$s'^2=\frac{\sum_{i=1}^{n}x_i^2-[(\sum_{i=1}^{n}x_i)^2/n]}{n-1} \text{ — დისპერსია,}$$

$$s'=\sqrt{\frac{\sum_{i=1}^{n}x_i^2-[(\sum_{i=1}^{n}x_i)^2/n]}{n-1}} \text{ — სტანდარტული გადახრა.}$$

მაგალითი. ვიპოვოთ დისპერსია და სტანდარტული გადახრა ევროპაში ექვს წელიწადში გაყიდული ავტომანქანების ერთობლივი ღირებულების შერჩევის მიხედვით (მონაცემები მოყვანილია მილიონ დოლარებში):

11.2 11.9 12.0 12.8 13.4 14.3.

ამოხსნა.

ნაბიჯი 1. ვიპოვოთ მნიშვნელობათა ჯამი:

$$\sum_{i=1}^{6}x_i=11.2+11.9+12.0+12.8+13.4+14.3=75.6;$$

ნაბიჯი 2. თითოეული მნიშვნელობა ავიყვანოთ კვადრატში და ვიპოვოთ მათი ჯამი:

$$\sum_{i=1}^{6}x_i^2=11.2^2+11.9^2+12.0^2+12.8^2+13.4^2+14.3^2=958.94;$$

ნაბიჯი 3. შევიტანოთ შესაბამის ფორმულაში და გამოვთვალოთ:

$$s'^2=\frac{\sum_{i=1}^{6}x_i^2-[(\sum_{i=1}^{6}x_i)^2/6]}{6-1}=\frac{958.94-[(75.6)^2/6]}{5}=1.28;$$

ე. ი. შერჩევის დისპერსიაა 1.28. სტანდარტული გადახრა კი იქნება $s'=\sqrt{1.28}\approx 1.33$.

შევნიშნავთ, რომ $\sum_{i=1}^{n}x_i^2$ არ არის იგივე, რაც $(\sum_{i=1}^{n}x_i)^2$. აღნიშვნაში $\sum_{i=1}^{n}x_i^2$ — იგულისხმება მნიშვნელობების ჯერ კვადრატში აყვანა და შემდეგ

აჯამეა, ხოლო სიმბოლო $(\sum_{i=1}^{n} x_i)^2$ გულისხმობს მნიშვნელობების ჯერ შეკრებას და შემდეგ კი მიღებული შედეგის კვადრატში აყვანას.

ცნობილია, რომ მცირე ინვესტორისათვის, როგორც საინვესტიციო ალტერნატივა, დიდი პოპულარობით სარგებლობს ინვესტიციების ჩადება საინვესტიციო ფონდებში – ტრესტებში. იმისათვის, რომ ინვესტორს გაუადვილდეს გადაწყვეტილების მიღება, თუ რომელ კონკრეტულ ტრესტში მოახდინოს ინვესტირება, ფინანსურ ჟურნალებში რეგულარულად ქვეყნდება ყოველი ცალკეული ტრესტის წლიური საპროცენტო შემოსავლის მონაცემები ბოლო ათი წლის განმავლობაში. პუბლიკაციებში ამასთან ერთად მითითებულია ყოველი ტრესტის რისკის დონე: მაღალი, საშუალო და დაბალი. რისკის დონეების კლასიფიკაცია კი ხდება საშუალო წლიური საპროცენტო შემოსავლის ისტორიული ცვალებადობის მიხედვით. რაც უფრო მაღალია წლიური საპროცენტო შემოსავლის ცვალებადობა, მით უფრო მაღალია რისკის დონე.

მაგალითი. მოცემულია ორი ტრესტის წლიური საპროცენტო შემოსავლის მონაცემები ბოლო 10 წლის განმავლობაში:

ტრესტი A	8.3	-6.2	20.9	-2.7	33.6	42.8	24.4	5.2	3.1	30.5
ტრესტი B	12.1	-2.8	6.4	12.2	27.8	25.3	18.2	10.7	-1.3	11.4

რომელი ტრესტი მიეკუთვნება რისკის უფრო მაღალ დონეს, A თუ B?

ამოხსნა. რადგან დისპერსია მონაცემთა ცვალებადობის ყველაზე თვალსაჩინო საზომია, ყოველი ტრესტის მონაცემების მიხედვით უნდა გამოვთვალოთ თითოეული მათგანის შესწორებული შერჩევითი დისპერსია, $s_A'^2$ და $s_B'^2$ და შევადაროთ ისინი ერთმანეთს. წინასწარ გამოვთვალოთ შერჩევითი საშუალოები. გვაქვს:

$$\bar{x}_A = (8.3-6.2+20.9-2.7+33.6+42.8+24.4+5.2+3.1+30.5)/10 = 16\%,$$

$$\bar{x}_B = (12.1-2.8+6.4+12.2+27.8+25.3+18.2+10.7-1.3+11.4)/10 = 12\%.$$

შესაბამისად,

$$s_A'^2 = (8.3^2+\cdots+10.5^2-10\cdot 16^2)/9 = (5083.36-2560)/9 = 280.34\,(\%)^2,$$

$$s_B'^2 = (12.1^2+\cdots+11.4^2-10\cdot 12^2)/9 = (2334.36-1440)/9 = 99.38\,(\%)^2.$$

როგორც ვხედავთ, $s_A' > s_B'$. ეს იმას ნიშნავს, რომ A ტრესტი მიეკუთვნება რისკის უფრო მაღალ დონეს (უფრო რისკიანია), ვიდრე B ტრესტი, იმავდროულად, $\bar{x}_A > \bar{x}_B$, ანუ A ტრესტის ამონაგები საშუალოდ უფრო მაღალია, ვიდრე B ტრესტისა. ეს ფაქტი სავსებით ეთანხმება ჩვენს ინტუიციას: ინვესტიცია, რომელიც დაკავშირებულია რისკის უფრო მაღალ დონესთან, უნდა იძლეოდეს უფრო მაღალ საშუალო ამონაგებსაც.

მაგალითი. დავუშვათ, რომ 10 წლის წინ თქვენ მოახდინეთ ტოლი თანხების ინვესტირება A და B ტრესტებში, ე. ი. შექმენით საინვესტიციო პორტფელი, რომელშიც A და B ტრესტებში ინვესტიციათა წონები ტოლია

და უდრის 1/2-ს. ახალი პორტფელი აღვნიშნოთ C-თი. შევადაროთ A , B და C პორტფელების რისკიანობა.

ამოხსნა. წინასწარ შემოვიღოთ *შემოსავლიანობის* ცნება: დავუშვათ, რომ საინვესტიციო თანხა წლის დასაწყისში უდრის P_0-ს და ამ თანხის ინვესტირებამ A და B ტრესტებში წლის ბოლოს მოგვცა, შესაბამისად, P_A და P_B თანხები. მაშინ A და B ტრესტების ამონაგები R_A და R_B შემდეგი ფორმულებით მოიცემა (%-ში):

$$R_A = \frac{P_A - P_0}{P_0} \cdot 100\%,$$

$$R_B = \frac{P_B - P_0}{P_0} \cdot 100\%.$$

ვთქვათ, ახლა A და B ტრესტებიდან თითოეულში ინვესტირებულია $\frac{1}{2}P_0$-ის ტოლი თანხა. მაშინ წლის ბოლოს C პორტფელისაგან მიღებული თანხაა $\frac{1}{2}P_A + \frac{1}{2}P_B$. ამიტომ მისი ამონაგები იქნება:

$$R_C = \frac{(\frac{1}{2}P_A + \frac{1}{2}P_B) - P_0}{P_0} = \frac{\frac{1}{2}(P_A - P_0) + \frac{1}{2}(P_B - P_0)}{P_0} = \frac{1}{2}R_A + \frac{1}{2}R_B.$$

გასაგებია, რომ წინა მაგალითში მოყვანილი მონაცემები წარმოადგენენ ბოლო 10 წლის დაკვირვებებს R_A-სა და R_B-ზე, ამიტომ ჩვენ შეგვიძლია შევქმნათ C პორტფელის ამონაგებთა მწკრივი. ის იქნება:

10.2 -4.5 13.65 4.75 30.7 34.1 21.3 7.95 0.9 20.95.

სტანდარტული გადახრის მეშვეობით შევადაროთ C პორტფელის რისკიანობა A და B ტრესტებში ინვესტიციათა რისკიანობას. საშუალოს მე-3 და მე-4 თვისებების საფუძველზე შეგვიძლია დავწეროთ, რომ:

$$\bar{x}_A = \frac{1}{2}(\bar{x}_A + \bar{x}_B) = 14\%.$$

ამიტომ, შესწორებული შერჩევითი დისპერსია იქნება:

$$s_C'^2 = (10.2^2 + \cdots + 20.95^2 - 10 \cdot 14^2)/9 = 159.45(\%)^2,$$

შესაბამისად, $s_C' = 12.63\%$.

როგორც ვხედავთ, რისკიანობის მიხედვით A , B და C ინვესტიციები დალაგდა შემდეგი მიმდევრობით:

$$s_B' < s_C' < s_A',$$

ხოლო მათი ამონაგები კი შემდეგი მიმდევრობით:

$$R_A > R_C > R_B.$$

დისპერსიისა და სტანდარტული გადახრის მოძებნის პროცედურა დაჯგუფებული მონაცემებისათვის ანალოგიურია დაჯგუფებული მონაცემებისათვის საშუალოს მოძებნის პროცედურის და ის იყენებს თითოეული კლასის შუაწერტილს.

მაგალითი. მოცემულია კვირის განმავლობაში შემთხვევით შერჩეული 20 მორბენალის მიერ გარბენილი მანძილების სიხშირული განაწილების ცხრილი:

კლასის საზღვრები	სიხშირე
5.5—10.5	1
10.5—15.5	2
15.5—20.5	3
20.5—25.5	5
25.5—30.5	4
30.5—35.5	3
35.5—40.5	2
	$n=20$

ვიპოვოთ დისპერსია და სტანდარტული გადახრა.

ამოხსნა.

ნაბიჯი 1. გავაკეთოთ შემდეგი სახის ცხრილი, ვიპოვოთ თითოეული კლასის შუაწერტილი და შევიტანოთ C სვეტში:

$$x_m=\frac{5.5+10.5}{2}=8,\quad \frac{10.5+15.5}{2}=13,\text{ და ა. შ.}$$

A კლასები	B სიხშირე (f)	C შუაწერტილი (x_m)	D $f\cdot x_m$	E $f\cdot x_m^2$
5.5—10.5	1	8		
10.5—15.5	2	13		
15.5—20.5	3	18		
20.5—25.5	5	23		
25.5—30.5	4	28		
30.5—35.5	3	33		
35.5—40.5	2	38		
	$n=20$			

ნაბიჯი 2. გავამრავლოთ თითოეული კლასის სიხშირე ამავე კლასის შუაწერტილზე და ნამრავლი ჩავწეროთ D სვეტში:

$$1\cdot 8=8,\ 2\cdot 13=26,\text{ და ა. შ. } 2\cdot 38=76.$$

ნაბიჯი 3. გავამრავლოთ თითოეული კლასის სიხშირე ამავე კლასის შუაწერტილის კვადრატზე და ნამრავლი ჩავწეროთ E სვეტში:

$$1\cdot 8^2=64,\ 2\cdot 13^2=338,\text{ და ა. შ. } 2\cdot 38^2=2888.$$

ნაბიჯი 4. ვიპოვოთ **B, D** და **E** სვეტების ჯამი. **B** სვეტის ჯამია n, **D** სვეტის ჯამია $\sum f\cdot x_m$, ხოლო **E** სვეტის ჯამია $\sum f\cdot x_m^2$.

საბოლოოდ ცხრილს ექნება ასეთი სახე:

A კლასები	B სიხშირე (f)	C შუაწერტილი (x_m)	D $f\cdot x_m$	E $f\cdot x_m^2$
5.5—10.5	1	8	8	64
10.5—15.5	2	13	26	338
15.5—20.5	3	18	54	972
20.5—25.5	5	23	115	2645
25.5—30.5	4	28	112	3136
30.5—35.5	3	33	99	3267
35.5—40.5	2	38	76	2888
	$n=20$		$\sum f\cdot x_m=490$	$\sum f\cdot x_m^2=13310$

ნაბიჯი 5. შევიტანოთ მიღებული მონაცემების შესაბამის ფორმულაში და გამოვთვალოთ დისპერსია:

$$s'^2=\frac{\sum\limits_{m=1}^{n} f\cdot x_m^2-[(\sum\limits_{m=1}^{n} f\cdot x_m)^2/n]}{n-1}=\frac{13310-[490^2/20]}{20-1}=68.7\,.$$

ნაბიჯი 5. სტანდარტული გადახრის გამოსათვლელად ამოვიღოთ კვადრატული ფესვი დისპერსიიდან:

$$s'=\sqrt{68.7}\approx 8.3\,.$$

ცხადია, რომ $s=0$, მაშინ და მხოლოდ მაშინ, როდესაც შერჩევის ყველა მნიშვნელობა ტოლია. საჭიროა აღინიშნოს, რომ სტანდარტული გადახრა არ წარმოადგენს გაფანტულობის მდგრად საზომს, ის გაცილებით უფრო მგრძნობიარეა ექსტრემალური დაკვირვებების მიმართ, ვიდრე საშუალო. მართლაც, განვიხილოთ მონაცემები: 34 46 40 57 50. მაშინ მათთვის საშუალო იქნება $\bar{x}=46$. გამოვთვალოთ სტანდარტული გადახრა s'. ჯერ გამოვთვალოთ საშუალოდან გადახრები, გვექნება: -9 0 6 11 4. შესაბამისად, $s'^2=(81+0+36+121+16)/4=63.5$. ამიტომ $s'=\sqrt{63.5}\approx 8.2$

ახლა მონაცემთა მოყვანილ სიმრავლეში მონაცემი 50 შევცვალოთ 200-ით, ანუ განვიხილოთ ახალი მონაცემები: 34 46 40 57 200. ამ შემთხვევაში საშუალო იქნება $\bar{x}=76$, ხოლო სტანდარტული გადახრა კი:

$$s'=\sqrt{(39^2+30^2+36^2+19^2+124^2)/4}=\sqrt{4383.5}\approx 66\,.$$

როგორც ვხედავთ, შედარებით თანაბრად განაწილებული მონაცემებიდან ერთ-ერთი მონაცემის მკვეთრ ცვლილებაზე სტანდარტული გადახრის რეაგირება უფრო ძლიერი აღმოჩნდა, ვიდრე საშუალოსი. ამ ცვლილებით მონაცემები გახდა მარჯვნივ ასიმეტრიული. მკვეთრად ასიმეტრიული, მაგალითად, მარჯვნივ ასიმეტრიული განაწილებებისათვის, მარჯვენა მხარეს გრძელი კუდით, შერჩევითი დისპერსია დიდია და არ იძლევა სასარგებლო ინფორმაციას გაფანტულობის შესახებ.

იმ შემთხვევაში, როდესაც ორ შერჩევას აქვს ერთი და იგივე ზომის ერთეული, მაშინ მათი დისპერსია და სტანდარტული გადახრა შეიძლება პირდაპირ შევადაროთ ერთმანეთს. მაგალითად, დავუშვათ, რომ ავტომობილების დილერს სურს შეადაროს ახალი ავტომობილების რომელიმე ნაწილის პირველად შეცვლამდე მათ მიერ გავლილი მანძილების (მილებში) სტანდარტული გადახრები. დილერმა დაადგინა, რომ რომელიღაც კონკრეტულ წელს აღნიშნული მანძილების სტანდარტული გადახრა "ჯეტასათვის" იყო 442 მილი, ხოლო "კადილაკისათვის" კი – 350 მილი. ამ შემთხვევაში, დილერს შეუძლია თქვას, რომ ავტომანქანების გარბენის გადახრა "ჯეტასათვის" უფრო დიდია. მაგრამ, როგორ უნდა მოიქცეს მენეჯერი, თუ მას სურს შეადაროს განსხვავებული სიდიდეების სტანდარტული გადახრები, მაგალითად, როგორიცაა სამი თვის განმავლობაში გამყიდველების მიერ განხორციელებულ გაყიდვათა რიცხვი და გადახდილი თანხების რაოდენობა.

სტატისტიკას, რომელიც სტანდარტული გადახრების შედარების საშუალებას იძლევა იმ შემთხვევაში, როცა ერთეულები სხვადასხვაა, ეწოდება **ვარიაციის კოეფიციენტი (coefficient of variation)**. *ვარიაციის კოეფიციენტი* არის სტანდარტული გადახრა გაყოფილი საშუალოზე. იგი გამოისახება პროცენტებში და აღინიშნება *CVar* სიმბოლოთი:

შერჩევის შემთხვევაში   პოპულაციის შემთხვევაში

$$CVar = \frac{s}{\bar{x}} \cdot 100\% \qquad\qquad CVar = \frac{\sigma}{\mu} \cdot 100\%.$$

მაგალითი. სამთვიანი პერიოდის განმავლობაში გაყიდული მანქანების საშუალო არის 87, ხოლო სტანდარტული გადახრა კი – 5. გადახდილი თანხების საშუალოა 5225 დოლარი, სტანდარტული გადახრა კი – 773 დოლარი. შევადაროთ ვარიაციები.

ამოხსნა. ვარიაციის კოეფიციენტები შესაბამისად იქნება:

$$CVar = \frac{5}{87} \cdot 100\% = 5.7\% \quad \text{გაყიდვების შემთხვევაში;}$$

$$CVar = \frac{773}{5225} \cdot 100\% = 14.8\% \quad \text{თანხების შემთხვევაში.}$$

ვინაიდან ვარიაციის კოეფიციენტი გადახდილი თანხების შემთხვევაში უფრო დიდია, ამიტომ თანხები უფრო ცვალებადია, ვიდრე გაყიდვები.

როგორც ზემოთ აღვნიშნეთ, მონაცემთა დისპერსია და სტანდარტული გადახრა გამოიყენება მონაცემთა ცვალებადობის ხარისხის დასადგენად. ასე რომ, უფრო დიდი დისპერსია ან სტანდარტული გადახრა მონაცემთა მნიშვნელობების უფრო დიდი გაბნეულობის მაჩვენებელია. მაგალითად, თუ ორ სიდიდეს, რომლებიც იზომება ერთი და იგივე ერთეულებში, აქვს ერთი და იგივე საშუალო, ვთქვათ, 70 და ერთი სიდიდის სტანდარტული გადახრაა 1.5, მაშინ როცა მეორის გადახრაა 10, მაშინ მეორე სიდიდის მონაცემთა მნიშვნელობები უფრო გაფანტულია, ვიდრე პირველი სიდიდის. **ჩებიშევის თეორემა**, რომელიც ეკუთვნის რუს მათემატიკოსს – პ. ლ. ჩებიშევს (1821 – 1894), ზუსტად განსაზღვრავს გაფანტულობის სიგანის პროპორციებს სტანდარტული გადახრის ტერმინებში.

ჩებიშევის თეორემა. მონაცემთა სიმრავლის იმ მონაცემების წილი, რომლებიც მოთავსებული არიან საშუალოდან k სტანდარტული გადახრის ფარგლებში, არის სულ ცოტა $1-1/k^2$, სადაც k არის 1-ზე დიდი რიცხვი (k არ არის აუცილებლად მთელი).

ფორმალურად ეს შედეგი ასე ჩამოყალიბდება: დაკვირვებათა მოცემული $x_1,...,x_n$ სიმრავლისათვის სამართლიანია შემდეგი უტოლობა

$$\frac{N\{x_i : |x_i-\bar{x}| \le ks\}}{n} \ge 1-\frac{1}{k^2}.$$

სხვა სიტყვებით: $[\bar{x}-ks, \bar{x}+ks]$ ინტერვალში მოხვედრილ მონაცემთა ფარდობითი სიხშირე (ანუ ამ ინტერვალში მოხვედრილ მონაცემთა რაოდენობის შეფარდება შერჩევის მოცულობასთან) არანაკლებია $(1-1/k^2)$-ზე.

დამტკიცება. შერჩევითი დისპერსიის განმარტების თანახმად გვაქვს:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{\{x_i:|x_i-\bar{x}|\le ks\}}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{\{x_i:|x_i-\bar{x}|>ks\}}(x_i-\bar{x})^2 \ge$$

$$\ge \frac{k^2s^2}{n}N\{x_i : |x_i-\bar{x}|>ks\}.$$

აქედან, თუ უტოლობის ორივე მხარე გავყოფთ k^2s^2-ზე, მივიღებთ

$$\frac{N\{x_i : |x_i-\bar{x}|>ks\}}{n} \le \frac{1}{k^2},$$

საიდანაც, თავის მხრივ, გვაქვს:

$$\frac{N\{x_i : |x_i-\bar{x}| \le ks\}}{n} = 1-\frac{N\{x_i : |x_i-\bar{x}|>ks\}}{n} \ge 1-\frac{1}{k^2}.$$

ეს თეორემა ამტკიცებს, რომ მონაცემთა მნიშვნელობების სულ ცოტა 3/4 ანუ 75% მდებარეობს (ვარდება) მონაცემთა სიმრავლის საშუალოდან ორი სტანდარტული გადახრის ფარგლებში (მიდამოში). ამ შედეგის სამართლიანობაში დასარწმუნებლად ჩებიშევის თეორემაში $1-1/k^2$ გამოსახულებაში ჩავსვათ $k=2$. მართლაც, $1-1/2^2=1-1/4=3/4=75\%$.

თუ დავუბრუნდებით მაგალითს, რომელშიც სიდიდის საშუალო იყო 70 და სტანდარტული გადახრა – 1.5, მაშინ ამ სიდიდის მონაცემთა მნიშვნელობების სულ ცოტა 3/4 ანუ 75% ვარდება 67-სა და 73-ს შორის. ეს საზღვრები მოძებნილია საშუალოსათვის ორი სტანდარტული გადახრის, შესაბამისად, გამოკლებითა და მიმატებით:

$$70-2\cdot1.5=70-3=67 \text{ და } 70+2\cdot1.5=70+3=73.$$

ანალოგიურად, მეორე სიდიდის მონაცემთა მნიშვნელობების სულ ცოტა 3/4 ანუ 75% თავსდება 50-სა და 90-ს შორის, ვინაიდან:

$$70-2\cdot10=70-20=50 \text{ და } 70+2\cdot10=70+20=90.$$

გარდა ამისა, ჩებიშევის თეორემა ამტკიცებს, რომ მონაცემთა მნიშვნელობების სულ ცოტა 8/9 ანუ 88.89% მოთავსებულია საშუალოდან სამი სტანდარტული გადახრის მიდამოში. ეს შედეგი მიიღება, თუ $1-1/k^2$ გამოსახულებაში ჩავსვამთ $k=3$. მართლაც, $1-1/3^2=1-1/9=8/9=88.89\%$.

პირველი სიდიდის შემთხვევაში, მონაცემთა მნიშვნელობების სულ ცოტა 8/9 ანუ 88.89% მოთავსებულია 65.5-სა და 74.5-ს შორის, რადგანაც:

$$70-3\cdot1.5=70-4.5=65.5 \text{ და } 70+3\cdot1.5=70+4.5=74.5.$$

მაგალითი. გარკვეულ საგრაფოში სახლების ფასების საშუალოა 50000 დოლარი, ხოლო სტანდარტული გადახრა კი – 10000 დოლარი. ვიპოვოთ ფასების გაბნევის დიაპაზონი, რომელშიც მოხვდება ამ საგრაფოში გაყიდული სახლების ფასების სულ ცოტა 75%.

ამოხსნა. ჩებიშევის თეორემის თანახმად მონაცემთა მნიშვნელობების 3/4 ან 75% მოთავსებულია საშუალოს ორი სტანდარტული გადახრის მიდამოში. ვინაიდან:

$$50000-2\cdot10000=50000-20000=30000, \text{ და}$$

$$50000+2\cdot10000=50000+20000=70000,$$

ამიტომ ამ საგრაფოში გაყიდული ყველა სახლის სულ ცოტა 75%-ის გაყიდვის ფასი მოთავსებული იქნება 30000 დოლარიდან 70000 დოლარამდე.

ჩებიშევის თეორემა შეიძლება გამოყენებულ იქნეს მონაცემთა მნიშვნელობების იმ მინიმალური პროცენტის დასადგენად, რომლებიც მოთავსებული არიან ორ მოცემულ მნიშვნელობას შორის.

მაგალითი. კომპანიაში ჩატარებულმა ადგილობრივმა გამოკითხვამ ცხადყო, რომ მგზავრობის ერთიბლივი დანახარჯების საშუალომ ერთ მილზე შეადგინა 0.25 დოლარი, ხოლო სტანდარტულმა გადახრამ კი – 0.02 დოლარი. ჩებიშევის თეორემის გამოყენებით ვიპოვოთ მონაცემთა მნიშვნელობების ის მინიმალური პროცენტი, რომლითაც მნიშვნელობები მოთავსებული არიან 0.20 დოლარსა და 0.30 დოლარს შორის?

ამოხსნა.

ნაბიჯი 1. გამოვაკლოთ საშუალო ზედა საზღვარს:

$$0.30-0.25=0.05;$$

ნაბიჯი 2. k-ს დასადგენად მიღებული სხვაობა გავყოთ სტანდარტულ გადახრაზე:

$$k=\frac{0.05}{0.02}=2.5;$$

ნაბიჯი 3. პროცენტის საპოვნელად გამოვიყენოთ ჩებიშევის თეორემა:

$$1-\frac{1}{k^2}=1-\frac{1}{2.5^2}=1-\frac{1}{6.25}=1-0.16=0.84 \text{ ანუ } 84\%.$$

შესაბამისად, მონაცემთა მნიშვნელობების სულ ცოტა 84% მოთავსებულია 0.20 დოლარსა და 0.30 დოლარს შორის.

ჩებიშევის თეორემა გამოიყენება ნებისმიერი განაწილებისათვის, მისი სახისგან განურჩევლად. მაგალითად, თუ განაწილება არის **ზარის ფორმის (bell-shaped)**, ანუ თუ გვაქვს ე. წ. **ნორმალური (normal)** განაწილება, მაშინ სამართლიანია ე. წ. **ემპირიული წესი (empirical rule)**:

მონაცემთა მნიშვნელობების დაახლოებით 68% მოთავსებულია საშუალოდან ერთი სტანდარტული გადახრის ფარგლებში;

მონაცემთა მნიშვნელობების დაახლოებით 95% მოთავსებულია საშუალოდან ორი სტანდარტული გადახრის ფარგლებში;

მონაცემთა მნიშვნელობების დაახლოებით 99.7% მოთავსებულია საშუალოდან სამი სტანდარტული გადახრის ფარგლებში.

მაგალითად, დავუშვათ, რომ ეროვნულ გამოცდებზე წარმატების მისაღწევად საჭირო ქულების საშუალოა 480, ხოლო სტანდარტული გადახრა – 90. თუ ვიგულისხმებთ, რომ ქულები ნორმალურადაა განაწილებული, მათი დაახლოებით 68% მოთავსებული იქნება 390-სა და 570-ს შორის (ვინაიდან 480-90=390 და 480+90=570). ქულების დაახლოებით 95% ჩავარდება 300-სა და 660-ს შორის (რადგანაც $480-2\cdot 90=300$ და $480+2\cdot 90=660$). და, ბოლოს, ქულათა დაახლოებით 99.7% მოხვდება შუალედში [210, 750] (ვინაიდან $480-3\cdot 90=210$ და $480+3\cdot 90=750$).

ასიმეტრიისა და ექსცესის შერჩევითი კოეფიციენტები. ასიმეტრიის შერჩევითი კოეფიციენტი ეწოდება სიდიდეს

$$\hat{a}_n=\frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}.$$

ასიმეტრიის კოეფიციენტი იძლევა სიხშირეთა ჰისტოგრამის სიმეტრიულობის (ან ასიმეტრიულობის) შესახებ დასკვნის გაკეთების საშუალებას მისი აგების გარეშე. კერძოდ, თუ $\hat{a}_n\neq 0$, შეგვიძლია დავასკვნათ, რომ სიხშირეთა ჰისტოგრამა არაა სიმეტრიული საშუალოს მიმართ; თუ $\hat{a}_n>0$ (შესაბამისად, $\hat{a}_n<0$), მაშინ სიხშირეთა განაწილება მარჯვნივ (შესაბამისად, მარცხნივ) ასიმეტრიულია.

ექსცესის შერჩევითი კოეფიციენტი ეწოდება სიდიდეს

$$\hat{e}_n=\frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4}-3.$$

ექსცესის კოეფიციენტი არის სიხშირეთა ჰისტოგრამის $\bar{x}$-ის მიდამოში ზევით გაწელილობის (ანუ წვეტიანობის) საზომი.

სტანდარტიზაცია

ცენტრალური ტენდენციისა და გაფანტულობის საზომებთან ერთად დამატებით განიხილავენ მდებარეობის ან დისლოკაციის საზომებს. ამ საზომებს განეკუთვნება *სტანდარტიზაცია, პროცენტილები, დეცილები* და *კვარტილები.* ისინი გამოიყენება მონაცემის მნიშვნელობის შედარებითი მდებარეობის დასადგენად მონაცემთა სიმრავლეში. მაგალითად, მონაცემის მნიშვნელობის მდებარეობას შეესაბამება 80-პროცენტილი, ეს იმას ნიშნავს, რომ ამ განაწილების მნიშვნელობების 80% მოცემული მნიშვნელობის ქვევითაა, ხოლო მნიშვნელობების 20% კი – ზევით. *მედიანა* არის მნიშვნელობა, რომელიც შეესაბამება 50-პროცენტილს, ვინაიდან მნიშვნელობების ნახევარი მის ქვევითაა, ხოლო მეორე ნახევარი კი—ზევით.

არსებობს ძველი გამოთქმა, რომლის თანახმადაც: "თქვენ ვერ შეადარებთ ერთმანეთს ვაშლებსა და ფორთოხლებს". მაგრამ სტატისტიკოსმა

ეს უნდა მოახერხოს გარკვეული ხარისხით. დავუშვათ, რომ სტუდენტმა მუსიკის გამოცდაში მოაგროვა 90 ქულა, ხოლო ინგლისურის გამოცდაში კი – 45 ქულა. ამ ქულების პირდაპირი შედარება შეუძლებელია, ვინაიდან ეს გამოცდები ვერ იქნება ექვივალენტური კითხვების რაოდენობისა და მნიშვნელოვნების თვალსაზრისით. მიუხედავად ამისა, ჩვენ შეგვიძლია თითოეული მათგანი შევადაროთ ერთმანეთის მსგავს შედარებით სტანდარტებს. ეს შედარება იყენებს შერჩევის საშუალოსა და სტანდარტულ გადახრას და მას ეწოდება **სტანდარტიზაცია** ან z **ქულა (standard score or z score)** და აღინიშნება z ასოთი.

მნიშვნელობის *სტანდარტიზაცია* ანუ z ქულა მიიღება მნიშვნელობისაგან საშუალოს გამოკლებითა და მიღებული შედეგის სტანდარტულ გადახრაზე გაყოფით:

$$z = \frac{\text{მნიშვნელობა - საშუალო}}{\text{სტანდარტული გადახრა}}.$$

შესაბამისად, შერჩევის შემთხვევაში გვაქვს: $z = \frac{x - \bar{x}}{s}$, ხოლო პოპულაციის შემთხვევაში, კი $z = \frac{x - \mu}{\sigma}$.

z ქულა გვიჩვენებს სტანდარტული გადახრების რა რაოდენობა თავსდება საშუალოდან ქვევით ან ზევით მდებარე მონაცემამდე.

მაგალითი 1. სტუდენტმა მოაგროვა 65 ქულა კალკულუსის ტესტში, რომლის საშუალოა 50, ხოლო სტანდარტული გადახრა კი – 10. იგივე სტუდენტმა 30 ქულა მოაგროვა ისტორიის ტესტში, რომლის საშუალოა 25, ხოლო სტანდარტული გადახრა – 5. შევადაროთ სტუდენტის შედარებითი პოზიციები ამ ორ ტესტში.

ამოხსნა. ვიპოვოთ თითოეულ ტესტში სტუდენტის მიერ მოგროვილი ქულის შესაბამისი z ქულა. კალკულუსის შემთხვევაში z ქულა იქნება:

$$z = \frac{x - \bar{x}}{s} = \frac{65 - 30}{10} = 1.5,$$

ხოლო ისტორიის ტესტის შემთხვევაში კი – $z = (30 - 25)/5 = 1.0$.

ვინაიდან, z ქულა კალკულუსის შემთხვევაში უფრო დიდია, სტუდენტის შედარებითი პოზიცია კალკულუსის ჯგუფში უფრო მაღალია, ვიდრე მისი შედარებითი პოზიცია ისტორიის ჯგუფში.

შევნიშნოთ, რომ თუ z ქულა დადებითია, მაშინ შესაბამისი ქულა საშუალოზე ზევითაა. თუ z ქულა ნულია, მაშინ შესაბამისი ქულა ემთხვევა საშუალოს, და თუ z ქულა უარყოფითია, მაშინ შესაბამისი ქულა საშუალოზე ქვევითაა.

მაგალითი 2. ვიპოვოთ ქვევით მოყვანილი ტესტების z ქულა და შევადაროთ ერთმანეთს:

Test A	$x = 38$	$\bar{x} = 40$	$s = 5$
Test B	$x = 94$	$\bar{x} = 100$	$s = 10$

ამოხსნა. A ტესტის შემთხვევაში z ქულა იქნება:

$$z = \frac{x - \bar{x}}{s} = \frac{38 - 40}{5} = -0.4,$$

ხოლო B ტესტის შემთხვევაში კი – $z = (94 - 100)/10 = -0.6$.

შესაბამისად, ქულა A ტესტის შემთხვევაში შედარებით მაღალია, ვიდრე ქულა B ტესტის შემთხვევაში.

იმ შემთხვევაში, როცა ცვლადი სიდიდის ყველა მონაცემი გადაყვანილია z ქულებში, მიღებულ განაწილებას საშუალო ექნება 0, ხოლო სტანდარტული გადახრა კი – 1. სინამდვილეში z ქულა არის სტანდარტულ გადახრათა ის რიცხვი, რომლის საშუალოზე მიმატებით ან გამოკლებით მიიღება რეალური ქულა. მაგალით 1–ში კალკულუსში მიღებული 65 ქულა ფაქტიურად არის 1. 5 სტანდარტული გადახრა (10) საშუალოს (50) ზევით: 65 = 50 + 1.5·10.

სკალირების მეთოდი

მას შემდეგ რაც ჩვენი ქვეყანა გადავიდა ერთიან ეროვნულ გამოცდებზე საზოგადოების დიდ ინტერესს იწვევს ე. წ. სკალირების მეთოდი. როგორც ცნობილია, ერთიანი ეროვნული გამოცდების დროს, გამოცდები თითოეულ საგანში რამდენიმე სესიად მიმდინარეობს, რაც განაპირობებს საგამოცდო ტესტების სხვადასხვა ვარიანტის გამოყენების აუცილებლობას. შესაძლოა, კონკრეტული საგნის ტესტის სხვადასხვა ვარიანტი, სირთულის ხარისხით, უმნიშვნელოდ განსხვავდებოდეს ერთმანეთისაგან. შესაბამისად, სკალირება გამოიყენება აბიტურიენტებისათვის თანაბარი საკონკურსო პირობების უზრუნველყოფის მიზნით.

აბიტურიენტის მიერ მიღებული შედეგებიდან განისაზღვრება ის მახასიათებლები (საშუალო, სტანდარტული გადახრა), რომლის საფუძველზე მოხდება წრფივი სკალირება. მართალია, სკალირებით მიღებული საკონკურსო ქულა განსხვავებული იქნება გამოცდებში მიღებული ქულისაგან, მაგრამ სწორედ ეს მეთოდი იძლევა როგორც ტესტების სხვადასხვა ვარიანტში, ისე სხვადსხვა საგანში მიღებული ქულების ერთმანეთთან შედარების საშუალებას და ამდენად მაქსიმალურად იცავს ცალკეული აბიტურიენტის ინტერესებს. გარდა ამისა, სკალირება ერთიან სისტემაში აქცევს აბიტურიენტების მიერ სხვადსხვა წლებში მიღებულ შეფასებებს, რაც თავის მხრივ, საშუალებას აძლევს შეფასებისა და გამოცდების ერთიან ცენტრს აბიტურიენტს მომდევნო წელსაც შეუნარჩუნოს წინა წელს მიღებული შეფასება.

სკალირება მსოფლიოში აპრობირებული უნივერსალური მეცნიერული მეთოდია, რომელიც მოიცავს ორ საფეხურს: პირველ ეტაპზე ხდება ერთი საგამოცდო ტესტის სხვადასხვა ვარიანტში აბიტურიენტის მიერ მიღებულ ქულებს შორის შესაბამისობის დადგენა (გაიგივება), ხოლო მეორე ეტაპზე – თითოეულ საგამოცდო საგანში გაიგივებული ქულების სტანდარტიზაცია და საერთო სკალაზე გადატანა.

სირთულის თვალსაზრისით ტესტის ვარიანტებს შორის უმნიშვნელო განსხვავებაც კი არათანაბარ პირობებში აყენებს აბიტურიენტებს. ამ უთანაბრობის აღმოსაფხვრელად მიმართავენ სწორედ თითოეული საგნის

საგამოცდო ტესტის სხვადასხვა ვარიანტის ერთმანეთთან შედარების მეთოდს, რომელიც გულისხმობს კუმულაციური ქულების (აბიტურიენტის რეიტინგი თანაბარ პირობებში მყოფ აბიტურიენტებს შორის) გამოთვლის გზით ერთი საგამოცდო ტესტის სხვადასხვა ვარიანტში აბიტურიენტების მიერ მიღებული ქულების ერთმანეთთან შედარების შესაძლებლობას. ამ პროცედურით ხდება საგამოცდო ტესტის სხვადასხვა ვარიანტში მიღებულ ქულებს შორის შესაბამისობის დადგენა, ანუ მათი გაიგივება, რაც საშუალებას იძლევა ტესტის ვარიანტებს შორის სირთულის ხარისხის მიხედვით არსებული უმნიშვნელო განსხვავების შემთხვევაშიც კი მაქსიმალურ სამართლიანობას მივაღწიოთ.

მოვიყვანოთ პირველი ეტაპის პროცედურული აღწერა. დავუშვათ, რომ ჩვენ გაგვაჩნია ერთი კონკრეტული საგნის საგამოცდო ტესტის რამდენიმე ვარიანტის მიხედვით აბიტურიენტების მიერ მიღებული ქულების მონაცემები. პირველ რიგში ამ საგნის ყოველი X ვარიანტისათვის და ამ ვარიანტის ყოველი x ქულისათვის უნდა ვიპოვოთ იმ ქულების პროცენტული რაოდენობა, რომელიც ამ ვარიანტში x ქულაზე ნაკლებია (ეს ფაქტიურად მოცემულ X ვარიანტში დაწერილი ქულების ვარიაციულ მწკრივში x ქულის შესაბამისი პროცენტული რანგია). ეს რიცხვი აღვნიშნოთ $f_X(x)$ სიმბოლოთი (ცხადია, რომ $f_X(101)=100$). თუ N_X სიმბოლოთი აღვნიშნავთ იმ აბიტურიენტების რაოდენობას, რომლებმაც გამოცდა ჩააბარეს X ვარიანტის მიხედვით, ხოლო N_X^x სიმბოლოთი კი – იმ აბიტურიენტების რაოდენობას, რომლებმაც X ვარიანტში მიიღეს x ქულაზე ნაკლები შეფასება, მაშინ ცხადია, რომ

$$f_X(x)=\frac{N_X^x}{N_X}\cdot 100\%,\ \ 0\le x\le 100,\ \ 0\le f_X(x)\le 100.$$

საბაზისოდ ავირჩიოთ ამ ტესტის ის ვარიანტი, რომელშიც საშუალო ქულა (ამ ვარიანტში აბიტურიენტების მიერ მიღებული ქულების საშუალო არითმეტიკული) ყველაზე მაღალია და ამ ვარიანტში აბიტურიენტების მიერ მიღებული ქულები დავტოვოთ უცვლელად. გარკვეულობისთვის ეს ვარიანტი აღვნიშნოთ A-თი. იმისათვის, რომ რომელიმე სხვა B ვარიანტის რომელიმე b ქულა გადავიყვანოთ საბაზისო A ვარიანტის შესაბამის ქულაში, საჭიროა A ვარიანტში მოვძებნოთ ისეთი მაქსიმალური a ქულა, რომლისთვისაც შესრულდება თანაფარდობა:

$$f_A(a)\le f_B(b)<f_A(a+1)$$

და b ქულის შესაბამის (ანუ მასთან გაიგივებულ) ქულად გამოვაცხადოთ:

$$\bar{b}=\min\{a+\frac{f_B(b)-f_A(a)}{f_A(a+1)-f_A(a)},100\}.$$

ცხადია, რომ b ქულასთან გაიგივებული $\bar{b}$ ქულა A ვარიანტში მოთავსებული იქნება $[a,a+1)$ ინტერვალში ($a\le\bar{b}<a+1$).

მოვიყვანოთ საილუსტრაციო მაგალითი: დავუშვათ, რომ B ვარიანტის $b=95$ ქულის შესაბამისი $f_B(b)=f_B(95)=88\%$, და დავუშვათ, რომ A ვარიანტში არსებობს ისეთი მაქსიმალური $a=85$ ქულა, რომლისთვისაც:

აშინ ზემოთმოყვანილი წესის თანახმად 95-თთან გაიგივებული ქულა იქნება:

$$\overline{95} = \min\{85 + \frac{88-88}{89-88}, 100\} = \min\{85, 100\} = 85 .$$

თუ კი ყველა დანარჩენ პირობებში $f_A(85) = 87.5\%$-ს, მაშინ მივიღებდით, რომ:

$$\overline{95} = \min\{85 + \frac{88-87.5}{89-87.5}, 100\} = \min\{85 + \frac{0.5}{1.5}, 100\} = \min\{85 + 0.33, 100\} = 85.33 .$$

შენიშვნა. თუ აღმოჩნდა, რომ

$$f_A(a+1) = f_A(a+2) = \cdots = f_A(\tilde{a}) ,$$

სადაც $\tilde{a} \le 101$ და ის უდიდესია ასეთ რიცხვებს შორის, მაშინ b-ს შესაბამისი ქულა A ვარიანტში იქნება:

$$\bar{b} = \min\{a + (\tilde{a} - a) \cdot \frac{f_H(b) - f_A(a)}{f_A(\tilde{a}) - f_A(a)}, 100\} .$$

გადავიდეთ ახლა მეორე ეტაპზე – სკალირებული ქულის დადგენა: სკალირების პროცედურა სხვადასხვა საგამოცდო საგანში მიღებული ქულების ერთიან სკალაზე განთავსების საშუალებას იძლევა. ამ მიზნით თითოეული საგამოცდო საგნისათვის უნდა გამოვთვალოთ საშუალო ქულა და სტანდარტული გადახრა, და შემდეგ თითოეული აბიტურიენტის სტანდარტიზებული ანუ ე. წ. z ქულა, რომელიც გვიჩვენებს, თუ რამდენი სტანდარტული ერთეულით მეტი ან ნაკლებია ესა თუ ის კონკრეტული ქულა შესაბამისი საგამოცდო საგნის საშუალო ქულაზე:

$$z = \frac{x - \mu}{\sigma} ,$$

სადაც x – კონკრეტულ საგანში გაიგივების შედეგად მიღებული აბიტურიენტის ქულაა, μ – მოცემული საგამოცდო საგნის გაიგივებული ქულების საშუალო ქულაა, ხოლო σ – მოცემული საგამოცდო საგნის გაიგივებული ქულების სტანდარტული გადახრა.

ყველა საგამოცდო საგანში მიღებული ქულების ერთიან სკალაზე განთავსება ხდება წრფივი გარდაქმნის საშუალებით. გამოცდების სკალირებული შედეგები დამოკიდებულია ამ გარდაქმნის კოეფიციენტების მნიშვნელობაზე. მიმდინარე წლის გამოცდების შედეგების შედარებას გასული წლის შედეგებთან (ანუ გასულ წელს გამოცდა ჩაბარებული აბიტურიენტების მონაწილეობას მიმდინარე წლის კონკურსში) აზრი აქვს მხოლოდ იმ შემთხვევაში, როცა წრფივი გარდაქმნის კოეფიციენტები ერთი და იგივეა. ამიტომ ყოველ წელს თითოეულ საგამოცდო საგანში აბიტურიენტის z ქულა სტანდარტულ სკალაზე გადაყავთ შემდეგი წრფივი გარდაქმნით:

$$\textit{სკალირებული ქულა} = 15 \times z + 150 .$$

ვინაიდან საშუალო z ქულა ნულია, ამიტომ თითოეულ საგანში აბიტურიენტების საშუალო სკალირებული ქულა იქნება 150, ხოლო ამ ქულების სტანდარტული გადახრა 15. ამიტომ, ე. წ. სამის სტანდარტული გადახრის წესის თანახმად, სკალირებული ქულა დიდი ალბათობით მოთავსებულია ინტერვალში [105, 195]

თავი IX
დაწყვილებული მონაცემები, კორელაცია

აქამდე ჩვენ ვიხილავდით მონაცემებს, რომლებიც წარმოადგენდა რაიმე ერთი მახასიათებლის დაკვირვებულ მნიშვნელობებს. მრავალი პრობლემის კვლევისას, კერძოდ, როდესაც შეისწავლება კავშირი ორ (ან რამდენიმე) მახასიათებელს შორის, ჩვენ უკვე საქმე გვაქვს ისეთ მონაცემებთან, რომლებიც წარმოადგენს ამ მახასიათებლების ერთდროულად დაკვირვებულ (გაზომილ) მნიშვნელობათა სიმრავლეს. მაგალითად, არის თუ არა ეროვნული გამოცდების ჩაბარების ხარისხი დამოკიდებული აბიტურიენტის სოციალურ მდგომარეობაზე, რელიგიურ კუთვნილებაზე, საცხოვრებელ ადგილზე ან სხვა ფაქტორზე. სადაზღვევო კომპანიის მენეჯერს აინტერესებს არის თუ არა ავტო-საგზაო შემთხვევათა რაოდენობა დამოკიდებული ავტომანქანის მფლობელის ასაკზე, სქესზე, მანქანის გამოშვების წელზე, მანქანის ტექნიკურ მახასიათებლებზე (მაქსიმალური სიჩქარე, ძრავის კუბატურა, სამუხრუჭე მანძილი, საჭის მდებარეობა) და სხვა.

ხშირ შემთხვევაში მკვლევარი დაინტერესებულია დაადგინოს ორ სიდიდეს შორის კავშირი არსებობს თუ არა. ამ მიზნის მისაღწევად, მკვლევარმა უნდა შეაგროვოს მონაცემები ორივე მახასიათებლებზე ერთდროული დაკვირვებების შედეგად და მიღებული მონაცემები დააწყვილოს ერთმანეთთან (უფრო ზუსტად, ორი დამკვირვებელი ერთდროულად ატარებს დაკვირვებებს და მიღებულ მონაცემებს აწყვილებს ერთმანეთთან). მაგალითად, თუ ჩვენ გვაინტერესებს კავშირი ჰაერის ტემპერატურასა და ატმოსფერულ წნევას შორის, მაშინ დროის ერთი და იმავე მომენტში პირველი დამკვირვებელი ზომავს ჰაერის ტემპერატურას, ხოლო მეორე – ატმოსფერულ წნევას. პირველი დაკვირვების ან გაზომვის ობიექტს (ამ შემთხვევაში – ტემპერატურას) უწოდებენ *დამოუკიდებელ სიდიდეს*, და აღნიშნავენ x-ით, ხოლო მეორე დაკვირვების ან გაზომვის ობიექტს (ამ შემთხვევაში – ატმოსფერულ წნევას) უწოდებენ *დამოკიდებულ სიდიდეს* და აღნიშნავენ y-ით. ამა თუ იმ ორ მახასიათებელს (ან ორ მაჩვენებელს) შორის კავშირის შესწავლის მიზნით ამ მახასიათებლების დაკვირვებული მნიშვნელობების სიმრავლე წარმოვადგინოთ წყვილების შემდეგი სიმრავლის სახით: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). ამ მონაცემების მიხედვით აგებენ ე. წ. *გაბნევის დიაგრამას* შემდეგი წესით: სიბრტყეზე, კოორდინატთა მართკუთხა სისტემაში, მონიშნავენ წერტილებს, რომელთა კოორდინატებია (x_i, y_i), $i = 1, 2, ..., n$.

გაბნევის გრაფიკი ან გაბნევის დიაგრამა წარმოადგენს მონაცემთა მნიშვნელობების დალაგებული წყვილების გრაფიკს (დიაგრამას), რომელიც გამოიყენება ორ სიდიდეს შორის არსებული კავშირის დასადგენად.

მაგალითი 1. მკვლევარს აინტერესებს დაადგინოს არსებობს თუ არა კავშირი ველოსიპედით წვიმიან ამინდში მომხდარ მსუბუქ ავარიათა რიცხვსა და სიკვდილით (ფატალური შედეგით) დამთავრებულ ავარიათა რიცხვს შორის. ქვემოთ მოყვანილია 10 წლის განმავლობაში წვიმიან ამინდში ველოსიპედით მომხდარი ავარიების რიცხვის ცხრილი ორივე შემთხვევაში:

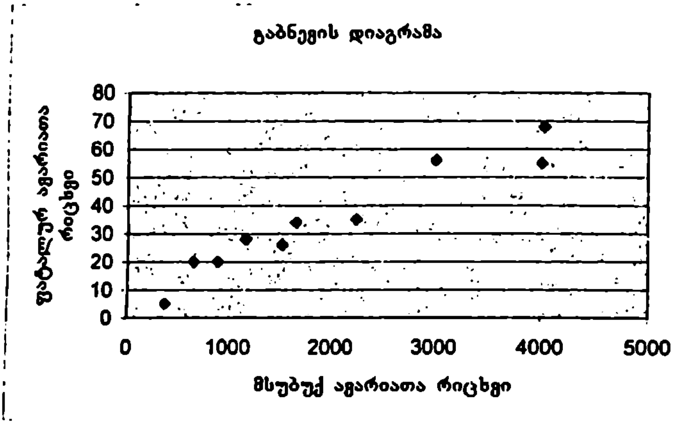
მსუბ. ავარიებ-ბის რიცხვი, x	376	650	884	1162	1513	1650	2236	3002	4028	4010
ფატ. ავარიებ-ბის რიცხვი, y	5	20	20	28	26	34	35	56	68	55

ავაგოთ გაბნევის დიაგრამა ამ მონაცემებისათვის.

ამოხსნა.

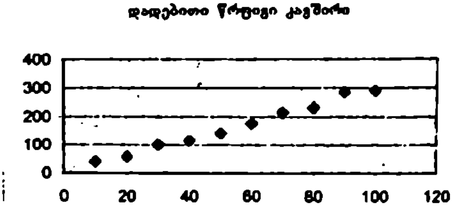
ნაბიჯი 1. სიბრტყეზე ავაგოთ მართკუთხა კოორდინატთა სისტემა და მოვნიშნოთ x და y ღერძები.

ნაბიჯი 2. ავაგოთ წერტილთა წყვილები საკოორდინატო სიბრტყეზე ისე როგორც ეს ქვემოთაა ნაჩვენები:

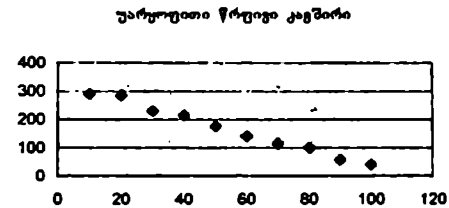


x და y სიდიდეებს შორის შესაძლებელია არსებობდეს რამოდენიმე განსხვავებული ტიპის დამოკიდებულება. ამ დამოკიდებულების გარკვევა (იდენტიფიცირება) შესაძლებელია საკოორდინატო სიბრტყეზე წერტილების განლაგების სტრუქტურაზე დაყრდნობით. ქვემოთ მოყვანილი იქნება წერტილების სტრუქტურული განლაგების სხვადასხვა ტიპები და შესაბამისი დამოკიდებულების სახეები.

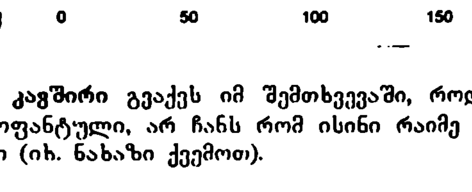
1. დადებითი წრფივი კავშირი არსებობს (გვაქვს) იმ შემთხვევაში, როდესაც წერტილები დაახლოებით ლაგდება (თავმოყრილია, კონცენტრირებულია) აღმავალი სწორი ხაზის ირგელივ (მიდამოში) და ერთდროულად ორივე x და y სიდიდის მნიშვნელობები ზრდადია (იხ. ნახაზი ქვემოთ). ასეთი კავშირი გვაქვს მაშინ, როდესაც x სიდიდის ზრდასთან ერთად იზრდება y სიდიდეც.



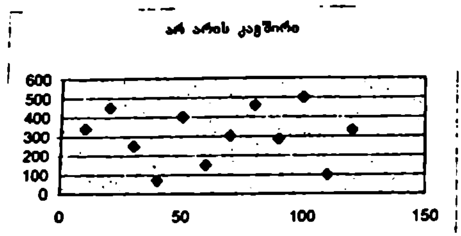
2. **უარყოფითი წრფივი კავშირი** არსებობს (გვაქვს) იმ შემთხვევაში, როდესაც წერტილები დაახლოებით ლაგდება (თავმოყრილია, კონცენტრირებულია) დაღმავალი სწორი ხაზის ირგვლივ (იხ. ნახაზი ქვემოთ). ასეთი კავშირი გვაქვს მაშინ, როდესაც x სიდიდის ზრდასთან ერთად კლებულობს y სიდიდე და პირიქით – x სიდიდის კლებასთან ერთად იზრდება y სიდიდე.



3. **არაწრფივი კავშირი** არსებობს (გვაქვს) იმ შემთხვევაში, როდესაც წერტილები დაახლოებით ლაგდება (თავმოყრილია, კონცენტრირებულია) არაწრფივი (მრუდი) ხაზის (წირის) ირგვლივ (იხ. ნახაზი ქვემოთ). ეს დამოკიდებულება აღიწერება შესაბამისი წირის ბუნებით.



4. **არ არის კავშირი** გვაქვს იმ შემთხვევაში, როდესაც წერტილები უწესრიგოდაა მიმოფანტული, არ ჩანს რომ ისინი რაიმე წირის ირგვლივაა კონცენტრირებული (იხ. ნახაზი ქვემოთ).



ზემოთ მოყვანილ მაგალითში x (მსუბუქ ავარიათა რიცხვი) და y (ფატალურ ავარიათა რიცხვი) სიდიდეებს შორის ადგილი აქვს წრფივ დადებით კავშირს (დამოკიდებულებას). სხვა სიტყვებით, როგორც კი სველ ასფალტზე ველოსიპედით მსუბუქ ავარიათა რიცხვი იწყებს ზრდას , მაშინ იმ ავარიების რიცხვი, სადაც დადგა სასიკვდილო შედეგი, აგრეთვე მატულობს.

წირს, რომელიც აღწერს კავშირს (დამოკიდებულებას) x და y სიდიდეების მნიშვნელობებს შორის *მისადაგების წირი* ეწოდება. ამ წირის პოვნის საკითხებს სწავლობს მათემატიკური სტატისტიკის ნაწილი, რომელსაც *რეგრესიული ანალიზი* ეწოდება. აქ მხოლოდ შევნიშნავთ, რომ წირის მოსაძებნად იყენებენ ე. წ. *უმცირეს კვადრატთა მეთოდს*, რომლის თანახმადაც იძებნება ისეთი წირი $y=f(x)$, რომლისთვისაც სრულდება პირობა:

გამოსახულება $\sum_{i=1}^{n}[y_i - f(x_i)]^2$ ***აღწევს თავის მინიმუმს,***

ანუ წირი, რომლიდანაც დამოუკიდებელი x სიდიდის დაკვირვებული x_i მნიშვნელობების შესაბამისი დაკვირვებული y_i მნიშვნელობების თეორიული $f(x_i)$ მნიშვნელობებისაგან ვერტიკალური გადახრების კვადრატების ჯამი მინიმალურია.

თუ მაგალითად, სიდიდეებს შორის გვაქვს წრფივი კავშირი, მაშინ ბუნებრივია მისადაგების წირი ვეძებოთ წრფივ ფუნქციებს შორის, ანუ $f(x)=ax+b$. წრფივი ფუნქციის მოძებნა ნიშნავს a და b კოეფიციენტების პოვნას. უმცირეს კვადრატთა მეთოდის თანახმად უნდა ვიპოვოთ ისეთი a და b კოეფიციენტები, რომ გამოსახულებამ – $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ მიიღოს თავისი მინიმალური მნიშვნელობა. ფუნქციის ექსტრემუმის პოვნის წესის თანახმად, აღნიშნული გამოსახულება უნდა გავაწარმოოთ როგორც a-ს, ისე b-ს მიმართ, მიღებული წარმოებულები გავუტოლოთ ნულს და ამოვხსნათ მიღებული განტოლებათა სისტემა a და b-ს მიმართ. საბოლოოდ, მივიღებთ, რომ

$$a=\frac{n(\sum_{i=1}^{n}x_iy_i)-(\sum_{i=1}^{n}x_i)\cdot(\sum_{i=1}^{n}y_i)}{n(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)^2} \text{ და } b=\frac{(\sum_{i=1}^{n}y_i)\cdot(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)\cdot(\sum_{i=1}^{n}x_iy_i)}{n(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)^2}.$$

დავალება 1. დაწერეთ და ამოხსენით ექსტრემუმის მოსაძებნი განტოლებათა სისტემა (გამოიყვანეთ ზემოთ დაწერილი ფორმულები a და b კოეფიციენტებისათვის).

იმ შემთხვევაში, როდესაც x და y სიდიდეებს შორის არსებობს წრფივი კავშირი $y = ax + b$, მაშინ ამბობენ რომ x და y სიდიდეებს შორის *წრფივი რეგრესიული კავშირია* და $y = ax + b$ წრფეს, სადაც a და b კოეფიციენტები გამოითვლება ზემოთმოყვანილი ფორმულებით, უწოდებენ *რეგრესიის წრფის განტოლებას.*

დავალება 2. გამოიყენეთ რეგრესიის წრფის განტოლება (ანუ გამოთვალეთ a და b კოეფიციენტები) მაგალით 1-ში.

ახდენს თუ არა გავლენას მტერიანი ქარიშხალი სასუნთქი გზების ჯანმრთელობაზე? ცნობილია, რომ ვაშინგტონის სამხრეთ აღმოსავლეთით სეზონურად ადგილი აქვს მტერიან ქარიშხლებს. რამოდენიმე მკვლევარმა გადაწყვიტა დაედგინა როგორ გავლენას ახდენს ეს ქარიშხალი ამ არეალში მცხოვრები ადამიანების სასუნთქი გზების ჯანმრთელობაზე. ამ მიზნით ისინი აკვირდებოდნენ იყო თუ არა კავშირი ქარიშხალის ამოვარდნის შემთხვევაში ჰაერში მტერისა და ქვიშის ნაწილაკების რაოდენობასა და სასუნთქი გზების სამკურნალო დაწესებულებებში გაუთვალისწინებელ მიმართვების რაოდენობასთან ვაშინგტონის სამხრეთ აღმოსავლეთით მდებარე სამ სახელმწიფო ჰოსპიტალში. ქვემოთ გადმოცემული კორელაციური ანალიზის მეთოდის გამოყენებით მკვლევარებმა დაასკვნეს, რომ მტვრის ქარიშხალი გავლენას ახდენდა ადგილობრივი მცხოვრებლების ჯანმრთელობის მდგომარეობაზე.

სტატისტიკური დასკვნების თეორიის ერთ-ერთი მნიშვნელოვანი მიმართულებაა იმის გარკვევა (დადგენა), არსებობს თუ არა რაიმე დამოკიდებულება (კავშირი) ორ ან მეტ რაოდენობრივ ან თვისებრივ სიდიდეებს შორის. მაგალითად, ბიზნესში მოღვაწე ადამიანს აინტერესებს იცოდეს მოცემულ თვეში გაყიდვების რაოდენობა არის თუ არა დამოკიდებული ფირმის მიერ ამ თვეში გაწეული რეკლამების რაოდენობასთან; პედაგოგს აინტერესებს დაადგინოს სტუდენტის მეცადინეობის საათების რაოდენობა როგორაა დაკავშირებული მის მიერ კონკრეტულ გამოცდაზე მიღებულ ქულასთან; ფსიქოლოგს აინტერესებს დაადგინოს მარტოობასა და დეპრესიას შორის კავშირი; სოციოლოგს აინტერესებს გამოიკვლიოს ამომრჩევლის მიერ გაკეთებული არჩევანის დამოკიდებულება წინაასარჩევნო დაპირებებზე; სამედიცინო სფეროს მუშაკებს აინტერესებთ პასუხი შემდეგი ტიპის კითხვებზე: "არის თუ არა დაკავშირებული გულის დაავადება კოფეინთან?" "არის თუ არა კავშირი ადამიანის ასაკსა და მის არტერიულ წნევას შორის?"; ზოოლოგს აინტერესებს არის თუ არა დაბადებისას კონკრეტული ცხოველის წონა დაკავშირებული ამ ცხოველის სიცოცხლის ხანგრძლივობასთან. ეს არის მხოლოდ რამოდენიმე იმ მრავალ კითხვათაგან, რომელზეც პასუხის გაცემა შესაძლებელია კორელაციური და რეგრესიული ანალიზის მეთოდების გამოყენებით. კორელაციური ანალიზი წარმოადგენს სტატისტიკურ მეთოდს, რომელიც გამოიყენება იმის დასადგენად არსებობს თუ არა დამოკიდებულება (კავშირი) სიდიდეებს შორის (ეს დამოკიდებულება შეიძლება იყოს დადებითი ან უარყოფითი). რეგრესიული ანალი-

ზი წარმოადგენს სტატისტიკურ მეთოდს, რომელიც გამოიყენება სიდიდეებს შორის კავშირის ბუნების აღსაწერად (ეს კავშირი შეიძლება იყოს წრფივი ან არაწრფივი).

ჩვენ უნდა გავცეთ პასუხი შემდეგ კითხვებზე:

1. არის თუ არა ორი ან მეტი სიდიდე დამოკიდებული (დაკავშირებული)?
2. თუ ეს ასეა, მაშინ რამდენად ძლიერია ეს კავშირი?
3. რა ტიპის კავშირი არსებობს?
4. რა ტიპის პროგნოზის გაკეთება შეიძლება ამ კავშირიდან გამომდინარე?

პირველ ორ კითხვაზე პასუხის გასაცემად, სტატისტიკაში გამოიყენება რიცხვითი საზომი, რათა დადგინდეს არის თუ არა ორი ან მეტი სიდიდე დაკავშირებული და აგრეთვე იმის დასადგენად, თუ რამდენად ძლიერია ეს დამოკიდებულება ამ სიდიდეებს შორის. ამ საზომს **კორელაციის კოეფიციენტი (correlation coefficient)** ეწოდება. მაგალითად, გულის ავადმყოფობას ხელს უწყობს მრავალი სიდიდე (ფაქტორი), რომელთა შორისაა მოწევა, მემკვიდრეობითობა, ასაკი, სტრესები, კვების რეჟიმი და სხვა. ამ სიდიდეებიდან ზოგიერთი უფრო მნიშვნელოვანია, ვიდრე დანარჩენი. შესაბამისად, ექიმი რომ დაეხმაროს პაციენტს, მან უნდა იცოდეს ამ ფაქტორებიდან რომელია უფრო მნიშვნელოვანი.

მესამე კითხვაზე პასუხის გასაცემად, მკვლევარმა უნდა დაადგინოს რა ტიპის დამოკიდებულებას აქვს ადგილი. არსებობს ორი ტიპის დამოკიდებულება: მარტივი და რთული (შედგენილი). *მარტივი დამოკიდებულების* შემთხვევაში შეისწავლება კავშირი მხოლოდ ორ სიდიდეს შორის. მაგალითად, მენეჯერს სურს იცოდეს გაყიდვების აგენტის მიერ კომპანიაში ნამუშევარი წლების რაოდენობა არის თუ არა რაიმე კავშირში ამ აგენტის მიერ გაყიდული საგნების რაოდენობასთან. ამ შემთხვევაში საქმე გვაქვს მარტივ დამოკიდებულებასთან, ვინაიდან აქ მონაწილეობს მხოლოდ ორი სიდიდე: წლების რაოდენობა და გაყიდვების რიცხვი.

რთული დამოკიდებულების შემთხვევაში საქმე გვაქვს ორზე მეტ სიდიდესთან. მაგალითად, პედაგოგს სურს გამოიკვლიოს დამოკიდებულება კოლეჯში სტუდენტის წარმატებულ სწავლასა და ისეთ ფაქტორებს შორის როგორიცაა: სტუდენტის მიერ სწავლაზე დახარჯული საათების რაოდენობა, სტუდენტის საშუალო ქულა და სტუდენტის ყოფაქცევა სკოლაში. ამ შემთხვევაში საქმე გვაქვს რთულ დამოკიდებულებასთან.

მარტივი დამოკიდებულება შეიძლება იყოს აგრეთვე დადებითი ან უარყოფითი. *დადებით დამოკიდებულებას* ადგილი აქვს თუ სიდიდეები ერთდროულად იზრდებიან ან კლებულობენ. მაგალითად, არსებობს კავშირი ადამიანის სიმაღლესა და წონას შორის, და ეს კავშირი დადებითია, ვინაიდან საზოგადოდ უფრო მაღალ ადამიანს უფრო მეტი წონა აქვს. *უარყოფით დამოკიდებულებას* ადგილი აქვს, თუ როცა ერთი სიდიდე იზრდება, მაშინ მეორე კლებულობს, და პირიქით. მაგალითად, თუ ჩვენ დავაკვირდებით 60 წელს გადაცილებული ადამიანების ძლიერებას, მაშინ დავინახავთ, რომ ასაკის ზრდასთან ერთად, საზოგადოდ, ადამიანის სიძლიერე კლებულობს.

და ბოლოს, მეოთხე კითხვა: რა ტიპის პროგნოზი შეიძლება გაკეთდეს? პროგნოზი კეთდება ადამიანის მოღვაწეობის ნებისმიერ სფეროში და ყოველდღიური მოთხოვნებიდან გამომდინარე. მაგალითად, ამინდის პროგნოზი, აქციათა კურსის პროგნოზი, გაყიდვების პროგნოზი, მოსავლის პროგნოზი, საწვავის ფასის პროგნოზი, სპორტული შეჯიბრების შედეგის პროგნოზი და ა. შ. ზოგიერთი პროგნოზი არის უფრო ზუსტი, ვიდრე დანარჩენი, გამომდინარე დამოკიდებულების ხარისხის სიძლიერისაგან. ანუ ორ სიდიდეს შორის დამოკიდებულების სიძლიერე უფრო ზუსტი პროგნოზის გაკეთების საშუალებას იძლევა. პროგნოზის გასაკეთებლად საჭიროა სიდიდეებს შორის კავშირის სახის დადგენა, რასაც სწავლობს `რეგრესიული ანალიზი და იგი ემყარება ე. წ. უმცირეს კვადრატთა მეთოდს.

მარტივი კორელაციისა და რეგრესიის შესწავლისას მკვლევარი აგროვებს (აკვირდება) ორი რიცხვითი ან თვისებრივი სიდიდის მონაცემებს იმის დასადგენად არსებობს თუ არა ამ სიდიდეებს შორის კავშირი. მაგალითად, თუ მკვლევარს სურს დაადგინოს, არსებობს თუ არა კავშირი სტუდენტის მიერ სწავლაზე დახარჯული საათების რაოდენობასა და მის მიერ გამოცდაზე მიღებულ ქულას შორის, მან უნდა მოახდინოს სტუდენტების შემთხვევითი შერჩევა, განსაზღვროს თითოეული სტუდენტის მიერ სწავლაზე დახარჯული დრო და მოიპოვოს თითოეული სტუდენტის მიერ გამოცდაზე მიღებული ქულა. ამ მონაცემებით გაკეთებულ ცხრილს ექნება შემდეგი სახე:

სტუდენტი	მეცადინეობის დრო, x	შეფასება, y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

ორი სიდიდე შეიძლება იყოს *დამოუკიდებელი და დამოკიდებული.* **დამოუკიდებელი (independent)** სიდიდე ის სიდიდეა, რომელის კონტროლი (შეცვლა, მისით მანიპულირება) შესაძლებელია. ამ შემთხვევაში, "მეცადინეობის დრო" დამოუკიდებელი სიდიდეა და მას აცხადებენ x სიდიდედ. **დამოკიდებული (dependent)** სიდიდე ის სიდიდეა, რომლის კონტროლი (მისით მანიპულირება) შეუძლებელია. სტუდენტის მიერ გამოცდაზე მიღებული ქულა დამოკიდებული სიდიდეა, და მას აცხადებენ y სიდიდედ. სიდიდეებს შორის ამ განსხვავების არსი იმაში მდგომარეობს, რომ სტუდენტის მიერ გამოცდაზე მიღებული ქულა დამოკიდებულია მის მიერ მეცადინეობაზე დახარჯულ საათების რაოდენობაზე. გარდა ამისა, შეიძლება ითქვას, რომ სტუდენტს შეუძლია არეგულიროს (აკონტროლოს) გამოცდისათვის მოსამზადებელი მეცადინეობის საათების რაოდენობა.

საჭიროა აღინიშნოს, რომ x და y სიდიდეების განსხვავება ყოველთვის არაა ცხადი და ხანდახან ღებულობენ ნებისმიერ გადაწყვეტილებას.

მაგალითად, თუ მკვლევარი სწავლობს ადამიანის ასაკის დამოკიდებულებას არტერიულ წნევაზე, მას საზოგადოდ შეუძლია ჩათვალოს რომ ასაკი გავლენას ახდენს წნევაზე. შესაბამისად, ამ შემთხვევაში სიდიდეს "ასაკი" შეიძლება დავარქვათ *დამოუკიდებელი სიდიდე*, ხოლო სიდიდეს "არტერიული წნევა" – *დამოკიდებული სიდიდე*. მეორეს მხრივ, თუ მკვლევარი სწავლობს კავშირს გარკვეული ობიექტის მიმართ მამაკაცის (ქმრის) შეხედულებასა (პოზიციასა) და მისი ცოლის შეხედულებას შორის იმავე ობიექტის მიმართ, მაშინ ძნელია იმის თქმა, თუ რომელი სიდიდეა დამოუკიდებელი და რომელი დამოკიდებული. ამ შემთხვევაში მკვლევარს შეუძლია ნებისმიერად დაყოს სიდიდეები: ერთ-ერთს დაარქვას დამოუკიდებელი, ხოლო მეორეს – დამოკიდებული.

დამოუკიდებელი და დამოკიდებული სიდიდეები შეიძლება გამოსახულ იქნეს გრაფიკულად და მას **გაბნევის დიაგრამა (scatter plot)** ეწოდება. დამოუკიდებელი სიდიდე, x გადაიზომება ჰორიზონტალურ ღერძზე, ხოლო დამოკიდებული სიდიდე, y გადაიზომება ვერტიკალურ ღერძზე.

გაბნევის დიაგრამა არის გრაფიკი რიცხვთა დალაგებული (x,y) წყვილებისა, რომელიც შედგება დამოუკიდებელი x სიდიდისაგან და დამოკიდებული y სიდიდისაგან.

გაბნევის დიაგრამა წარმოადგენს დამოუკიდებელ და დამოკიდებულ სიდიდეებს შორის დამოკიდებულების ვიზუალური გამოსახვის საშუალებას. თითეული ღერძის გრადუირება შეიძლება იყოს განსხვავებული, ხოლო ღერძზე კოორდინატები განისაზღვრება შესაბამისი სიდიდის უდიდესი და უმცირესი მნიშვნელობების მიხედვით.

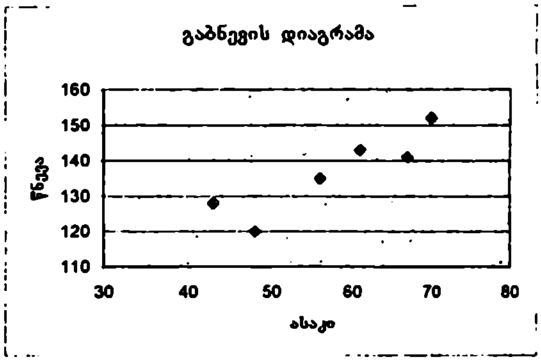
მაგალითი2 ავაგოთ გაბნევის დიაგრამა მონაცემებისათვის, რომელიც მიიღება 6 შემთხვევით შერჩეული პიროვნების ასაკსა და არტერიული წნევას შორის დამოკიდებულების შესწავლისას. მონაცემების ცხრილია:

პერსონა	ასაკი, x	წნევა, y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152 .

ამოხსნა.

ნაბიჯი 1. დავხაზოთ საკოორდინატო სიბრტყე და მოვნიშნოთ x და y ღერძები.

ნაბიჯი 2. მოვნიშნოთ თითოეული წერტილი გრაფიკზე როგორც ეს ქვემოთაა ნაჩვენები:



როგორც აღინიშნა სტატისტიკოსი იყენებს ე. წ. *კორელაციის კოეფიციენტს*, რათა გაზომოს ორ სიდიდეს შორის დამოკიდებულების ხარისხი. არსებობს სხვადასხვა სახის კორელაციის კოეფიციენტი. ჩვენ მოვიყვანთ ე. წ. **პირსონის მომენტთა ნამრავლის კორელაციის კოეფიციენტს (Pearson product moment correlation coefficient – PPMC)**, რომელიც დაკავშირებულია ცნობილი სტატისტიკოსის კარლ პირსონის სახელთან (პირსონი იყო პიონერი ამ ტიპის კვლევებში).

კორელაციის კოეფიციენტი, გამოთვლილი მონაცემთა შერჩევის მიხედვით, წარმოადგენს ორ სიდიდეს შორის წრფივი დამოკიდებულების ხარისხისა და მიმართულების საზომს. შერჩევითი კორელაციის კოეფიციენტი აღინიშნება r ასოთი, ხოლო პოპულაციის მიხედვით გამოთვლილი კორელაციის კოეფიციენტი აღინიშნება ბერძნული ρ (რო) ასოთი.

თავდაპირველად განვმარტოთ შერჩევითი კოვარიაციის კოეფიციენტი. ორ x და y სიდიდეს შორის **კოვარიაციის კოეფიციენტი** აღინიშნება $\mathrm{cov}(x,y)$ სიმბოლოთი და განიმარტება შემდეგი თანაფარდობით:

$$\mathrm{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),$$

სადაც $\bar{x}$ და $\bar{y}$ არის შესაბამისად x და y სიდიდეების შერჩევითი საშუალო ($\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$).

შერჩევითი კორელაციის კოეფიციენტი მოიცემა შემდეგი თანაფარდობით:

$$r=\frac{\mathrm{cov}(x,y)}{s_x s_y}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

კოვარიაციის კოეფიციენტის გამოსათვლელად მოხერხებულია ვისარგებლოთ შემდეგი ექვივალენტური და გამარტივებული ფორმულით:

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\,\overline{y}.$$

ხშირად იყენებენ კორელაციის კოეფიციენტის შემდეგ ექვივალენტურ გამოსახულებასაც (რომელიც მოხერხებულია იმ შემთხვევაში, როდესაც თითოეული სიდიდის საშუალო და სტანდარტული გადახრა უკვე ნაპოვნია):

$$r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right).$$

ამ ფორმულის მიხედვით, თავდაპირველი მონაცემების კორელაციის კოეფიციენტი შეიძლება წარმოვადგინოთ, როგორც გარდაქმნილი (სტანდარტიზებული)

$$\left(\frac{x_i-\bar{x}}{s_x},\frac{y_i-\bar{y}}{s_y}\right),\quad i=1,\dots,n,$$

მონაცემების კოვარიაციის კოეფიციენტი (შევნიშნოთ, რომ თუ x და y სიდიდეებიდან ერთ-ერთის შერჩევითი საშუალო ნულია, მაშინ კოვარიაციის კოეფიციენტი $x\cdot y$ სიდიდის საშუალოს ტოლია: $\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i$, გარდაქმნილი მონაცემების საშუალოები კი ნულის ტოლია).

ჩამოვთვალოთ კორელაციის კოეფიციენტის თვისებები:

1. $-1\le r\le 1$. ეს თვისება მარტივად გამომდინარეობს მათემატიკაში კარგად ცნობილი კოში-ბუნიაკოვსკის უტოლობიდან, რომლის თანახმადაც:

$$|\operatorname{cov}(x,y)|=\left|\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right|\le\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}=s_xs_y,$$

შესაბამისად, $|r|=|\operatorname{cov}(x,y)|/(s_xs_y)\le(s_xs_y)/(s_xs_y)=1$. აღსანიშნავია, რომ r-ის დადებითი მნიშვნელობა მიუთითებს სიდიდეებს შორის პოზიტიური დამოკიდებულების არსებობაზე, ხოლო უარყოფითი მნიშვნელობა კი – უარყოფითზე.

2. $r=-1$ ან $r=1$ მაშინ და მხოლოდ მაშინ, როცა x და y სიდიდეებს შორის არსებობს წრფივი კავშირი, ანუ მოიძებნება ისეთი a და b მუდმივები, რომ $y_i=ax_i+b$, ანუ $|r|=1$ მაშინ და მხოლოდ მაშინ, როცა წერტილები გაბნევის დიაგრამაზე ზუსტად განლაგდება რაიმე წრფეზე. ამასთანავე, თუ $a>0$, მაშინ $r=1$ და თუ $a<0$, მაშინ $r=-1$.

3. კორელაციის კოეფიციენტი უგანზომილებო სიდიდეა. ამასთან მისი სიდიდე უცვლელი რჩება, როცა x ან y სიდიდის ან ორივეს ერთად

საზომი ერთეული იცვლება. იგი უცვლელი რჩება მაშინაც, როცა იცვლება ათვლის წერტილი, ანუ სხვა სიტყვებით, რომ ვთქვათ, კორელაციის კოეფიციენტი უცვლელი რჩება სიდიდეების წრფივი გარდაქმნისას.

4. კორელაციის კოეფიციენტი ზომავს სიდიდეებს შორის წრფივი დამოკიდებულების სიძლიერეს. სხვა, არაწრფივი, თუნდაც დეტერმინისტული კავშირი კორელაციაში არ აისახება. კორელაცია არ წარმოადგენს გაბნევის დიაგრამის მეშვეობით აღმოჩენილი ყველა ტიპის კავშირების ზოგად საზომს.

5. რაც უფრო კონცენტრირებულია მონაცემები რაიმე წრფის მიდამოში, მით უფრო დიდია კორელაციის კოეფიციენტი. ამასთანავე, თუ შესაბამისი წრფის დახრილობა დადებითია (წრფე აბსცისთა ღერძის დადებით მიმართულებასთან ადგენს მახვილ კუთხეს), კორელაციის კოეფიციენტი დადებითია (თუ წრფე აბსცისთა ღერძის დადებით მიმართულებასთან ადგენს ბლაგვ კუთხეს, მაშინ კორელაციის კოეფიციენტი უარყოფითია).

6. ორ სიდიდეს შორის კავშირი არსებობს, თუ ადგილი აქვს თანაფარდობას: $|r| \geq 2/\sqrt{n}$, სადაც n შერჩევის მოცულობაა.

7. თუ მოცემულია მონაცემთა ორი სიმრავლე $x_1, \ldots, x_n$ და $y_1, \ldots, y_n$, მაშინ მონაცემთა ახალი $z_i = x_i \pm y_i$ $(i = 1, \ldots, n)$ სიმრავლისათვის შერჩევითი დისპერსია ასე გამოითვლება:

$$s_z^2 = s_x^2 + s_y^2 \pm 2\operatorname{cov}(x, y)$$

(თუ $\operatorname{cov}(x, y) = 0$, მაშინ $s_z^2 = s_x^2 + s_y^2$). ანალოგიურად,

$$s_z'^2 = s_x'^2 + s_y'^2 \pm 2\operatorname{cov}'(x, y),$$

სადაც

$$\operatorname{cov}'(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{n-1}\operatorname{cov}(x, y)$$

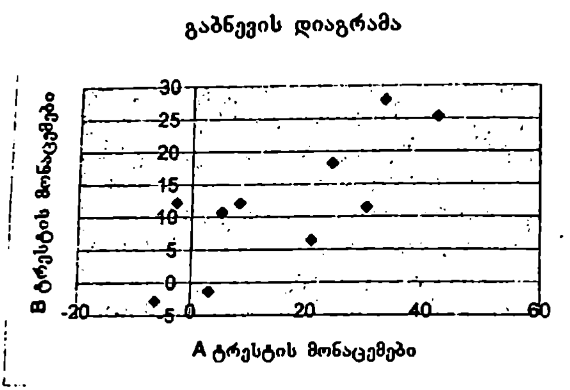
(გავიხსენოთ, რომ $\bar{z} = \bar{x} \pm \bar{y}$).

მაგალითი 3. მოცემულია ორი ტრესტის წლიური საპროცენტო შემოსავლის მონაცემები ბოლო 10 წლის განმავლობაში:

ტრესტი A	8.3	-6.2	20.9	-2.7	33.6	42.8	24.4	5.2	3.1	30.5
ტრესტი B	12.1	-2.8	6.4	12.2	27.8	25.3	18.2	10.7	-1.3	11.4

ავაგოთ გაბნევის დიაგრამა და გამოვთვალოთ კორელაციის კოეფიციენტი.

ამოხსნა. A ტრესტის მონაცემები გადავზომოთ x ღერძზე, B ტრესტისა კი y ღერძზე.



ამ შემთხვევაში $\bar{x}=16$, $\bar{y}=12$, $s_x=17.65$ და $s_y=10.51$. ამიტომ გვაქვს:

$$\operatorname{cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{x}\,\overline{y}=$$

$$=\frac{1}{10}\cdot[8.3\cdot12.1+(-6.2)\cdot(-2.8)+\cdots+30.5\cdot11.4]-16\cdot12=71.48,$$

შესაბამისად,

$$r=\frac{\operatorname{cov}(x,y)}{s_xs_y}=\frac{71.48}{17.65\cdot10.51}\approx0.39.$$

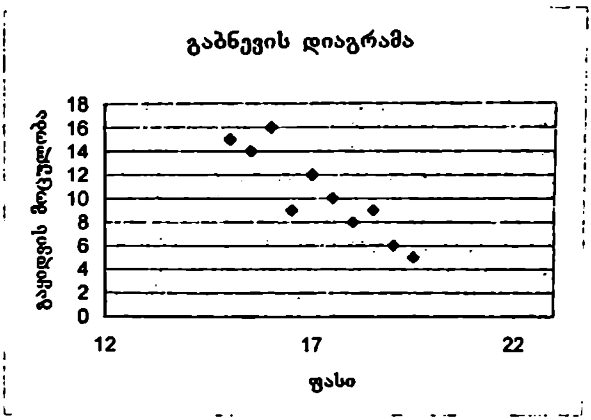
კორელაციის კოეფიციენტის ეს მნიშვნელობა გვაეწმუნებს, რომ არსებობს სუსტად გამოხატული პოზიტიური წრფივი კავშირი A და B ტრესტების ამონაგებს შორის, რასაც გაბნევის დიაგრამაც ადასტურებს.

მაგალითი 4. ახალი პროდუქციის ფასის დასადგენად კომპანიამ შეარჩია 10 ქალაქი, რომელთაც არსებითად ერთი და იგივე ყიდვის პოტენციალი გააჩნია და გამოიტანა პროდუქცია ბაზარზე ყოველ მათგანში სხვადასხვა ფასად. კომპანიის მიერ მიღებული შედეგები მოყვანილია ცხრილში:

ქალაქის ნომერი	1	2	3	4	5	6	7	8	9	10
ფასი, x	15.0	15.5	16.0	16.5	17.0	17.5	18.0	18.5	19.0	19.5
გაყიდვის მოცულობა, y	15	14	16	9	12	10	8	9	6	5

დავადგინოთ გაყიდვის მოცულობის ფასზე დამოკიდებულების სახე.

ამოხსნა. ავაგოთ გაბნევის დიაგრამა



როგორც დიაგრამა გვიჩვენებს, ჩვენს წინაშეა უარყოფითი წრფივი კავშირის ტენდენცია: მონაცემები თავმოყრილი არიან უარყოფითი დახრილობის წრფის მიდამოში. რადგანაც კორელაციის კოეფიციენტი ზომავს ორ სიდიდეს შორის არსებული წრფივი კავშირის სიძლიერეს, უნდა მოველოდეთ, რომ კორელაციის კოეფიციენტი საკმაოდ ახლოს იქნება -1-თან. გამოვთვალოთ კორელაციის კოეფიციენტი. გვაქვს:

$$\bar{x}=17.25,\quad s_x=1.44,\quad \bar{y}=10.4,\quad s_y=3.56,\quad \mathrm{cov}(x,y)=-4.75 \text{ და}$$

$$r=\frac{\mathrm{cov}(x,y)}{s_x s_y}=\frac{-4.75}{5.13}\approx-0.93.$$

კორელაციის კოეფიციენტის ეს მნიშვნელობა ადასტურებს გაბნევის დიაგრამის დახმარებით მიღებულ დასკვნას. მაგრამ ეს მნიშვნელობა არაფერს გვეუბნება იმ წრფის შესახებ, რომლის მიდამოშიაც თავმოყრილია დაკვირვებული წყვილები. ამ წრფის მოსაძებნად იყენებენ უმცირეს კვადრატთა მეთოდს, რომლის თანახმადაც ეძებენ ისეთ წრფეს $y=\bar{a}x+\bar{b}$, რომლის კოეფიციენტებისთვისაც სრულდება პირობა:

$$\min_{a,b\in(-\infty,+\infty)}\sum_{i=1}^{n}(y_i-ax_i-b)^2=\sum_{i=1}^{n}(y_i-\bar{a}x_i-\bar{b})^2,$$

ანუ, წრფე რომლიდანაც y_i-ების ვერტიკალური გადახრების კვადრატების ჯამი მინიმალურია. ამ წრფის კოეფიციენტების მოსაძებნად უნდა ვიპოვოთ ორი ცვლადის $\varphi(a,b)=\sum_{i=1}^{n}(y_i-ax-b)^2$ ფუნქციის მინიმუმის წერტილები, რისთვისაც, თავის მხრივ, უნდა ამოიხსნას განტოლებათა სისტემა:

$$\frac{\partial}{\partial a}\varphi(a,b)=0 \text{ და } \frac{\partial}{\partial b}\varphi(a,b)=0,$$

სადაც $\frac{\partial}{\partial a}\varphi(a,b)$ და $\frac{\partial}{\partial b}\varphi(a,b)$ სიმბოლოებით აღნიშნულია $\varphi(a,b)$ ფუნქციის კერძო წარმოებულები შესაბამისად a-სა და b-ს მიმართ.

ძნელი დასანახი არ არის, რომ უკანასკნელი სისტემის ამონახსნია:

$$\tilde{a} = \frac{\mathrm{cov}(x,y)}{s_x^2} \quad \text{და} \quad \tilde{b} = \bar{y} - \tilde{a}\bar{x}.$$

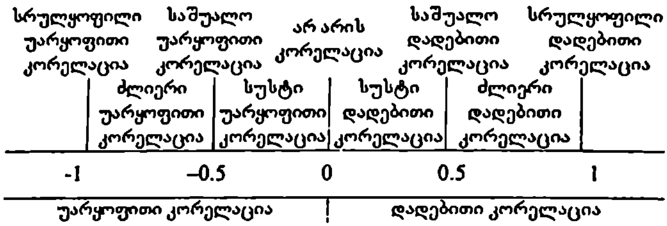
ამ მაგალითში მოყვანილი მონაცემებისათვის გვექნება:

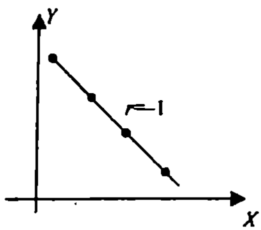
$$\tilde{a} = \frac{-4.75}{1.44^2} = -\frac{4.75}{2.07} \approx -2.29 \quad \text{და} \quad \tilde{b} = 10.4 + 2.29 \cdot 17.25 \approx 49.90,$$

ამიტომ საძებნი წრფის განტოლება იქნება: $y = -2.29x + 49.9$.

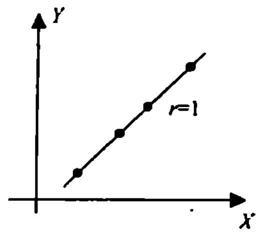
ეს განტოლება კომპანიას საშუალებას აძლევს ახალი პროდუქციის ამა თუ იმ ფასის შემთხვევაში გააკეთოს პროგნოზი გაყიდვის შესაძლო მოცულობის შესახებ.

კორელაციის კოეფიციენტის მნიშვნელობა მოთავსებულია -1-სა და 1-ს შორის. თუ ორ სიდიდეს შორის ადგილი აქვს *მკაცრად დადებით წრფივ დამოკიდებულებას,* მაშინ r-ის მნიშვნელობა ახლოსაა +1-თან. თუ ორ სიდიდეს შორის ადგილი აქვს *მკაცრად უარყოფით წრფივ დამოკიდებულებას,* მაშინ r-ის მნიშვნელობა ახლოსაა -1-თან. როდესაც ორ სიდიდეს შორის არა გვაქვს წრფივი დამოკიდებულება, ან ამ სიდიდეებს შორის სუსტი კავშირია, მაშინ r-ის მნიშვნელობა ახლოსაა 0-თან. ქვემოთ მოყვანილია კორელაციის ხარისხის სიტყვიერი დახასიათებისათვის მიღებული ტერმინოლოგია და შესაბამისი დიაგრამები:

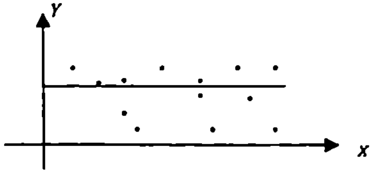




სრულყოფილი უარყოფითი კორელაცია



სრულყოფილი დადებითი კორელაცია



არ არის კორელაცია

ძლიერი უარყოფითი კორელაცია

ძლიერი დადებითი კორელაცია

კორელაციის კოეფიციენტის პირდაპირი გამოთვლისათვის (როცა წინასწარ არ არის გამოთვლილი სიდიდეების საშუალო და სტანდარტული გადახრა) იყენებენ შემდეგ ფორმულას:

$$r=\frac{n(\sum_{i=1}^{n}x_i y_i)-(\sum_{i=1}^{n}x_i)\cdot(\sum_{i=1}^{n}y_i)}{\sqrt{[n(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)^2]\cdot[n(\sum_{i=1}^{n}y_i^2)-(\sum_{i=1}^{n}y_i)^2]}},$$

სადაც n მონაცემთა წყვილების რაოდენობაა.

ეს ფორმულა საკმაოდ რთულად გამოიყურება. გამოთვლების გამარტივების მიზნით ჩვენ ვიხელმძღვანელებთ ქვემოთ მოყვანილი ცხრილით.

მაგალითი 5. გამოვთვალოთ კორელაციის კოეფიციენტის მნიშვნელობა იმ მონაცემებისათვის, რომელიც მიიღებული იყო 6 შემთხვევით შერჩეული პიროვნების ასაკსა და არტერიული წნევას შორის დამოკიდებულების შესწავლისას.

ამოხსნა.

ნაბიჯი 1. გავაკეთოთ ცხრილი ისე როგორც ეს ქვემოთაა ნაჩვენები:

პერსონა	ასაკი, x	წნევა, y	xy	x^2	y^2
A	43	128			
B	48	120			
C	56	135			
D	61	143			
E	67	141			
F	70	152			

ნაბიჯი 2. გამოვთვალოთ xy, x^2 და y^2 გამოსახულებათა მნიშვნელობები და შევიტანოთ ცხრილის შესაბამის სვეტებში.

შევკრიბოთ სვეტებში მოთავსებული მნიშვნელობები. დასრულებულ ცხრილს ექნება შემდეგი სახე:

პერსონა	ასაკი, x	წნევა, y	xy	x^2	y^2
A	43	128	5.504	1.849	16.384
B	48	120	5.760	2.304	14.400
C	56	135	7.560	3.136	18.225
D	61	143	8.723	3.721	20.449
E	67	141	9.447	4.489	19.881
F	70	152	10.640	4.900	23.104
	$\sum x = 345$	$\sum y = 819$	$\sum xy = 47.634$	$\sum x^2 = 20.399$	$\sum y^2 = 112.443$

ნაბიჯი 3. მიღებული მნიშვნელობები შევიტანოთ კორელაციის კოეფიციენტის ფორმულაში და გამოვთვალოთ r:

$$r = \frac{6 \cdot 47.634 - 345 \cdot 819}{\sqrt{(6 \cdot 20.399 - 345^2) \cdot (6 \cdot 112.443 - 819^2)}} = 0.897.$$

დასკვნა. კორელაციის კოეფიციენტის მიღებული მნიშვნელობა გვეუბნება, რომ ადამიანის ასაკსა და არტერიულ წნევას შორის არსებობს მკაცრად დადებითი წრფივი დამოკიდებულება.

ალბათობის თეორია

თავი I
ალბათობის თეორიის ელემენტები

ალბათობის თეორიის საგანი.

ალბათობის თეორია წარმოადგენს მათემატიკის დარგს, რომელიც შეისწავლის ისეთი ექსპერიმენტების (მოვლენების) მათემატიკურ მოდელებს, რომელთა შედეგები ცალსახად არ განისაზღვრება ცდის პირობებით. ასეთ ექსპერიმენტებს ეწოდებათ *შემთხვევითი ექსპერიმენტები.* შემთხვევითი ექსპერიმენტებია: მონეტის აგდების შედეგი, მიზანში სროლისას მიზნის დაზიანება ან არდაზიანება, ხელსაწყოს მუშაობის ხანგრძლივობა, სატელეფონო სადგურში გამოძახებათა რიცხვი, ავტო-საგზაო შემთხვევათა რაოდენობა, არაბეჯითი სტუდენტის მიერ გამოცდის ჩაბარების შედეგი, ვალუტის კურსი რომ არ ჩამოვა გარკვეულ ნიშნულამდე და სხვა.

საჭიროა აღინიშნოს, რომ ალბათობის თეორია იკვლევს არა ნებისმიერ შემთხვევით ექსპერიმენტს, არამედ მხოლოდ ისეთ ექსპერიმენტებს, რომლებიც ხასიათდებიან *სტატისტიკური მდგრადობის ანუ სიხშირეთა მდგრადობის* თვისებით. ეს თვისება ხასიათდება შემდეგნაირად. განვიხილოთ შემთხვევითი ექსპერიმენტი და ვიგულისხმოთ, რომ ამ ექსპერიმენტის ჩატარების იდენტური პირობების შენარჩუნება და მისი განმეორებითი ჩატარება შესაძლებელია (თუ ფიზიკურად არა, აზრობრივად მაინც) ნებისმიერ სასურველ რაოდენობა რიცხვჯერ. აღვნიშნოთ A-თი ამ ექსპერიმენტის ერთერთი შესაძლო შედეგი. გავიმეოროთ ეს ექსპერიმენტი n-ჯერ და აღვნიშნოთ $\mu_n(A)$-თი A შედეგის (ხდომილების) მოხდენათა რიცხვი ამ n ექსპერიმენტში. მაშინ შეფარდებას $\mu_n(A)/n$ ეწოდება A ხდომილების *ფარდობითი სიხშირე,* ხოლო სიხშირეთა მდგრადობის თვისება მდგომარეობს შემდეგში:

დიდი n-ებისათვის A ხდომილების ფარდობითი სიხშირე მხოლოდ მცირედ ირხევა (n-ის ცვლილებისას) გარკვეული მუდმივი მნიშვნელობის ირგვლივ. რაც უნდა გავიგოთ შემდეგნაირად: თუ ჩავატარებთ ექსპერიმენტების რამოდენიმე სერიას, მაშინ *სტატისტიკური მდგრადობის თვისება* გულისხმობს, რომ: სიხშირეები $\mu_{n_i}(A)/n_i$ (სადაც n_i არის i-ურ სერიაში ექსპერიმენტების რიცხვია, ხოლო $\mu_{n_i}(A)$ — კი A ხდომილების მოხდენათა რაოდენობა ამ სერიაში) ახლოს იქნებიან ერთმანეთთან (მცირედ განსხვავდებიან ერთმანეთისაგან ან რაიმე საშუალო სიდიდიდან) ყოველთვის როგორც კი n_i რიცხვები იქნებიან საკმაოდ დიდები. მაგალითად, ვ. ფელერის წიგნში მოყვანილია მონეტის აგდების 10 სერიაში ($i=1,...10$), სადაც თითეულ სერიაში $n_i=1000$ ექსპერიმენტია, გერბის მოსვლათა μ_{n_i}("გ") რიცხვის შემდეგი მონაცემები: 501, 485, 509, 536, 485, 488, 500, 497, 494, 484. ცხადია, რომ აქ ფარდობითი სიხშირეები ახლოსაა ერთმანეთთან, და მაშასადამე, ექსპერიმენტს, რომელიც მდგომარეობს სიმეტრიული მონეტის აგდებაში, გააჩნია სიხშირის მდგრადობის თვისება.

შემდეგ ცხრილში მოგვყავს ის შედეგები, რომლებიც ექსპერიმენტალურად მიღებული იყო სხვადასხვა მკვლევარების მიერ მე-18 საუკუნიდან მოყოლებული სიმეტრიული მონეტის n-ჯერ აგდებისას გერბთა მოსვლის m/n ფარდობითი სიხშირისათვის:

მკვლევარი	აგდების რაოდენობა n	ფარდობითი სიხშირე m/n
ბიუფონი	4040	0.507
დე მორგანი	4092	0.5005
ჯევონსი	20480	0.5068
რომანოვსკი	80640	0.4923
კ. პირსონი	24000	0.5005
ფელერი	10000	0.4979

ეს ცხრილი გვიჩვენებს, რომ ფარდობითი სიხშირეები ახლოსაა 0.5-თთან.

მოვიყვანოთ კიდევ ერთი მაგალითი. გ. კრამერის მიერ მოყვანილი მონაცემები 1935 წელს შვეციაში დაბადებული ახალშობილების შესახებ (სადაც n – ახალშობილთა რაოდენობაა, ხოლო m/n – ვაჟების დაბადების ფარდობითი სიხშირე) ასე გამოიყურება:

თვეები	I	II	III	IV	V	VI	
n	7280	6957	7883	7884	7892	7609	
m/n	0.515	0.510	0.510	0.529	0.522	0.518	
თვეები	VII	VIII	IX	X	XI	XII	სულ
n	7585	7393	7203	6903	6552	7132	88273
m/n	0.523	0.514	0.515	0.509	0.518	0.527	0.517

მიუხედავად იმისა, რომ ახალშობილთა საერთო რაოდენობა იცვლება წლის განმავლობაში, ვაჟების დაბადების ფარდობითი სიხშირე საკმაოდ მდგრადად მერყეობს 0.517 – საშუალო მნიშვნელობის ირგვლივ.

ასეთი ტიპის სტატისტიკური კანონზომიერებები აღმოჩენილ იქნა უკვე მე-18 საუკუნეში დემოგრაფიულ მონაცემებში – ახალშობილთა სტატისტიკის შესწავლისას, სიკვდილიანობის სტატისტიკის შესწავლისას, უბედურ შემთხვევათა სტატისტიკის შესწავლისას და ა. შ. (რაც, თავის მხრივ, საკმაოდ ეფექტურად გამოიყენებოდა სადაზღვევო კომპანიების საქმიანობაში). მოგვიანებით, მე-19 საუკუნის ბოლოს და მე-20 საუკუნის დასაწყისში ახალი სტატისტიკური კანონზომიერებები აღმოჩენილ იქნა ფიზიკაში, ქიმიაში, ბიოლოგიაში, ეკონომიკაში და სხვა მეცნიერებებში.

ამ კანონზომიერებებს მივყავართ ალბათობის სტატისტიკური განმარტებისაკენ.

განმარტება. რიცხვს, რომლის ირგვლივაც ირხევა A ხდომილების ფარდობითი სიხშირე, ეწოდება A *ხდომილების ალბათობა* და აღინიშნება $P(A)$ სიმბოლოთი.

მათემატიკური დასაბუთება იმისა, რომ ფარდობითი სიხშირე ახლოსაა ალბათობასთან (ფარდობით სიხშირეთა მიმდევრობის ზღვარია ალბათობა და შესაბამისად, იგი ერთადერთია) მოყვანილია იაკობ ბერნულის ცნობილ თეორემაში, რომელიც ალბათობის თეორიაში ცნობილია აგრეთვე *დიდ რიცხვთა კანონის* სახელწოდებით და იგი წარმოადგენს ერთ-ერთ ფუნდამენტურ თეორემას.

აღსანიშნავია, რომ ყველა შესაძლო ექსპერიმენტი შეიძლება გაიყოს სამ კატეგორიად:

I. ექსპერიმენტები სრული მდგრადობით, სადაც საერთოდ არაა განუზღვრელობა;

II. ექსპერიმენტები სადაც არა გვაქვს სრული მდგრადობა, მაგრამ არის სტატისტიკური მდგრადობა;

III. ექსპერიმენტები სადაც სტატისტიკური მდგრადობაც კი არა გვაქვს.

პირველ ჯგუფს მიეკუთვნება უმრავლესობა ექსპერიმენტების, რომლებიც აღიწერებიან საბუნებისმეტყველო მეცნიერებების (ფიზიკა, ქიმია) კლასიკური კანონებით და მათი შესწავლა ხდება ალბათობის თეორიის გამოყენების გარეშე. მესამე კატეგორიისათვის ალბათობის თეორია გამოუსადეგარია. მეორე ჯგუფი წარმოადგენს სწორედ ალბათობის თეორიის გამოყენების არეს. თვითონ ალბათობის თეორიის მიერ გადასაწყვეტი ამოცანები იყოფა ორ დიდ ჯგუფად. პირველი ჯგუფის ამოცანებს შეიძლება ვუწოდოთ ამოცანები რთული ხდომილებების ალბათობების გამოთვლაზე, როცა ცნობილია მარტივი ხდომილებების ალბათობები. მაგალითად, თუ ცნობილია, რომ გერბის მოსვლის ალბათობა $P\{\text{"გ"}\}=1/2$, ვიპოვოთ ალბათობები იმისა, რომ ზემოთ მოყვანილ მაგალითში $\mu_{1000}(\text{"გ"})$ ტოლია 501-ის, ან 485-ის, ან და ა. შ. 484-ის.

ამოცანების მეორე ჯგუფი გარკვეული აზრით პირველი ჯგუფის ამოცანების შებრუნებულია. აქ ექსპერიმენტების სერიის შედეგების საფუძველზე რაიმე გზით უნდა შევაფასოთ მარტივი ხდომილებების ალბათობები. ვინაიდან, მარტივი ხდომილებების ალბათობები უცნობია იმ ამოცანებში, რომლებიც პრაქტიკაში წარმოიშობა, ამიტომ ამოცანების ეს ჯგუფი განსაკუთრებით საინტერესოა გამოყენებების თვალსაზრისით. ალბათობის თეორიის იმ ნაწილს, რომელიც ახდენს მეორე ჯგუფის ამოცანების ზუსტ დასმას და იკვლევს მათი ამოხსნის მეთოდებს, მათემატიკური სტატისტიკა ეწოდება.

ელემენტარულ ხდომილებათა სივრცე.

ალბათობის მათემატიკური თეორიის პრაქტიკული ღირებულება და მნიშვნელობა წარმოჩინდება ისეთ ნამდვილ თუ წარმოსახვით ცდებთან და მოვლენებთან დაკავშირებით, როგორიცაა მაგალითად, მონეტის ერთჯერადი აგდება, მონეტის აგდება 100-ჯერ, ორი სათამაშო კამათლის გაგორება, კარტის დარიგება, "რულეტკის" თამაში, ადამიანის ან რადიოაქტიური ატომის სიცოცხლის ხანგრძლივობაზე დაკვირვება, ადამიანების გარკვეული შემთხვევითი ჯგუფის შერჩევა და მათში ცაციების რაოდენობის დათვლა, ორი სახეობის მცენარის შეჯვარება და შედეგზე დაკვირვება, სატელეფონო სადგურის დაკავებული ხაზების ან სატელეფონო გამოძახებათა რიცხ-

ვის განსაზღვრა, ელექტრონული სისტემების შემთხვევითი ხმაურები, სამრეწველო პროდუქციის ხარისხის შერჩევითი კონტროლი, ავტოსაგზაო შემთხვევათა რაოდენობა, ახალშობილის სქესი, ორმაგი ვარსკვლავების რიცხვი ცის გარკვეულ უბანზე, სითხეში ჩავარდნილი მტვრის მცირე ნაწილაკის მდებარეობა, ვალუტის კურსი, ფასების ინდექსი. ჯერ-ჯერობით ყველა ამ მოვლენის აღწერა საკმაოდ ბუნდოვანია, და, იმისათვის, რომ თეორიას მიეცეს ზუსტი აზრი, ჩვენ უნდა შევთანხმდეთ იმაზე, თუ რა გვესმის ჩვენ განსახილველი ცდის ან დაკვირვების *შესაძლებელი შედეგების* ქვეშ.

მონეტის აგდების შედეგად არაა აუცილებელი მოვიდეს გერბი ან საფასური: მონეტა შეიძლება დადგეს წიბოზე ან გაგორდეს შორს (დაიკარგოს). მიუხედავად ამისა, ჩვენ ვთანხმდებით განვიხილოთ გერბი და საფასური როგორც მონეტის აგდების ორად-ორი შესაძლებელი შედეგი. ეს შეთანხმება ამარტივებს თეორიას და არ ახდენს გავლენას მის შესაძლო გამოყენებებზე. ხშირად ასეთი შეანხმებები აუცილებელია. შეუძლებელია უშეცდომოდ გაიზომოს რაიმე ატომის არსებობის ხანგრძლივობა ან რომელიმე პირის სიცოცხლის ხანგრძლივობა. მიუხედავად ამისა, მეცნიერული მიზნებისათვის სასარგებლოა ეს სიდიდეები ჩაითვალოს ზუსტ რიცხვებად. მაგრამ ამასთანავე, იბადება კითხვა: რომელი რიცხვი შეიძლება და რომელი – არა წარმოადგენდეს ადამიანის სიცოცხლის ხანგრძლივობას? არსებობს თუ არა მაქსიმალური ასაკი, რომლის ზემოთაც სიცოცხლე შეუძლებელია, თუ ასაკი შეიძლება იყოს ნებისმიერი?

ჩვენ ცხადია არ შევუდგებით იმის მტკიცებას, რომ ადამიანს შეუძლია იცოცხლოს 1000 წელი, თუმცა ჩვეულებრივი სადაზღვევო პრაქტიკა არ აწესებს სიცოცხლის ხანგრძლივობის ასეთ საზღვარსაც. იმ ფორმულების თანახმად, რომელზეც დაფუძნებულია თანამედროვე სიკვდილიანობის ცხრილები იმ ადამიანების წილი, რომლებმაც იცოცხლეს 1000 წლამდე არის 10 ხარისხად (-10^{36}) რიგის. ბიოლოგიის თვალსაზრისით ეს მტკიცებულება აზრს მოკლებულია, მაგრამ ის არ ეწინააღმდეგება ცდას: საუკუნის განმავლობაში იბადება არაუმეტეს 10^{10} ადამიანი და იმისათვის, რომ ზემოთ მოყვანილი მტკიცებულობა სტატისტიკურად უარეყოთ საჭირო იქნებოდა 10 ხარისხად 10^{35} საუკუნე, რაც აჭარბებს დედამიწის ასაკს 10 ხარისხად 10^{34}-ჯერ. ცხადია, ასეთი მცირე შანსები თავსებადია შეუძლებელი შედეგის ჩვენს წარმოდგენასთან და შეიძლებოდა გვეფიქრა, რომ მისი გათვალისწინება აბსურდულია, თუმცა სინამდვილეში იგი ამარტივებს ბევრ ფორმულებს. გარდა ამისა, თუ ჩვენ გამოვრიცხავდით, რომ შეუძლებელია 1000 წლამდე ცხოვრება, მაშინ წავაწყდებოდით უფრო დიდ სირთულეებს, ვინაიდან მაშინ ჩვენ უნდა დაგვეშვა რომ არსებობს მაქსიმალური ასაკი. მაგრამ დაშვება, რომ ადამიანმა შეიძლება იცოცხლოს x წლამდე, მაგრამ არ შეიძლება იცოცხლოს x წელი და 2 წამი, არაფრით არაა უკეთესი, ვიდრე ასაკის ზედა ზღვარის არარსებობა.

ნებისმიერი თეორია აუცილებლად გულისხმობს ზოგიერთ გამარტივებას. ჩვენი პირველი გამარტივება ეხება “ცდის” ან “დაკვირვების” შესაძლო შედეგებს. მათემატიკური თეორიის შესასწავლი ობიექტები შეიძლება იყვნენ მხოლოდ ეს შესაძლებელი შედეგები. თუ ჩვენ გვინდა ავაგოთ ცდის აბსტრაქტული მოდელი, ჩვენ თავიდან უნდა დავადგინოთ რას წარმ-

ოადგენს გამარტივებული (იდეალიზირებული) ცდის შესაძლო შედეგი. ტერმინოლოგიის ერთიანობისათვის ექსპერიმენტის (ცდის) ან დაკვირვების შედეგებს უწოდებენ *ხდომილებებს.*

განვიხილოთ ექსპერიმენტი, რომლის ყველა შესაძლო შედეგები ამოიწურება N სხვადასხვა მნიშვნელობით $\omega_1, \omega_2, ..., \omega_N$. ეს მნიშვნელობები არ არის აუცილებლად რიცხვითი და მათი ფიზიკური ბუნება არ არის არსებითი.

განმარტება 1. ექსპერიმენტის ცალკეულ შესაძლო შედეგებს *ელემენტარული ხდომილებები* ეწოდება, ხოლო მათ ერთობლიობას – *ელემენტარულ ხდომილებათა სივრცე* და აღინიშნება Ω ასოთი: $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$.

მოვიყვანოთ მაგალითები:

I. მონეტის ერთხელ აგდებისას – Ω={გ, ს};

II. მონეტის ორჯერ აგდებისას, ან ორი მონეტის ერთდროულად აგდებისას – Ω={გგ, გს, სგ, სს};

III. მონეტის სამჯერ აგდებისას, ან სამი მონეტის ერთდროულად აგდებისას – Ω= {გგგ, გგს, გსგ, სგგ, გსს, სგს, სსგ, სსს};

IV. მონეტის n-ჯერ აგდებისას $\Omega = \{\omega : \omega = (a_1, ..., a_n), a_i =$ გ ან ს$\}$ და შედეგების საერთო რაოდენობა ტოლია 2^n-ის;

V. ერთი სათამაშო კამათლის გაგორებისას – Ω={1, 2, 3, 4, 5, 6};

VI. ვთქვათ, თავიდან ვაგდებთ მონეტას. თუ მოვა გერბი, მაშინ ვაგორებთ სათამაშო კამათელს; ხოლო თუ მოვა საფასური, მაშინ კიდევ ერთხელ ვაგდებთ მონეტას. ამ შემთხვევაში Ω = {გ1, გ2, გ3, გ4, გ5, გ6, სგ, სს};

VII. ორი სათამაშო კამათლის გაგორებისას – Ω={(1,1); (1,2); (1,6); (2,1); (2,2); ; (2,6); ; (6,1); ; (6,6)} ანუ

$$\Omega = \{(i, j) : i, j = 1, 2, ..., 6\};$$

VIII. პროდუქციის ვარგისიანობის დადგენისას – Ω={"ვარგისი", "უვარგისი"};

IX. სატელეფონო სადგურში გამოძახებათა რაოდენობა – Ω= {0, 1, 2, };

X. ძაბვა ქსელში – Ω={[0, 220]}.

განმარტება 2. ელემენტარულ ხდომილებათა სივრცის ნებისმიერ ქვესიმრავლეს *ხდომილება* ეწოდება.

ცხადია, ელემენტარული ხდომილებები აგრეთვე ხდომილებებია, ისინი წარმოადგენენ ელემენტარულ ხდომილებათა სივრცის ერთელემენტიან ქვესიმრავლეებს. ყველა დანარჩენ ქვესიმრავლეს (მათ შორის ცარიელი სიმრავლისა და თვითონ სივრცის ჩათვლით) ხდომილებას უწოდებენ. ზოგჯერ (იმის აღსანიშნავად, რომ ქვესიმრავლეში ერთზე მეტი ელემენტია) ხმარობენ აგრეთვე *შედგენილი ან რთული* ხდომილების ცნებასაც. ჩვენ ვისარგებლებთ უბრალოდ ხდომილების ცნებით.

მონეტის ორჯერ აგდებისას (იხ. მაგალითი II) ხდომილების მაგალითებია: ა). ერთჯერ მაინც მოვიდა გერბი (ანუ მოვიდა ერთი ან მეტი, მაშასადამე, ორი, გერბი). იგი წარმოადგენს სიმრავლეს – {გს,სგ,გგ}; ბ). გერბი მოვიდა არაუმეტეს ერთისა (ანუ მოვიდა ერთი ან ნაკლები, მაშასადამე,

ნული – არცერთი, გერბი). იგი წარმოადგენს სიმრავლეს – {გს,სგ,სს}; გ). გერბი მოვიდა ზუსტად ერთჯერ (ანუ პირველად მოვიდა გერბი და მეორედ კი საფასური ან პირიქით). ეს არის შემდეგი სიმრავლე – {გს,სგ}; და ა. შ. აღსანიშნავია, რომ ამ შემთხვევაში სულ გვექნება $2^4=16$ ხდომილება (როგორც ცნობილია n ელემენტიანი სიმრავლის ყველა შესაძლო ქვესიმრავლეთა რაოდენობაა 2^n ქვესიმრავლე).

განმარტება 3. თუ ექსპერიმენტის კონკრეტული შედეგი ეკუთვნის რაიმე ხდომილებას, მაშინ ამბობენ რომ ეს *ხდომილება მოხდა*, ხოლო რომელსაც არ ეკუთვნის – ის *ხდომილება არ მოხდა*.

ალბათობის თეორიაში ხდომილებები აღინიშნება დიდი ლათინური ასოებით: $A,B,C,D,\ldots$. ხდომილებას $A=\Omega$ უწოდებენ *აუცილებელ ხდომილებას*, ვინაიდან ის აუცილებლად ხდება (ის შეუძლებელია არ მოხდეს, რადგან ექსპერიმენტის ყველა შედეგი მას ეკუთვნის); ხოლო ხდომილებას, რომელიც არ შეიცავს არც ერთ ელემენტარულ ხდომილებას აღნიშნავენ $\emptyset$ სიმბოლოთი და უწოდებენ *შეუძლებელ ხდომილებას*, ვინაიდან მისი მოხდენა შეუძლებელია (რადგან ექსპერიმენტის არც ერთი შედეგი მას არ ეკუთვნის).

შემოვიღოთ აღნიშვნები: A={მონეტის ორჯერ აგდებისას ერთჯერ მაინც მოვიდა გერბი}={გს,სგ,გგ}; B={ მონეტის ორჯერ აგდებისას გერბი მოვიდა არაუმეტეს ერთისა}={გს,სგ,სს}; C={მონეტის ორჯერ აგდებისას გერბი მოვიდა ზუსტად ერთჯერ}={ გს,სგ }: D={მონეტის ორჯერ აგდებისას ორივეჯერ მოვიდა საფასური}={სს}; E={მონეტის ორჯერ აგდებისას საფასური მოვიდა არაუმეტეს ერთისა}={გგ,სგ,გს}. ამ აღნიშვნებში, თუ მონეტის ორჯერ აგდებისას საფასური მოვიდა მხოლოდ მეორედ აგდებისას, მაშინ შეგვიძლია ვთქვათ, რომ მოხდა A, B, C და E ხდომილებები, ხოლო D ხდომილება კი არ მოხდა.

თუ A ხდომილების მოხდენას მოსდევს B ხდომილების მოხდენა (სიმრავლეთა თეორიის ენაზე ეს ნიშნავს, რომ A ხდომილება ნაწილია, ქვესიმრავლეა B ხდომილების), მაშინ ჩვენ დავწერთ, რომ $A\subset B$ და ვიტყვით, რომ A ხდომილება *იწვევს* B ხდომილებას. გასაგებია, რომ ნებისმიერი A ხდომილება იწვევს აუცილებელ ხდომილებას – $A\subset\Omega$. თუ A ხდომილება იწვევს B ხდომილებას და იმავდროულად B ხდომილება იწვევს A ხდომილებას, მაშინ ვიტყვით, რომ A და B ხდომილები ერთმანეთის *ტოლია* და დავწერთ $A=B$.

წინა აბზაცის აღნიშვნებში: C ხდომილება იწვევს A, B და E ხდომილებებს; D ხდომილება იწვევს B ხდომილებას; A და E ხდომილებები ერთმანეთის ტოლია.

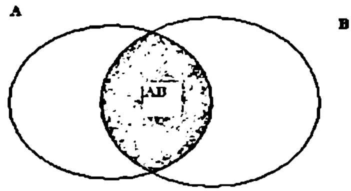
ოპერაციები ხდომილებებზე.

ხდომილებათა მოცემული სისტემის საშუალებით შესაძლებელია ახალი ხდომილებების აგება, ისევე როგორც სიმრავლეთა მოცემული სისტემის საშუალებით იგება ახალი სიმრავლეები მათი გაერთიანებებით, თანაკვეთებითა და დამატებებით.

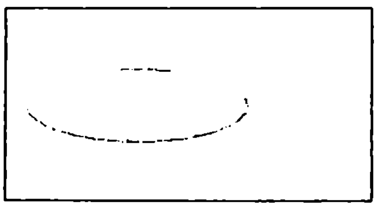
ორი A და B ხდომილების *გაერთიანება (ან ჯამი)* ეწოდება ისეთ ხდომილებას, რომელიც ხდება მაშინ, როცა ამ ხდომილებებიდან ერთი მა-

ინც ხდება და აღინიშნება სიმბოლოთი $A\cup B$ (ან $A+B$). სქემატურად ეს ასე გამოისახება:

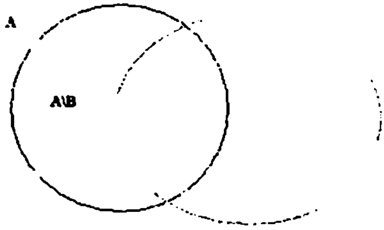
ორი A და B ხდომილების თანაკვეთა *(ან ნამრავლი)* ეწოდება ისეთ ხდომილებას, რომელიც ხდება მაშინ, როცა ეს ხდომილებები ერთდროულად ხდება და აღინიშნება სიმბოლოთი $A\cap B$ (ან AB). სქემატურად ეს ასე გამოისახება:



A ხდომილების *საწინააღმდეგო* ხდომილება ეწოდება ისეთ ხდომილებას, რომელიც ხდება მაშინ, როცა A არ ხდება და აღინიშნება სიმბოლოთი $\bar{A}$. სქემატურად, თუ წარმოვიდგენთ, რომ ელემენტარულ ხდომილებათა სივრცე მართკუთხედია, ხოლო A ხდომილება – წრე, მაშინ საწინააღმდეგო ხდომილება იქნება ოთხკუთხედის გაფერადებული ნაწილი:

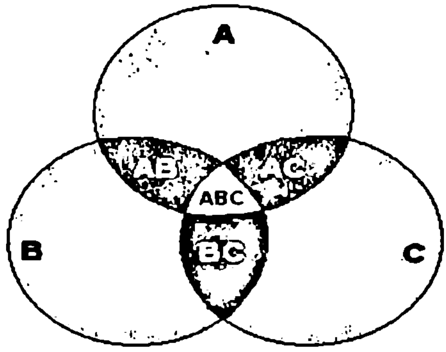


ორი A და B ხდომილების *სხვაობა* ეწოდება ისეთ ხდომილებას, რომელიც ხდება მაშინ, როცა ხდება A მაგრამ არ ხდება B და აღინიშნება სიმბოლოთი $A\setminus B$. სქემატურად ეს ასე გამოისახება:



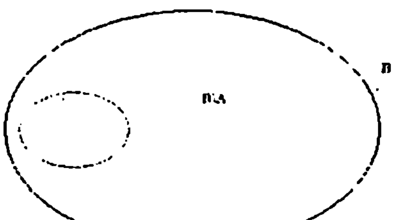
ცხადია, რომ ორი A და B ხდომილების სხვაობა აგრეთვე შეიძლება წარმოდგეს, როგორც A ხდომილებისა და B ხდომილების საწინააღმდეგო $\overline{B}$ ხდომილების თანაკვეთა: $A \setminus B = A \cap \overline{B}$.

გასაგებია, რომ მას შემდეგ რაც ჩვენ განვმარტეთ ორი ხდომილების გაერთიანება და თანაკვეთა, ბუნებრივად შესაძლებელია ხდომილებათა ნებისმიერი რაოდენობის გაერთიანებისა და თანაკვეთის განმარტება. ასე მაგალითად, სქემატურად სამი A, B და C ხდომილებისათვის თანაკვეთა ABC იქნება:



ორ A და B ხდომილებას ეწოდება *არათავსებადი (უთავსებადი, შეუთავსებელი)*, თუ მათი ერთდროულად მოხდენა შეუძლებელია, ან რაც იგივეა მათი თანაკვეთა არის შეუძლებელი ხდომილება: $A \cap B = \emptyset$. სიმრავლეთა თეორიის ენაზე ეს ნიშნავს, რომ ეს ორი სიმრავლე თანაუკვეთია.

ცხადია, რომ რაიმე A ხდომილება და მისი საწინააღმდეგო $\overline{A}$ ხდომილება უთავსებადია – $A \cap \overline{A} = \emptyset$. გარდა ამისა, $A \cup \overline{A} = \Omega$. თუ A ხდომილება იწვევს B ხდომილებას ($A \subset B$), მაშინ ცხადია, რომ $A \cap B = A$, $A \cup B = B$, $A \setminus B = \emptyset$, ხოლო სხვაობა $B \setminus A$ სქემატურად ასე გამოისახება:



იმისათვის, რომ დავინახოთ რა განსხვავებაა და რა აქვთ საერთო სიმრავლეთა თეორიისა და ალბათობის თეორიის ტრადიციულ ტერმინებს, ქვემოთ ჩვენ მოვიყვანთ შესაბამის ცხრილს:

აღნიშვნები	სიმრავლეთა თეორიის ინტერპრეტაცია	ალბათობის თეორიის ინტერპრეტაცია
ω	ელემენტი, წერტილი	შედეგი, ელემენტარული ხდომილება
Ω	წერტილთა სიმრავლე	ელემენტარულ ხდომილებათა სივრცე, აუცილებელი ხდომილება
A	წერტილთა სიმრავლე	ხდომილება (თუ შედეგი $\omega \in A$, მაშინ ამბობენ, რომ მოხდა A ხდომილება)
$\bar{A}=\Omega \setminus A$	A სიმრავლის დამატება, ე.ი. იმ წერტილების სიმრავლე, რომლებიც არ შედიან A-ში	ხდომილება, რომელიც მდგომარეობს A-ს არ მოხდენაში
$A \cup B$	A და B სიმრავლეების გაერთიანება, ე.ი. სიმრავლე იმ წერტილების, რომლებიც შედიან ან A-ში ან B-ში	ხდომილება, რომელიც მდგომარეობს A და B ხდომილებებიდან ერთის მაინც მოხდენაში
$A \cap B$	A და B სიმრავლეების თანაკვეთა, ე.ი. სიმრავლე იმ წერტილების, რომლებიც შედიან როგორც A, ისე B ხდომილებაში	ხდომილება, რომელიც მდგომარეობს A და B ხდომილებების ერთდროულ მოხდენაში
$\varnothing$	ცარიელი სიმრავლე	შეუძლებელი ხდომილება
$A \cap B=\varnothing$	A და B სიმრავლეები არ იკვეთებიან	A და B ხდომილებები არათავსებადია (მათი ერთდროულად მოხდენა შეუძლებელია)
$A+B$	სიმრავლეთა ჯამი, ე.ი. თანაუკვეთი სიმრავლეების გაერთიანება	ხდომილება, რომელიც მდგომარეობს ორი უთავსებადი ხდომილებიდან ერთ-ის მოხდენაში
$A \setminus B$	A და B სიმრავლეების სხვაობა, ე. ი. სიმრავლე იმ წერტილების, რომლებიც შედიან A-ში, მაგრამ არ შედიან B-ში	ხდომილება, რომელიც მდგომარეობს A ხდომილების მოხდენაში და B ხდომილების არ მოხდენაში
$A \Delta B$	სიმრავლეების სიმეტრიული სხვაობა, ე.ი. სიმრავლე $(A \setminus B) \cup (B \setminus A)$	ხდომილება, რომელიც მდგომარეობს A და B ხდომილებებიდან ერთის მოხდენაში, მაგრამ არა ორივეს ერთდროულად
$\bigcup_{n=1}^{\infty} A_n$	$A_1, A_2, \ldots$ სიმრავლეების გაერთიანება, ე.ი. იმ წერტილების სიმრავლე, რომლებიც შედიან ერთერთში მაინც	ხდომილება, რომელიც მდგომარეობს $A_1, A_2, \ldots$ ხდომილებებიდან ერთერთის მაინც მოხდენაში
$\sum_{n=1}^{\infty} A_n$	$A_1, A_2, \ldots$ სიმრავლეების ჯამი, ე.ი. გაერთიანება წყვილ-წყვილად თანაუკვეთი სიმრავლეების	ხდომილება, რომელიც მდგომარეობს $A_1, A_2, \ldots$ უთავსებადი ხდომილებებიდან ერთის მოხდენაში
$\bigcap_{n=1}^{\infty} A_n$	$A_1, A_2, \ldots$ ხდომილებების თანაკვეთა, ე.ი. სიმრავლე იმ წერტილების, რომლებიც შედიან ყველა ხდომილებაში	ხდომილება, რომელიც მდგომარეობს $A_1, A_2, \ldots$ ხდომილებების ერთდროულ მოხდენაში

ალბათობის განმარტება.

განმარტება 1. თუ ელემენტარული ხდომილებათა სივრცის ნებისმიერ ელემენტარულ ხდომილებას ω_i შეესაბამება გარკვეული რიცხვები $p_i = P(\omega_i)$, რომლებიც აკმაყოფილებენ პირობებს: $0 \le p_i \le 1$ და $\sum_{i=1}^{N} p_i = 1$, მაშინ ამ რიცხვებს ეწოდებათ ω_i ელემენტარული ხდომილებების *ალბათობები* (P – არის პირველი ასო ინგლისური სიტყვის "Probability", რომელიც ნიშნავს ალბათობას).

მონეტის ერთჯერ აგდებისას (Ω={გ, ს}) ცხადია უნდა დავუშვათ, რომ: P{გ}$= p$, P{ს}$=1-p$, სადაც $0 \le p \le 1$. შემთხვევას, როცა $p = 1/2$ ეწოდება *სიმეტრიული მონეტის აგდება*. მონეტის ორჯერ აგდებისას (Ω={გგ, გს, სგ, სს}) – P{გგ}$= p_1$, P{გს}$= p_2$, P{სგ}$= p_3$, P{სს}$= p_4$, სადაც ყველა $0 \le p_i \le 1$ და $\sum_{i=1}^{4} p_i = 1$. შემთხვევას, როცა ყველა $p_i = 1/4$, ჩვენ ვუწოდებთ *სიმეტრიული მონეტის ორ დამოუკიდებელ აგდებას*.

განმარტება 2. A ხდომილების ალბათობა $P(A)$ ეწოდება მასში შემავალი ელემენტარული ხდომილებების ალბათობების ჯამს:

$$P(A) = \sum_{\omega_i \in A} P(\omega_i).$$

აქ და ყველგან შემდგომში სიმბოლოთი $\omega \in A$ აღინიშნება ის ფაქტი, რომ ω ელემენტარული ხდომილება *ეკუთვნის* A ხდომილებას.

ცხადია, რომ $P(\Omega) = 1$ და $P(\varnothing)=0$.

თუ ყველა ელემენტარული ხდომილება ტოლალბათურია ანუ ყველა $P(\omega_i) = 1/N$, $i = 1,2,...,N$, მაშინ ამბობენ, რომ გვაქვს *კლასიკური ალბათური მოდელი* (პრაქტიკულ სიტუაციაში ეს ნიშნავს, რომ შემთხვევითი ექსპერიმენტის ყველა შედეგი ერთნაირად მოსალოდნელია და ყველას განხორციელების ერთი და იგივე შანსი აქვს). ამ შემთხვევაში განმარტება 2 გვაძლევს, რომ კლასიკურ მოდელში ხდომილების ალბათობა დაემთხვევა *ალბათობის კლასიკურ განმარტებას*, რომელიც შემოღებული იყო 1812 წელს *ლაპლასის* მიერ:

ალბათობის კლასიკურ განმარტება. თუ შემთხვევითი ექსპერიმენტის შესაძლო შედეგთა რაოდენობა სასრულია და ცალკეულ შედეგს აქვს განხორციელების თანაბარი შანსი, მაშინ ხდომილების ალბათობა ტოლია მასში შემავალ ელემენტარულ ხდომილებათა რაოდენობა გაყოფილი ყველა შესაძლო ელემენტარულ ხდომილებათა რაოდენობაზე:

$$P(A) = \frac{|A|}{|\Omega|},$$

სადაც $|B|$ – აღნიშნავს B ხდომილებაში შემავალ ელემენტარულ ხდომილებათა რაოდენობას.

მოცემულ ხდომილებაში შემავალ ელემენტარულ ხდომილებებს უწოდებენ *ხელშემწყობ* ელემენტარულ ხდომილებებს. შესაბამისად, *ალბათობის კლასიკური განმარტება* ასე გამოითქმის: კლასიკურ მოდელში ხდომი-

ლების ალბათობა ტოლია ხელშემწყობ ელემენტარულ ხდომილებათა რაოდენობა გაყოფილი ყველა შესაძლო ელემენტარულ ხდომილებათა რაოდენობაზე.

ამოცანა 1. სიმეტრიული მონეტის ორი დამოუკიდებელი აგდებისას ვიპოვოთ ალბათობა იმისა, რომ გერბი მოვა ერთხელ მაინც.

ამოხსნა. ამ შემთხვევაში $\Omega=\{$გგ, გს, სგ, სს$\}$; $P\{$გგ$\}=P\{$გს$\}=$ $=P\{$სგ$\}=P\{$სს$\}=1/4$; $A=\{$გგ, გს, სგ$\}$ და $P(A)=3/4$.

ამოცანა 2 ("ბედნიერ" ბილეთებზე). 25 საგამოცდო ბილეთიდან 5 "ბედნიერია", ხოლო დანარჩენი 20 – "არა ბედნიერი". რომელ სტუდენტს აქვს "ბედნიერი" ბილეთის აღების მეტი ალბათობა: ვინც პირველი იღებს ბილეთს, თუ ვინც მეორე იღებს ბილეთს?

ამოხსნა. ავღნიშნოთ პირველი სტუდენტის მიერ აღებული ბილეთის ნომერი i ასოთი, ხოლო მეორე სტუდენტის მიერ აღებული ბილეთის ნომერი j ასოთი. დავუშვათ, რომ "ბედნიერი" ბილეთების ნომრებია: 1, 2, 3, 4, 5. მაშინ ცხადია, რომ $\Omega=\{(i,j):i,j=1,2,...,25,i\neq j\}$, $|\Omega|=25\cdot 24=600$ და ბუნებრივია ჩავთვალოთ, რომ ყველა ელემენტარული ხდომილება ტოლალბათურია: $P(i,j)=1/600$.

შემოვიღოთ ხდომილებები:

$A=\{$პირველმა სტუდენტმა აიღო კარგი ბილეთი$\}$,

$B=\{$მეორე სტუდენტმა აიღო კარგი ბილეთი$\}$,

მაშინ ამ ხდომილებებს აქვთ შემდეგი სახე:

$A=\{(i,j):i=1,...,5;j=1,...,25;i\neq j\}$ და $B=\{(i,j):i=1,...,25;j=1,...,5;i\neq j\}$.

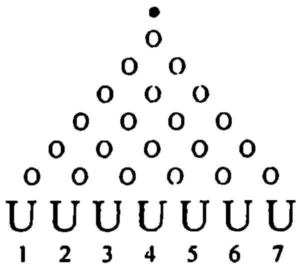
ადვილი დასანახია, რომ $|A|=|B|=5\cdot 24=120$. ამიტომ ალბათობის კლასიკური განმარტების თანახმად:

$$P(A)=P(B)=\frac{120}{600}=\frac{1}{5}.$$

ე. ი. ორივე მოსწავლეს აქვს კარგი ბილეთის აღების ერთი და იგივე ალბათობა.

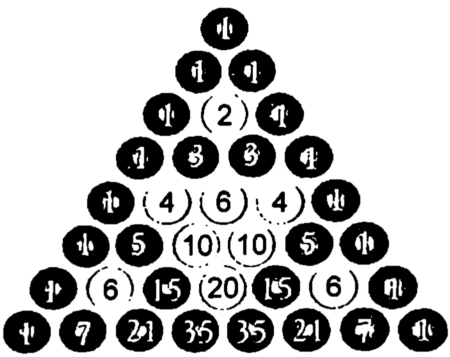
დავალება. წინა ამოცანაში ვიპოვოთ ალბათობა იმისა, რომ მესამე სტუდენტი ამოიღებს "ბედნიერ" ბილეთს.

ამოცანა 3 (გალტონის დაფა). გვაქვს დაფაზე სამკუთხედის ფორმით დალაგებული რგოლები, ისე რომ წვეროში ერთი რგოლია, მეორე რიგში წინასგან თანაბარ მანძილებზე ორი რგოლი, მესამე რიგში ზედა ორი რგოლიდან თანაბარ მანძილებზე სამი რგოლი და ა.შ. ბოლოში არის ექვსი რგოლი. მე-7 სტრიქონში კი არის ბოლო 6 რგოლიდან თანაბარ მანძილებზე 7 ღრმული. ზედა რგოლზე აგდებენ ბურთს და მას შეუძლია იგორაოს თანაბარი ალბათობით ან მარჯვნივ ან მარცხნივ რგოლიდან რგოლზე, რაც საბოლოოდ სრულდება რომელიმე ღრმულში ჩავარდნით. როგორია ალბათობა იმისა, რომ ბურთი ჩავარდება მეხუთე ღრმულში?



ამოხსნა. როგორც ვხედავთ არსებობს პირველ და მე-7 ღრმულებში ბურთის ჩავარდნის ერთადერთი გზა (ტრაექტორია), მეორე და მე-6 ღრმულებში ბურთის ჩავარდნის – ექვს-ექვსი გზა, მესამე და მეხუთე ღრმულებში – თხუთმეტ-თხუთმეტი გზა და ბოლოს, მეოთხე ღრმულში – ბურთის ჩავარდნის 20 გზა. გზების (შედეგების) სრული რაოდენობაა 1+6+15+20+15+6+1=64 და ყველა ეს შედეგი თანაბრად მოსალოდნელია, ვინაიდან თითოეული ტრაექტორიის გავლისას ბურთი განიცდის ექვს დაჯახებას რგოლებზე და ყოველი დაჯახებისას ის თანაბარი ალბათობებით გადაადგილდება ან მარჯვნივ ან მარცხნივ. თანაბარალბათური 64 შედეგიდან მეხუთე ღრმულში ჩავარდნას ხელს უწყობს 15 შედეგი და შესაბამისად, საძებნი ალბათობა იქნება 15/64.

ქვემოთ მოყვანილია გალტონის დაფაზე რგოლების 7 რიგის შემთხვევაში თითოეულ პოზიციაზე ბურთის მოხვედრის შესაძლო გზების რიცხვი.



ალბათობის კლასიკური განმარტება გამოსადეგია ამოცანათა მხოლოდ ძალიან ვიწრო კლასისათვის, სადაც ყველა შესაძლო შედეგთა სიმრავლე სასრულია და თითოეული შესაძლო შედეგი თანაბრად შესაძლებელია. უმრავლეს რეალურ ამოცანაში ეს პირობა ირღვევა, და შესაბამისად, კლასიკური განმარტების გამოყენება შეუძლებელია. ასეთ შემთხვევაში საჭიროა ხდომილების ალბათობის სხვა გზით განმარტება. ამ მიზნით, წინასწარ შემოვიღოთ A ხდომილების **ფარდობითი სიხშირის W(A)** ცნება, როგორც ცდათა იმ რიცხვის შეფარდება, რომელშიც დაფიქსირდა (განხორც-

იელდა) A ხდომილება, ჩატარებული ექსპერიმენტების საერთო რაოდენობასთან:

$$W(A)=\frac{M}{N},$$

სადაც N – ცდათა საერთო რიცხვია, M – A ხდომილების მოხდენათა რიცხვი.

ექსპერიმენტების დიდმა რაოდენობამ აჩვენა, რომ თუ ცდები ტარდება ერთი და იგივე პირობებში, მაშინ ცდათა (დაკვირვებათა) დიდი რიცხვისათვის, ფარდობითი სიხშირე უმნიშვნელოდ იცვლება, მერყეობს (ირხევა) რა გარკვეული მუდმივი რიცხვის ირგვლივ. ეს რიცხვი შეიძლება ჩაითვალოს განსახილველი ხდომილების ალბათობად.

განმარტება. ხდომილების **სტატისტიკურ ალაბათობად** ითვლება ამ ხდომილების ფარდობითი სიხშირე ან მასთან ახლოს მყოფი რიცხვი (მათემატიკურად ზუსტი ფორმულირება ასეთია: $P(A)=\lim\limits_{N\to\infty}W_N(A)$).

შენიშვნა. იმისათვის, რომ არსებობდეს A ხდომილების სტატისტიკური ალბათობა, საჭიროა:

ა). ექსპერიმენტების უსასრულოდ დიდი რიცხვის ჩატარების შესაძლებლობა;

ბ). სხვადასხვა სერიებში A ხდომილების მოხდენის ფარდობითი სიხშირეების მდგრადობა საკმარისად დიდი რაოდენობა ცდებისათვის.

გეომეტრიული ალბათობა.

ალბათობის კლასიკური განმარტების ერთ-ერთი ნაკლი ისაა, რომ მისი გამოყენება არ შეიძლება იმ ექსპერიმენტებისათვის, რომლებსაც გააჩნიათ შედეგების უსასრულო რაოდენობა. ასეთ შემთხვევაში შესაძლებელია ვისარგებლოთ **გეომეტრიული ალბათობის ცნებით.**

დავუშვათ რომ L მონაკვეთზე შემთხვევით აგდებენ წერტილს. რაც იმას ნიშნავს, რომ წერტილი აუცილებლად დაეცემა L მონაკვეთზე და ამასთანავე, თანაბარი შესაძლებლობებით ის შესაძლებელია დაემთხვეს ამ მონაკვეთის ნებისმიერ წერტილს. ამავე დროს, ალბათობა იმისა, რომ წერტილი დაეცემა L მონაკვეთის ნებისმიერ ნაწილზე არაა დამოკიდებული ამ ნაწილის განლაგებაზე მონაკვეთის შიგნით და პროპორციულია მისი სიგრძის. მაშინ ალბათობა იმისა, რომ აგდებული წერტილი დაეცემა l მონაკვეთზე, რომელიც არის ნაწილი L მონაკვეთის, გამოითვლება ფორმულით:

$$P=|l|/|L|, \tag{1}$$

სადაც $|l|$ – l მონაკვეთის სიგრძეა, ხოლო $|L|$ – L მონაკვეთის სიგრძე.

ანალოგიურად, შესაძლებელია ამოცანის დასმა წერტილისათვის, რომელსაც ვაგდებთ ბრტყელ S არეზე და ალბათობა იმისა, რომ ის მოხვდება ამ არის s ნაწილში განიმარტება როგორც:

$$P=|s|/|S|, \tag{2}$$

სადაც $|s|$ – s არის ფართობია, ხოლო $|S|$ – S არის ფართობი.

სამგანზომილებიან შემთხვევაში ალბათობა იმისა, რომ შემთხვევითი გზით V სხეულში ჩავარდნილი წერტილი აღმოჩნდება ამ სხეულის v ნაწილში, გამოითვლება ფორმულით:

$$P = |v| / |V|, \qquad (3)$$

სადაც $|v|$ – v სხეულის მოცულობაა, ხოლო $|V|$ – V სხეულის მოცულობა.

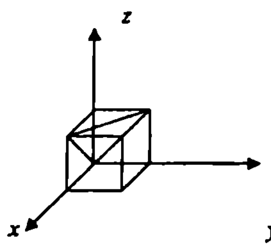
ამოცანა 1. ვიპოვოთ ალბათობა იმისა, რომ წრეში შემთხვევით ჩაგდებული წერტილი არ ჩავარდება ამ წრეში ჩახაზულ წესიერ ექვსკუთხედში.

ამოხსნა. დავუშვათ, რომ წრის რადიუსია R, მაშინ მასში ჩახაზული წესიერი ექვსკუთხედის გვერდიც იქნება R. ამასთანავე, წრის რადიუსია $|S| = \pi R^2$, ხოლო ექვსკუთხედის ფართობია – $|s| = \frac{3\sqrt{3}}{2} R^2$. ამიტომ საძიებელი ალბათობა იქნება:

$$P = \frac{|S| - |s|}{|S|} = \frac{\pi R^2 - \frac{3\sqrt{3}}{2} R^2}{\pi R^2} = \frac{\pi - 3\sqrt{3}}{2\pi} \approx 0{,}174.$$

ამოცანა 2. AB მონაკვეთზე შემთხვევით აგდებენ სამ წერტილს C, D და M. ვიპოვოთ ალბათობა იმისა, რომ AC, AD და AM მონაკვეთებისაგან შეიძლება აიგოს სამკუთხედი?

ამოხსნა. ავღნიშნოთ AC, AD და AM მონაკვეთების სიგრძეები შესაბამისად x, y და z-ით და ელემენტარულ ხდომილებათა სივრცის როლში განვიხილოთ სამგანზომილებიანი სივრცის წერტილთა სიმრავლე კოორდინატებით (x, y, z). თუ ჩავთვლით, რომ AB მონაკვეთის სიგრძე ტოლია 1-ის, მაშინ ელემენტარულ ხდომილებათა სივრცე იქნება კუბი, რომლის წიბოა ერთი. ამავე დროს, ხელშემწყობ ელემენტარულ ხდომილებათა სიმრავლე (სამკუთხედის აქსიომის თანახმად)შედგება იმ წერტილებისაგან, რომელთა კოორდინატებისათვის სრულდება სამკუთხედის უტოლობები: : $x + y > z$, $x + z > y$, $y + z > x$. ეს კი წარმოადგენს კუბის ნაწილს, რომელიც მოჭრილია მისგან სიბრტყეებით: $x + y = z$, $x + z = y$, $y + z = x$



(ერთ-ერთი ამ სიბრტყიდან, კერძოდ $x + y = z$, მოყვანილია ნახაზზე). ყოველი ასეთი სიბრტყე კუბიდან მოჭრის პირამიდას, რომლის მოცულობა ტოლია

$$\frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6}.$$

შესაბამისად, კუბის დარჩენილი ნაწილის მოცულობა იქნება

$$|v| = 1 - 3\cdot\frac{1}{6} = \frac{1}{2}.$$

ამიტომ საძიებელი ალბათობა, განმარტების თანახმად, იქნება

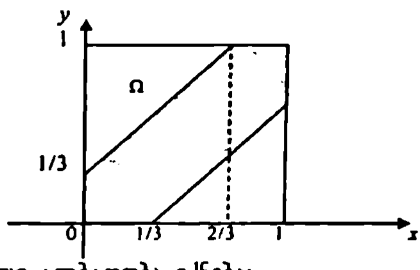
$$|P| = \frac{|v|}{|V|} = \frac{1}{2} : 1 = \frac{1}{2}.$$

ამოცანა 3 (შეხვედრის ამოცანა). ორი პირი შეთანხმდა გარკვეულ ადგილზე შეხვდეს ერთმანეთს 6-დან 7 საათამდე. თითოეული მათგანი შემთხვევით მომენტში მიდის დათქმულ ადგილას და მეორეს ელოდება 20 წუთის მანძილზე (შემდეგ კი მიდის). ვიპოვოთ ალბათობა იმისა, რომ ეს პირები შეხვდებიან ერთმანეთს?

ამოხსნა. ავღნიშნოთ, ერთ-ერთი პირის დათქმულ ადგილზე მოსვლის დრო $6+x$-ით, ხოლო მეორე პირის – $6+y$-ით (სადაც x და y გამოსახულია საათებში). ელემენტარულ ხდომილებათა სივრცის როლში შეგვიძლია ავიღოთ იმ (x,y) წერტილთა სიმრავლე, რომლებიც ეკუთვნიან ერთეულოვან კვადრატს. ამ შემთხვევაში ჩვენთვის საინტერესო წერტილთა (ხელშემწყობ ელემენტარულ ხდომილებათა) სიმრავლე იქნება ერთეულოვანი კვადრატის იმ წერტილთა სიმრავლე, რომელთა კოორდინატები ერთმანეთისაგან დაშორებულია არაუმეტეს $20/60=1/3$-ით:

$$A=\{(x,y);|x-y|\le 1/3\}\cap\Omega.$$

ვინაიდან, $|x-y|\le 1/3 \Leftrightarrow -1/3\le y-x\le 1/3 \Leftrightarrow x-1/3\le y\le x+1/3$. ამიტომ ადვილი გასაგებია, რომ A სიმრავლე იქნება ერთეულოვანი კვადრატის შიგნით $y=x-1/3$ და $y=x+1/3$ წრფეებს შორის მოქცეული გაფერადებული არე (როცა $0\le x\le 1/3$, მაშინ $y=x-1/3$ წრფის ნაცვლად ქვედა საზღვრის როლში იქნება x ღერძი: $0\le y\le x+1/3$, ხოლო როცა $2/3\le x\le 1$, მაშინ ზედა საზღვარი $y=x+1/3$ წრფის ნაცვლად იქნება $y=1$ წრფე: $x-1/3\le y\le 1$).



ამიტომ საძიებელი ალბათობა იქნება:

$$P(A)=\frac{|A|}{|\Omega|}=\frac{1-2/3\cdot 2/3}{1}=\frac{5}{9}.$$

დავალება 1. იპოვეთ ალბათობა იმისა, რომ ერთეულოვან გვერდიან კვადრატში შემთხვევით შერჩეული წერტილი ამ კვადრატის უახლოესი გვერდიდან დაშორებული იქნება არაუმეტეს 0,15-ით?

დავალება 2. ორი სიგნალი მიმღებ მოწყობილობაზე T დროის მანძილზე შემთხვევით მომენტში მიიღება. მოწყობილობა მათ განარჩევს, თუ

ისინი t დროით მაინც არიან დაცილებული ერთმანეთს. იპოვეთ ალბათობა იმისა, რომ ორივე სიგნალი მიღებული იქნება?

კომბინატორიკის ელემენტები.

ძალიან ბევრ შემთხვევაში ელემენტარულ ხდომილებათა სივრცისა და ხდომილების ალბათობის განსაზღვრა პირდაპირი გადათვლით შეუძლებელია. ხდომილებათა ალბათობების გამოთვლის დროს ხშირად საჭირო ხდება კომბინატორიკის ფორმულების გამოყენება. **კომბინატორიკა** – არის მეცნიერება, რომელიც სწავლობს კომბინაციებს, რომელთა შედგენაც შესაძლებელია გარკვეული სასრული სიმრავლის ელემენტებიდან განსაზღვრული წესების გამოყენებით. განვსაზღვროთ ძირითადი კომბინაციები.

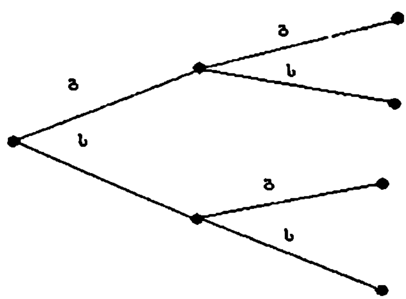
პირველ რიგში ჩამოვაყალიბოთ კომბინატორული ანალიზის ძირითადი პრინციპი, რომელსაც **გამრავლების პრინციპი** ეწოდება:

თუ ასარჩევია m ობიექტი და არსებობს პირველი ობიექტის არჩევის n_1 შესაძლებლობა, პირველი ობიექტის არჩევის შემდეგ არსებობს მეორე ობიექტის არჩევის n_2 შესაძლებლობა, და ა. შ., $m-1$ ობიექტის არჩევის შემდეგ არსებობს m-ური ობიექტის არჩევის n_m შესაძლებლობა, მაშინ არსებობს ამ m ობიექტის ამ მიმდევრობით არჩევის $n_1 \times n_2 \times \cdots \times n_m$ შესაძლებლობა.

მაგალითად, თუ მამაკაცს აქვს 4 პერანგი და 2 პიჯაკი, მაშინ ამ მამაკაცს აქვს პერანგისა და პიჯაკის შერჩევის $4 \times 2 = 8$ შესაძლებლობა (ვარიანტი).

ნამრავლის პრინციპთან დაკავშირებით ხშირად სასარგებლოა ხის მსგავსი (ხისებრი) დიაგრამის ანუ *დენდროგრამის* გამოყენება.

მაგალითად, დენდროგრამით გამოსახული მონეტის ორჯერ აგდების შესაბამისი შედეგების სიმრავლე იქნება:



განმარტება 1. გადანაცვლებები – ეს არის კომბინაციები, რომლებიც შედგენილია მოცემული n ელემენტიანი სიმრავლის ყველა n ელემენტისაგან და ერთმანეთისაგან განსხვავდება მხოლოდ ელემენტების განლაგების რიგით. n ელემენტიანი სიმრავლის ყველა შესაძლო გადანაცვლებათა რიცხვია:

$$P_n = n!.$$

მართლაც, თუ ამ ამოცანას შევხედავთ როგორც n ობიექტიდან n ობიექტის არჩევის ამოცანას, მაშინ არსებობს პირველი ობიექტის არჩევის n შესაძლებლობა, პირველი ობიექტის არჩევის შემდეგ არსებობს მეორე ობიექტის არჩევის $n-1$ შესაძლებლობა, და ა. შ., მე-n ობიექტის არჩევის ერთადერთი შესაძლებლობა. შესაბამისად, გამრავლების პრინციპის თანახმად:

$$P_n = n\cdot(n-1)\cdots1 = n!$$

მაგალითი 1. რამდენი განსხვავებული სია შეიძლება შედგენილ იქნეს 7 სხვადასხვა გვარისაგან?

ამოხსნა. $P_7 = 7! = 2\cdot3\cdot4\cdot5\cdot6\cdot7 = 5040.$

განმარტება 2. წყობები – ეს არის m ელემენტიანი კომბინაციები n განსხვავებული ელემენტის მქონე სიმრავლიდან, რომლებიც ერთმანეთისაგან განსხვავდებიან ან ელემენტების შემადგენლობით ან ელემენტების განლაგების რიგით. სხვა სიტყვებით, წყობა არის n ელემენტიანი სიმრავლის m ელემენტიან, დალაგებულ ქვესიმრავლეთა რაოდენობა. ყველა შესაძლო წყობების რიცხვია:

$$A_n^m = n!/(n-m)!\,.$$

ამ ამოცანას ჩვენ შეგვიძლია შევხედოთ, როგორც n ობიექტიდან m ობიექტის შერჩევის ამოცანას. რადგანაც, არსებობს პირველი ობიექტის არჩევის n შესაძლებლობა, პირველი ობიექტის არჩევის შემდეგ არსებობს მეორე ობიექტის არჩევის $n-1$ შესაძლებლობა, და ა. შ., მე-m ობიექტის არჩევის $n-m+1$ შესაძლებლობა, ამიტომ გამრავლების პრინციპის საფუძველზე გვაქვს:

$$A_n^m = n(n-1)(n-2)\ldots(n-m+1) = n!/(n-m)!$$

მაგალითი 2. შეჯიბრებაში მონაწილე 10 სპორტსმენიდან პირველ სამ ადგილზე გასული სამი გამარჯვებული რამდენ სხვადასხვანაირად შეიძლება განთავსდეს დასაჯილდოებელ კვარცხლბეკზე?

ამოხსნა. $A_{10}^3 = 10\cdot9\cdot8 = 720.$

განმარტება 3. ჯუფდებები – ეს არის n ელემენტიანი სიმრავლის დაულაგებელი m ელემენტიანი ქვესიმრავლეები (ანუ ისეთი ერთობლიობები, რომლებიც ერთმანეთისაგან განსხვავდებიან მხოლოდ ელემენტების შემადგენლობით). ჯუფდებათა რიცხვი ტოლია:

$$C_n^m = \frac{n!}{m!(n-m)!}\,.$$

ვინაიდან, n ელემენტიანი სიმრავლის ყველა m ელემენტიანი, დალაგებული ქვესიმრავლის მისაღებად, ჩვენ შეგვიძლია ავიღოთ n ელემენტიანი სიმრავლის ყველა დაულაგებელი m ელემენტიანი ქვესიმრავლე (რომელთა რაოდენობაა C_n^m) და თითოეულში მოვახდინოთ ყველანაირი გადანაცვლებები (ამის შესაძლებლობაა $m!$), ამიტომ ადგილი აქვს თანაფარდობას:

$$C_n^m\cdot m! = A_n^m\,.$$

აქედან ცხადია, რომ

$$A_n^m = \frac{C_n^m}{m!} = \frac{n!}{m!(n-m)!}$$

მაგალითი 3. შესარჩევ შეჯიბრებაში მონაწილეობს 10 ადამიანი, რომელთაგან ფინალში გადის სამი. ფინალისტების რამდენი განსხვავებული სამეული შეიძლება გამოვლინდეს?

ამოხსნა. წინა მაგალითისაგან განსხვავებით, აქ ფინალისტების რიგს (დალაგებას) მნიშვნელობა არა აქვს. ამიტომ ვიყენებთ ჯუფდების ფორმულას. შესაბამისად, საძიებელი რიცხვია:

$$C_{10}^3 = \frac{10!}{3!\cdot 7!} = \frac{8\cdot 9\cdot 10}{6} = 120.$$

ალბათობის გამოთვლა კომბინატორიკის გამოყენებით.

განვიხილოთ მაგალითები, რომლებიც დაკავშირებულია ყუთიდან, რომელშიც M რაოდენობის ბურთია, სხვადასხვა გზებით n რაოდენობის ბურთის ამოღებასთან.

ამორჩევა დაბრუნებით. ამ შემთხვევაში ექსპერიმენტის ყოველ ნაბიჯზე ყუთიდან ამოღებული ბურთი უკან ბრუნდება. ვიგულისხმოთ, რომ ბურთები გადანომრილია რიცხვებით ერთიდან M-მდე. შესაბამისად, ელემენტარულ ხდომილებათა სივრცე შეიძლება წარმოდგენილ იქნეს იქნება როგორც რიცხვთა n-ეულები $a_1,\dots,a_n$, სადაც a_i აღნიშნავს i-ურ ნაბიჯზე ამოღებული ბურთის ნომერს. განასხვავებენ ორი ტიპის ამორჩევებს: **დალაგებული ამორჩევები** და **დაულაგებელი ამორჩევები**. პირველ შემთხვევაში ამორჩევები, რომლებიც შედგებიან ერთი და იგივე ელემენტებისგან, მაგრამ განსხვავდებიან ამ ელემენტების დალაგების რიგით, ცხადდება **განსხვავებულად**. მეორე შემთხვევაში ელემენტების დალაგების რიცხვი არ მიიღება მხედველობაში და ასეთი ამორჩევები ცხადდება **იდენტურად**. იმისათვის, რომ განვასხვავოთ ეს ამორჩევები დალაგებული ამორჩ-ევები აღინიშნება სიმბოლოთი $(a_1,\dots,a_n)$, ხოლო დაულაგებელი ამორჩევები – $[a_1,\dots,a_n]$.

ამგვარად, **დაბრუნებით დალაგებული ამორჩევის** შემთხვევაში ელემენტარულ ხდომილებათა სივრცეს აქვს შემდეგი სტრუქტურა

$$\Omega = \{\omega : \omega = (a_1,\dots,a_n) : a_i = 1,\dots,M;\ i = 1,\dots,n\}$$

და განსხვავებულ შედეგთა (ამორჩევათა) რაოდენობა, რომელსაც კომბინატორიკაში უწოდებენ *წყობებს* M-დან n *დაბრუნებებით*, ტოლია $|\Omega| = M^n$.

თუ კი ჩვენ ვიხილავთ **დაულაგებელ ამორჩევებს დაბრუნებით**, რომელსაც კომბინატორიკაში უწოდებენ *ჯუფდებას* M-დან n, მაშინ

$$\Omega = \{\omega : \omega = [a_1,\dots,a_n] : a_i = 1,\dots,M;\ i = 1,\dots,n\}.$$

გასაგებია, რომ განსხვავებულ დაულაგებელ ამორჩევათა რიცხვი ნაკლებია ვიდრე დალაგებულების რიცხვი. ვაჩვენოთ, რომ $|\Omega| = C_{M+n-1}^n$, სადაც $C_k^l := \dfrac{k!}{l!(k-l)!}$ – არის *ჯუფდებათა რიცხვი* k-*დან* l.

დამტკიცებას ჩვენ ჩავატარებთ მათემატიკური ინდუქციის პრინციპით. აღვ-ნიშნოთ $N(M,n)$-ით ჩვენთვის საინტერესო შედეგების რიცხვი. ცხადია, რომ თუ $k \le M$, მაშინ $N(k,1)=k=C_k^1$. დავუშვათ ახლა, რომ $N(k,n)=C_{k+n-1}^n$, $k \le M$ და ვაჩვენოთ, რომ ეს ფორმულა ძალაში დარჩება n-ის $n+1$-ით შეცვლისას. შევნიშნათ, რომ დაულაგებელი $[a_1,\dots,a_{n+1}]$ შერჩევების განხილვისას შეიძლება ვიგულისხმოთ, რომ მისი ელემენტები დალაგებულია კლების მიხედვით: $a_1 \le a_2 \le \cdots \le a_{n+1}$.

ადვილი დასანახია, რომ იმ დაულაგებელი $a_1 \le a_2 \le \cdots \le a_{n+1}$ შერჩევების რაოდე-ნობა, რომელთათვისაც $a_1 = 1$ ტოლია $N(M,n)$-ის, რომელთათვისაც $a_1 = 2$ – $N(M-1,n)$-ის და ა. შ. ამიტომ

$$N(M,n+1) = N(M,n) + N(M-1,n) + \cdots + N(1,n) =$$
$$= C_{M+n-1}^n + C_{M-1+n-1}^n + \cdots C_n^n = (C_{M+n}^{n+1} - C_{M+n-1}^{n+1}) +$$
$$+ (C_{M-1+n}^{n+1} - C_{M-1+n-1}^{n+1}) + \cdots + (C_{n+1}^{n+1} - C_n^n) + C_n^n = C_{M+n}^{n+1},$$

სადაც ჩვენ ვისარგებლეთ ბინომის კოეფიციენტების შემდეგი თვისებით:

$$C_k^{l-1} + C_k^l = C_{k+1}^l.$$

ამორჩევა დაბრუნების გარეშე. ვიგულისხმოთ, რომ $n \le M$ და ამოღებული ბურთი უკან არ ბრუნდება. ამ შემთხვევაშიც აგრეთვე განიხილება ორი შესაძლებლობა დაკავშირებული ამორჩევის დალაგება – დაულაგებლობასთან.

დაბრუნების გარეშე დალაგებული შერჩევის შემთხვევაში, რომელსაც კომბინატორიკაში უწოდებენ *წყობებს M-დან n დაბრუნებების გარეშე*, შედეგების სიმრავლეს აქვს სახე:

$$\Omega = \{\omega : \omega = (a_1,\dots,a_n) : a_k \ne a_l,\ k \ne l;\ a_i = 1,\dots,M;\ i = 1,\dots,n\},$$

ხოლო ამ სიმრავლის ელემენტთა რაოდენობა ტოლია

$|\Omega| = M(M-1)\cdots(M-n+1)$. ეს რიცხვი აღინიშნება სიმბოლოთი $(M)_n$ ან A_M^n და ეწოდება *წყობათა რიცხვი M-დან n*.

დაულაგებელი ამორჩევისას დაბრუნების გარეშე, რომელსაც კომბინატორიკაში უწოდებენ *ჯუფდებას M-დან n დაბრუნებების გარეშე*, შედეგთა სიმრავ-ლეს აქვს სახე

$$\Omega = \{\omega : \omega = [a_1,\dots,a_n] : a_k \ne a_l,\ k \ne l;\ a_i = 1,\dots,M;\ i = 1,\dots,n\}$$

და მისი ელემენტების რაოდენობა ტოლია $|\Omega| = C_M^n$. მართლაც, დაულაგებელი $[a_1,\dots,a_n]$ ერთობლიობიდან, რომელიც შედგება განსხვავებული ელემენტებისაგან, შეიძლება მივიღოთ ზუსტად $n!$ რაოდენობა დალაგებული ერთობლიობები. ამიტომ $|\Omega| \cdot n! = A_M^n$ და მაშასადამე, $|\Omega| = A_M^n / n! = C_M^n$.

საბოლოოდ ჩვენ გვაქვს ყუთიდან, რომელშიც M რაოდენობის ბურთია n რაოდენობის ბურთის ამოღების რიცხვის შემდეგი ცხრილი:

	დალაგებული	დაულაგებელი
დაბრუნებით	M^n	C^n_{M+n-1}
დაბრუნების გარეშე	A^n_M	C^n_M

მაგალითად, $M=3$ და $n=2$-ის შემთხვევაში შესაბამის ელემენტარულ ხდომილებათა სივრცეებს ექნებათ შემდეგი სახის სტრუქტურები:

			შერჩევა
	(1,1)(1,2)(1,3) (2,1)(2,2)(2,3) (3,1)(3,2)(3,3)	[1,1][2,2][3,3] [1,2][1,3][2,3]	დაბრუნებით
	(1,2)(1,3) (2,1)(2,3) (3,1)(3,2)	[1,2] [1,3] [2,3]	დაბრუნების გარეშე
ერთობლიობა	დალაგებული	დაულაგებელი	

ნაწილაკების განლაგება დანაყოფებში. განვიხილოთ საკითხი ელემენტარულ ხდომილებათა სივრცის სტრუქტურის შესახებ n ნაწილაკის (ბურთის) M დანაყოფში (ყუთში) განლაგების ამოცანაში. ასეთი ამოცანებთან საქმე გვაქვს სტატისტიკურ ფიზიკაში, როცა სწავლობენ n ნაწილაკის (ეს შეიძლება იყოს პროტონები, ნეიტრონები, და სხვა) M მდგომარეობაში (ეს შეიძლება იყოს ენერგეტიკული დონეები) განაწილების საკითხებს. ვიგულისხმოთ, რომ დანაყოფები გადანომრილია ნომრებით $1,2,\dots,M$ და თავიდან დავუშვათ, რომ ნაწილაკები გარჩევადია (განსხვავებულებია, აქვთ ნომრები $1,2,\dots,n$). მაშინ n ნაწილაკის M დანაყოფში განაწილების სრულიად აღიწერება დალაგებული ერთობლიობით $(a_1,\dots,a_n)$, სადაც a_i – წარმოადგენს იმ დანაყოფის ნომერს, რომელშიც მოხვდა ნაწილაკი ნომრით i. თუ კი განვიხილავთ განურჩეველ ნაწილაკებს, მაშინ მათი განაწილება M დანაყოფში სრულიად აღიწერება დაულაგებელი ერთობლიობით $[a_1,\dots,a_n]$, სადაც a_i – იმ დანაყოფის ნომერია, რომელშიც მოხვდა ნაწილაკი i-ურ ნაბიჯზე.

მაგალითები:

I. კურსზე, რომელზეც სამი ჯგუფია, ჯგუფხელების არჩევის ყველა შესაძ-ლო ვარიანტების რიცხვია $n_1 \cdot n_2 \cdot n_3$, სადაც n_i – i-ურ ჯგუფში სტუდენტების რაოდენობაა (ვიყენებთ ნამრავლის პრინციპს);

II. რაოდენობა ყველა შესაძლო კომბინაციების, რამდენნაირადაც შესაძლე-ბელია m მგზავრი განვათავსოთ n ვაგონში ტოლია n^m (დალაგებული ამორჩე-ვა დაბრუნებით, სადაც $M=n$ და $n=m$);

III. m ადამიანის დაბადების დღეების ყველა შესაძლო კომბინაცია (იმ პირობით, რომ თითეული დაბადების დღე არის რომელიმე 365 დღიდან) ტოლია 365^m (საქმე გვაქვს ამორჩევასთან დაბრუნებით, ამასთანავე ამორჩევები ითვლება დალაგებულად, სადაც $M=365$ და $n=m$);

IV. რაოდენობა ყველა შესაძლო კომბინაციების, რამდენნაირადაც შესაძლ-ებელია 5 ბურთი განვათავსოთ 5 ყუთში, ისე რომ ერთ ყუთში იყოს ერთი ბურთი, ტოლია 5! (ნამრავლის პრინციპის თანახმად);

V. პარტიების რაოდენობა, რომელიც უნდა შედგეს n მონაწილისაგან შემდგარ წრიულ საჭადრაკო ტურნირში ტოლია $C_n^2=n(n-1)/2$ (დაულაგებელი ამორჩევა დაბრუნების გარეშე).

ამოცანა 1 (დამთხვევებზე). ვიპოვოთ ალბათობა იმისა, რომ: ა). m შემთხ-ევვით არჩეული ადამიანის დაბადების დღეები არ დაემთხვევა ერთმანეთს (იმ პირობით, რომ ყველა დღე ტოლალბათურია); ბ). m შემთხვევით არჩეულ ადამიანში მოიძებნება ორი მაინც ისეთი, რომელთა დაბადების დღეები დაემთხვევა ერთმანეთს.

ამოხსნა. ა). ელემენტარულ ხდომილებათა სივრცე შეესაბამება დალაგებულ ამორჩევებს დაბრუნებით, სადაც $M=365$ და $n=m$, ანუ $|\Omega|=365^m$; ხოლო ხელშემწყობი ელემენტარული ხდომილებები კი დალაგებულ ამორჩევებს დაბრუნების გარეშე, სადაც აგრეთვე $M=365$ და $n=m$, ამიტომ მათი რაოდენობაა – A_{365}^m. შესაბამისად, ალბათობის კლასიკური განმარტების თანახმად, გვაქვს

$$P(m)=\frac{A_{365}^m}{365^m}=(1-\frac{1}{365})(1-\frac{2}{365})\cdots(1-\frac{m-1}{365}).$$

ბ). საწინააღმდეგო ხდომილების ალბათობის გამოთვლის წესის თანახმად კი გვაქვს, რომ

$$Q(m)=1-\frac{A_{365}^m}{365^m}.$$

მოვიყვანოთ ამ ალბათობის მნიშვნელობების ცხრილი ზოგიერთი m-ის შემთხვევაში:

m	4	16	22	23	40	64	70
Q(m)	0.01636	0.28360	0.47569	0.50730	0.89123	0.99711	0.99916

საინტერესოა აღინიშნოს, რომ (მოლოდინის საწინააღმდეგოდ!) კლასის მოსწავლეთა რაოდენობა, რომელშიც 1/2-ის ტოლი ალბათობით მოიძებნება ორი მოსწავლე მაინც ერთი და იგივე დაბადების დღეებით, არც ისე დიდია: იგი ტოლია მხოლოდ 23-ის.

ამოცანა 2 (მოგება ლატარეაში). გვაქვს M ბილეთი გადანომრილი რიცხვ-ებით ერთიდან M-მდე, რომელთაგან n ბილეთი ნომრებით ერთიდან n-მდე მომგებიანია ($M\geq 2n$). ვყიდულობთ n ცალ ბილეთს. რას უდრის ალბათობა იმისა, რომ ამ n ბილეთიდან ერთი მაინც იქნება მომგებიანი (აღვნიშნოთ ეს ხდომილება A-თი)?

ამოხსნა. ვინაიდან ბილეთების ამოღების (ყიდვის) თანმიმდევრობას არა აქვს მნიშვნელობა ნაყიდ ბილეთებში მომგებიანი ბილეთების არსებობის ან არ არსებობის თვალსაზრისით, ამიტომ ელემენტარულ ხდომილებათა სივრცეს ექნება შემდეგი სტრუქტურა:

$$\Omega = \{\omega : \omega = [a_1,...,a_n] : a_k \neq a_l, k \neq l : a_i = 1,...,M\}.$$

შესაბამისად, ჩვენს მიერ ზემოთმოყვანილი ცხრილის თანახმად $|\Omega| = C_M^n$.

ხდომილებას (ავღნიშნოთ იგი B_0-ით), რომ ნაყიდ ბილეთებში არ არის მო-მგებიანი ბილეთები, ექნება შემდეგი სტრუქტურა:

$$B_0 = \{\omega : \omega = [a_1,...,a_n] : a_k \neq a_l, k \neq l; a_i = n+1,...,M\} \text{ და } |B_0| = C_{M-n}^n.$$

ამიტომ

$$P(B_0) = \frac{C_{M-n}^n}{C_M^n} = \frac{A_{M-n}^n}{A_M^n} = (1-\frac{n}{M})(1-\frac{n}{M-1})\cdots(1-\frac{n}{M-n+1}).$$

შესაბამისად, საძებნი ალბათობა იქნება

$$P(B) = 1 - P(B_0) = 1-(1-\frac{n}{M})(1-\frac{n}{M-1})\cdots(1-\frac{n}{M-n+1}).$$

თუ მაგალითად, $M = n^2$ და $n \to \infty$, მაშინ $P(B_0) \to e^{-1}$ (აქ e ნეპერის რიცხვია) და $P(B) \to 1-e^{-1} \approx 0{,}632$, სადაც კრებადობის სიჩქარე საკმაოდ დიდია: უკვე როცა $n = 10$ – $P(B) = 0{,}670$.

დავალება. წინა ამოცანის პირობებში ვიპოვოთ ალბათობა იმისა, რომ ნაყიდი n ბილეთიდან ზუსტად m ($m \le n$) იქნება მომგებიანი.

ამოცანა 3 (ურთიერთობის უპირატესობაზე). დავუშვათ კლასში, რომელშიც 10 მოსწავლეა ტარდება გამოკითხვა, სადაც თითოეულმა მოსწავლემ ანკეტაში უნდა მიუთითოს ის სამი ამხანაგი, რომელსაც აძლევს უპირატესობას ცხრა ამხანაგიდან. A იყოს ხდომილება, რომ ერთერთი მოსწავლე დასახელებულ იქნა ყველა შესაძლო ცხრა ანკეტაში. რას უდრის მისი ალბათობა, თუ ანკეტის შევსება იყო შემთხვევითი, ანუ ანკეტის შევსების ნებისმიერი კომბინაცია ტოლალბათურია.

ამოხსნა. ცალკეული მოსწავლისათვის ანკეტის შევსების სხვადასხვა კომბინაციათა რაოდენობა ტოლია C_9^3-ის, ხოლო 10 ანკეტის შევსების ვარიანტების რაოდენობა პირველი მაგალითის ანალოგიურად ტოლია $(C_9^3)^{10}$-ის. ვინაიდან ერთი ანკეტა შეიძლება შევსებულ იქნეს ნებისმიერად, ხოლო დანარჩენ ცხრა ანკეტაში ერთი პასუხი დაფიქსირებულია, ხოლო დანარჩენი ორი პასუხი კი შეიძლება ნებისმიერად ამოირჩეს რვა შესაძლებელი პასუხიდან, ამიტომ ელემენტარული ხდომილებების რაოდენობა A-ში ტოლია $|A| = C_9^3 \cdot (C_8^2)^9$ (თუ წყვილის პირველ კომპონენტს შეუძლია მიიღოს m განსხვავებული მნიშვნელობა, ხოლო მეორე კომპონენტს კი პირველისგან დამოუკიდებლად – n განსხვავებული მნიშვნელობა, მაშინ ასეთი წყვილების რაოდენობა ნამრავლის პრინციპის თანახმად იქნება – $m \cdot n$). აქედან გვაქვს

$$P(A)=\frac{C_9^1\cdot(C_8^2)^9}{(C_9^3)^{10}}=(\frac{C_8^2}{C_9^3})^9=\frac{1}{3^9}.$$

დავალება. წინა ამოცანაში იპოვეთ ალბათობა იმისა, რომ ერთერთი სტუდენტი დასახელებული იქნება k-ჯერ ($k\le 9$).

ამოცანა 4 (ორ "ტუზზე"). განვიხილოთ პრეფერანსის თამაში, როდესაც კარტის შეკვრის მაღალი 32 კარტი შემთხვევით ნაწილდება (რიგდება) სამ მოთამაშეს შორის, ისე რომ თითეული ღებულობს 10 კარტს და ორი კარტი ინახება "საყიდლებში". როგორია ალბათობა იმისა, რომ "საყიდლებში" აღმოჩნდება ორი "ტუზი"?

ამოხსნა. ორი კარტის სხვადასხვა კომბინაციების რაოდენობა, რომელიც შეიძლება აღმოჩნდეს "საყიდლებში" ტოლია $C_{32}^2=496$. კარტის შეკვრაში ოთხი "ტუზია" და სხვადასხვა კომბინაციების რაოდენობა, რომელიც მოგვცემდა ორ "ტუზს" ტოლია $C_4^2=6$. შესაბამისად, საძიებელი ალბათობა ტოლია

$$C_4^2\big/C_{32}^2=6\big/496=0{,}012.$$

დავალება. დავუშვათ წინა ამოცანაში ერთერთმა მოთამაშემ, ნახა რა თავისი კარტები, იცის, რომ მას "ტუზი" არა აქვს. შეიცვლება თუ არა მაშინ ალბათობა იმისა, რომ "საყიდლებში" ორი "ტუზია"? გამოთვალეთ ეს ალბათობა.

ამოცანა 5 (მხედველობით მოძებნაზე). დავუშვათ გვაქვს N ცალი შემთხვევით დალაგებული გეომეტრიული ფიგურა, რომელთა შორის M ცალი მართკუთხედია ($M\le N$). მოითხოვება მოიძებნოს ყველა მართკუთხედი, თუ ძებნა წარმ-ოებს ელემენტების (ფიგურების) სათითაოდ სკანირებით ფიქსაციის მოცულობით ერთი ელემენტი, ამასთანავე ხდება დაკვირვებული ელემენტის პოზიციის დამახსოვრება და მას თავიდან აღარ ვუბრუნდებით. ვიპოვოთ ალბათობა იმისა, რომ დამკვირვებელი შეძლებს აღმოაჩინოს ყველა M მართკუთხედი არაუმეტეს n დაკვირვებისას ($n=M,\dots,N$)?

ამოხსნა. ყველა შესაძლო ელემენტარულ ხდომილებათა რაოდენობა იქნება $|\Omega|=C_N^n$. ხელშემწყობი ელემენტარული ხდომილებები ისეთი $[a_1,\dots,a_n]$ ერთობლიობებია, რომლებშიც M ადგილას განთავსებულია მართკუთხედები (ამის შესაძლებლობათა რაოდენობაა C_M^M), ხოლო დანარჩენ $N-M$ ადგილას კი არა მართკუთხედები (ამის შესაძლებლობათა რაოდენობაა C_{N-M}^{n-M}). ამიტომ პირველი მაგალითის ანალოგიურად ხელშემწყობი ელემენტარული ხდომილებების რაოდენობა იქნება $C_M^M\cdot C_{N-M}^{n-M}$ შესაბამისად, საძიებელი ალბათობა ტოლია

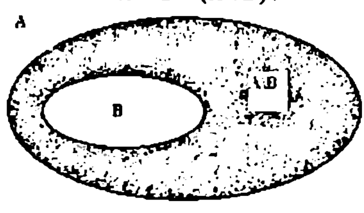
$$P_M(n)=\frac{C_M^M\cdot C_{N-M}^{n-M}}{C_N^n}=\frac{C_{N-M}^{n-M}}{C_N^n}=\frac{C_n^M}{C_N^M}.$$

თავი II
რთული ხდომილების ალბათობები

გავიხსენოთ, რომ A ხდომილების ალბათობა ეწოდება მასში შემავალი ელემენტარული ხდომილებების ალბათობების ჯამს:
$$P(A)=\sum_{\omega_i\in A}P(\omega_i).$$
აქედან ჩვენ შეგვიძლია მივიღოთ ე. წ. **ალბათობათა შეკრების კანონი**: თუ A და B ხდომილებები უთავსებადია ($A\cap B=\varnothing$), მაშინ ხდომილებათა ჯამის ალბათობა ალბათობები ჯამის ტოლია
$$P(A+B)=P(A)+P(B). \qquad (1)$$
მართლაც, გვაქვს:
$$P(A+B)=\sum_{\omega_i\in(A+B)}P(\omega_i)=\sum_{\omega_i\in A}P(\omega_i)+\sum_{\omega_i\in B}P(\omega_i)=P(A)+P(B).$$
შედეგი 1. საწინააღმდეგო ხდომილების ალბათობა ტოლია ერთს გამოკლებული თვითონ ამ ხდომილების ალბათობა:
$$P(\bar{A})=1-P(A).$$
დამტკიცება. ვინაიდან $A\cap\bar{A}=\varnothing$ და $A\cup\bar{A}=\Omega$, ამიტომ ალბათობათა შეკრების კანონის თანახმად გვაქვს:
$$1=P(\Omega)=P(A\cup\bar{A})=P(A)+P(\bar{A}),$$
საიდანაც ვრწმუნდებით შედეგის სისწორეში.

შედეგი 2 (სხვაობის ალბათობის ფორმულა). თუ B ხდომილება იწვევს A ხდომილებას ($B\subset A$), მაშინ A და B ხდომილებების სხვაობის ალბათობა ალბათობათა სხვაობის ტოლია: $P(A\setminus B)=P(A)-P(B)$.

დამტკიცება. ამ შემთხვევაში A ხდომილება შეიძლება წარმოდგენილ იქნეს როგორც უთავსებადი B და $A\setminus B$ ხდომილებების ჯამი:
$$A=B+(A\setminus B).$$



ამიტომ ალბათობათა შეკრების კანონის თანახმად გვაქვს:
$$P(A)=P(B)+P(A\setminus B),$$
საიდანაც ჩანს შედეგის მართებულობა.

ცხადია, რომ ალბათობათა შეკრების კანონის განზოგადოება შესაძლებელია წყვილ-წყვილად უთავსებად ხდომილებათა ნებისმიერი რაოდენობისათვის:

თუ $A_1, A_2, \dots, A_n$ ხდომილებათა წყვილ-წყვილად უთავსებადი სისტემაა ($A_i\cap A_j = =\varnothing$, როცა $i\neq j$0, მაშინ:

$$P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i). \qquad (2)$$

ვისარგებლოთ მათემატიკური ინდუქციის მეთოდით. (2) ფორმულა $n=2$-ის შემთხვევაში სამართლიანია (იხ. (1)). დავუშვათ, რომ იგი სამართლიანია n-სათვის და ვაჩვენოთ, რომ სამართლიანი იქნება $n+1$-ის შემთხვევაში. გვაქვს:

$$P(\sum_{i=1}^{n+1} A_i) = P(\sum_{i=1}^{n} A_i + A_{n+1}) \overset{(1)}{=} P(\sum_{i=1}^{n} A_i) + P(A_{n+1}) \overset{(2)}{=} \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) = \sum_{i=1}^{n+1} P(A_i).$$

ამით (2) თანაფარდობა დამტკიცებულია.

თეორემა 1. თუ A და B ნებისმიერი ხდომილებებია, მაშინ

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \qquad (3)$$

დამტკიცება. განმარტების თანახმად

$$P(A \cup B) = \sum_{\omega_i \in (A \cup B)} P(\omega_i) \quad \text{და} \quad P(A) + P(B) = \sum_{\omega_i \in A} P(\omega_i) + \sum_{\omega_i \in B} P(\omega_i).$$

მაგრამ ჯამში $P(A)+P(B)$ ალბათობები $P(\omega_i)$, როცა $\omega_i \in A \cap B$, გათვალისწინებულია ორჯერ, ხოლო ამ $P(\omega_i)$-ების ჯამი არის $P(A \cap B)$. ამიტომ თუ $P(A)+P(B)$ ჯამს გამოვაკლებთ $P(A \cap B)$-ს, მივიღებთ სწორედ საძიებელ $P(A \cup B)$ ალბათობას.

დავალება. დაამტკიცეთ თეორემა 1 ალბათობათა შეკრების კანონისა და სხვაობის ალბათობის ფორმულის გამოყენებით.

სამი ხდომილების შემთხვევაში (თუ ვისარგებლებთ ე. წ. დე მორგანის კანონით $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$) გვექნება:

$$P(A \cup B \cup C) = P((A \cup B) \cup C) \overset{(3)}{=} P(A \cup B) + P(C) - P[(A \cup B) \cap C] \overset{(3)}{=}$$

$$\overset{(3)}{=} P(A) + P(B) - P(A \cap B) + P(C) - P[(A \cap C) \cup (B \cap C)] \overset{(3)}{=}$$

$$\overset{(3)}{=} P(A) + P(B) + P(C) - P(A \cap B) - \{P(A \cap C) + P(B \cap C) - P[(A \cap C) \cap (B \cap C)]\} =$$

$$= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

ზოგად შემთხვევაში სამართლიანია შემდეგი თანაფარდობა:

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) + \sum_{i \neq j} P(A_i \cap A_j) - \sum_{i \neq j \neq k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(\bigcap_{i=1}^{n} A_i).$$

პირობითი ალბათობის ფორმულა.

ალბათობის თეორიაში ალბათობის ცნებასთან ერთად შემოდის ე. წ. პირობითი ალბათობის ცნება. იმისათვის, რომ დავინახოთ მისი საჭიროება მოვიყვანოთ შემდეგი მარტივი მაგალითი: დავუშვათ, რომ ადამიანს დაავიწყდა ტელეფონის ნომრის ორი ციფრი. ვიპოვოთ ალბათობა იმისა, რომ შემთხვევით აკრეფილი ორი ციფრით მოხერხდება სასურველ აბონენტთან დაკავშირება?

ცხადია, რომ ამ შემთხვევაში $\Omega = \{(i,j) : i, j = 0,1,\ldots,9\}$ და ნამრავლის პრინციპის თანახმად $|\Omega| = 10 \cdot 10 = 100$. ვინაიდან სასურველ აბონენტთან შეერთებას ხელს უწყობს ციფრების ერთადერთი $(i.j)$ წყვილი, ამიტომ საძიებელი ალბათობა ტოლია $1/100$-ის.

ვთქვათ, ადამიანს გაახსენდა რომ ეს ციფრები იყო სხვადასხვა. რა იქნება მაშინ სასურველ აბონენტთან დაკავშირების ალბათობა? ამ შემთხვევაში ელემენტარულ ხდომილებათა სივრცე იქნება $\Omega=\{(i,j):i=0,1,...,9;\ j=0,1,...,9;i\neq j\}$. შესაბამისად, $|\Omega|=10\cdot 9=90$, ხოლო საძიებელი ალბათობა იქნება 1/90. როგორც ვხედავთ დამატებითი ინფორმაციის გაჩენამ შეცვალა (კერძოდ, გაზარდა) ალბათობა.

მოვიყვანოთ კიდევ ერთი მაგალითი. ვიპოვოთ ალბათობა იმისა, რომ ერთი კამათლის ერთჯერ გაგორებისას მოვა სამის ჯერადი ქულა, თუ ცნობილია, რომ მოვიდა ლუწი ქულა? ამ შემთხვევაში $\Omega=\{2,4,6\}$, $|\Omega|=3$, ხოლო ხელშემწყობი ელემენტარული ხდომილება ერთადერთია (კერძოდ, 6-იანის მოსვლა) და საძიებელი ალბათობა იქნება 1/3.

შევხედოთ იგივე მაგალითს სხვანაირად. დავუშვათ, შემთხვევითი მოვლენა მდგომარეობს ერთი კამათლის ერთჯერ გაგორებაში. შემოვიღოთ ხდომილობები: A -- მოვიდა სამის ჯერადი ქულა და B — მოვიდა ლუწი ქულა. ნათელია, რომ ამ დროს $A\cap B$ იქნება — მოვიდა სამის ჯერადი ლუწი ქულა. ცხადია, რომ ამ შემთხვევაში: $\Omega=\{1,2,3,4,5,6\}$; $A=\{3,6\}$; $B=\{2,4,6\}$; $A\cap B=\{6\}$, ხოლო ალბათობის კლასიკური განმარტების თანახმად კი: $P(A)=2/6=1/3$; $P(B)=3/6=1/2$ და $P(A\cap B)=1/6$. ადვილი დასანახია, რომ ზემოთ გამოთვლილი ალბათობა 1/3 (რომელსაც ბუნებრივია დავარქვათ A ხდომილების ალბათობა პირობაში, რომ ადგილი ჰქონდა B ხდომილებას) ფორმალურად შეგვეძლო მიგვეღო შემდეგნაირად:

$$1/3=\frac{1/6}{1/2}=\frac{P(A\cap B)}{P(B)}.$$

ამა თუ იმ მოვლენის ანალიზის დროს ხშირად იბადება კითხვა რა გავლენას ახდენს რაიმე A ხდომილების მოხდენის შესაძლებლობაზე რაიმე სხვა B ხდომილების მოხდენა. უმარტივესი მაგალითებია, როცა B ხდომილების მოხდენა აუცილებლად იწვევს A ხდომილების განხორციელებას, ე. ი. $B\subset A$, ან პირიქით, B ხდომილების მოხდენა გამორიცხავს A ხდომილების განხორციელების შესაძლებლობას, ე. ი. $A\cap B=\varnothing$. ალბათობის თეორიაში A და B ხდომილებებს შორის კავშირის დასახასიათებლად შემოდის A ხდომილების **პირობითი ალბათობის ცნება** B ხდომილების მიმართ.

განმარტება 1. *A ხდომილების პირობითი ალბათობა პირობაში, რომ ადგილი ჰქონდა B ხდომილებას* აღინიშნება $P(A\mid B)$ სიმბოლოთი და განიმარტება შემდეგნაირად

$$P(A\mid B)=P(A\cap B)/P(B), \tag{1}$$

თუ $P(B)\neq 0$.

ზემოთ მოყვანილი უკანასკნელი მაგალითი გარკვეული აზრით შეიძლება ჩაითვალოს პირობითი ალბათობის განმარტების მოტივაციად. მოვიყვანოთ ახლა მოსაზრება, რომელიც გაამართლებს A ხდომილების პირობითი ალბათობის (პირობაში, რომ მოხდა B ხდომილება) $P(A\mid B)$-ს განმარტებას (1) თანაფარდობით. დავუშვათ, რომ ექსპერიმენტის შედეგად შესაძლებელია მოხდეს A და B ხდომილები. A ხდომილების ***პირობითი სიხ-***

მორე, პირობაში რომ მოხდა B ხდომილება ეუწოდოთ A ხდომილების სიხშირეს გამოთვლილს არა ყველა ექსპერიმენტის მიმართ, არამედ იმ ექსპერიმენტების მიმართ, რომლებშიც ადგილი ჰქონდა B ხდომილებას. სხვა სიტყვებით, რომ ეთქვათ, თუ n — ყველა ექსპერიმენტის რაოდენობაა, $\mu_n(B)$ – B ხდომილების მოხდენათა რიცხვია, $\mu_n(A\cap B)$– $A\cap B$ ხდომილების მოხდენათა რიცხვია (ანუ იმ ექსპერიმენტების რაოდენობა, სადაც ერთდროულად მოხდა A და B), მაშინ პირობითი სიხშირე არის

$$\frac{\mu_n(A\cap B)}{\mu_n(B)}=\frac{\mu_n(A\cap B)/n}{\mu_n(B)/n}.$$

n-ის დიდი მნიშვნელობებისათვის ამ გამოსახულების მარცხენა მხარეს შეიძლება შევხედოთ როგორც B ხდომილების მოხდენის პირობაში A ხდომილების მოხდენის პირობითი ალბათობის $P(A\,|\,B)$ მიახლოებით მნიშვნელობას, შეფარდება $\mu_n(A\cap B)/n$ — არის $P(A\cap B)$ ალბათობის მიახლოებითი მნიშვნელობა, ხოლო შეფარდება $\mu_n(B)/n$ კი არის $P(B)$ ალბათობის მიახლოებითი მნიშვნელობა. ეს მსჯელობა უდევს სწორედ საფუძვლად პირობითი ალაბათობის (1) განმარტებას.

პირობით ალბათობებს გააჩნიათ ჩვეულებრივი ალბათობების ანალოგიური თვისებები:

I. ნებისმიერი A და B ხდომილებებისათვის ($P(B)\neq 0$):

$$0\le P(A\,|\,B)\le 1;$$

II. A და B უთავსებადია ($A\cap B=\varnothing$), მაშინ $P(A\,|\,B)=0$;

III. თუ B იწვევს A-ს ($B\subset A$), მაშინ $P(A\,|\,B)=1$.

თეორემა 1. თუ $A_1,A_2,\dots,A_n$ წყვილ-წყვილად უთავსებადი ხდომილებებია, მაშინ ნებისმიერი B ხდომილებისათვის ($P(B)\neq 0$) სამართლიანია ტოლობა:

$$P[(\sum_{i=1}^{n}A_i)\,|\,B]=\sum_{i=1}^{n}P(A_i\,|\,B).$$

დამტკიცება. ცხადია, რომ $(\sum_{i=1}^{n}A_i)\cap B=\sum_{i=1}^{n}(A_i\cap B)$. გარდა ამისა, ვინაიდან ხდომილებები $A_1,A_2,\dots,A_n$ წყვილ-წყვილად უთავსებადი ხდომილებებია, წყვილ-წყვილად უთავსებადი იქნება ხდომილებები $A_1\cap B, A_2\cap,\dots, A_n\cap B$. ამიტომ პირობითი ალბათობის განმარტების თანახმად, ალბათობათა შეკრების კანონის გათვალისწინებით, ვწერთ:

$$P[(\sum_{i=1}^{n}A_i)\,|\,B]=\frac{P[(\sum_{i=1}^{n}A_i)\cap B]}{P(B)}=\frac{P[\sum_{i=1}^{n}(A_i\cap B)]}{P(B)}=\frac{\sum_{i=1}^{n}P(A_i\cap B)}{P(B)}=$$

$$=\sum_{i=1}^{n}\frac{P(A_i\cap B)}{P(B)}=\sum_{i=1}^{n}P(A_i\,|\,B).$$

ამოცანა 1. განვიხილოთ ოჯახები, სადაც ორ-ორი ბავშვია. როგორია ალბათობა იმისა, რომ ოჯახში ორივე ბავშვი ვაჟია პირობაში, რომ: ა). უფროსი ბავშვი – ვაჟია; ბ). ერთი ბავშვი მაინც – ვაჟია?

ამოხსნა. აქ ელემენტარულ ხდომილებათა სივრცე ასეთია

$$\Omega = \{ვვ, ვქ, ქვ, ქქ\},$$

სადაც "ვ" აღნიშნავს ვაჟს, ხოლო "ქ" – ქალს. ჩავთვალოთ, რომ ოთხივე შედეგი ტოლალბათურია. შემოვიღოთ ხდომილებები: A – იყოს ხდომილება, რომ უფროსი ბავშვი – ვაჟია, ხოლო B – იყოს ხდომილება, რომ უმცროსი ბავშვი – ვაჟია. მაშინ $A \cap B$ – იქნება ხდომილება, რომ ორივე ბავშვი ვაჟია, ხოლო $A \cup B$ – კი იქნება ხდომილება, რომ ერთი ბავშვი მაინც ვაჟია. შესაბამისად, საძებნი ალბათობები იქნება: ა). $P(A \cap B | A)$ და ბ). $P(A \cap B | A \cup B)$. ადვილი დასანახია, რომ:

$$P(A \cap B | A) = \frac{P[(A \cap B) \cap A]}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{1/4}{1/2} = \frac{1}{2},$$

$$P(A \cap B | A \cup B) = \frac{P[(A \cap B) \cap (A \cup B)]}{P(A \cup B)} = \frac{P(A \cap B)}{P(A \cup B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

ამოცანა 2 (საუკეთესოს შერჩევაზე). მოცემულია m ობიექტი გადანომრილი რიცხვებით $1,2,\ldots,m$, ამასთანავე ისე, რომ ვთქვათ, ობიექტი №1 კლასიფიცირდება როგორც "საუკეთესო", . , ობიექტი №m კი როგორც "ყველაზე უარესი". იგულისხმება რომ ობიექტები შემოდიან დროის მომენტებში $1,2,\ldots,m$ შემთხვევითი რიგით (ანუ ყველა შესაძლო $m!$ გადანაცვლება ტოლალბათურია). დამკვირვებელს შეუძლია ორი მათგანის შედარებით თქვას რომელია უკეთესი და რომელი უარესი. საჭიროა საუკეთესოს შერჩევა იმ პირობით რომ ობიექტები წარმოიდგინება სათითაოდ და უკუგდებულის დამახსოვრება ხდება დამკვირვებლის მიერ. არ შეიძლება საუკეთესოდ მიჩნეულ იქნეს ის ობიექტი, რომელიც დაკვირვებული ობიექტებიდან ერთზე მაინც უარესია ან რომელიც უკვე იქნა უკუგდებული. ვთქვათ, დამკვირვებელმა ობიექტი შეარჩია k-ურ ნაბიჯზე ($k \le m$), ანუ დათვალიერებული ობიექტებიდან უკანასკნელი აღმოჩნდა ყველა წინაზე უკეთესი და ამიტომ მოხდა მისი შერჩევა. როგორია ალბათობა იმისა, რომ ამორჩეული ობიექტი იქნება საუკეთესო მთელი ერთობლიობიდან როგორც უკვე განხილულ, ისე ჯერ კიდევ განუხილავ ობიექტებს შორის?

ამოხსნა. შემოვიღოთ ხდომილებები: A იყოს ხდომილება, რომ k-ური ობიექტი საუკეთეოა ყველა არსებულ m ობიექტს შორის და B იყოს ხდომილება, რომ k-ური ობიექტი საუკეთეოა დაკვირვებულ k ობიექტს შორის. გასაგებია, რომ მოსაძებნია პირობითი ალბათობა $P(A | B)$.

ვინაიდან $A \subset B$, ამიტომ $A \cap B = A$ და $P(A \cap B) = P(A)$. შესაბამისად, პირობითი ალბათობის განმარტების თანახმად

$$P(A | B) = P(A) / P(B)$$

ვინაიდან ობიექტების ყველა შესაძლო გადანაცვლებები ტოლალბათურია, ამიტომ ალბათობის კლასიკური განმარტების თანახმად ადვილი დასანახია, რომ

$$P(B)=\frac{(k-1)!}{k!}=\frac{1}{k} \text{ და } P(A)=\frac{(m-1)!}{m!}=\frac{1}{m}.$$

შესაბამისად, $P(A \mid B)=k/m$.

სტრატეგია. შეიძლება დამტკიცდეს, რომ საუკეთესო ობიექტის ამორჩევის ოპტიმალური სტრატეგია მოწყობილია შემდეგნაირად. ავღნიშნოთ სიმბოლოთი m^* ისეთი ნატურალური რიცხვი, რომლისთვისაც სამართლიანია უტოლობა

$$\frac{1}{m^*}+\cdots+\frac{1}{m-1} \le 1 < \frac{1}{m^*-1}+\cdots+\frac{1}{m-1}.$$

საუკეთესო ობიექტის არჩევის ოპტიმალური სტრატეგია მდგომარეობს იმაში, რომ დავაკვირდეთ და უკუვაგდოთ პირველი m^*-1 ობიექტი და შემდეგ გავაგრძელოთ დაკვირვება ისეთ t^* მომენტამდე, როცა პირველად გამოჩნდება ყველა წინამორბედზე უკეთეს ობიექტი.

მაგალითად, თუ $m=1,\dots,10$, მაშინ m^*-ის შესაბამისი მნიშვნელობებია:

m	1	2	3	4	5	6	7	8	9	10
m-ოპტიმალური	1	1	1	1	2	2	2	3	3	4

საკმაოდ დიდი m-ისათვის $m^* \approx m/e$ (სადაც e – ნეპერის რიცხვია, $e \cong 2.718$) და საუკეთესო ობიექტის არჩევის ალბათობა დაახლოებით ტოლია $1/e \cong 0.368$ (თუმცა, ერთი შეხედვით ბუნებრივია მოგვეჩვენებოდა, რომ განსახილველი ობიექტების m რაოდენობის ზრდასთან ერთად, საუკეთესო ობიექტის არჩევის ალბათობა უნდა წასულიყო ნულისაკენ). ე. ი. საუკეთესო ობიექტის არჩევის ოპტიმალური სტრატეგია მდგომარეობს იმაში, რომ უნდა უკუვაგდოთ ობიექტების საერთო რაოდენობის დაახლოებით მესამედი და შემდეგ ავირჩიოთ პირველი ისეთი ობიექტი, რომელიც ყველა წინაზე უკეთესია.

ნამრავლის ალბათობის ფორმულა.

ვიგულისხმოთ, რომ $P(B)\neq 0$, მაშინ პირობითი ალბათობის ფორმულიდან

$$P(A \mid B)=P(A\cap B)/P(B)$$

შეგვიძლია დავწეროთ, რომ

$$P(A\cap B)=P(B)\cdot P(A \mid B). \tag{1}$$

ანალოგიურად, თუ ვიგულისხმებთ, რომ $P(A)\neq 0$, მაშინ $P(B \mid A)$ პირობითი ალბათობის ფორმულიდან მივიღებთ, რომ

$$P(A\cap B)=P(A)\cdot P(B \mid A).$$

ე. ი. **ორი ხდომილების ნამრავლის ალბათობა** *ტოლია ერთ-ერთის ალბათობა გამრავლებული მეორის პირობით ალბათობაზე პირობაში, რომ მოხდა პირველი.*

ცხადია, რომ სამი ხდომილების შემთხვევაში (თუ $P(B)\neq 0$ და $P(C)\neq 0$) გვექნება:

$$P(A\cap B\cap C)=P[(A\cap B)\cap C]\overset{(1)}{=}P(A\cap B)\cdot P[(A\cap B) \mid C]\overset{(1)}{=}$$

$$\stackrel{(1)}{=} P(A)\cdot P(A\,|\,B)\cdot P[(A\cap B)\,|\,C].$$

ანალოგიურად, $A_1, A_2, ..., A_n$ ხდომილებებისათვის გვექნება, რომ:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\cdot P(A_1\,|\,A_2)\cdots P(A_1 \cap \cdots \cap A_{n-1}\,|\,A_n). \qquad (2)$$

დავალება 1. მიუთითეთ რა შემთხვევაშია სამართლიანი (2) თანაფარდობა და დაამტკიცეთ იგი.

ამოცანა 1 "ბედნიერ" ბილეთებზე). 25 საგამოცდო ბილეთიდან 5 "ბედნიერია", ხოლო დანარჩენი 20 – "არა ბედნიერი". რომელ სტუდენტს აქვს "ბედნიერი" ბილეთის აღების მეტი ალბათობა: ვინც პირველი იღებს ბილეთს, თუ ვინც მეორე იღებს ბილეთს?

ამოხსნა. ეს ამოცანა ჩვენ უკვე ამოვხსენით ალბათობის კლასიკური განმარტების გამოყენებით. ამოვხსნათ ახლა იგი ელემენტარულ ხდომილებათა სივრცის ახლებური შემოტანითა და ნამრავლის ალბათობის ფორმულის გამოყენებით. წინასწარ შემოვიღოთ შემდეგი ხდომილებები: A იყოს ხდომილება, რომ პირველმა სტუდენტმა აიღო ბედნიერი ბილეთი, ხოლო B იყოს ხდომილება, რომ მეორე სტუდენტმა აიღო ბედნიერი ბილეთი. მაშინ ცხადია, რომ ელემენტარულ ხდომილებათა სივრცე შედგება ოთხი ხდომილებისაგან

$$\Omega = \{A\cap B, A\cap \overline{B}, \overline{A}\cap B, \overline{A}\cap \overline{B}\}.$$

ალბათობის კლასიკური განმარტების თანახმად $P(A) = 5/25 = 1/5$, ხოლო $P(\overline{A}) = 20/25 = 4/5$. მეორეს მხრივ, ამოცანის შინაარსიდან გამომდინარე, თუ ცნობილია, რომ პირველმა სტუდენტმა აიღო ბედნიერი ბილეთი, მაშინ ალბათობა იმისა რომ მეორე სტუდენტი აიღებს ბედნიერ ბილეთს ისევ შეიძლება გამოვითვალოთ ალბათობის კლასიკური განმარტებით: ამ შემთხვევაში ყველა შესაძლო შედეგთა რაოდენობაა 24, ხოლო ხელშემწყობ ელემენტარულ ხდომილებათა რაოდენობა კი მხოლოდ 4 (რადგან ერთი ბედნიერი ბილეთი უკვე აღებულია) და შესაბამისად,

$$P(B\,|\,A) = 4/24 = 1/6.$$

ანალოგიურად, $P(\overline{B}\,|\,A) = 20/24 = 5/6$, $P(B\,|\,\overline{A}) = 5/24 = 5/24$ და $P(\overline{B}\,|\,\overline{A}) = 19/24$. ამიტომ ორი ხდომილების ნამრავლის ალბათობის ფორმულის თანახმად გვაქვს:

$$P(A\cap B) = P(A)\cdot P(B\,|\,A) = 1/5\cdot 1/6 = 1/30;\ \ P(A\cap \overline{B}) = 1/5\cdot 5/6 = 1/6;$$

$$P(\overline{A}\cap B) = 4/5\cdot 5/24 = 1/6 \ \text{ და } \ P(\overline{A}\cap \overline{B}) = 4/5\cdot 19/24 = 19/30$$

(შევნიშნავთ, რომ ჩვენ აქ მხოლოდ სისრულისათვის გამოვთვალეთ ყველა შესაძლო პირობითი და ნამრავლის ალბათობები).

ცხადია, რომ $B = (A\cap B) + (\overline{A}\cap B)$. ამიტომ ალბათობათა შეკრების კანონის თანახმად

$$P(B) = P(A\cap B) + P(\overline{A}\cap B) = 1/30 + 1/6 = 15\ (= P(A)).$$

ამოცანა 2. ყუთში m ბურთია, მათ შორის n თეთრია. ვიპოვოთ ალბათობა იმისა, რომ ყუთიდან ორი ბურთის მიმდევრობით დაბრუნების გარეშე ამოღებისას: ა). პირველი ბურთი თეთრია; ბ). მეორე ბურთი თეთრია; გ). ორივე ბირთვი თეთრია.

ამოხსნა. A_i იყოს ხდომილება, რომ i-ური ბურთი თეთრია ($i=1,2$). მაშინ ალბათობის კლასიკური განმარტების თანახმად:

ა). $P(A_1)=n/m$.

გარდა ამისა,

$$P(A_2\mid A_1)=(n-1)/(m-1) \text{ და } P(A_2\mid \overline{A_1})=n/(m-1).$$

ამიტომ ნამრავლის ალბათობის ფორმულის თანახმად:

გ). $P(A_1\cap A_2)=P(A_1)\cdot P(A_2\mid A_1)=n(n-1)/m(m-1)$.

ანალოგიურად,

$$P(\overline{A_1}\cap A_2)=P(\overline{A_1})\cdot P(A_2\mid \overline{A_1})=n(m-n)/m(m-1).$$

ამიტომ:

ბ). $P(A_2)=P[(A_1\cap A_2)+(\overline{A_1}\cap A_2)]=P(A_1\cap A_2)+P(\overline{A_1}\cap A_2)=n/m$.

დავალება 2. დავუშვათ, რომ შესამოწმებელი ჯგუფის 1% ავადმყოფია, ხოლო დანარჩენი 99% კი ჯანმრთელია. ადამიანების შერჩევა ხდება შემთხვევით და ამიტომ

$$P(\text{ავადმყოფი})=1\%=0.01 \text{ და } P(\text{ჯანმრთელი})=99\%=0.99.$$

ვიგულისხმოთ, რომ იმ შემთხვევაში, როცა ტესტირება უტარდება ადამიანს, რომელსაც არა აქვს ავადმყოფობა, მაშინ 1%-ია ალბათობა იმისა, რომ მივიღოთ მცდარი დადებითი შედეგი, ე.ი.

$$P(\text{დადებითი}\mid\text{ჯანმრთელი})=1\% \text{ და } P(\text{უარყოფითი}\mid\text{ჯანმრთელი})=99\%.$$

და ბოლოს, დავუშვათ, რომ იმ შემთხვევაში, როცა ტესტირება უტარდება ავადმყოფ ადამიანს, მაშინ 1%-ია ალბათობა იმისა, რომ მივიღოთ მცდარი უარყოფითი შედეგი, ე.ი.

$$P(\text{უარყოფითი}\mid\text{ავადმყოფი})=1\% \text{ და } P(\text{დადებითი}\mid\text{ავადმყოფი})=99\%.$$

გამოთვალეთ ალბათობა იმისა, რომ: ა). ადამიანი ჯანმრთელია, ხოლო ტესტმა აჩვენა უარყოფითი შედეგი; ბ). ადამიანი ავადმყოფია, ხოლო ტესტმა აჩვენა დადებითი შედეგი; გ). ადამიანი ჯანმრთელია, ხოლო ტესტმა აჩვენა დადებითი შედეგი; გ). ადამიანი ავადმყოფია, ხოლო ტესტმა აჩვენა უარყოფითი შედეგი.

ამოცანა 3. ყუთს აქვს n უჯრა. ალბათობა იმისა რომ ბურთი არის ამ უჯრებიდან ერთ-ერთში ტოლია p-ჰი. იპოვეთ ალბათობა იმისა, რომ ბურთი არის i-ურ უჯრაში, თუ ცნობილია, რომ ბურთი თითეულ უჯრაში შესაძლებელია იყოს თანაბარი ალბათობებით?

ამოხსნა. A_i იყოს ხდომილება, რომ ბურთი არის i-ურ უჯრაში. A იყოს ამ ხდომილებების გაერთიანება $A=\bigcup_{i=1}^{n}A_i$. პირობის თანახმად

$$P(A)=p \text{ და } P(A_i\mid A)=1/n.$$

ამიტომ

$$P(A_i)=P(A_i\cap A)=P(A)P(A_i\mid A)=p\cdot 1/n=p/n.$$

დამოკიდებული და დამოუკიდებელი ხდომილებები.

ალბათობის თეორიაში ორ A და B ხდომილებას ეწოდება **დამოუკიდებელი**, თუ ერთ-ერთი მათგანის მოხდენა არ ცვლის მეორე მათგანის მოხდენის ალბათობას. წინააღმდეგ შემთხვევაში ამ ხდომილებებს ეწოდე-

ბათ **დამოკიდებული**. იმ შემთხვევაში, როცა ერთ-ერთი ხდომილების ალბათობა არანულოვანია, ვთქვათ, $P(B) \neq 0$, მაშინ გვაქვს შემდეგი

განმარტება 1. A ხდომილებას ეწოდება B ხდომილებისაგან *დამოუკიდებელი*, თუ

$$P(A \mid B) = P(A), \tag{1}$$

ხოლო თუ $P(A \mid B) \neq P(A)$, მაშინ გვაქვს *დამოკიდებული* ხდომილებები.

თუ გავიხსენებთ პირობითი ალბათობის განმარტებას

$$P(A \mid B) = P(A \cap B) / P(B),$$

მაშინ (1) თანაფარდობიდან მივიღებთ, რომ

$$P(A \cap B) = P(A) \cdot P(B). \tag{2}$$

პირიქით, იმ შემთხვევაში, როცა $P(B) \neq 0$, (2) თანაფარდობიდან მიიღება (1) თანაფარდობა.

შენიშვნა. ზოგიერთ სახელმძღვანელოში ხდომილებათა დამოუკიდებლობა განიმარტება (2) თანაფარდობით ((1) და (2) ექვივალენტურია, თუ $P(B) \neq 0$). მას აქვს ის უპირატესობა, რომ მისი გამოყენება შესაძლებელია მაშინაც, როცა ხდომილებების ალბათობები ნულოვანია (მაგალითად, შეუძლებელი ხდომილება დამოუკიდებელია ნებისმიერი ხდომილებისაგან $P(A \cap \emptyset) = P(\Omega) = 0 = P(A) \cdot P(\emptyset)$) და გარდა ამისა, იგი სიმეტრიულია A და B-ს მიმართ (ასეთ შემთხვევაში, თუ A დამოუკიდებელია B-საგან, მაშინ B დამოუკიდებელია A-საგან), მაგრამ (1) თანაფარდობის უპირატესობა ის არის, რომ იქიდან ჩანს აღნიშნული განმარტების შინაარსი.

ცხადია, რომ აუცილებელი ხდომილება დამოუკიდებელია ნებისმიერი ხდომილებისაგან:

$$P(A \mid \Omega) = \frac{P(A \cap \Omega)}{P(\Omega)} = \frac{P(A)}{1} = P(A),$$

რაც ბუნებრივია იმის გამო, რომ ამ შემთხვევაში პირობა არ წარმოადგენს დამატებით ინფორმაციას (ჩვენ ისედაც ვიცოდით, რომ აუცილებელი ხდომილება ეს ის ხდომილებაა, რომელიც ყოველთვის ხდება) და ამიტომ ალბათობა არც უნდა შეცვლილიყო.

ვინაიდან A და B ხდომილებების დამოუკიდებლობა ნიშნავს, რომ ინფორმაცია A-ს მოხდენის შესახებ არ ცვლის B-ს ალბათობას, ბუნებრივია, რომ ინფორმაციამ A-ს არ მოხდენის შესახებ აგრეთვე არ უნდა შეცვალოს B-ს მოხდენის ალბათობა. მართლაც სამართლიანია შემდეგი

თეორემა 1. თუ A და B ხდომილებები დამოუკიდებელია, მაშინ ხდომილებები $\overline{A}$ და B აგრეთვე დამოუკიდებელია.

დამტკიცება. ადვილი დასანახია, რომ $\overline{A} \cap B = B \setminus (A \cap B)$ (ვინაიდან $\overline{A} \cap B = B \setminus (A \cap B)$), ამიტომ სხვაობის ალბათობის ფორმულის გამოყენებით თეორემის პირობებში ვღებულობთ, რომ

$$P(\overline{A} \cap B) = P(B) - P(A \cap B) \overset{(2)}{=} P(B) - P(A)P(B) = (1 - P(A))P(B) = P(\overline{A})P(B).$$

შედეგი. თუ A და B ხდომილებები დამოუკიდებელია, მაშინ დამოუკიდებელია $\overline{A}$ და $\overline{B}$.

მაგალითი 1. კარტების ნაკრებიდან (რომელშიც 36 კარტია) შემთხვევით იღებენ ერთ კარტს. განვიხილოთ ხდომილებები: A იყოს ხდომილება, რომ ამოღებული კარტი "აგურისა", ხოლო B იყოს ხდომილება, რომ ამოღებული კარტი "მეფეა". არიან თუ არა ეს ხდომილებები დამოუკიდებელი?

ცხადია, რომ ამ შემთხვევაში

$$|\Omega| = 36, \; P(A) = 9/36 = 1/4, \; P(B) = 4/36 = 1/9 \text{ და}$$

$$P(A \cap B) = 1/36 = 1/4 \cdot 1/9 = P(A) \cdot P(B).$$

ე.ი. ეს ხდომილებები დამოუკიდებელია.

მაგალითი 2. დავუშვათ, ვაგორებთ ორ სათამაშო კამათელს. განვიხილოთ ხდომილებები: A — პირველ კამათელზე მოვიდა კენტი ქულა, B — მეორე კამათელზე მოვიდა კენტი ქულა, C — ორივე კამათელზე მოსულ ქულათა ჯამი კენტია. გავარკვიოთ ამ ხდომილებების დამოუკიდებლობის საკითხი.

ცხადია, რომ $P(A) = P(B) = 3/6 = 1/2$, ხოლო $P(A \cap B) = 3 \cdot 3/36 = 1/4$. ამიტომ A და B ხდომილებები დამოუკიდებლია.

დავალება 1. შეამოწმეთ, რომ $P(C) = 1/2$.

შევნიშნოთ, რომ A და B ხდომილებიდან ერთ-ერთის მოხდენის პირობაში C ხდომილება ხდება მაშინ და მხოლოდ მაშინ, როცა შესაბამისად, ან პირველ ან მეორე კამათელზე მოვიდა ლუწი ქულა, ანუ გვაქვს თანაფარდობები:

$$A \cap C = A \cap \overline{B} \text{ და } B \cap C = \overline{A} \cap B.$$

ვინაიდან, თეორემა 1-ის ძალით, ხდომილებები A და $\overline{B}$ და $\overline{A}$ და B აგრეთვე დამოუკიდებლებია, ამიტომ

$$P(A \cap \overline{B}) = P(A)P(\overline{B}) = 1/4 \text{ და } P(\overline{A} \cap B) = 1/4.$$

შესაბამისად,

$$P(A \cap C) = P(A \cap \overline{B}) = 1/4 \text{ და } P(B \cap C) = P(\overline{A} \cap B) = 1/4.$$

ეს თანაფარდობები კი, $P(C) = 1/2$ ალბათობის გათვალისწინებით, ნიშნავს, რომ დამოუკიდებლებია A და C და B და C ხდომილებათა წყვილებიც.

განმარტება 2. ხდომილებათა ერთობლიობას $A_1, A_2, \ldots, A_n$ ეწოდება **წყვილ-წყვილად დამოუკიდებელი** თუ ნებისმიერი ორი ხდომილება ამ ერთობლიობიდან დამოუკიდებელია, ანუ

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j), \quad \forall i \neq j.$$

წინა მაგალითში ჩვენ ვნახეთ, რომ ხდომილებები A, B და C წყვილ-წყვილად დამოუკიდებელია.

განმარტება 3. ხდომილებათა ერთობლიობას $A_1, A_2, \ldots, A_n$ ეწოდება **ერთობლივად დამოუკიდებელი** თუ ნებისმიერი $k \leq n$ რაოდენობისათვის და ერთმანეთისაგან განსხვავებული $i_1, i_2, \ldots, i_k$ ინდექსებისათვის, რომელთაგან თითოეული იცვლება ერთიდან n-მდე:

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdots P(A_{i_k}).$$

ცხადია, რომ თუ ხდომილებები ერთობლივად დამოუკიდებელია, მაშინ ისინი იქნებიან წყვილ-წყვილად დამოუკიდებელი. პირიქით, კი საზოგადოდ სწორი არ არის. ამის მაგალითად გამოდგება მაგალითი 2.

დავალება 2. შეამოწმეთ, რომ ხდომილებები მაგალითი 2-დან არ არიან ერთობლივად დამოუკიდებელი.

მაგალითი 3. დავუშვათ, ვაგდებთ სამ მონეტას. შემოვიღოთ ხდომილებები:

A_1 – პირველი და მეორე მონეტა დაეცა ერთი და იგივე მხარეზე;

A_2 – მეორე და მესამე მონეტა დაეცა ერთი და იგივე მხარეზე;

A_3 – პირველი და მესამე მონეტა დაეცა ერთი და იგივე მხარეზე.

ადვილი შესამოწმებელია, რომ აქედან ნებისმიერი ორი ხდომილება დამოუკიდებელია, ხოლო სამივე ეროად დამოკიდებულია, ვინაიდან თუ ჩვენ გვეცოდინება რომ მაგალითად, A_1 და A_2 მოხდა, მაშინ ჩვენ ზუსტად ვიცით, რომ A_3 აგრეთვე მოხდა.

დავალება 3. შეამოწმეთ, რომ ხდომილებები A_1, A_2 და A_3 წყვილ-წყვილად დამოუკიდებელია.

თეორემა 2. თუ $A_1, A_2, ..., A_n$ ხდომილებები ერთობლივად დამოუკიდებელია, მაშინ ხდომილებები $\overline{A_1}, A_2, ..., A_n$ აგრეთვე ერთობლივად დამოუკიდებლებია.

დავალება 4. დაამტკიცეთ თეორემა 2.

ამოცანა 1 (დაბადების დღეებზე). ვიპოვოთ ალბათობა იმისა, კონკრეტული სკოლის რომ 150 მოსწავლიდან ერთი მაინც დაბადებულია მოცემულ ფიქსირებულ დღეს, მაგალითად პირველ სექტემბერს?

ამოხსნა. შემოვიღოთ ხდომილებები:

$$A_i = \{i-\text{ური სტუდენტი დაბადებულია 1.09}\}, \quad i = 1,2,...,150;$$

$$A = \{\text{ერთი მაინც 150 სტუდენტიდან დაბადებულია 1.09}\}.$$

ნათელია, რომ A_i ხდომილებები ერთობლივად დამოუკიდებელია და $P(A_i) = 1/365$. გარდა ამისა, $A = \bigcup_{i=1}^{150} A_i$ და მაშასადამე, საპოვნელია დამოუკიდებელ ხდომილებათა გაერთიანების ალბათობა. გადავიდეთ საწინააღმდეგო ხდომილებაზე და ვისარგებლოთ დე-მორგანის კანონით. მაშინ გვაქვს:

$$P(A) = 1 - P(\overline{A}) = 1 - P(\overline{\bigcup_{i=1}^{150} A_i}) = 1 - P(\bigcap_{i=1}^{150} \overline{A_i}) = 1 - P(\overline{A_1}) \cdots P(\overline{A_n}).$$

რამდენადაც $P(\overline{A_i}) = 1 - 1/365$, თუ ვისარგებლებთ ნიუტონის ბინომის ფორმულით $(1+x)^n = \sum_{j=1}^{n} C_n^j x^j$, ვღებულობთ

$$P(A) = \frac{150}{365} - C_{150}^2 (\frac{1}{365})^2 + C_{150}^3 (\frac{1}{365})^3 - C_{150}^4 (\frac{1}{365})^4 + \cdots.$$

ვინაიდან, $150/365 \cong 0,41$, ხოლო

$$C_{150}^4\left(\frac{1}{365}\right)^4 < \left(\frac{150}{365}\right)^4\frac{1}{24} \cong \frac{(0,41)^4}{24} < 0.005,$$

ამიტომ მწკრივის ნიშანცვლადობის გამო, თუ გადავაგდებთ მწკრივის წევრებს დაწყებული მე-5 წევრიდან (მწკრივების ზოგადი თეორიიდან გამომდინარე), შესაძლებელია ვამტკიცოთ, რომ

$$P(A) \cong 0.41 - \frac{(0.41)^2}{2} + \frac{(0.41)^3}{6} \cong 0.41 - 0.08 + 0.01 = 0.34.$$

სრული ალბათობის ფორმულა.

ხდომილებათა ერთობლიობას $A_1, A_2, ..., A_n$ ეწოდება **ხდომილებათა სრული სისტემა,** თუ ეს ხდომილებები წყვილ-წყვილად უთავსებადია და მათი გაერთიანება ემთხვევა აუცილებელ ხდომილებას: $A_i \cap A_j = \emptyset$, როცა $i \neq j$ და $\bigcup_{i=1}^{n} A_i = \Omega$. სხვა სიტყვებით, ელემენტარულ ხდომილებათა სივრცე დაყოფილია (დახლეჩილია) თანაუკვეთ ნაწილებად. ქვემოთ მოყვანილ ნახაზზე ელემენტარულ ხდომილებათა სივრცე წარმოდგენილა მართკუთხედის სახით და ხდომილებათა სრული სისტემა შედგება ხუთი თანაუკვეთი A, B, C, D, E ხდომილებისაგან.

ხდომილებათა სრული სისტემაა ნებისმიერი A ხდომილება და მისი საწინააღმდეგო $\bar{A}$ ხდომილება, ვინაიდან $A \cap \bar{A} = \emptyset$ და $A \cup \bar{A} = \Omega$.

ქვემოთ ჩვენ მოვიყვანთ ფორმულას, რომელსაც *სრული ალბათობის ფორმულა* ეწოდება და რომელიც წარმოადგენს ძირითად საშუალებას რთული ხდომილებების ალბათობების გამოსათვლელად პირობითი ალბათობების საშუალებით.

თეორემა 1. თუ $A_1, A_2, ..., A_n$ ხდომილებათა სრული სისტემაა ისეთი, რომ მისი თითოეული ხდომილების ალბათობა არანულოვანია ($P(A_i) > 0$, $i = 1, 2, ..., n$), მაშინ ნებისმიერი B ხდომილების ალბათობა გამოითვლება ფორმულით

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i), \qquad (1)$$

რომელსაც **სრული ალბათობის ფორმულა** ეწოდება.

დამტკიცება. დე-მორგანის ფორმულის თანახმად

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} A_i\right) = \bigcup_{i=1}^{n} (B \cap A_i).$$

თუ ახლა გავითვალისწინებთ, რომ ვინაიდან $A_1, A_2, \dots, A_n$ ხდომილებები წყვილ-წყვილად უთავსებადია, მითუმეტეს წყვილ-წყვილად უთავსებადი იქნებიან ხდომილებები $B \cap A_1, B \cap A_2, \dots, B \cap A_n$, ამიტომ ხდომილებათა ჯამის ალბათობისა და ნამრავლის ალბათობის ფორმულების გამოყენებით გვექნება

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i).$$

ამოცანა 1 (ასოების ამოცნობა). გვაქვს ასოების ორი ერთობლიობა:

I={К, И, З, Н} და II={З, Ч, Н}

შემთხვევით ვირჩევთ ერთ ერთობლიობას და არჩეული ერთობლიობიდან ერთ ასოს. ამორჩეულ ასოს ძალიან მცირე დროის განმავლობაში ვუჩვენებთ დამკვირვებელს (ისე რომ მას არ შეუძლია მთლიანად აღიქვას ასო). როგორია ასოს სწორად გამოცნობის ალბათობა, თუ დამკვირვებლის პასუხია "Н", როცა ის ასოს გამოსახულებაში დაინახავს ვერტიკალურ ხაზს და პასუხია "З", როცა ასოს გამოსახულებაში ვერტიკალური ხაზი არ არის?

ამოხსნა. შემოვიღოთ ხდომილებები:

$A_i = \{$ამორჩეულია i – ური ერთობლიობა, $i = 1,2\}$;

"К", "И", "З", "Н" და "Ч" – იყოს ხდომილება, რომ წარმოდგენილია შესაბამისად К, И, З, Н და Ч ასოები;

$B = \{$დამკვირვებელმა სწორად უპასუხა$\}$.

ამოცანის პირობებში ცხადია:

$$P(A_1) = P(A_2) = 1/2;$$

$$P(\text{“К”} \mid A_1) = P(\text{“И”} \mid A_1) = P(\text{“З”} \mid A_1) = P(\text{“Н”} \mid A_1) = 1/4. \quad P(\text{“Ч”} \mid A_1) = 0;$$

$$P(\text{“З”} \mid A_2) = P(\text{“Ч”} \mid A_2) = P(\text{“Н”} \mid A_2) = 1/3. \quad P(\text{“К”} \mid A_2) = P(\text{“И”} \mid A_2) = 0.$$

ვინაიდან A_1 და A_2 ქმნიან ხდომილებათა სრულ სისტემას, ამიტომ თითოეული ასოს სწორად ამოცნობის ალბათობა შეგვიძლია გამოვთვალოთ სრული ალბათობის ფორმულით:

$$P(\text{“К”}) = P(A_1)\, P(\text{“К”} \mid A_1) + P(A_2)\, P(\text{“К”} \mid A_2) = 1/8;$$

$$P(\text{“И”}) = P(A_1)\, P(\text{“И”} \mid A_1) + P(A_2)\, P(\text{“И”} \mid A_2) = 1/8;$$

$$P(\text{“З”}) = P(A_1)\, P(\text{“З”} \mid A_1) + P(A_2)\, P(\text{“З”} \mid A_2) = 7/24;$$

$$P(\text{“Н”}) = P(A_1)\, P(\text{“Н”} \mid A_1) + P(A_2)\, P(\text{“Н”} \mid A_2) = 7/24;$$

$$P(\text{“Ч”}) = P(A_1)\, P(\text{“Ч”} \mid A_1) + P(A_2)\, P(\text{“Ч”} \mid A_2) = 1/6.$$

ამოცანის პირობებში ცხადია აგრეთვე, რომ სწორი პასუხის პირობითი ალბათობები სხვადასხვა ასოების წარმოდგენის შემთხვევაში შესაბამისად იქნება:

$$P(B \mid \text{“К”}) = P(B \mid \text{“И”}) = P(B \mid \text{“Ч”}) = 0 \text{ და } P(B \mid \text{“З”}) = P(B \mid \text{“Н”}) = 1.$$

ვინაიდან, "К", "И", "З", "Н" და "Ч" აგრეთვე ხდომილებათა სრული სისტემაა, ამიტომ სრული ალბათობის ფორმულა გვაძლევს სწორი პასუხის ალბათობას:

$$P(B) = P(\text{“К”})\, P(B \mid \text{“К”}) + P(\text{“Н”})\, P(B \mid \text{“Н”}) + P(\text{“И”})\, P(B \mid \text{“И”}) +$$

$$+P(\text{"3"})P(B|\text{"3"})+P(\text{"Ч"})P(B|\text{"Ч"})=P(\text{"H"})+P(\text{"3"})=7/12.$$

ამოცანა 2 (მოთამაშის გაკოტრებაზე). განვიხილოთ ე. წ. "გერბი-საფასურის" თამაში: თუ მონეტის აგდებისას მოვა მოთამაშის მიერ წინასწარ დასახელებული მონეტის მხარე, მაშინ იგი იგებს 1 ლარს, წინააღმდეგ შემთხვევაში კი აგებს 1 ლარს. ვთქვათ, მოთამაშის საწყისი კაპიტალი შეადგენს x ლარს და მისი მიზანია მიიყვანოს ეს თანხა a ლარამდე. თამაში გრძელდება მანამ სანამ მოთამაშე არ მიიყვანს თავის თანხას წინასწარ განსაზღვრულ a ლარამდე, ან იგი არ გაკოტრდება (ანუ წააგებს მის ხელთ არსებულ მთელ x ლარს). როგორია ალბათობა იმისა, რომ მოთამაშე გაკოტრდება?

ამოხსნა. ეს ალბათობა დამოკიდებული იქნება საწყის x კაპიტალზე. აღვნიშნოთ იგი $p(x)$ სიმბოლოთი. ცხადია, რომ იგი განმარტებულია ნებისმიერი $0 \le x \le a$ და ამასთანავე, $P(0)=1$ და $P(a)=0$. შემოვიღოთ ხდომილებები:

$$A_1=\{\text{მოთამაშემ მოიგო პირველ ნაბიჯზე}\},$$

$$B=\{\text{მოთამაშე, რომელსაც გააჩნია საწყისი კაპიტალი } x, \text{ გაკოტრდება}\}.$$

ამოცანის პირობებში გვაქვს:

$$P(A_1)=P(\overline{A_1})=1/2,\quad P(B|A_1)=p(x+1) \text{ და } P(B|\overline{A_1})=p(x-1)\quad (1\le x\le a-1).$$

ვინაიდან, A_1 და $\overline{A_1}$ ხდომილებათა სრული სისტემაა, ამიტომ სრული ალბათობის ფორმულა $p(x)$ ალბათობისათვის გვაძლევს შემდეგ განტოლებას:

$$p(x)=\frac{1}{2}p(x+1)+\frac{1}{2}p(x-1),\quad 1\le x\le a-1$$

(ამ ტიპის განტოლებებს მათემატიკაში *რეკურენტულ განტოლებებს* უწოდებენ). შეიძლება შემოწმდეს, რომ ამ განტოლების ამოხსნას აქვს სახე:

$$p(x)=bx+c,$$

სადაც b და c — ნებისმიერი მუდმივებია. ამ კოეფიციენტების მოსაძებნად უნდა ვისარგებლოთ სასაზღვრო პირობებით $P(0)=1$ და $P(a)=0$. მაშინ მივიღებთ, რომ

$$c=1 \text{ და } ab+c=0,$$

საიდანაც, $b=-1/a$ და საბოლოოდ $p(x)=1-x/a$, $0\le x\le a$.

განვიხილოთ რეალური სიტუაცია, რომელიც გვიჩვენებს ერთი შეხედვით მოულოდნელ განსხვავებას $P(A|B)$ და $P(B|A)$ პირობით ალბათობებს შორის. იმისათვის, რომ გამოვავლინოთ სერიოზული ავადმყოფობის მქონე ადამიანები ადრეულ სტადიაზე, ხდება ადამიანების დიდი ჯგუფის ტესტირება. მიუხედავად წინასწარი შემოწმების სარგებლობისა, ამ მიდგომას გააჩნია უარყოფითი მხარე: თუ ადამიანს სინამდვილეში არ გააჩნია ავადმყოფობა და საწყისმა ტესტმა აჩვენა დადებითი შედეგი (დაუდგინა ავადმყოფობა), ის იქნება სტრესულ მდგომარეობაში (რაც თავის მხრივ უარყოფითად მოქმედებს მის ცხოვრებაზე) სანამ უფრო წარმატებული ტესტი არ აჩვენებს, რომ ის ჯანმრთელია. ამ პრობლემის მნიშვნელობა შესაძლებელია კარგად გავიგოთ პირობითი ალბათობების ტერმინებში.

ორი ხდომილების ნამრავლის ალბათობის ფორმულაში მოყვანილი დავალება 2-ის მონაცემებში გამოვთვალოთ ალბათობა იმისა, რომ ტესტი აჩვენებს დადებით შედეგს. სრული ალბათობის ფორმულის თანახმად:

$$= P(\text{ჯანმრთელი})P(\text{დადებითი} \mid \text{ჯანმრთელი}) +$$

$$+ P(\text{ავადმყოფი})P(\text{დადებითი} \mid \text{ავადმყოფი}) = 0.99 \cdot 0.01 + 0.01 \cdot 0.99 = 0.0198 .$$

როგორც ცნობილია, მაგალითის პირობებში

$$P(\text{დადებითი} \mid \text{ავადმყოფი}) = 99\% .$$

გამოვთვალოთ ახლა შებრუნებული პირობითი ალბათობა, რისთვისაც ვისარგებლოთ პირობითი ალბათობის განმარტებითა და ნამრავლის ალბათობის ფორმულებით. მაშინ ზემოთ მიღებული

$$P(\text{დადებითი}) = 0.0198 = 1.98\%$$

შედეგის თანახმად:

$$P(\text{ავადმყოფი} \mid \text{დადებითი}) = \frac{P(\text{ავადმყოფი} \cap \text{დადებითი})}{P(\text{დადებითი})} =$$

$$= \frac{P(\text{ავადმყოფი})\ P(\text{დადებითი} \mid \text{ავადმყოფი})}{1.98\%} = \frac{1\% \cdot 99\%}{1.98\%} = 50\% .$$

როგორც ვხედავთ, პირობითი ალბათობა იმისა რომ ტესტი მოგვცემს დადებით შედეგს, პირობაში რომ ადამიანი ავადმყოფია ტოლია 99%-ის, მაშინ როდესაც პირობითი ალბათობა იმისა რომ ადამიანი ავადმყოფია, პირობაში რომ ტესტმა მოგვცა დადებითი შედეგი არის მხოლოდ 50%. აქ შერჩეული მონაცემების შემთხვევაში უკანასკნელი შედეგი შეიძლება ჩაითვალოს მიუღებელად: ნახევარი ადამიანების, რომელთა ტესტირებამ აჩვენა დადებითი შედეგი, ფაქტიურად არის მცდარი დადებითი.

ბაიესის ფორმულა.

ვიგულისხმოთ, რომ A და B ხდომილებები ისეთია, რომ $P(A) > 0$ და $P(B) > 0$. მაშინ $P(A \mid B)$ და $P(B \mid A)$ პირობით ალბათობების განმარტებიდან:

$$P(A \cap B) = P(A)P(B \mid A) = P(B)P(A \mid B),$$

საიდანაც მიიღება ე. წ. *ბაიესის ფორმულა:*

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}.$$

თუ $A_1, A_2, ..., A_n$ ხდომილებათა სრული სისტემაა ისეთი, რომ $P(A_i) > 0$, $i = 1,2,...,n$, მაშინ ბაიესის ფორმულიდან სრული ალბათობის ფორმულის გამოყენებით ვღებულობთ ე. წ. **ბაიესის თეორემას:**

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{n} P(A_j)P(B \mid A_j)}.$$

შევნიშნოთ, რომ ორივე ამ ფორმულაში ერთი პირობითი ალბათობა იცვლება შებრუნებული პირობითი ალბათობებით, რომლებიც ხშირ შემთხვევაში შედარებით მარტივად გამოითვლება (ან პირდაპირ მოცემულია) და მათი კომბინაციით ითვლება პირდაპირი პირობითი ალბათობა.

ბაიესის ფორმულას შეიძლება მიეცეს შემდეგი ინტერპრეტაცია: დავუშვათ, სამეცნიერო გამოკვლევის დაწყებამდე ჩვენ გვაქვს n სხვადასხვა ვარაუდი (ჰიპოთეზა) $A_1, A_2, \dots, A_n$ შესასწავლი ობიექტის ბუნების შესახებ, ამასთანავე ჩვენ მათ მივაწერთ ალბათობებს $P(A_1), P(A_2), \dots, P(A_n)$ (ამ ალბათობებს უწოდებენ **აპრიორულ ალბათობებს**). შემდეგ ჩვენ ვატარებთ ექსპერიმენტს (ან დაკვირვებას), რომლის შედეგადაც შეიძლება მოხდეს ან არ მოხდეს B ხდომილება (ე. ი. მოხდეს $\bar{B}$ ხდომილება). თუ მოხდა B ხდომილება, ვახდენთ თითოეული ჰიპოთეზის სამართლიანობის შესახებ ჩვენი რწმენის გადაფასებას ვცვლით რა $P(A_i)$ ალბათობებს $P(A_i \mid B)$ ალბათობებით (ამ ალბათობებს ეწოდება **აპოსტერიორული ალბათობები**). ასე ჩვენ ვაგრძელებთ, სანამ რომელიმე $i = i_0$-სათვის A_{i_0} ხდომილების აპოსტერიორული ალბათობა არ გახდება თითქმის ერთის ტოლი. მაშინ A_{i_0} ჰიპოთეზა ფაქტიურად სამართლიანია. თუ კი გადაწყვეტილების მიღება საჭიროა N ექსპერიმენტის ჩატარების შემდეგ, ხოლო ამ მომენტისათვის აპოსტერიორული ალბათობებიდან არც ერთი არ არის ერთთან საკმაოდ ახლოს, მაშინ გადაწყვეტილება მიიღება იმ ჰიპოთეზის სასარგებლოდ, რომლის აპოსტერიორული ალბათობაც მაქსიმალურია.

მოკლედ, რომ ვთქვათ: სტატისტიკურ გამოყენებებში $A_1, A_2, \dots, A_n$ ხდომილებებს, რომლებიც ქმნიან ხდომილებათა სრულ სისტემას, ხშირად "ჰიპოთეზებს" უწოდებენ, $P(A_i)$ -- ალბათობებს A_i ხდომილებების აპრიორულ (ცდამდე) ალბათობებს. პირობით ალბათობას $P(A_i \mid B)$ კი ეძლევა B ხდომილების მოხდენის შემდეგ A_i ჰიპოთეზის აპოსტერიორული (ცდის შემდგომი) ალბათობის ინტერპრეტაცია.

ამოცანა 1. ყუთში მოთავსებულია ორი მონეტა: A_1 – სიმეტრიული მონეტა გერბის მოსვლის ალბათობით 1/2, და A_2 – არასიმეტრიული მონეტა გერბის მოსვლის ალბათობით 1/3. შემთხვევით ვიღებთ ერთ მონეტას და ვაგდებთ. დავუშვათ, რომ მოვიდა გერბი. როგორია ალბათობა იმისა, რომ ამოღებული მონეტა იყო სიმეტრიული?

ამოხსნა. ამ შემთხვევაში ელემენტარულ ხდომილებათა სივრცე იქნება:

$$\Omega = \{(A_1, \text{გ}), (A_1, \text{ს}), (A_2, \text{გ}), (A_2, \text{ს})\},$$

სადაც მაგალითად, $(A_1, \text{გ})$ – ნიშნავს, რომ ამოვიღეთ A_1 მონეტა და მისი აგდების შედეგად მოვიდა გერბი. ამოცანის პირობებში გვაქვს:

$$P(A_1) = P(A_2) = 1/2, \quad P(\text{გ} \mid A_1) = 1/2 \ \text{ და } \ P(\text{გ} \mid A_2) = 1/3.$$

შესაბამისად, ნამრავლის ალბათობის ფორმულის გამოყენებით გამოვითვლით:

$$P\{(A_1, \text{გ})\} = 1/4, \quad P\{(A_1, \text{ს})\} = 1/4, \quad P\{(A_2, \text{გ})\} = 1/6 \ \text{ და } \ P\{(A_2, \text{ს})\} = 1/3.$$

ამიტომ ბაიესის ფორმულის თანახმად

$$P(A_1 \mid \text{გ}) = \frac{P(A_1)P(\text{გ} \mid A_1)}{P(A_1)P(\text{გ} \mid A_1) + P(A_2)P(\text{გ} \mid A_2)} = \frac{3}{5}.$$

ცხადია, აგრეთვე რომ $P(A_2\,|\,გ)=2/5$.

ამოცანა 2 (კეთილ გამომცდელზე I). ვთქვათ, ჩვენ ჩასაბარებელი გვაქვს გამოცდა და შეგვიძლია ავირჩიოთ ნებისმიერი სამი გამომცდელიდან. დავუშვათ, ჩვენთვის ცნობილია, რომ ერთ-ერთი სამი გამომცდელიდან (უცნობია რომელი) – "კეთილია" და ალბათობა იმისა, რომ მასთან ჩააბარო გამოცდა ტოლია 0.4-ის, ხოლო დანარჩენი ორი გამომცდელი "ავია" და მათთან გამოცდის ჩაბარების ალბათობა ტოლია 0,1-ის. ჩვენ შემთხვევით ავირჩიეთ გამომცდელი და წარმატებით ჩავაბარეთ გამოცდა. როგორია ალბათობა იმისა, რომ ჩვენ ავირჩიეთ "კეთილი" გამომცდელი?

ამოხსნა. შემოვიღოთ შემდეგი ხდომილებები: A – ამორჩეული გამომცდელი "კეთილია" (მაშინ $\overline{A}$ – იქნება ხდომილება, რომ ამორჩეული გამომცდელი "ავია") და B – გამოცდა ჩაბარებულია (შესაბამისად, $\overline{B}$ – გამოცდა არაა ჩაბარებული). ამოცანის პირობებში გვაქვს:

$$P(A)=1/3, \qquad P(\overline{A})=1-P(A)=2/3;$$

$$P(B\,|\,A)=0.4, \qquad P(\overline{B}\,|\,A)=1-P(B\,|\,A)=0{,}6;$$

$$P(B\,|\,\overline{A})=0{,}1, \qquad P(\overline{B}\,|\,\overline{A})=1-P(B\,|\,\overline{A})=0{,}9.$$

ცნობილია, რომ მოხდა B ხდომილება და გამოსათვლელია პირობითი ალბათობა $P(A\,|\,B)$. ვინაიდან, A და $\overline{A}$ ხდომილებები ქმნიან სრულ სისტემას, ბაიესის ფორმულის თანახმად საძიებელი ალბათობა იქნება:

$$P(A\,|\,B)=\frac{P(A)P(B\,|\,A)}{P(A)P(B\,|\,A)+P(\overline{A})P(B\,|\,\overline{A})}=\frac{2}{3}.$$

დავალება. ამოცანა 2-ის პირობებში გამოთვალეთ ალბათობა იმისა, რომ არჩეულ იქნა "ავი" გამომცდელი, თუ ცნობილია, რომ გამოცდა ჩაბარებულ იქნა წარმატებით?

ამოცანა 3 (კეთილ გამომცდელზე II). დავუშვათ, რომ გამომცდელთან, რომელთანაც წარმატებით ჩაიარა გამოცდამ (იხ. ამოცანა 2) გამოსაცდელად რიგ-რიგობით მივიდა კიდევ ორი მოსწავლე. ჯერ გამოცდა ვერ ჩააბარა მეორე მოსწავლემ, შემდეგ მივიდა მესამე და მანაც ვერ ჩააბარა გამოცდა. ამ ფაქტის შემდეგ რომელი ჰიპოთეზაა უფრო დასაჯერებელი: ეს გამომცდელი "კეთილია" თუ "ავი"?

ამოხსნა. ავღნიშნოთ $P_i(A)$ (შესაბამისად, $P_i(\overline{A})$) სიმბოლოთი ალბათობა (აპოსტერიორული) იმისა, რომ ეს გამომცდელი "კეთილია" (შესაბამისად, "ავია") მას შემდეგ რაც გამოცდილ იქნა i-ური სტუდენტი, $i=1,2,3$. ჩვენ უკვე დავადგინეთ, რომ $P_1(A)=2/3$. შესაბამისად,

$$P_1(\overline{A})=1-P_1(A)=1/3.$$

მეორე მოსწავლის თვალსაზრისით ეს ალბათობები წარმოადგენენ ორი შესაძლო ჰიპოთეზის აპრიორულ ალბათობებს. ამიტომ, ბაიესის ფორმულის თანახმად, მეორე სტუდენტის ჩაჭრის შემდეგ აპოსტერიორული ალბათობები იქნება:

$$P_2(A)=\frac{P(\overline{B}\,|\,A)P_1(A)}{P(\overline{B}\,|\,A)P_1(A)+P(\overline{B}\,|\,\overline{A})P_1(\overline{A})}=\frac{4}{7} \quad \text{და} \quad P_2(\overline{A})=1-P_2(A)=\frac{3}{7}.$$

ანალოგიურად, ახლა მიღებული ალბათობები უკვე იქნება აპრიორული ალბათობები მესამე მოსწავლისათვის, და ამიტომ საძიებელი აპოსტერიორული ალბათობები, მას შემდეგ რაც მესამე მოსწავლემ ვერ ჩააბარა გამოცდა, გამოითვლება ისევ ბაიესის ფორმულით:

$$P_3(A)=\frac{P(\overline{B}\mid A)P_2(A)}{P(\overline{B}\mid A)P_2(A)+P(\overline{B}\mid \overline{A})P_2(\overline{A})}=\frac{8}{17} \quad \text{და} \quad P_3(\overline{A})=1-P_3(A)=\frac{9}{17}>P_3(A).$$

როგორც ვხედავთ, ექსპერიმენტის (გამოცდის) დაწყების წინ აპრიორული ალბათობა იმისა, რომ არჩეული გამომცდელი "კეთილია" ტოლი იყო 1/3-ის. ექსპერიმენტების შემდეგ ამ ხდომილების აპოსტერიორული ალბათობა გაიზარდა და გახდა 8/17. მიუხედავად ამისა, თუ სამი ექსპერიმენტის შემდეგ მოსალოდნელია გადაწყვეტილება ამ გამომცდელის შესახებ, მაშინ უფრო სარწმუნოა ჩავთვალოთ იგი "ავად" (ვინაიდან $P_3(\overline{A})>P_3(A)$).

განმეორებითი ცდები. ბერნულის ფორმულა.

განვიხილოთ ერთი და იგივე ექსპერიმენტების სერია, რომლებიც ტარდება ერთი და იგივე პირობებში ერთმანეთისაგან დამოუკიდებლად (ნებისმიერი ექსპერიმენტის შედეგი დამოუკიდებელია დანარჩენი ექსპერიმენტების შედეგებისაგან). ამასთანავე, ყოველ კონკრეტულ ექსპერიმენტში (ელემენტარული ხდომილებების როლში) ჩვენ განვიხილავთ მხოლოდ ორ შედეგს: გარკვეული A ხდომილების მოხდენა (რომელსაც პირობითად "*წარმატებას*" უწოდებენ) და მისი არ მოხდენა – $\overline{A}$ (ე. ი. A ხდომილების საწინააღმდეგო ხდომილების მოხდენა, რომელსაც "*მარცხს*" უწოდებენ). ასე რომ $A+\overline{A}=\Omega$. A ხდომილების მოხდენის ალბათობა ნებისმიერი ექსპერიმენტისათვის მუდმივია და ტოლია $P(A)=p$, სადაც $0<p<1$. შესაბამისად, $P(\overline{A})=1-P(A)=1-p=q$ $(p+q=1)$.

დავუშვათ, ჩატარდა n დამოუკიდებელი ექსპერიმენტი, რომლებსაც ჩვენ განვიხილავთ როგორც ერთ რთულ ექსპერიმენტს. ყოველი ექსპერიმენტის შედეგს ჩვენ წარმოვადგენთ n-ეულების სახით, სადაც თითოეულ ადგილზე დავწერთ ან A-ს ან $\overline{A}$-ს იმის მიხედვით მოხდა A თუ $\overline{A}$. მაგალითად, ორი ექსპერიმენტის შემთხვევაში შესაძლებელია $2^2=4$ შედეგი: $AA, A\overline{A}, \overline{A}A, \overline{A}\overline{A}$ (A ხდომილება მოხდა ორჯერ, A ხდომილება მოხდა პირველ და არ მოხდა მეორე ექსპერიმენტში, A ხდომილება არ მოხდა პირველ და მოხდა მეორე ექსპერიმენტში, A ხდომილება არ მოხდა ორჯერ). სამი ექსპერიმენტის შემთხვევაში მოსალოდნელია $2^3=8$ შედეგი:

$$AAA, AA\overline{A}, A\overline{A}A, \overline{A}AA, A\overline{A}\overline{A}, \overline{A}A\overline{A}, \overline{A}\overline{A}A, \overline{A}\overline{A}\overline{A}.$$

და ა. შ. n ექსპერიმენტის ყველა შესაძლო შედეგს (სულ იქნება 2^n შედეგი) შეესაბამება n ასოს მიმდევრობა A, $\overline{A}$ იმ რიგით რა მიმდევრობითაც შეგვხვდება ეს ხდომილებები n ექსპერიმენტში, მაგალითად, $\overline{A}\overline{A}AA\cdots\overline{A}$.

ვინაიდან ექსპერიმენტები დამოუკიდებელია, ამიტომ n ექსპერიმენტის თითოეული შესაძლო შედეგის ალბათობა გამოითვლება შესაბამის ექსპერიმენტებში A და $\overline{A}$ ხდომილებების ალბათობების გადამრავლებით. ასე

მაგალითად, ხუთი დაწერილი შედეგისათვის (იმის გათვალისწინებით, რომ ყოველ ექსპერიმენტში $P(A)=p$ და $P(\bar{A})=q$) მივიღებთ ალბათობას:

$$P(A)P(\bar{A})P(\bar{A})P(A)\cdots P(\bar{A})=pqqp\cdots q.$$

ცხადია, რომ თუ დაწერილ შედეგში ასო A შეგვხვდა x, და შესაბამისად, ასო $\bar{A}$ გვხვდება $(n-x)$-ჯერ, მაშინ ასეთი შედეგის ალბათობა იქნება: $p^x q^{n-x}$. დამოუკიდებლად იმისგან რა თანმიმდევრობითაა განლაგებული n-ეულში x ასო A და $n-x$ ასო $\bar{A}$. სამი ექსპერიმენტის რვა შესაძლო შედეგისათვის ამ გზით დათვლილი ალბათობები იქნება:

$$P(AAA)=p^3.\ \ P(AA\bar{A})=P(A\bar{A}A)=P(\bar{A}AA)=p^2q.$$

$$P(A\overline{AA})=P(\bar{A}A\bar{A})=P(\overline{AA}A)=pq^2 \text{ და } P(\overline{AAA})=q^3$$

აღვნიშნოთ $P_3(i)$ სიმბოლოთი ალბათობა იმისა, რომ სამ ექსპერიმენტში A ხდომილება შეგვხვდა (მოხდა) ზუსტად i-ჯერ. მაშინ შედეგს სამ ექსპერიმენტში A ხდომილება არც ერთხელ არ შეგვხვდა (A ხდომილება მოხდა 0-ჯერ) აქვს ალბათობა $P_3(0)=P(\overline{AAA})=q^3$. A ხდომილება მოხდა ზუსტად ერთჯერ განხორციელდება თუ მოხდა რომელიმე შემდეგი სამი ვარიანტიდან: $AA\bar{A}$ ან $A\bar{A}A$ ან $\bar{A}AA$. რომელთაგან თითოეულის ალბათობაა pq^2. ამიტომ ალბათობათა ჯამის კანონის თანახმად:

$$P_3(1)=P(A\overline{AA})+P(\bar{A}A\bar{A})+P(\overline{AA}A)=3q^2p.$$

ანალოგიურად,

$$P_3(2)=P(AA\bar{A})+P(A\bar{A}A)+P(\bar{A}AA)=3qp^2$$

და, ბოლოს, $P_3(3)=P(AAA)=p^3$

ვნახოთ, რისი ტოლია ამ ალბათობების ჯამი. გვაქვს:

$$P_3(0)+P_3(1)+P_3(2)+P_3(3)=q^3+q^2p+qp^2+p^3=(q+p)^3=1^3=1,$$

რაც ბუნებრივიცაა ასეც უნდა ყოფილიყო, ვინაიდან ჩვენ განვიხილეთ ალბათობების ჯამი იმ ხდომილებების, რომლებიც ქმნიან ხდომილებათა სრულ სისტემას: A ხდომილება სამ ექსპერიმენტში აუცილებლად მოხდება ან 0-ჯერ, ან 1-ჯერ, ან 2-ჯერ, ან 3-ჯერ.

თუ ახლა $P_n(x)$ სიმბოლოთი აღვნიშნავთ ალბათობას იმისა, რომ n ექსპერიმენტში A ხდომილება (წარმატება) შეგვხვდა (მოხდა) x-ჯერ, მაშინ ანალოგიური მსჯელობით, მივიღებთ ე. წ. **ბერნულის ფორმულას:**

$$P_n(x)=C_n^x p^x q^{n-x}=\frac{n!}{x!(n-x)!}p^x q^{n-x}=\frac{n(n-1)\cdots(n-x+1)}{x!}p^x q^{n-x}. \qquad (1)$$

მართლაც, n ექსპერიმენტის ისეთი შედეგთა რაოდენობა, რომლებიც ჩაიწერებიან x ასო A და $n-x$ ასო $\bar{A}$-ს სხვადასხვა კომბინაციით, ტოლი იქნება ჯუფდებათა რიცხვის n-დან x, ვინაიდან ნებისმიერი ასეთი n-ეული სავსებით განისაზღვრება, თუ n ადგილიდან ამოვარჩევთ ზუსტად x ადგილს ასო A-სათვის, ხოლო დანარჩენ $n-x$ ადგილს დავტოვებთ ასო $\bar{A}$-სათვის. მაგრამ x ნომრის ამორჩევა n ადგილიდან შესაძლებელია სწორედ C_n^x სხვადასხვა გზით, რადგანაც ჯგუფები შედგენილი ნომ-

რებისაგან, რიგითობისაგან დამოუკიდებლად. უნდა განსხვავდებოდნენ ერთი მაინც ელემენტით.

მაგალითი 1. ყუთში 3 თეთრი და 5 შავი ბურთია. ყუთიდან შემთხვევით დაბრუნებით იღებენ 4 ბურთს. იპოვეთ ალბათობა იმისა, რომ ამოღებულ ბურთებიდან თეთრი ბურთების რაოდენობა მეტი იქნება შავი ბურთების რაოდენობაზე?

ამოხსნა. თუ წარმატებად ჩავთვლით თეთრი ბურთის ამოღებას, მაშინ პირობის თანახმად ერთ ცდაში წარმატების ალბათობა იქნება $p=3/8$. საძიებელი ხდომილების ხელშემწყობ უთავსებად შემთხვევებს წარმოადგენს ოთხ ცდაში 3 ან 4 თეთრი ბურთის ამოღება. რომელთა ალბათობები ბერნულის ფორმულის თანახმად შესაბამისად ტოლია:

$$P_4(3)=C_4^3\cdot(3/8)^3\cdot(1-3/8)^{4-3}=135/1024 \quad \text{და}$$

$$P_4(4)=C_4^4\cdot(3/8)^4\cdot(5/8)^0=81/4096.$$

შესაბამისად, საძიებელი ალბათობა ალბათობათა შეკრების კანონის თანახმად იქნება:

$$135/1024+81/4096=621/4096.$$

ალბათობების ერთობლიობას $P_n(x)$, როცა $x=0,1,...,n$, ე. ი. ($P_n(0)$, $P_n(1),...,P_n(n)$ ალბათობებს) ეწოდება **ალბათობების ბინომიალური განაწილება**. რადგანაც ეს ალბათობები შეესაბამება უთავსებად ხდომილებებს, რომლებიც ქმნიან სრულ სისტემას, ამიტომ გასაგებია, რომ:

$$\sum_{x=0}^{n}P_n(x)=1,$$

რაც, მეორეს მხრივ, ადვილად მოწმდება ნიუტონის ბინომის ფორმულის გამოყენებითაც, რადგანაც ბერნულის ფორმულაში მონაწილეობენ სწორედ $(q+p)^n$ ნიუტონის ბინომის კოეფიციენტები (აქედან მოდის სახელწოდებაც "ბინომიალური განაწილება"):

$$\sum_{x=0}^{n}P_n(x)=\sum_{x=0}^{n}C_n^x p^x q^{n-x}=(q+p)^n=1^n=1.$$

ხშირ შემთხვევაში საჭიროა გამოითვალოს ალბათობა იმისა, რომ A ხდომილება n ექსპერიმენტში შეგვხვდება არაუმეტეს x-ჯერ. ამ ალბათობას ეწოდებენ ბინომიალური განაწილების **კუმულატიურ ანუ დაგროვილ ალბათობას**. აღვნიშნოთ იგი $\overline{P}_n(x)$ სიმბოლოთი. მაშინ ალბათობათა შეკრების კანონის თანახმად კუმულატიური ალბათობა ასე გამოითვლება:

$$\overline{P}_n(x)=P_n(0)+P_n(1)+\cdots+P_n(x)=\sum_{i=0}^{x}P_n(i).$$

თუ ექსპერიმენტების n რიცხვი საკმაოდ დიდია, მაშინ $P_n(x)$ და $\overline{P}_n(x)$ ალბათობების გამოთვლა ხდება სპეციალური "ასიმპტოტური" ფორმულებით. თუ n მცირეა, მაშინ შეგვიძლია გამოვიყენოთ მარტივი თანაფარდობა, რომელიც აკავშირებს ბინომური განაწილების ორ მომდევნო $P_n(x)$ და $P_n(x+1)$ წევრს:

$$\frac{P_n(x+1)}{P_n(x)}=\frac{(n-x)p}{(x+1)q}. \qquad (2)$$

თუ ნაპოვნია $P_n(x)$, მაშინ უკანასკნელი თანაფარდობიდან ადვილად გამოვითვლით $P_n(x+1)$-ს.

დამოუკიდებელი ექსპერიმენტების სერიის ზემოთაღწერილი სქემა პირველად განხილული და შესწავლილი იყო შვეიცარიელი მათემატიკოსის იაკობ ბერნულის (1654-1705) მიერ და ამიტომ იგი ატარებს **ბერნულის სქემის** სახელს.

მაგალითი 2. დავუშვათ ვამოწმებთ დეფექტურობაზე საქონლის პარტიას, რომელიც შედგება 30 ნაწარმისაგან. ცნობილია, რომ დეფექტური პროდუქციის წილი შეადგენს 5%-ს. როგორია საქონლის ამ პარტიაში დეფექტური პროდუქციის ამა თუ იმ რიცხვის ალბათობები?

ამოხსნა. ამ შემთხვევაში ექსპერიმენტების რიცხვია $n=30$. სადაც ალბათობის კლასიკური განმარტების თანახმად $p=5/100=0.05$ (შესაბამისად, $q=0.95$). ვისარგებლოთ (1) და (2) ფორმულებით. გვაქვს:

$$P_{30}(0)=C_{30}^{0}0.05^{0}0.95^{30-0}=0.95^{30}=0.2146.$$

გარდა ამისა,

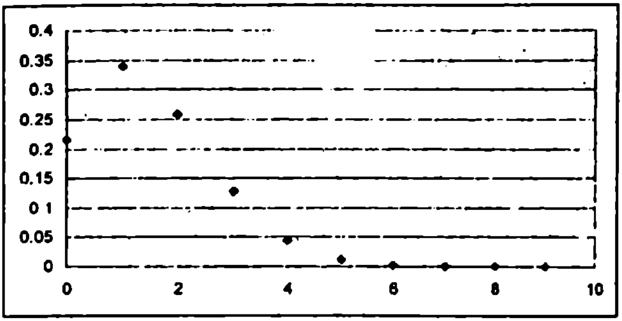
$$P_{30}(x+1)=\frac{30-x}{x+1}\frac{p}{q}P_{30}(x)=\frac{30-x}{x+1}\frac{0.05}{0.95}P_{30}(x)=\frac{30-x}{19(x+1)}P_{30}(x),$$

საიდანაც როცა $x=0$: $P_{30}(0+1)=\dfrac{30-0}{19\cdot(0+1)}\cdot 0.2146=0.3389$;

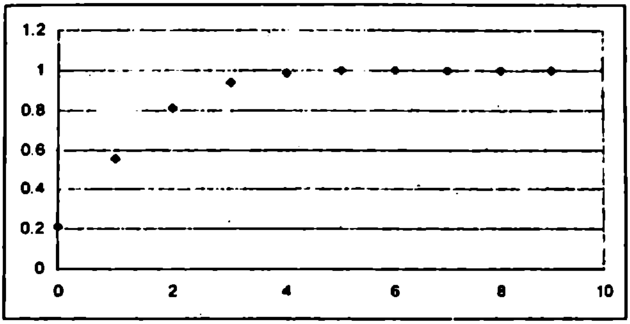
როცა $x=1$: $P_{30}(1+1)=\dfrac{30-1}{19\cdot(1+1)}\cdot 0.3389=0.2586$ და ა.შ. საბოლოოდ გვექნება შემდეგი ცხრილი:

დეფექტური ნაწარმის რიცხვი x	ალბათობა $P_n(x)$	კუმულატიური ალბათობა $\overline{P}_n(x)$
0	0.2146	0.2146
1	0.3389	0.5535
2	0.2586	0.8122
3	0.1270	0.9392
4	0.0451	0.9844
5	0.0124	0.9967
6	0.0027	0.9994
7	0.0005	0.9999
8	0.0001	0.999998
9	0.000001	0.999999

ამ ცხრილის შესაბამისი ალბათობების განაწილების გრაფიკი იქნება:



კუმულატიური ალბათობების შესაბამისი განაწილების გრაფიკი იქნება:



დავალება 1. ექსპერიმენტი მდგომარეობს სამი სათამაშო კამათლის გაგორებაში. იპოვეთ ალბათობა იმისა, რომ ექსპერიმენტის 10-ჯერ გამეორებისას, ზუსტად 4 ექსპერიმენტში მოვა ზუსტად ორ-ორი "6"?

დავალება 2. რამდენი შემთხვევითი ციფრი უნდა ავიღოთ, რომ ციფრი "5" მოვიდეს ერთხელ მაინც არანაკლებ 0.9-ის ტოლი ალბათობით?

განმარტება. იმ k_0 რიცხვს, რომლის შესაბამისი ალბათობა $P_n(k_0)$ უდიდესია $P_n(0)$, $P_n(1) ..... P_n(n)$ ალბათობებს შორის **უალბათესი რიცხვი** ეწოდება.

უალბათესი რიცხვი გვიჩვენებს n დამოუკიდებელ ცდაში წარმატებათა რა რაოდენობაა ყველაზე უფრო მოსალოდნელი.

განვიხილოთ ფარდობა

$$a_k = P_n(k+1) / P_n(k) = (n-k)p/[(k+1)q].$$

ადვილი დასანახია, რომ $a_k > 1$, როცა $k < np-q$; $a_k = 1$, როცა $k = np-q$ და $a_k < 1$, როცა $k > np-q$. საიდანაც ცხადია, რომ უალბათესი რიცხვი წარმოადგენს შემდეგი უტოლობის მთელ ამონახსნს:

$$np - q \le k_0 \le np + p.$$

უალბათესი რიცხვი შეიძლება იყოს ერთი ან ორი. [illegible] ამ უტოლობის საზღვრები მთელი რიცხვებია, თუ [illegible]. მაგალითად, მონეტის 11-ჯერ აგდებისას გერბის მოსვლის უალბათესი რიცხვი იქნება

$$11\cdot1/2-1/2\le k_0\le 11\cdot1/2+1/2 \Leftrightarrow 5\le k_0\le 6$$

ერთეულების მთელი ამონახსნი, ანუ 5 და 6. რაც იმას ნიშნავს, რომ მონეტის 11-ჯერ აგდებისას გერბის 5-ჯერ და 6-ჯერ მოსვლის ალბათობები ერთმანეთის ტოლია და ყველა დანარჩენ ალბათობებზე მეტია.

პუასონის ფორმულა.

გამოთვლების ჩატარება ბერნულის ფორმულის გამოყენებით, ცდათა დიდი რიცხვის შემთხვევაში, მოითხოვს ძალიან დიდ ძალისხმევას. მიახლოებითი გამოთვლების ჩასატარებლად შესაძლებელია უფრო მოხერხებული ფორმულის მიღება, თუ კი ცდათა დიდი რიცხვის შემთხვევაში ცალკეულ ცდაში A ხდომილების მოხდენის p ალბათობა მცირეა, ხოლო ნამრავლი $np=\lambda$ ინარჩუნებს მუდმივ მნიშვნელობას ექსპერიმენტების სხვადასხვა სერიაში (ანუ A ხდომილების მოხდენის საშუალო რიცხვი უცვლელი რჩება ექსპერიმენტების სხვადასხვა სერიაში). ბერნულის ფორმულა შეგვიძლია გადავწეროთ შემდეგი სახით:

$$p_n(k)=\frac{n(n-1)(n-2)...(n-k+1)}{k!}p^k(1-p)^{n-k}=\frac{n(n-1)...(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

გამოვთვალოთ მიღებული გამოსახულების ზღვარი, როცა $p\to0$ და $n\to\infty$, ისე რომ $np\to\lambda$. ადვილი დასანახია, რომ:

$$p_n(k)=\frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1\cdot\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{n-k}\right)=\frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}=\frac{\lambda^k}{k!}$$

მიღებულ ფორმულას: $\lim\limits_{np\to\lambda}p_n(k)=\lambda^k e^{-\lambda}/k!$ **პუასონის ფორმულა** ეწოდება. იგი საშუალებას იძლევა ვიპოვოთ n დამოუკიდებელ ცდაში A ხდომილების k-ჯერ მოხდენის ალბათობა (როცა n საკმაოდ დიდია, ხოლო p საკმაოდ მცირე, ამასთანავე $np=\lambda<15$) პუასონის მიახლოებითი ფორმულით: $p_n(k)\approx\lambda^k e^{-\lambda}/k!$

აღსანიშნავია, რომ პუასონის ფორმულით სარგებლობისას, განსხვავებით ბერნულის ფორმულის შემთხვევისაგან, ჩვენ არ გვჭირდება მის გამოსახულებაში სიდიდეების (მონაცემების) შეტანა კონკრეტული ამოცანიდან, არამედ უბრალოდ ვსარგებლობთ პუასონის განაწილების ცხრილებით. ქვემოთ მოყვანილია ამ ცხრილის ერთი ფრაგმენტი:

	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$
p(0)	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066
p(1)	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
p(2)	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
p(3)	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
p(4)		0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
p(5)				0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
p(6)							0.0001	0.0002	0.0003

თავი III
შემთხვევითი სიდიდის განაწილება

ალბათობის თეორიაში შემთხვევითი ხდომილების ცნებასთან ერთად გამოიყენება გარკვეული აზრით უფრო მოხერხებული *შემთხვევითი სიდიდის* ცნება. ცვლად სიდიდეს, რომლის მნიშვნელობები დამოკიდებულია შემთხვევითი ექსპერიმენტის ან მოვლენის შესაძლო შედეგებზე, შემთხვევითი სიდიდეს უწოდებენ. შემთხვევითი სიდიდის მაგალითებია: სათამაშო კამათლის გაგორებისას მოსულ ქულათა რიცხვი; მონეტის განმეორებითი აგდებისას მონეტის რომელიმე მხარის გამოჩენათა რიცხვი; გასროლათა რაოდენობა მიზანში პირველად მოხვედრამდე; მანძილი სამიზნის ცენტრიდან დაზიანების წერტილამდე; სხვადასხვა დროს გარკვეულ პროდუქციაზე მოთხოვნათა რაოდენობა; სითხეში ჩაძირული მტვრის მცირე ნაწილაკის (რომელსაც ვაკვირდებით მიკროსკოპში) მდებარეობა და ა. შ.

განმარტება. შემთხვევითი ექსპერიმენტის ელემენტარულ ხდომილებათა სივრცეზე განსაზღვრულ რიცხვით ფუნქციას **შემთხვევითი სიდიდე** ეწოდება. შემთხვევით სიდიდეს ეწოდება **დისკრეტული ტიპის** თუ ის ღებულობს ცალკეულ, იზოლირებულ შესაძლო მნიშვნელობებს. შემთხვევით სიდიდეს ეწოდება **უწყვეტი ტიპის** თუ მისი შესაძლო მნიშვნელობების სიმრავლე მთლიანად ავსებს რაიმე სასრულ ან უსასრულო რიცხვით შუალედს.

დისკრეტული ტიპის შემთხვევითი სიდიდე ღებულობს სასრულ ან თვლად რაოდენობა განსხვავებულ მნიშვნელობებს, ხოლო უწყვეტი ტიპის შემთხვევითი სიდიდის მნიშვნელობათა რაოდენობა კონტინიუმის სიმძლავრისაა.

შემთხვევით სიდიდეებს აღნიშნავენ დიდი ლათინური ასოებით: $X, Y, Z, \ldots$ (ან პატარა ბერძნული ასოებით $\xi, \eta, \varsigma, \ldots$), ხოლო შემთხვევითი სიდიდის შესაძლო მნიშვნელობებს აღნიშნავენ პატარა ლათინური ასოებით: $x_i, y_i, z_i, \ldots$.

მაგალითი 1. შემთხვევითი სიდიდე იყოს მონეტის სამჯერ აგდებისას მოსულ გერბთა რიცხვი. ამ შემთხვევაში ელემენტარულ ხდომილებათა სივრცე რვა ელემენტიანი სიმრავლეა:

$$\Omega = \{\text{გგგ}, \text{გგს}, \text{გსგ}, \text{სგგ}, \text{გსს}, \text{სგს}, \text{სსგ}, \text{სსს}\}$$

და, შესაბამისად, საძიებელი შემთხვევითი სიდიდე იქნება Ω-ზე განსაზღვრული შემდეგი რიცხვითი ფუნქცია:

$$X(\text{გგგ}) = 3;\ \ X(\text{გგს}) = X(\text{გსგ}) = X(\text{სგგ}) = 2;$$

$$X(\text{გსს}) = X(\text{სგს}) = X(\text{სსგ}) = 1 \ \text{ და } \ X(\text{სსს}) = 0.$$

ცხადია ეს შემთხვევითი სიდიდე დისკრეტული ტიპისაა, ის ღებულობს იზოლირებულ მნიშვნელობებს, მაგალითად, 1-სა და 2-ს შორის ის არ ღებულობს არცერთ მნიშვნელობას.

მაგალითი 2. შემთხვევითი სიდიდე იყოს ორი სათამაშო კამათლის გაგორებისას მოსულ ქულათა ჯამი. ამ შემთხვევაში ელემენტარულ ხდომილებათა სივრცე შედგება 36 ელემენტარული ხდომილებისაგან:

$$\Omega = \{(i, j) : i, j = 1, 2, \ldots, 6\},$$

ხოლო შემთხვევითი სიდიდე ცალკეულ ელემენტარულ ხდომილებას (i,j) (სადაც i – პირველ კამათელზე მოსული ქულაა, ხოლო j – მეორე კამათელზე მოსული ქულა) შეუსაბამებს: $X(i,j)=i+j$ (პირველ და მეორე კამათელზე მოსული ქულების ჯამი). მაგალითად, $X(1,3)=X(2,2)=X(3,1)=4$. აღნიშნული შემთხვევითი სიდიდის შესაძლო მნიშვნელობებია: 2, 3, 12. ის ასევე დიკრეტული ტიპისაა.

ზემოთ ჩამოთვლილი მაგალითებიდან უწყვეტი ტიპის შემთხვევითი სიდიდეა მანძილი სამიზნის ცენტრიდან დაზიანების წერტილამდე და მტერის ნაწილაკის მდებარეობა სითხეში. თითოეული მათგან ნებისმიერ ორ მიღებულ მნიშვნელობას შორის არ გამოტოვებს არცერთ მნიშვნელობას.

შემთხვევითი სიდიდე მოცემულია თუ ჩვენ ვიცით ექსპერიმენტის ამა თუ იმ შედეგს რა რიცხვი შეესაბამება. მაგრამ, იმისათვის რომ ალბათურად დავახასიათოთ შემთხვევითი სიდიდე, ჩვენ კიდევ უნდა ვიცოდეთ თუ რამდენად ხშირად ანუ რა ალბათობებით ღებულობს ეს შემთხვევითი სიდიდე თავის ამა თუ იმ მნიშვნელობას. შესაბამისობას, შემთხვევითი სიდიდის შესაძლო მნიშვნელობებსა და მათ შესაბამის ალბათობებს შორის, დისკრეტული ტიპის შემთხვევითი სიდიდის **განაწილების კანონი** ეწოდება. შემთხვევითი სიდიდის განაწილების კანონი შეიძლება მოცემული იყოს *ცხრილის, ფორმულის ან გრაფიკის* სახით.

ცხრილს, რომელშიც ჩამოთვლილია შემთხვევითი სიდიდის შესაძლო მნიშვნელობები და მათი შესაბამისი ალბათობები, დისკრეტული ტიპის შემთხვევითი სიდიდის **განაწილების მწკრივი** ეწოდება:

x_i	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

შევნიშნოთ, რომ ხდომილება, რომ შემთხვევითი სიდიდე მიიღებს ერთ-ერთ მნიშვნელობას თავისი შესაძლო მნიშვნელობებიდან, წარმოადგენს აუცილებელ ხდომილებას და ამიტომ: $\sum_i p_i = 1$ (ჩვენ არ ვუთითებთ შესაკრებთა რაოდენობას, ის შეიძლება იყოს როგორც სასრული, ისე უსასრულო).

ამოცანა 1. ორი მსროლელი თითოჯერ ესვრის სამიზნეს. მათ მიერ სამიზნის დაზიანების (მიზანში მოხვედრის) ალბათობებია შესაბამისად 0.6 და 0.7. შემთხვევითი სიდიდე X იყოს დაზიანებულ სამიზნეთა რაოდენობა. შევადგინოთ მისი განაწილების მწკრივი.

ამოხსნა. ცხადია, რომ X შემთხვევითმა სიდიდემ შეიძლება მიიღოს შემდეგი მნიშვნელობები: 0 (ვერც ერთმა მსროლელმა ვერ დააზიანა სამიზნე), 1 (მხოლოდ ერთმა მსროლელმა დააზიანა სამიზნე) და 2 (ორივე მსროლელმა დააზიანა სამიზნე). ვიპოვოთ შესაბამისი ალბათობები.

ბუნებრივია შეგვეძლია ვიგულისხმოთ რომ პირველი და მეორე მსროლელის სროლის შედეგები ერთმანეთისაგან დამოუკიდებელია. შემოვიღოთ ხდომილებები: A – პირველმა მსროლელმა დააზიანა სამიზნე და B - მეორე მსროლელმა დააზიანა სამიზნე. მოცემულია, რომ $P(A)=0.6$ და

$P(B)=0.7$. შესაბამისად, $P(\overline{A})=0.4$ და $P(\overline{B})=0.3$. გარდა ამისა, A და B დამოუკიდებელი ხდომილებებია. დამოუკიდებელი ხდომილებებია აგრეთვე: $\overline{A}$ და B, $\overline{A}$ და B , $\overline{A}$ და $\overline{B}$.

ადვილი დასანახია, რომ ხდომილება – ვერც ერთმა მსროლელმა ვერ დააზიანა სამიზნე იქნება $\overline{A}\cap\overline{B}$, ხდომილება – მხოლოდ ერთმა მსროლელმა დააზიანა სამიზნე იქნება $(A\cap\overline{B})\cup(\overline{A}\cap B)$ და ხდომილება – ორივე მსროლელმა დააზიანა სამიზნე იქნება $A\cap B$. გასაგებია, რომ $(A\cap\overline{B})$ და $(\overline{A}\cap B)$ უთავსებადი ხდომილებებია $(A\cap\overline{B})\cap(\overline{A}\cap B)=\varnothing$.

ამიტომ, დამოუკიდებელ ხდომილებათა ნამრავლის ალბათობისა და უთავსებად ხდომილებათა ჯამის ალბათობის ფორმულების თანახმად გვექნება:

$$P(X=0)=P(\overline{A}\cap\overline{B})=P(\overline{A})\cdot P(\overline{B})=0.4\cdot 0.3=0{,}12;$$

$$P(X=1)=P\{(A\cap\overline{B})\cap(\overline{A}\cap B)\}=P(A\cap\overline{B})+P(\overline{A}\cap B)=$$

$$=P(A)P(\overline{B})+P(\overline{A})P(B)=0.6\cdot 0.3+0{,}4\cdot 0{,}7=0{,}46;$$

$$P(A\cap B)=P(A)\cdot P(B)=0.6\cdot 0.7=0{,}42$$

შესაბამისად, X შემთხვევითი სიდიდის განაწილების მწკრივი იქნება:

x_i	0	1	2
p_i	0.12	0.46	0.42

გრაფიკულად დისკრეტული შემთხვევითი სიდიდის განაწილების კანონი შეიძლება წარმოვადგინოთ *განაწილების მრავალკუთხედის სახით,* რომელიც წარმოადგენს ტეხილს სიბრტყეზე, რომელიც მიიღება საკოორდინატო სიბრტყეზე იმ წერტილების შეერთებით, რომელთა კოორდინატებია (x_i, p_i).

თუ მოცემულია დისკრეტული ტიპის შემთხვევითი სიდიდე X და რაიმე რიცხვითი g ფუნქცია, მაშინ $g(X)$ ისევ იქნება დისკრეტული ტიპის შემთხვევითი სიდიდე, რომლის განაწილების მწკრივის პირველ სტრიქონში იქნება $g(x_i)$ რიცხვები ($g(X)$ შემთხვევითი სიდიდის შესაძლო მნიშვნელობები), ხოლო მეორე სტრიქონში იგივე p_i ალბათობები, რაც გვქონდა X შემთხვევითი სიდიდის განაწილების მწკრივში, ვინაიდან:

$$P\{g(X)=g(x_i)\}=P(X=x_i)=p_i,$$

ანუ გვექნება განაწილების მწკრივი:

$g(x_i)$	$g(x_1)$	$g(x_2)$		$g(x_n)$	
p_i	p_1	p_2		p_n	

შევნიშნოთ, რომ შესაძლებელია X-ის რომელიმე ორი განსხვავებუ-ლი $x_j \neq x_k$ მნიშვნელობისათვის $g(x_j)=g(x_k)$, მაშინ $g(X)$-ის განაწილების მწკრივში მხოლოდ ერთ ადგილას დავწერთ $g(x_j)$-ს და ქვეშ მივუწერთ შესაბამისი ალბათობის როლში (p_j+p_k) სიდიდეს. მაგალითად, თუ X შემთხვევითი სიდიდის განაწილების მწკრივია:

x_i	-3	-1	0	1	2
p_i	0.15	0.12	0.2	0.18	0.35

მაშინ X^2-ის (ამ შემთხვევაში $g(x)=x$) განაწილების მწკრივი იქნება:

x_i^2	0	1	4	9
p_i	0.2	0.3	0.35	0.15

აქ $P(X^2=1)=P\{(X=-1)\cup(X=1)\}=P(X=-1)+P(X=-1)=0.12+0.18=0.3$.

ჰიპერგეომეტრიული განაწილება. დავუშვათ, რომ ყუთში N ბურთია და მათ შორის M თეთრია. შემთხვევით, დაბრუნების გარეშე ყუთიდან ვი-ღებთ n ბურთს. ვიპოვოთ ალბათობა იმისა, რომ ამოღებულ n ბურთს შორის ზუსტად k ცალი იქნება თეთრი?

ავღნიშნოთ μ_n-ით ამოღებულ n ბურთს შორის თეთრი ბურთების რაოდენობა. ჩვენ გვაინტერესებს $P(\mu_n=k)$ ალბათობა. ვისარგებლოთ ალბ-ათობის კლასიკური განმარტებით. გასაგებია, რომ ყველა შესაძლო შედეგ-თა რაოდენობა დაემთხვევა N ელემენტიანი სიმრავლის n ელემენტიან ქვესიმრავლეთა რაოდენობას, ანუ $P(\Omega)=C_N^n$. ჩვენთვის საინტერესო n ელ-ემენტიანი ქვესიმრავლეები უნდა შედგებოდნენ ზუსტად k ცალი თეთრი და $n-k$ ცალი შავი ბურთებისაგან. k ცალი თეთრი ბურთი შეიძლება შე-ირჩეს C_M^k სხვადასხვა გზით, ხოლო $n-k$ ცალი შავი ბურთი კი – C_{N-M}^{n-k} სხვადასხვანაირად. ნამრავლის წესის თანახმად ხელშემწყობ ელემენტარ-ულ ხდომილებთა რაოდენობა იქნება $C_M^k \cdot C_{N-M}^{n-k}$. შესაბამისად, კლასიკური განმარტების საფუძველზე გვაქვს:

$$P(N;M;n;k) := P(\mu_n=k)=\frac{C_M^k \cdot C_{N-M}^{n-k}}{C_N^n},\quad k=0,1,\dots,n.$$

რიცხვთა ამ მიმდევრობას **ჰიპერგეომეტრიული განაწილება** ეწოდება.

ამოცანა. აუდიტორიაში მყოფი 15 სტუდენტიდან 5 ვაჟია. ვიპოვოთ ალბათობა იმისა, რომ შემთხვევით შერჩეულ 6 სტუდენტს შორის 3 ვაჟია?

ამოხსნა. თუ მივუსადაგებთ ზემოთ განხილულ სქემას, გასაგებია, რომ: $N=15$, $M=5$, $n=6$ და $k=3$. ამიტომ საძიებელი ალბათობა იქნება:

$$P(15;5;6;3)=\frac{C_{10}^{3}C_{15-10}^{6-3}}{C_{15}^{6}}=\frac{C_{10}^{3}C_{5}^{3}}{C_{15}^{6}}=\frac{120\cdot 10}{5005}\approx 0.239$$

დავუშვათ, რომ ვატარებთ დამოუკიდებელი ორშედეგიანი ცდების სერიას ერთ-ერთი შედეგის (პირობითად მას ეუწოდოთ "წარმატება") პირველად მიხდენამდე. ცალკეულ ცდაში "წარმატების" ალბათობა იყოს p (მეორე შედეგის ალბათობა იქნება $1-p\equiv q$). შემთხვევითი სიდიდე იყოს ჩატარებული ცდების რაოდენობა. მაშინ ცხადია, რომ ეს შემთხვევითი სიდიდე მიიღებს მნიშვნელობას k ალბათობით pq^{k-1}, $k=1,2,\ldots$.

გეომეტრიული განაწილება. დისკრეტულ X შემთხვევით სიდიდეს, რომელიც ღებულობს ნატურალურ k მნიშვნელობებს ალბათობებით

$$P(X=k)=pq^{k-1},$$

სადაც $0<p<1$ $(q=1-p)$, გეომეტრიული კანონით განაწილებული შემთხვევითი სიდიდე ეწოდება.

უსასრულოდ კლებადი გეომეტრიული პროგრესიის წევრთა ჯამის ფორმულის გამოყენებით ადვილი შესამოწმებელია, რომ ამ ალბათობების ჯამი 1-ის ტოლია:

$$\sum_{k=1}^{\infty}pq^{k-1}=p\cdot\sum_{k=1}^{\infty}q^{k-1}=p\cdot\frac{1}{1-q}=p\cdot\frac{1}{p}=1.$$

პუასონის განაწილება. განვიხილოთ დისკრეტული შემთხვევითი სიდიდე X, რომელიც ღებულობს მხოლოდ მთელ არაუარყოფით მნიშვნელობებს $(0,1,2,\ldots,m,\ldots)$. რომელთა მიმდევრობა შემოუსაზღვრელია. ასეთ შემთხვევით სიდიდეს ეწოდება **პუასონის კანონით** განაწილებული, თუ ალბათობა იმისა, რომ ის მიიღებს მნიშვნელობას m, გამოისახება ფორმულით

$$p(X=m)=\frac{a^{m}}{m!}e^{-a},$$

სადაც a — გარკვეული დადებითი სიდიდეა, რომელსაც პუასონის კანონის (განაწილების) **პარამეტრი** ეწოდება. თუ ვისარგებლებთ e^{x} ფუნქციის გაშლით ხარისხოვან მწკრივად $(e^{x}=\sum_{m=0}^{\infty}x^{m}/m!)$, ადვილად დავინახავთ, რომ ამ ალბათობების ჯამი 1-ის ტოლია. მართლაც,

$$\sum_{m=0}^{\infty}p(X=m)=e^{-a}\sum_{m=0}^{\infty}\frac{a^{m}}{m!}=e^{-a}\cdot e^{a}=1$$

განვიხილოთ ტიპიური ამოცანა, რომელსაც მივყავართ პუასონის განაწილებამდე. დავუშვათ, რომ აბსცისთა ღერძზე შემთხვევით განაწილდებიან წერტილები, ამასთანავე მათი განაწილება აკმაყოფილებს შემდეგ პირობებს:

1). ალბათობა იმისა, რომ გარკვეული რაოდენობის წერტილები მოხ-

ედება l სიგრძის ინტერვალში დამოკიდებულია მხოლოდ ინტერვალის სიგრძეზე და არაა დამოკიდებული აბსცისთა ღერძზე მის მდებარეობაზე (ე. ი. წერტილები განაწილებულია ერთნაირი საშუალო სიმკვრივით);

2). წერტილები ნაწილდებიან ერთმანეთისაგან დამოუკიდებლად: ალბათობა იმისა, რომ წერტილთა რაიმე რაოდენობა მოხვდება მოცემულ ინტერვალში არ არის დამოკიდებული წერტილთა რაოდენობაზე, რომლებიც მოხვდნენ ნებისმიერ სხვა ინტერვალში;

3). პრაქტიკულად შეუძლებელია ორი ან მეტი წერტილის დამთხვევა.

მაშინ შემთხვევითი სიდიდე X – l სიგრძის ინტერვალში მოხვედრილ წერტილთა რაოდენობა – განაწილებულია პუასონის კანონით, სადაც a – არის l სიგრძის ინტერვალზე მოსულ წერტილთა საშუალო რიცხვი.

შენიშვნა. ვინაიდან პუასონის ფორმულა გამოსახავს ბინომიალურ განაწილებას ცდათა დიდი რიცხვისა და ხდომილების მცირე ალბათობის შემთხვევაში, ამიტომ პუასონის კანონს ხშირად უწოდებენ **იშვიათ მოვლენათა კანონს**.

პუასონის განაწილება წარმოადგენს კარგ მათემატიკურ მოდელს იშვიათ ხდომილებათა აღსაწერად: დროის ფიქსირებულ შუალედში მომხდარ ხდომილებათა რაოდენობა ხშირად ემორჩილება პუასონის განაწილებას. მაგალითად, შეიძლება გამოდგეს გეიგერის მთვლელის მიერ t დროში რეგისტრირებული რადიაქტიური დაშლის შედეგად α ნაწილაკების რაოდენობა, სატელეფონო სადგურში t დროის განმავლობაში რეგისტრირებულ გამოძახებათა რაოდენობა. როგორც ჩვენ უკვე ვნახეთ, წარმატებების მცირე ალბათობისა და ცდათა რიცხვის საკმაოდ დიდი რაოდენობის შემთხვევაში პუასონის განაწილება გვევლინება ბინომური განაწილების მიახლოებად.

შემთხვევითი სიდიდის განაწილების ფუნქცია და განაწილების სიმკვრივე.

შემთხვევითი სიდიდის განაწილება – ეს არის ფუნქცია, რომელიც ცალსახად განსაზღვრავს ალბათობას იმისა, რომ: შემთხვევითმა სიდიდემ მიიღო მოცემული მნიშვნელობა ან შემთხვევითი სიდიდე ეკუთვნის გარკვეულ მოცემულ ინტერვალს. თუ შემთხვევითი სიდიდე ღებულობს სასრულ რაოდენობა მნიშვნელობებს, მაშინ განაწილება მოიცემა ფუნქციით $P(X=x)$, რომელიც X შემთხვევითი სიდიდის ყველა შესაძლო x მნიშვნელობას შეუსაბამებს ალბათობას იმისა, რომ $X=x$ (ანუ განაწილების კანონით):

$$P(X\in\langle a,b\rangle)=\sum_{x\in\langle a,b\rangle}P(X=x),$$

სადაც a და b ($a<b$) ნებისმიერი ნამდვილი რიცხვებია, ხოლო $\langle a,b\rangle$ – ნებისმიერი ტიპის ინტერვალია (როგორც ღია, ისე ნახევრად ღია და ჩაკეტილი).

თუ შემთხვევითი სიდიდე ღებულობს უსასრულოდ ბევრ მნიშვნელობას (რაც შესაძლებელია მხოლოდ მაშინ, როცა ელემენტარულ ხდომილებათა სივრცე, რომელზეც განმარტებულია შემთხვევითი სიდიდე შედგება უსასრულო რაოდენობა ელემენტარული ხდომილებებისაგან), მაშინ განაწილება მოიცემა $P(a\le X<b)$ ალბათობების ერთობლიობით რიცხვთა ნები-

სმიერი a,b წყვილისათვის, $a<b$. განაწილება შესაძლებელია მოცემულ იქნეს ე. წ. *განაწილების ფუნქციით:*

$$F(x):=P(X<x),$$

რომელიც ნებისმიერი ნამდვილი x რიცხვისთვის განსაზღვრავს ალბათობას იმისა, რომ X შემთხვევითი სიდიდე მიიღებს x-ზე ნაკლებ მნიშვნელობებს. ადვილი დასანახია, რომ:

$$P(a\le X<b)=F(b)-F(a).$$

მართლაც, ხდომილებათა სხვაობის ალბათობის ფორმულის თანახმად გვექნება:

$$P(a\le X<b)=P\{(X<b)\setminus(X<a)\}=P(X<b)-P(X<a)=F(b)-F(a).$$

ეს თანაფარდობა გვიჩვენებს თუ როგორ შეიძლება განაწილების ფუნქციის საშუალებით გამოვთვალოთ განაწილება და პირიქით, როგორ გამოვთვალოთ განაწილების ფუნქცია განაწილების საშუალებით:

$$F(x)=P(X<x)=P(-\infty\le X<x).$$

განაწლების ფუნქციის თვისებები:

1). ნებისმიერი x-სათვის $0\le F(x)\le 1$;

2). განაწილების ფუნქცია არაკლებადია;

3). განაწილების ფუნქცია უწყვეტია მარცხნიდან (თუ განაწილების ფუნქციას განვმარტავთ როგორც: $F(x):=P(X\le x)$, მაშინ ის იქნება მარჯვნიდან უწყვეტი);

4). $F(x)=\sum\limits_{x_k<x}P\{X=x_k\}$ (შესაბამისად, $F(x)=\sum\limits_{x_k\le x}P\{X=x_k\}$);

5). $P\{X=x_k\}=F(x_k+0)-F(x_k)$ (შესაბამისად,

$$P\{X=x_k\}=F(x_k)-F(x_k-0),\ \text{როცა}\ F(x):=P(X\le x)).$$

განაწილების ფუნქცია შეიძლება იყოს ან დისკრეტული, ან უწყვეტი, ან მათი კომბინაცია. დისკრეტული განაწილების ფუნქცია შეესაბამება დისკრეტულ შემთხვევით სიდიდეს, რომელიც ღებულობს სასრულ რაოდენობა მნიშვნელობებს ან მნიშვნელობებს ისეთი სიმრავლიდან, რომლის ელემენტების გადანომვრა-აც შეიძლება ნატურალური რიცხვებით (ასეთ სიმრავლეებს, მათემატიკაში, *თვლად სიმრავლეებს* უწოდებენ). დისკრეტულ განაწილების ფუნქციას აქვს საფეხურა კიბის სახე.

მაგალითი 1. საქონლის პარტიაში დეფექტურ ნაწარმთა რიცხვი X ღებულობს მნიშვნელობა 0-ს ალბათობით – 0.3; მნიშვნელობა 1-ს ალბათობით –0.4; მნიშვნელობა 2-ს ალბათობით – 0.2 და მნიშვნელობა 3-ს ალბათობით – 0.1 ანუ X-ის განაწილების მწკრივს აქვს სახე:

x_i	0	1	2	3
p_i	0.3	0.4	0.2	0.1

გამოვთვალოთ X-ის განაწილების ფუნქცია და ავაგოთ მისი გრაფიკი.

თუ $x\le 0$, მაშინ $F(x)=P(X<x)=P(\varnothing)=0$;

თუ $0<X\le 1$, მაშინ $F(x)=P(X<x)=P(X=0)=0.3$;

თუ $1<X\le 2$, მაშინ $F(x)=P(X<x)=P\{(X=0)\cup(X=1)\}=$

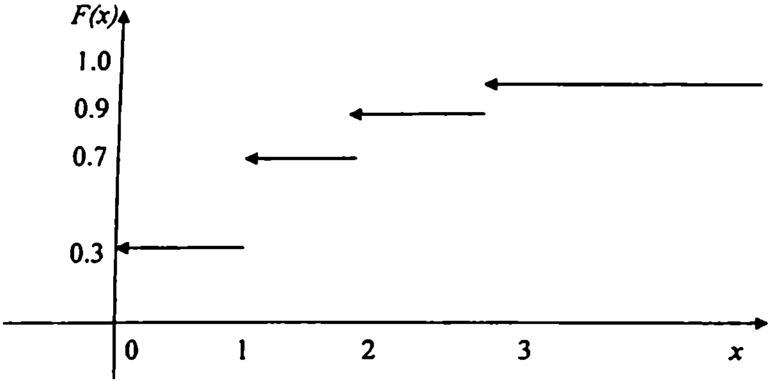
$$=P(X=0)+P(X=1)=0.3+0.4=0.7;$$

თუ $2<X\le 3$, მაშინ $F(x)=P(X<x)=P\{(X=0)\cup(X=1)\cup(X=2)\}=$

$$=P(X=0)+P(X=1)+P(X=2)=0.3+0.4+0.2=0.9;$$

და ბოლოს, თუ $x\ge 3$, მაშინ $F(x)=P(X<x)=P(\Omega)=1$.

შესაბამისად, განაწილების ფუნქციის გრაფიკს ექნება შემდეგი სახე:



უწყვეტ განაწილების ფუნქციას ნახტომები არა აქვს. ის მონოტონურად იზრდება არგუმენტის ზრდასთან ერთად 0-დან (როცა $x\to-\infty$) 1-მდე (როცა $x\to+\infty$). შემთხვევით სიდიდეს, რომელსაც აქვს უწყვეტი განაწილების ფუნქცია, უწოდებენ **უწყვეტ შემთხვევით სიდიდეს**.

თუ უწყვეტი შემთხვევითი სიდიდის განაწილების ფუნქცია $F(x)$ წარმოებადია, მაშინ მის წარმოებულს შემთხვევითი სიდიდის განაწილების სიმკვრივე ეწოდება და აღინიშნება $f(x)$-ით:

$$f(x)=\frac{dF(x)}{dx}.$$

შემთხვევითი სიდიდის განაწილების სიმკვრივიდან შეგვიძლია აღვადგინოთ განაწილების ფუნქცია:

$$F(x)=\int_{-\infty}^{x} f(y)dy.$$

ვინაიდან ნებისმიერი განაწილების ფუნქციისათვის:

$$\lim_{x\to-\infty}F(x)=0,\quad \lim_{x\to+\infty}F(x)=1,$$

ამიტომ, ნიუტონ-ლეიბნიცის ფორმულის გამოყენებით, გვაქვს:

$$\int_{-\infty}^{+\infty} f(x)dx=1.$$

მაგალითი 2. მოცემულია შემთხვევითი სიდიდის განაწილების ფუნქცია:

$$F(x)=\begin{cases} 0, & x<a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x>b. \end{cases}$$

(სადაც a და b ($a<b$) ნებისმიერი ნამდვილი რიცხვებია). ვიპოვოთ შესაბამისი განაწილების სიმკვრივე. ცხადია, რომ განმარტების თანახმად:

$$f(x)=\begin{cases} 0, & x<a, \\ \dfrac{1}{b-a}, & a<x<b, \\ 0, & x>b \end{cases}$$

რაც შეეხება $x=a$ და $x=b$ წერტილებს, აქ $F(x)$ ფუნქციას წარმოებული არა აქვს და იქ შეგვიძლია $f(x)$ განვმარტოთ ნებისმიერად, ვთქვათ, $f(a)=f(b)=0$. შემთხვევით სიდიდეს, რომელსაც აქვს აღნიშნული განაწილების სიმკვრივე, ეწოდება **თანაბარად განაწილებული $[a,b]$ მონაკვეთზე.**

შერეული ტიპის განაწილების ფუნქციები გვხვდება, როცა დაკვირვებები რომელიღაც მომენტში წყდება. მაგალითად, იმ სტატისტიკური მონაცემების ანალიზის დროს, რომლებიც მიღება ობიექტის საიმედობაზე გამოცდის (შემოწმების) ისეთი გეგმის გამოყენებისას, რომელიც გულისხმობს გამოცდის შეწყვეტას გარკვეული დროის ამოწურვის შემდეგ ან იმ ტექნიკური ნაწარმის მონაცემების ანალიზის დროს, რომლებსაც დასჭირდათ საგარანტიო შეკეთება.

მაგალითი 3. დავუშვათ, რომ ნათურის მუშაობის დრო არის შემთხვევითი სიდიდე განაწილების ფუნქციით $F(t)$, ხოლო ნათურის გამოცდა გრძელდება ნათურის მწყობრიდან გამოსვლამდე (გადაწვამდე), თუ ეს მოხდება გამოცდის დაწყებიდან არაუმეტეს 100 საათის განმავლობაში, ანუ $t_0=100$ სთ მომენტამდე. ვთქვათ, $G(t)$ — არის ამ გამოცდის დროს ნათურის ნორმალურად მუშაობის დროის განაწილების ფუნქცია. მაშინ ცხადია, რომ:

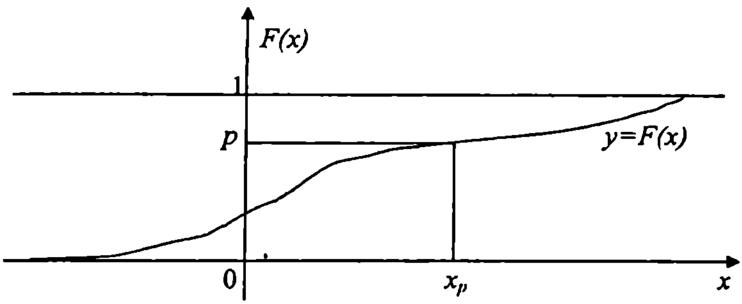
$$G(t)=\begin{cases} F(t), & t \le 100 \\ 1, & t>100. \end{cases}$$

$G(t)$ ფუნქციას აქვს ნახტომი t_0 წერტილში, ვინაიდან შესაბამისი შემთხვევითი სიდიდე ღებულობს t_0 მნიშვნელობას ალბათობით $1-F(t_0)>0$.

შემთხვევით სიდიდეთა მახასიათებლები. ალბათურ-სტატისტიკურ მეთოდებში გამოიყენება შემთხვევით სიდიდეთა სხვადასხვა მახასიათებლები, რომლებიც გამოისახება განაწილების ფუნქციითა და განაწილების სიმკვრივით.

კვანტილი. შემოსავლების დიფერენცირების აღწერისას, შემთხვევითი სიდიდის პარამეტრების შესაძლო (საიმედობის) საზღვრების დადგენისას და სხვა მრავალ შემთხვევაში გამოიყენება ე. წ. "*p* ***რიგის კვანტილ-***

ის" ცნება ($0<p<1$), რომელიც აღინიშნება x_p სიმბოლოთი. p რიგის კვანტილი ეწოდება შემთხვევითი სიდიდის იმ მნიშვნელობას, რომლისთვისაც განაწილების ფუნქცია ღებულობს მნიშვნელობას p ან ადგილი აქვს "ნახტომს" მნიშვნელობიდან, რომელიც ნაკლებია p-ზე მნიშვნელობისაკენ რომელიც მეტია p-ზე. შეიძლება მოხდეს, რომ ეს პირობა სრულდება x ყველა მნიშვნელობისათვის გარკვეული ინტერვალიდან (ე. ი. განაწილების ფუნქცია მუდმივია ამ ინტერვალზე და ტოლის p-სი), მაშინ x ყველა ასეთ მნიშვნელობას უწოდებენ p რიგის კვანტილს. უწყვეტი განაწილების ფუნქციების შემთხვევაში, როგორც წესი, არსებობს ერთადერთი p რიგის x_p კვანტილი, ამასთანავე $F(x_p)=p$.



მაგალითი 4. ვიპოვოთ p რიგის x_p კვანტილი თანაბარი განაწილების ფუნქციისათვის.

p ($0<p<1$) რიგის x_p კვანტილი უნდა ვეძებოთ როგორც $F(x)=p$ განტოლების ამონახსნი. თანაბარი განაწილების ფუნქციის შემთხვევაში ეს განტოლება მიიღებს სახეს:

$$\frac{x-a}{x-b}=p,$$

საიდანაც ცხადია, რომ:

$$x_p=a+p(b-a)=a(1-p)+bp.$$

როცა $p=0$, მაშინ ნებისმიერი $x\le a$ წარმოადგენს $p=0$ რიგის კვანტილს, ხოლო $p=1$ რიგის კვანტილი იქნება ნებისმიერი $x\ge b$ რიცხვი.

დისკრეტული განაწილების შემთხვევაში, როგორც წესი, არ არსებობს x_p, რომელიც აკმაყოფილებს $F(x_p)=p$ განტოლებას. უფრო ზუსტად, თუ დისკრეტულ შემთხვევით სიდიდეს აქვს n მნიშვნელობა $x_1<x_2<\cdots<x_n$ (ამასთანავე შესაბამისი ალბათობებია $P(X=x_i)=p_i$, $i=1,2,...,n$), მაშინ $F(x_p)=p$ განტოლებას x_p-ს მიმართ გააჩნია ამონახსნი p-ს მხოლოდ n მნიშვნელობისათვის, კერძოდ, როცა:

$$p=p_1,$$

$$p=p_1+p_2,$$

$$p = p_1 + p_2 + \cdots + p_n.$$

p-ს ჩამოთვლილი n მნიშვნელობისათვის $F(x_p) = p$ განტოლების x_p ამონახსნი არაერთადერთია, კერძოდ,

$$F(x) = p_1 + p_2 + \cdots + p_i$$

ყველა ისეთი x-სათვის, რომლისთვისაც სრულდება უტოლობა $x_i < x \le x_{i+1}$. ე. ი. x_p — ნებისმიერი რიცხვია $(x_i, x_{i+1}]$ ინტერვალიდან. ყველა დანარჩენი p-სათვის (0,1) ინტერვალიდან ადგილი აქვს "ნახტომს" p-ზე ნაკლები მნიშვნელობიდან p-ზე მეტი მნიშვნელობისაკენ. კერძოდ, თუ

$$p_1 + p_2 + \cdots + p_i < x < p_1 + p_2 + \cdots + p_i + p_{i+1},$$

მაშინ $x_p = x_{i+1}$.

მდებარეობის მახასიათებლები მიუთითებენ განაწილების "ცენტრზე". სტატისტიკაში დიდი მნიშვნელობა აქვს $p = 1/2$ რიგის კვანტილს. მას X შემთხვევითი სიდიდის ან მისი განაწილების $F(x)$ ფუნქციის **მედიანა** ეწოდება და აღინიშნება $\tilde{x}$. ე. ი. $\tilde{x} = x_{1/2}$. ისევე როგორც გეომეტრიაში სამკუთხედის მედიანა წვეროს მოპირდაპირე გვერდს ყოფს ორ ტოლ ნაწილად, მათემატიკურ სტატისტიკაში მედიანა შუაზე ყოფს შემთხვევითი სიდიდის განაწილებას: ტოლობა $F(x_{1/2}) = 1/2$ ნიშნავს, რომ ალბათობა იმისა, რომ შემთხვევითი სიდიდე მიიღებს მნიშვნელობას $x_{1/2}$-ის მარცხნივ და ალბათობა იმისა, რომ შემთხვევითი სიდიდე მიიღებს მნიშვნელობას $x_{1/2}$-ის მარჯვნივ ან თვითონ $x_{1/2}$-ის ტოლ მნიშვნელობას ერთმანეთის ტოლია და ორივე 1/2-ია, ანუ

$$P(X < 1/2) = P(X \ge 1/2) = 1/2.$$

თუ დისკრეტულ შემთხვევით სიდიდეს აქვს n მნიშვნელობა დალაგებული ზრდადობის მიხედვით $x_1 < x_2 < \cdots < x_n$ და თითოეული ეს მნიშვნელობა მიიღება ერთი და იგივე $1/n$-ის ტოლი ალბათობით, მაშინ გასაგებია, რომ როცა n კენტია მედიანა იქნება ზუსტად შუაში (ცენტრში) მდგომი მნიშვნელობა:

$$\tilde{x} = x_{(n+1)/2},$$

ხოლო როდესაც n ლუწია, მაშინ მედიანის როლში იღებენ ბოლოებიდან თანაბრად დაშორებული ორი (შუაში მდგომი ორი) მნიშვნელობის საშუალო არითმეტიკულს:

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}.$$

მედიანა განეკუთვნება შემთხვევითი სიდიდის ცენტრალური ტენდენციის მდგრადი საზომების ჯგუფს. საზოგადოდ მონაცემთა რაიმე რიცხვითი მახასიათებლის **მდგრადობა** ნიშნავს, რომ დაკვირვებათა მცირე რაოდენობის ცვლილებას შეზღუდული გავლენა აქვს მასზე, მიუხედავად იმისა, თუ როგორია ამ ცვლილების სიდიდე.

შემთხვევითი სიდიდის ერთ-ერთი რეალური შინაარსის მქონე მახასიათებელია *მოდა*. **მოდა** ეწოდება შემთხვევითი სიდიდის იმ მნიშვნელობას (ან მნიშვნელობებს), რომელიც შეესაბამება განაწილების სიმკვრივის ლოკალურ მაქსიმუმს უწყვეტი შემთხვევითი სიდიდის შემთხვევაში ან ალბათობის ლოკალურ მაქსიმუმს დისკრეტული შემთხვევითი სიდიდის შემთხვევაში.

თუ x_0 უწყვეტი ტიპის შემთხვევითი სიდიდის (რომლის განაწილების სიმკვრივეა $f(x)$) მოდაა, მაშინ გასაგებია, რომ $\frac{df(x_0)}{dx}=0$.

შემთხვევით სიდიდეს შეიძლება ჰქონდეს ბევრი მოდა. მაგალითად, თანაბარი განაწილების შემთხვევაში ნებისმიერი x წერტილი $a<x<b$ წარმოადგენს მოდას. თუმცა ეს გამონაკლისია. უმეტესობა შემთხვევით სიდიდეებს, რომლებიც გამოიყენებიან ალბათურ-სტატისტიკური მეთოდების საშუალებით გადაწყვეტილებების მიღებისას ან სხვა გამოყენებითი ხასიათის კვლევებში, გააჩნიათ ერთი მოდა. იმ შემთხვევით სიდიდეებს (შესაბამისად, იმ განაწილებებს ან განაწილებების სიმკვრივეებს), რომელთაც აქვთ ერთი მოდა – უწოდებენ **უნიმოდალურს**.

ორგანზომილებიანი შემთხვევითი სიდიდე.

ერთგანზომილებიან შემთხვევით სიდიდეებთან ერთად, რომელთა შესაძლო მნიშვნელობები განისაზღვრება ერთი რიცხვით, ალბათობის თეორიაში განიხილება მრავალგანზომილებიანი შემთხვევითი სიდიდეებიც. ასეთი შემთხვევითი სიდიდის ნებიმიერი მნიშვნელობა წარმოადგენს რამოდენიმე რიცხვის დალაგებულ ერთობლიობას. ამ ცნების გეომეტრიულ ილუსტრაციას წარმოადგენს n –განზომილებიანი სივრცის წერტილები, რომელთა თითოეული კოორდინატა წარმოადგენს შემთხვევით სიდიდეს (დისკრეტულს ან უწყვეტს), ანუ n –განზომილებიან ვექტორს. ამიტომ, მრავალგანზომილებიან შემთხვევით სიდიდეებს ასევე შემთხვევით ვექტორებსაც უწოდებენ. სიმარტივისათვის ჩვენ განვიხილავთ ორგანზომილებიან შემთხვევით ვექტორებს. ჯერ განვიხილოთ დისკრეტული შემთხვევა.

დისკრეტული ორგანზომილებიანი (X, Y) შემთხვევითი სიდიდის **განაწილების კანონს** (ანუ X და Y შემთხვევითი სიდიდეების **ერთობლივ განაწილების კანონს**) აქვს ორგანზომილებიანი ცხრილის სახე, რომელიც გვაძლევს შესაძლო მნიშვნელობების ცალკეული კომპონენტების ჩამონათვალს და იმ $p(x_i, y_j)$ ალბათობებს, რა ალბათობებითაც მიიღება მნიშვნელობა (x_i, y_j):

Y	X					
	x_1	x_2	...	x_i	...	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	...	$p(x_i, y_1)$	...	$p(x_n, y_1)$
...	...	...	...	...	...	...
y_j	$p(x_1, y_j)$	$p(x_2, y_j)$	...	$p(x_i, y_j)$	...	$p(x_n, y_j)$
...	...	...	...	...	...	...
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	...	$p(x_i, y_m)$	...	$p(x_n, y_m)$

ამასთანავე, ცხრილის ყველა უჯრაში მდგომი ალბათობების ჯამი 1-ის ტოლია. თუ ცნობილია ორგანზომილებიანი შემთხვევითი სიდიდის განაწილების კანონი, ჩვენ შეგვიძლია ვიპოვოთ მისი შემადგენელი ცალკეული შემთხვევითი სიდიდის განაწილების კანონი. მართლაც, ხდომილება $X = x_1$ წარმოადგენს ჯამს არათავსებადი $(X = x_1, Y = y_1)$, $(X = x_1, Y = y_2)$,..., $(X = x_1, Y = y_m)$ ხდომილებების, ამიტომ $p(X = x_1) = p(x_1, y_1) + p(x_1, y_2) +...+ p(x_1, y_m)$ (მარჯვენა მხარეს წერია $X = x_1$ სვეტის შესაბამისი ალბათობების ჯამი). ანალოგიურად შეგვიძლია ვიპოვოთ X –ის დანარჩენი მნიშვნელობების ალბათობები. იმისათვის, რომ განვსაზღვროთ Y –ის შესაძლო მნიშვნელობების ალბათობები, საჭიროა შეიკრიბოს $Y = y_j$ სვეტის შესაბამისი ალბათობები.

მაგალითი 1. მოცემულია ორგანზომილებიანი შემთხვევითი სიდიდის განაწილების კანონი:

Y	X		
	-2	3	6
-0.8	0.1	0.3	0.1
-0.5	0.15	0.25	0.1

ვიპოვოთ ცალკეული შემთხვევითი სიდიდის განაწილების კანონი.

ამოხსნა. ცხრილში მოყვანილი ალბათობების სვეტების მიხედვით შეკრებით მივიღებთ X–ის განაწილების მწკრივს:

X	-2	3	6
p	0.25	0.55	0.2

ალბათობების შეკრებით სტრიქონების მიხედვით, მივიღებთ Y–ის განაწილების მწკრივს:

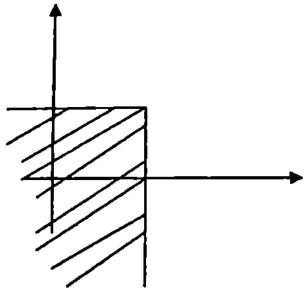
Y	-0.8	-0.5
p	0.5	0.5

გადავიდეთ უწყვეტი ტიპის შემთხვევითი სიდიდეების განხილვაზე.

განმარტება 1. ორგანზომილებიანი (X, Y) შემთხვევითი სიდიდის **განაწილების ფუნქცია** (ან X და Y შემთხვევითი სიდიდეების **ერთობლივი განაწილების ფუნქცია**) ეწოდება $(X \le x, Y \le y)$ ხდომილების ალბათობას:

$$F(x, y) = p(X \le x, Y \le y).$$

y



ეს ნიშნავს, რომ (X, Y) წერტილი მოხვდება დაშტრიხულ არეში, თუ მართი კუთხის წვერო მოთავსებულია წერტილში (x, y).

შევნიშნავთ, რომ ერთობლივი განაწილების ფუნქცია განიმარტება როგორც დისკრეტული, ისე უწყვეტი შემთხვევითი სიდიდეებისათვის. ჩამოვაყალიბოთ მისი თვისებები:

1). $0 \le F(x, y) \le 1$ (ვინაიდან $F(x, y)$ ალბათობაა).

2). $F(x, y)$ არის თითოეული არგუმენტის მიმართ არაკლებადი, მარჯვნიდან უწყვეტი ფუნქცია.

3). ადგილი აქვს ზღვრულ თანაფარდობებს: $F(-\infty, y) = 0$; $F(x, -\infty) = 0$; $F(-\infty, -\infty) = 0$; $F(\infty, \infty) = 1$.

4). $F(x, \infty) = F_1(x)$; $F(\infty, y) = F_2(y)$.

შემთხვევით სიდიდეებს ეწოდება **დამოუკიდებელი**, თუ

$$F(x, y) = F_1(x)\, F_2(y).$$

დისკრეტული შემთხვევითი სიდიდეები დამოუკიდებელია მაშინ და მხოლოდ მაშინ, როცა

$$p(x_i, y_j) = p(x_i)\, p(y_j),\ \forall i, j.$$

განმარტება 2. უწყვეტი ორგანზომილებიანი შემთხვევითი სიდიდის **ერთობლივი განაწილების სიმკვრივე (ანუ ორგანზომილებიანი სიმკვრივე)** ეწოდება ერთობლივი განაწილების ფუნქციის შერეულ მეორე რიგის კერძო წარმოებულს:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

შენიშვნა. ორგანზომილებიანი განაწილების სიმკვრივე წარმოადგენს შემთხვევითი წერტილის Δx და Δy გვერდების მქონე მართკუთხედში მოხვედრის ალბათობის ამ მართკუთხედის ფართობთან შეფარდების ზღვარს, როცა $\Delta x \to 0$, $\Delta y \to 0$.

ორგანზომილებიანი განაწილების სიმკვრივის თვისებები:

1). $f(x, y) \ge 0$ (წერტილის მართკუთხედში მოხვედრის ალბათობა არაუარყოფითია, ამ მართკუთხედის ფართობი არაუარყოფითია, და, შესაბამისად, მათი შეფარდების ზღვარი არაუარყოფითია).

2). $F(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(x, y)\,dx\,dy.$

3). $\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(x,y)dxdy = 1$ (ვინაიდან ეს არის აუცილებელი ხდომილების, კერძოდ, წერტილის Oxy სიბრტყეზე მოხვედრის, ალბათობა).

4). წერტილის სიბრტყის არეში მოხვედრის ალბათობა ტოლია:

$$p((X,Y)\subset D) = \iint\limits_{D} f(x,y)dxdy.$$

5). ორგანზომილებიანი განაწილების სიმკვრივიდან ერთ-ერთი შემთხვევითი სიდიდის განაწილების სიმკვრივის პოვნა:

$$f_1(x) = \frac{dF_1(x)}{dx} = \frac{dF(x,\infty)}{dx} = \frac{d\left(\int\limits_{-\infty}^{x}\int\limits_{-\infty}^{\infty} f(x,y)\right)}{dx} = \int\limits_{-\infty}^{\infty} f(x,y)dy,$$

ანალოგიურად, $f_2(y) = \int\limits_{-\infty}^{\infty} f(x,y)dx.$

6). უწყვეტი შემთხვევითი სიდიდეები დამოუკიდებელია მაშინ და მხოლოდ მაშინ, როცა

$$f(x,y)=f_1(x)f_2(y).$$

მაგალითი 2. მოცემულია X და Y დამოუკიდებელი შემთხვევითი სიდიდეების განაწილების კანონი:

X	-2	3	6
p	0.2	0.5	0.3

Y	-0.8	-0.5
p	0.4	0.6

ვიპოვოთ $Z=\max\{X,Y\}$ შემთხვევითი სიდიდის განაწილების კანონი.

ამოხსნა. Z შემთხვევითი სიდიდის შესაძლო მნიშვნელობებია: -0.8; -0.5; 3 და 6. გამოვთვალოთ შესაბამისი ალბათობები. გვაქვს:

$$P(Z=-0.8) = P\{(X=-2)\cap(Y=-0.8)\} = P(X=-2)\cdot P(Y=-0.8) = 0.2\cdot 0.4 = 0.08;$$

$$P(Z=-0.5) = P\{(X=-2)\cap(Y=-0.5)\} = P(X=-2)\cdot P(Y=-0.5) = 0.2\cdot 0.6 = 0.12;$$

$$P(Z=3) = P\{[(X=3)\cap(Y=-0.8)]\cup[(X=3)\cap(Y=-0.5)]\} =$$
$$= P[(X=3)\cap(Y=-0.8)] + P[(X=3)\cap(Y=-0.5)] = 0.5\cdot 0.4 + 0.5\cdot 0.6 = 0.5;$$

$$P(Z=6) = P\{[(X=6)\cap(Y=-0.8)]\cup[(X=6)\cap(Y=-0.5)]\} =$$
$$= P[(X=6)\cap(Y=-0.8)]\cdot P[(X=6)\cap(Y=-0.5)] = 0.3\cdot 0.4 + 0.3\cdot 0.6 = 0.3.$$

თავი IV
შემთხვევით სიდიდეთა რიცხვითი მახასიათებლები

შემთხვევითი სიდიდის მათემატიკური ლოდინი.

ხშირად შემთხვევითი სიდიდის დასახასიათებლად უფრო მოხერხებულია რიცხვითი მახასიათებლები, ნაცვლად ფუნქციონალურისა (როგორიცაა განაწილების კანონი, განაწილების ფუნქცია ან განაწილების სიმკვრივე უწყვეტ შემთხვევაში). შემთხვევითი სიდიდის რიცხვით მახასიათებლებს შორის პირველ რიგში გამოყოფენ ისეთებს, რომელთა "ირგვლივ" ("გარშემოც") ლაგდება (ჯგუფდება) შემთხვევითი სიდიდის შესაძლო მნიშვნელობები. ერთერთ ასეთ რიცხვით მახასიათებლს წარმოადგენს შემთხვევითი სიდიდის *მათემატიკური ლოდინი*, რომელსაც მისი არსიდან გამომდინარე (რასაც ჩვენ ქვემოთ დავინახავთ) შემთხვევითი სიდიდის *საშუალო მნიშვნელობასაც* ეძახიან.

ალბათობის თეორიის ძალიან ბევრ საკითხში მოსახერხებელია შემოვიტანოთ მათემატიკური ლოდინის ცნება. როცა მოთამაშემ უნდა მიიღოს განსაზღრული თანხა, თუ მოხდება გარკვეული შემთხვევითი ხდომილება, რომლის ალბათობა ცნობილია, მაშინ მისი მათემატიკური ლოდინი არის ის თანხა, რომელიც სამართლიანად უნდა შემოუტანოს მას იმან, ვინც იყიდის მისგან მოგების შანსებს. მაგალითად, მოთამაშემ უნდა გააგოროს ერთხელ სათამაშო კამათელი და მიიღოს მოგება 6 ლარი, თუ მოვა ციფრი 4. ადვილი დასანახია, რომ მისი მათემატიკური ლოდინი ტოლია 1 ლარის, ე. ი. იმ თანხის (6 ლარის), რომელიც შეიძლება მიიღოს მოთამაშემ, ნამრავლი სასურველი შედეგის ალბათობაზე (1/6ზე).

მართლაც, დავუშვათ, რომ ბანკომატი გვთავაზობს გავაგოროთ კამათელი და ყოველ მსურველს აძლევს შესაძლებლობას დადოს სანაძლეო მის მიერ შერჩეულ წახნაგზე (ქულაზე), რათა მოგების შემთხვევაში მიიღოს 6 ლარი. თუ 6 სხვადასხვა მოთამაშე დადებს ფსონს შესაბამისად 6 სხვადასხვა წახნაგზე, მაშინ ბანკომატმა ნებისმიერ შემთხვევაში უნდა გადაიხადოს 6 ლარი, ვინაიდან იქნება ერთი და მხოლოდ ერთი მოგებული. იმისათვის რომ თამაში იყოს სამართლიანი, საჭიროა რომ 6 მოთამაშიდან თითოეულმა შეიტანოს ბანკომატში 1 ლარი, რადგანაც არ არსებობს არანაირი საფუძველი იმისათვის, რომ რომელიმე მათგანმა გადაიხადოს სხვაზე მეტი ან ნაკლები, ვინაიდან სათამაშო კამათლის ექვსივე წახნაგი ტოლალბათურია. აქედან ჩვენ ვასკვნით, რომ მათემატიკური ლოდინი თითოეული მოთამაშისათვის შეადგენს 1 ლარს.

განმარტება 1. $X:\Omega \to R^1$ დისკრეტული შემთხვევითი სიდიდის მათემატიკური ლოდინი აღინიშნება EX სიმბოლოთი (E არის პირველი ასო ინგლისური სიტყვისა *Expectation*, რომელიც ნიშნავს – ლოდინი, მოსალოდნელობა) და ეწოდება რიცხვს:

$$EX = \sum_{\omega \in \Omega} X(\omega)P(\omega), \qquad (1)$$

ე. ი. შემთხვევითი სიდიდის მათემატიკური ლოდინი წარმოადგენს შემთხვევითი სიდიდის მნიშვნელობების შეწონილ ჯამს წონებით, რომლებიც ტოლია შესაბამისი ელემენტარული ხდომილებების ალბათობების.

შევნიშნავთ, რომ მათემატიკური ლოდინის აღსანიშნავად ასევე გამოიყენება სიმბოლო MX (M არის პირველი ასო რუსული სიტყვისა Математическое ожидание).

მაგალითი 1. გამოვთვალოთ სათამაშო კამათელზე მოსული ქულათა რიცხვის მათემატიკური ლოდინი. (1) თანაფარდობიდან გამომდინარე გვაქვს:

$$EX = 1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5.$$

თეორემა1. თუ შემთხვევითი სიდიდე ღებულობს მნიშვნელობებს $x_1, x_2, \ldots, x_n$, მაშინ სამართლიანია თანაფარდობა:

$$EX = \sum_{i=1}^{n} x_i P\{X = x_i\}, \qquad (2)$$

ე. ი. შემთხვევითი სიდიდის მათემატიკური ლოდინი წარმოადგენს შემთხვევითი სიდიდის მნიშვნელობების შეწონილ ჯამს წონებით, რომლებიც ტოლია ალბათობის იმისა, რომ შემთხვევითი სიდიდე ღებულობს გარკვეულ მნიშვნელობებს.

თუ შემოვიღებთ აღნიშვნებს $P\{X = x_i\} = p_i$, $i = 1, 2, \ldots, m$ მაშინ (2) თანაფარდობა ასე გადაიწერება

$$EX = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n. \qquad (3)$$

განსხვავებით (1) თანაფარდობისაგან, სადაც აჯამვა ხდება უშუალოდ ელემენტარული ხდომილებების მიმართ, ხდომილება

$$\{X = x_i\} = \{\omega : X(\omega) = x_i\}$$

შეიძლება შედგებოდეს რამოდენიმე ელემენტარული ხდომილებისაგან. ხშირ შემთხვევაში (2) თანაფარდობით განიმარტება მათემატიკური ლოდინი, თუმცა მათემატიკური ლოდინის თვისებების შესამოწმებლად უფრო მოხერხებულია (1) თანაფარდობა.

თეორემა 1-ის დამტკიცება. დავაჯგუფოთ (1) თანაფარდობაში შემთხვევითი სიდიდის ერთი და იგივე მნიშვნელობიანი წევრები:

$$EX = \sum_{i=1}^{n}\Big(\sum_{\omega: X(\omega)=x_i} X(\omega)P(\omega)\Big).$$

ვინაიდან მუდმივი გადის ჯამის ნიშნის გარეთ, ამიტომ

$$\sum_{\omega: X(\omega)=x_i} X(\omega)P(\omega) = \sum_{\omega: X(\omega)=x_i} x_i P(\omega) = x_i \sum_{\omega: X(\omega)=x_i} P(\omega).$$

მეორეს მხრივ, ალბათობის განმარტების თანახმად:

$$\sum_{\omega: X(\omega)=x_i} P(\omega) = P(X = x_i).$$

ორი უკანასკნელი თანაფარდობის გაერთიანება გვაძლევს:

$$EX = \sum_{i=1}^{n}\Big(x_i \sum_{\omega: X(\omega)=x_i} P(\omega)\Big) = \sum_{i=1}^{n} x_i P\{X = x_i\}.$$

მათემატიკური ლოდინის ცნება ალბათურ-სტატისტიკურ თეორიაში შეესაბამება სიმძიმის ცენტრის ცნებას მექანიკაში. რიცხვითი ღერძის $x_1, x_2, \ldots, x_n$ წერტილებში განვათავსოთ შესაბამისად $P\{X = x_1\}$, $P\{X = x_2\}, \ldots, P\{X = x_n\}$ მასები. მაშინ (2) თანაფარდობა გვიჩვენებს, რომ მატერიალური წერტილების ამ სისტემის სიმძიმის ცენტრი ემთხვევა მათემატიკურ ლოდინს. ეს, თავის მხრივ, გვიჩვენებს განმარტება 1ის ბუნებრიობას.

იმისათვის, რომ გასაგები გახდეს მათემატიკური ლოდინის შინაარსი, დავუშვათ, რომ ჩავატარეთ n დაკვირვება (ექსპერიმენტი) X შემ-

თხევევით სიდიდეზე და ეთქვათ, რომ მან n_1 ჯერ მიიღო მნიშვნელობა x_1, n_2 ჯერ – მნიშვნელობა x_2, და ა. შ. n_m ჯერ – მნიშვნელობა x_m. ცხადია $n_1+n_2+\cdots+n_m = =n$, ხოლო შემთხვევითი სიდიდის მიერ მიღებული მნიშვნელობების საშუალო არითმეტიკული $\bar{x}$ გამოითვლება ფორმულით

$$\bar{x}=\frac{x_1n_1+x_2n_2+\cdots+x_mn_m}{n},$$

ანუ,

$$\bar{x}=x_1\frac{n_1}{n}+x_2\frac{n_2}{n}+\cdots+x_m\frac{n_m}{n}. \qquad (4)$$

აქ $\frac{n_1}{n}$ არის x_1-ის განხორციელების ფარდობითი სიხშირე, $\frac{n_2}{n}$ არის x_2-ის განხორციელების ფარდობითი სიხშირე და ა. შ. $\frac{n_m}{n}$ არის x_m-ის განხორციელების ფარდობითი სიხშირე. თუ დავუშვებთ, რომ დაკვირვებათა რაოდენობა საკმარისად დიდია, მაშინ ფარდობითი სიხშირე ახლოსაა ხდომილების ალბათობასთან

$$\frac{n_1}{n}=p_1, \frac{n_2}{n}=p_2, \ldots, \frac{n_m}{n}=p_m.$$

თუ ახლა (4) თანაფარდობაში ფარდობით სიხშირეებს შევცვლით შესაბამისი ალბათობებით და გავითვალისწინებთ (3) თანაფარდობას, მივიღებთ, რომ

$$\bar{x}=x_1p_1+x_2p_2+\cdots+x_np_n=EX.$$

ე. ი. შემთხვევითი სიდიდის მათემატიკური ლოდინი დაახლოებით ტოლია ამ შემთხვევითი სიდიდის დაკვირვებული მნიშვნელობების საშუალო არითმეტიკულის.

ცხადია, რომ არაა აუცილებელი შემთხვევითი სიდიდის მათემატიკური ლოდინი ტოლი იყოს მისი რომელიმე შესაძლო მნიშვნელობის.

თუ დისკრეტული ტიპის შემთხვევითი სიდიდე თანაბარი კანონითაა განაწილებული ანუ ის ყველა თავის მნიშვნელობას $x_1, x_2, \ldots, x_n$ ღებულობს თანაბარი (ერთი და იგივე) ალბათობებით ($p_1=p_2=\cdots=p_n= =1/n$), მაშინ მათემატიკური ლოდინი ზუსტად ემთხვევა მისი მნიშვნელობების საშუალო არითმეტიკულს:

$$EX=x_1\frac{1}{n}+x_2\frac{1}{n}+\cdots+x_n\frac{1}{n}=\frac{x_1+x_2+\cdots+x_n}{n}.$$

განმარტება 2. თუ დისკრეტული ტიპის შემთხვევითი სიდიდის შესაძლო მნიშვნელობათა სიმრავლე თვლადია, მაშინ

$$EX=x_1p_1+x_2p_2+\cdots+x_np_n+\cdots=\sum_{i=1}^{\infty}x_ip_i,$$

თუ ცნობილია, რომ შესაბამისი მწკრივი აბსოლუტურად კრებადია –

$$\sum_{i=1}^{\infty}|x_i|\,p_i<\infty,$$

სადაც $p_i:=P\{X=x_i\}, i=1,2,\ldots$ და $p_1+p_2+\cdots=1$.

განმარტება 3. უწყვეტი ტიპის X შემთხვევითი სიდიდის მათემატიკური ლოდინი ეწოდება რიცხვს

$$EX = \int_{-\infty}^{+\infty} xf(x)dx,$$

სადაც $f(x)$ არის X შემთხვევითი სიდიდის განაწილების სიმკვრივე, თუ ცნობილია, რომ

$$\int_{-\infty}^{+\infty} |x| f(x)dx < \infty.$$

თეორემა 2. თუ X და Y ერთი და იგივე ელემენტარულ ხდომილებათა სივრცეზე განმარტებული შემთხვევითი სიდიდებია, ხოლო $c = const$ რაიმე მუდმივია, მაშინ:

ა). $Ec = c$;

ბ). $E(X+Y) = EX + EY$ და $E(X-Y) = EX - EY$;

გ). $E(cX) = cEX$; დ). $E(X - EX) = 0$ და

ე). $E(X-c)^2 = E(X-EX)^2 + (c - EX)^2$.

დამტკიცება. ა). ამ შემთხვევაში საქმე გვაქვს მუდმივ შემთხვევით სიდიდესთან $X(\omega) = c$, ანუ ფუნქცია $X(\omega)$ ასახავს ელემენტარულ ხდომილებათა სივრცეს Ω ერთადერთ c წერტილში. ვინაიდან მუდმივი მამრავლი შეგვიძლია გამოვიტანოთ ჯამის ნიშნის გარეთ, ამიტომ გვაქვს:

$$Ec = \sum_{\omega \in \Omega} cP(\omega) = c\sum_{\omega \in \Omega} P(\omega) = cP(\Omega) = c.$$

ბ). თუ ჯამის (შესაბამისად, სხვაობის) ყველა წევრი წარმოიდგინება ორ შესაკრებად (შესაბამისად, სხვაობად), მაშინ მთელი ჯამიც წარმოიდგინება ორი ჯამის ჯამად (შესაბამისად, სხვაობად), რომელთაგან პირველი შედგება თითოეული წევრის პირველი შესაკრებებისაგან (შესაბამისად, საკლებებისაგან), ხოლო მეორე – მეორე შესაკრებებისაგან (შესაბამისად, მაკლებებისაგან). ამიტომ

$$E(X+Y) = \sum_{\omega \in \Omega} [X(\omega) + Y(\omega)]P(\omega) = \sum_{\omega \in \Omega} X(\omega)P(\omega) + \sum_{\omega \in \Omega} Y(\omega)P(\omega) = EX + EY,$$

და ანალოგიურად, $E(X-Y) = EX - EY$.

ე. ი. შემთხვევით სიდიდეთა ჯამის (შესაბამისად, სხვაობის) მათემატიკური ლოდინი მათი მათემატიკური ლოდინების ჯამის (შესაბამისად, სხვაობის) ტოლია.

გ). $E(cX) = \sum_{\omega \in \Omega} cX(\omega)P(\omega) = c\sum_{\omega \in \Omega} X(\omega)P(\omega) = cEX$.

დ). ბ) და ა) პუნქტების თანახმად გვაქვს

$$E(X - EX) \overset{ბ)}{=} EX - E(EX) \overset{ა)}{=} EX - EX = 0.$$

ე). ვინაიდან

$$(X-c)^2 = [(X - EX) + (EX - c)]^2 =$$
$$= (X-EX)^2 + 2(X-EX)(EX-c) + (EX-c)^2.$$

ამიტომ

$$E(X-c)^2 \overset{ბ)}{=} E(X-EX)^2 + E[2(X-EX)(EX-c)] + E(EX-c)^2 \overset{ა,გ)}{=}$$

$$\overset{გ,ა)}{=} E(X-EX)^2 + 2(EX-c)E(X-EX) + (EX-c)^2 \overset{დ)}{=} E(X-EX)^2 + (EX-c)^2.$$

შედეგი 1. ბ) და გ) პუნქტების გაერთიანება გვაძლევს, რომ შემთხვევით სიდიდეთა წრფივი კომბინაციის მათემატიკური ლოდინი ტოლია მათი მათემატიკური ლოდინების წრფივი კომბინაციის:

$$E(aX+bY)=aEX+bEY\,,$$

სადაც a,b მუდმივებია.

დავალება. დაამტკიცეთ შედეგი 1 პირდაპირი გზით, მათემატიკური ლოდინის განმარტების გამოყენებით.

შედეგი 2. ვინაიდან ე) პუნქტის თანაფარდობის მარჯვენა მხარეში მეორე შესაკრები ყოველთვის არაუარყოფითია და ნულია მხოლოდ მაშინ, როცა $c=EX$, ამიტომ გამოსახულება $E(X-c)^2$ თავის მინიმუმს c-ს მიმართ აღწევს როცა $c=EX$:

$$\min_{c\in(-\infty,+\infty)} E(X-c)^2=E(X-EX)^2\,.$$

შემთხვევითი სიდიდის ფუნქციის მათემატიკური ლოდინი. ხშირად მოცემულია რაიმე X შემთხვევითი სიდიდის განაწილების კანონი და გვაინტერესებს $Y=g(X)$ შემთხვევითი სიდიდის მათემატიკური ლოდინი, სადაც $g(x)$ ნამდვილი x ცვლადის რაიმე ფუნქციაა. ამისათვის ჯერ შეიძლება დავადგინოთ Y შემთხვევითი სიდიდის განაწილება და შემდეგ გამოვთვალოთ მისი მათემატიკური ლოდინი განმარტება 1-ის გამოყენებით. მაგრამ უფრო მოხერხებულია $Y=g(X)$ შემთხვევითი სიდიდის მათემატიკური ლოდინი გამოვთვალოთ X შემთხვევითი სიდიდის განაწილების ტერმინებში.

ვთქვათ, X შემთხვევითი სიდიდის განაწილების კანონია:

x_i	x_1	x_2	$\cdots$	x_n	
p_i	p_1	p_2	$\cdots$	p_n	

მაშინ

$$Eg(X)=\sum_{i=1}^{m} g(x_i)P\{X=x_i\}=\sum_{i=1}^{m} g(x_i)p_i\,. \tag{5}$$

ამ ფაქტის შესამოწმებლად ვისარგებლოთ მათემატიკური ლოდინის განმარტებით და დავაჯგუფოთ ის წევრები, რომლებიც შეესაბამებიან $X(\omega)$-ს ერთი და იგივე მნიშვნელობას:

$$Eg(X)=\sum_{\omega\in\Omega} g(X(\omega))P(\omega)=\sum_{i=1}^{m}\Big(\sum_{\omega:X(\omega)=x_i} g(X(\omega))P(\omega)\Big).$$

თუ კი აქ მუდმივ თანამამრავლს გავიტანთ ჯამის ნიშნის გარეთ და ვისარგებლებთ ალბათობის განმარტებით, მიგიღებთ შესამოწმებელ თანაფარდობას:

$$Eg(X)=\sum_{i=1}^{m}\Big(\sum_{\omega:X(\omega)=x_i} g(x_i)P(\omega)\Big)=\sum_{i=1}^{m}\Big(g(x_i)\sum_{\omega:X(\omega)=x_i} P(\omega)\Big)=\sum_{i=1}^{m} g(x_i)P\{X=x_i\}.$$

მაგალითი 2. დავუშვათ, $g(x)=x^3-4x$ და მოცემულია X შემთხვევითი სიდიდის განაწილების კანონი:

2	1	0	2
0.1	0.3	0.4	0.2

დავადგინოთ $Y=g(X)$ შემთხვევითი სიდიდის განაწილების კანონი და გამოვთვალოთ მისი მათემატიკური ლოდინი.

ცხადია, რომ $g(-2)=g(0)=g(2)=0$ და $g(-1)=3$. ამიტომ Y შემთხვევითი სიდიდის შესაძლო მნიშვნელობებია 0 და 3. დავადგინოთ მისი განაწილების კანონი. ამისათვის გამოვთვალოთ ალბათობები: $P\{Y=0\}$

და $P\{Y=3\}$. რადგან ხდომილებები $\{X=-2\},\{X=0\}$ და $\{X=2\}$ უთავსებადია, ამიტომ ალბათობათა შეკრების წესის თანახმად გვექნება:

$$P\{Y=0\}=P\{\{X=-2\}\cup\{X=0\}\cup\{X=2\}\}=$$

$$=P\{X=-2\}+P\{X=0\}+P\{X=2\}=0.1+0.4+0.2=0.7.$$

გარდა ამისა, $P\{Y=3\}=P\{X=-1\}=0.3$. ამიტომ $Y=g(X)$ შემთხვევითი სიდიდის განაწილების კანონს აქვს სახე:

0	3
0.7	0.3

ხოლო (3) თანაფარდობის ძალით $EY=0\cdot 0.7+3\cdot 0.3=0.9$.

ახლა გამოვთვალოთ $Y=g(X)$ შემთხვევითი სიდიდის მათემატიკური ლოდინი (5) თანაფარდობის საშუალებით. გვექნება:

$$EY=g(-2)\cdot 0.1+g(-1)\cdot 0.3+g(0)\cdot 0.4+g(2)\cdot 0.2=$$

$$=0\cdot 0.1+3\cdot 0.3+0\cdot 0.4+0\cdot 0.2=0.9.$$

მაგალითი 3. გამოვთვალოთ პუასონის განაწილების მათემატიკური ლოდინი (გავიხსენოთ, რომ პუასონის განაწილებას პარამეტრით λ აქვს სახე: $P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}$, $k=0,1,2,\ldots$).

თუ ვისარგებლებთ წარმოდგენით $e^x=\sum_{m=0}^{\infty}\frac{x^m}{m!}$, მაშინ გვექნება:

$$EX=\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=1}^{\infty}k\frac{\lambda^k}{k!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda.$$

ე. ი. პუასონის განაწილების λ პარამეტრი წარმოადგენს ამ განაწილების მათემატიკურ ლოდინს (საშუალო მნიშვნელობას).

ჩვენ ზემოთ ვნახეთ როგორაა დამოკიდებული მათემატიკური ლოდინი ათვლის წერტილის შეცვლაზე და სხვა ზომის ერთეულზე გადასვლაზე (გადასვლა $Y=ax+b$), ასევე შემთხვევითი სიდიდის ფუნქციაზე. მიღებული შედეგები მუდმივად გამოიყენება ტექნიკურ-ეკონომიკურ ანალიზში, ორგანიზაციების საფინანსო-სამეურნეო მოქმედებების შეფასებისას, საგარეო-ეკონომიკურ გათვლებში ერთი ვალუტიდან მეორეზე გადასვლისას, ნორმატიულ-ტექნიკურ დოკუმენტაციაში და ა. შ. განხილული შედეგები საშუალებას იძლევა გამოყენებულ იქნეს ერთი და იგივე გამოსათვლელი ფორმულები მასშტაბების სხვადასხვა პარამეტრების და გადახრების დროს.

განმარტება 4. თუ X შემთხვევითი სიდიდის შესაძლო მნიშვნელობებია $x_1,x_2,\ \ldots,x_n$, ხოლო B რაიმე ხდომილებაა ($P(B)>0$), მაშინ X შემთხვევითი სიდიდის პირობითი მათემატიკური ლოდინი B ხდომილების მიმართ აღინიშნება სიმბოლოთი $E(X\mid B)$ და განიმარტება შემდეგნაირად:

$$E(X\mid B)=\sum_{i=1}^{n}x_iP(X=x_i\mid B). \qquad (6)$$

ცხადია, რომ $E(X\mid\Omega)=EX$ გარდა ამისა, ადვილი შესამოწმებელია, რომ:

$$E(X\mid B)=\frac{1}{P(B)}E(I_BX).$$

განმარტება 5. X შემთხვეეითი სიდიდის Y შემთხვევით სიდიდეზე **რეგრესიის მრუდი (ფუნქცია)** ეწოდება ფუნქციას $R(y) = E(X \mid Y = y)$.

რეგრესიის ფუნქციის აზრი იმაში მდგომარეობს, რომ Y შემთხვევითი სიდიდის მნიშვნელობების პირობით განისაზღვრება X შემთხვევითი სიდიდის პირობითი საშუალო მნიშვნელობა.

განვიხილოთ შემთხვევითი სიდიდე $R(Y)$ (რეგრესიის ფუნქციაში ჩასმულია Y შემთხვევითი სიდიდე). ამ შემთხვევით სიდიდეს X შემთხვევითი სიდიდის **პირობით მათემატიკურ ლოდინს** უწოდებენ Y პირობით და $E(X \mid Y)$ სიმბოლოთი აღნიშნავენ.

შემთხვევით სიდიდეთა დამოუკიდებლობა.

ისევე როგორც ხდომილებათა დამოუკიდებლობა, შემთხვევით სიდიდეთა დამოუკიდებლობა წარმოადგენს ალბათობის თეორიის ერთ-ერთ საბაზო ცნებას, რომელიც საფუძვლად უდევს გადაწყვეტილებების მიღების პრაქტიკულად ყველა ალბათურ-სტატისტიკურ მეთოდს.

განმარტება 1. ერთი და იგივე ელემენტარულ ხდომილებათა სივრცეზე განმარტებულ დისკრეტულ X და Y შემთხვევით სიდიდეებს ეწოდება **დამოუკიდებელი**, თუ ნებისმიერ ნამდვილი a და b რიცხვებისათვის დამოუკიდებელია ხდომილებები $\{\omega : X(\omega) = a\}$ და $\{\omega : Y(\omega) = b\}$.

ცხადია, რომ თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, ხოლო a, b რაიმე რიცხვებია, მაშინ შემთხვევითი სიდიდეები $X + a$ და $Y + b$ აგრეთვე დამოუკიდებლებია.

მართლაც, ხდომილებები $\{X + a = c\}$ და $\{Y + b = d\}$ ემთხვევა შესაბამისად ხდომილებებს $\{X = c - a\}$ და $\{Y = d - b\}$, ამიტომ ისინი დამოუკიდებლებია.

დავალება. აჩვენეთ, რომ თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, ხოლო a_1, b_1, a_2, b_2 რაიმე რიცხვებია, მაშინ შემთხვევითი სიდიდეები $a_1 X + b_1$ და $a_2 Y + b_2$ აგრეთვე დამოუკიდებლებია.

განმარტება 2. ერთი და იგივე ელემენტარულ ხდომილებათა სივრცეზე განმარტებულ დისკრეტულ $X, Y, Z, \ldots$ შემთხვევითი სიდიდეებს ეწოდება **ერთობლივად დამოუკიდებელი**, თუ ერთობლივად დამოუკიდებელია ხდომილებები $\{X = a\}, \{Y = b\}, \{Z = c\}, \ldots$.

მაგალითი 1. შემთხვევითი სიდიდეები, რომლებიც განიმარტებიან დამოუკიდებელ ცდათა სქემაში სხვადასხვა ცდის შედეგების მიხედვით, თვითონაც დამოუკიდებლებია. ეს გამოდის იქიდან, რომ ხდომილებები, რომელთა საშუალებითაც განიმარტება შემთხვევით სიდიდეთა დამოუკიდებლობა, განისაზღვრებიან სხვადასხვა ცდების შედეგების მიხედვით, და მაშასადამე, ისინი დამოუკიდებლები არიან თვითონ დამოუკიდებელ ცდათა სქემის განმარტების თანახმად.

გადაწყვეტილებების მიღების ალბათურ-სტატისტიკურ მეთოდებში მუდმივად გამოიყენება შემდეგი ფაქტი: თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, ხოლო $g(X)$ და $h(Y)$ შემთხვევითი სიდიდეებია, რომლებიც მიიღებიან X და Y შემთხვევით სიდიდების ჩასმით ნამდვილი ცვლადის g და h ფუნქციებში, მაშინ $g(X)$ და $h(Y)$ აგრეთვე დამოუკიდებელი შემთხვევითი სიდიდეებია. მაგალითად, თუ X და Y დამოუკიდებელია, მაშინ დამოუკიდებელია X^2 და $2Y + 3$, $\ln|X|$ და 3^Y, და ა. შ.

როგორც უკვე აღვნიშნეთ ალბათურ-სტატისტიკური მეთოდების უმრავლესობა, რომლებიც გამოიყენება პრაქტიკაში, დაფუძნებულია შემთხვევითი სიდიდეების დამოუკიდებლობის ცნებაზე, რამდენადაც, დაკვირვებების, გაზომვების, ექსპერიმენტების, ანალიზებისა და ცდების შედეგები ჩვეულებრივ მოდელირდება დამოუკიდებელი შემთხვევითი სიდიდეებით. ხშირად ითვლება, რომ დაკვირვებები ხორციელდება დამოუკიდებელი ცდების სქემის მიხედვით. მაგალითად, ორგანიზაციების საფინანსო-სამეურნეო ქმედებების შედეგები, მუშახელის გამომუშავება, შესამოწმებელი ნაწარმის (რომელიც ამორჩეულია ტექნოლოგიური პროცესის სტატისტიკური რეგულირების დროს) საკონტროლო პარამეტრების გაზომვის შედეგები (მონაცემები), მარკეტინგული გამოკითხვების დროს გამოკითხული მომხმარებლების პასუხები და სხვა ტიპის მონაცემები, რომლებიც გამოიყენება გადაწყვეტილებების მიღების დროს, ჩვეულებრივ განიხილება როგორც დამოუკიდებელი შემთხვევითი სიდიდეები (ვექტორები ან ელემენტები). შემთხვევით სიდიდეთა ცნების ასეთი პოპულარობის მიზეზი მდგომარეობს იმაში, რომ ამ მომენტისათვის კვლევის შესაბამისი თეორია გაცილებით წინაა წასული დამოუკიდებელი შემთხვევითი სიდიდეებისათვის, ვიდრე დამოკიდებულებისათვის.

თეორემა 1. თუ X და Y დამოუკიდებელი შემთხვევით სიდიდებია, მაშინ მათი ნამრავლის მათემატიკური ლოდინი თითოეულის მათემატიკური ლოდინების ნამრავლის ტოლია, ე. ი.

$$E(XY)=EX\cdot EY$$

დამტკიცება. დავუშვათ, რომ X შემთხვევითი სიდიდის მნიშვნელობებია $x_1,x_2,\ldots,x_n$, ხოლო Y შემთხვევითი სიდიდე კი ღებულობს მნიშვნელობებს y_1, $y_2,\ldots,y_n$. XY ნამრავლის მათემატიკური ლოდინის მოცემ ჯამში დავაჯგუფოთ ის წევრები, რომლებშიც X და Y ღებულობენ ფიქსირებულ მნიშვნელობებს, მუდმივი მამრავლები გავიტანოთ ჯამის ნიშნის გარეთ და გავიხსენოთ ალბათობის განმარტება. მივიღებთ:

$$E(XY)=\sum_{\omega\in\Omega}X(\omega)Y(\omega)P(\omega)=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\omega:X(\omega)=x_i,Y(\omega)=y_j}X(\omega)Y(\omega)P(\omega)=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\omega:X(\omega)=x_i,Y(\omega)=y_j}x_iy_jP(\omega)=\sum_{i=1}^{n}\sum_{j=1}^{m}x_iy_j\sum_{\omega:X(\omega)=x_i,Y(\omega)=y_j}P(\omega)=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}x_iy_jP\{X=x_i,Y=y_j\}.$$

ვინაიდან X და Y დამოუკიდებელი შემთხვევით სიდიდებია, ამიტომ

$$P\{X=x_i,Y=y_j\}=P\{X=x_i\}P\{Y=y_j\}.$$

მეორეს მხრივ, თუ ვისარგებლებთ ჯამის სიმბოლოს შემდეგი თვისებით:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_j=\sum_{i=1}^{n}a_i\cdot\sum_{j=1}^{m}b_j\,,$$

საბოლოოდ, თეორემა 19.1ის ძალით, მივიღებთ, რომ

$$E(XY)=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_iP\{X=x_i\})(y_jP\{Y=y_j\})=$$

$$= \sum_{i=1}^{n} x_i P\{X = x_i\} \cdot \sum_{j=1}^{m} y_j P\{Y = y_j\} = EX \cdot EY \quad \blacksquare$$

აღსანიშნავია, რომ თეორემა I-ის შებრუნებული თეორემა არაა მართებული. მოვიყვანოთ შესაბამისი მაგალითი.

მაგალითი 2. დავუშვათ, რომ ელემენტარულ ხდომილებათა სივრცე შედგება სამი ტოლალბათური ელემენტარული ხდომილებისაგან $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$. განვმარტოთ X და Y შემთხვევითი სიდიდეები შემდეგნაირად:

$$X(\omega_1) = 1, X(\omega_2) = 0, X(\omega_3) = -1;$$

$$Y(\omega_1) = 1, Y(\omega_2) = 0, Y(\omega_3) = 1.$$

მაშინ გასაგებია, რომ $XY = X$, $E(XY) = EX = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + (-1) \cdot \frac{1}{3} = 0$.

შესაბამისად, $E(XY) = EX \cdot EY$ მეორეს მხრივ,

$$P\{X = 0\} = P\{Y = 0\} = P\{X = 0, Y = 0\} = P(\omega_2) = 1/3,$$

მაშინ როდესაც X და Y შემთხვევითი სიდიდეები რომ იყვნენ დამოუკიდებლები $\{X = 0, Y = 0\}$ ხდომილების ალბათობა უნდა ყოფილიყო $\frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$.

შემთხვევითი სიდიდის დისპერსია.

მათემატიკური ლოდინი გვიჩვენებს თუ რომელი წერტილის (მნიშვნელობის) ირგვლივ ჯგუფდება (ლაგდება) შემთხვევითი სიდიდის მნიშვნელობები. ხშირ შემთხვევაში საჭიროა შევაფასოთ შემთხვევითი სიდიდის მნიშვნელობების ცვლილების გაზომვა მათემატიკური ლოდინის მიმართ. განვიხილოთ ორი დისკრეტული ტიპის შემთხვევითი სიდიდე განაწილების შემდეგი კანონებით:

X	-3	1
P	1/4	3/4

Y	-90	45
P	1/3	2/3

გამოვთვალოთ თითოეულის მათემატიკური ლოდინი:

$$EX = (-3) \cdot 1/4 + 1 \cdot 3/4 = 0 \text{ და } EY = (-90) \cdot 1/3 + 45 \cdot 2/3 = 0.$$

როგორც ვხედავთ ორივე შემთხვევით სიდიდეს აქვს ერთი და იგივე მათემატიკური ლოდინი, მაგრამ მათი განაწილებები განსხვავდებიან იმით, რომ X შემთხვევითი სიდიდის შესაძლო მნიშვნელობები გაცილებით ახ-ლოსაა მათემატიკურ ლოდინთან (ამ შემთხვევაში ნულთან), ვიდრე Y შემ-თხვევითი სიდიდის მნიშვნელობები.

შემთხვევითი სიდიდის მნიშვნელობების მათემატიკური ლოდინის მიმართ გაფანტულობის ერთ-ერთ მნიშვნელოვან საზომს წარმოადგენს შემთხვევითი სიდიდის დისპერსია. ჩვენ უკვე ვნახეთ, რომ გამოსახულება $E(X-c)^2$ აღწევს მინიმუმს c-ს მიმართ როცა $c = EX$. ამიტომ შემთხვევითი სიდიდის მნიშვნელობების გაფანტულობის საზომად ბუნებრივია ავიღოთ $E(X - EX)^2$.

განმარტება 1. X შემთხვევითი სიდიდის დისპერსია (აღინიშნება DX-ით, D არის პირველი ასო ინგლისური სიტყვისა – Dispersion) ეწოდება $(X - EX)^2$ შემთხვევითი სიდიდის მათემატიკურ ლოდინს

$$DX = E(X - EX)^2 . \qquad (1)$$

მათემატიკური ლოდინის თვისებების გამოყენებით დისპერსია შესაძლებელია გადაიწეროს სხვა ფორმით:

$$DX = E(X - EX)^2 = E[X^2 - 2X \cdot EX + (EX)^2] =$$
$$= EX^2 - 2EX \cdot EX + (EX)^2 = EX^2 - (EX)^2 , \qquad (2)$$

სადაც, EX^2-ს ეწოდება X შემთხვევითი სიდიდის **მეორე რიგის მომენტი** (თვითონ დისპერსიას უწოდებენ აგრეთვე – **მეორე რიგის ცენტრალურ მომენტს**).

თუ დისკრეტული ტიპის X შემთხვევითი სიდიდის განაწილების კანონია

x_i	x_1	x_2	...	x_n	
p_i	p_1	p_2	...	p_n	

მაშინ მათემატიკური ლოდინისა და შემთხვევითი სიდიდის ფუნქციებიდან მათემატიკური ლოდინის გამოსათვლელი ფორმულების თანახმად დისპერსიის გამოსათვლელ ფორმულებს (1) და (2) ფორმულების მიხედვით ექნება შესაბამისად შემდეგი სახე:

$$DX = \sum_{i=1}^{n} (x_i - \sum_{j=1}^{n} x_j p_j)^2 p_i , \qquad (3)$$

$$DX = \sum_{i=1}^{n} x_i^2 p_i - (\sum_{j=1}^{n} x_j p_j)^2 \qquad (4)$$

მაგალითი 1. დისკრეტული ტიპის X შემთხვევითი სიდიდის განაწი-ლების კანონია

x_i	-1	0	1	2	3
p_i	0.1	0.15	0.3	0.25	0.2

გამოვთვალოთ მისი დისპერსია.

ვინაიდან დისპერსიის გამოსათვლელად გვაქვს ორი (3) და (4) ფორმულები, შესაბამისად, გვექნება დისპერსიის გამოთვლის ორი ხერხი. მოხერხებულია ეს გამოთვლები ჩაიწეროს ცხრილების სახით.

დისპერსიის გამოთვლის პირველი ხერხი:

i	x_i	p_i	$x_i p_i$	$(x_i - EX)^2$	$(x_i - EX)^2 p_i$
1	-1	0.10	-0.1	5.29	0.5290
2	0	0.15	0	1.69	0.2535
3	1	0.30	0.3	0.09	0.0270
4	2	0.25	0.5	0.49	0.1225
5	3	0.20	0.6	2.89	0.5780
			$EX = \sum_{i=1}^{5} x_i p_i = 1.3$		$DX = \sum_{i=1}^{5} (x_i - EX)^2 p_i = 1.51$

დისპერსიის გამოთვლის მეორე ხერხი:

i	x_i	p_i	$x_i p_i$	x_i^2	$x_i^2 p_i$
1	-1	0.10	-0.1	1	0.1
2	0	0.15	0	0	0
3	1	0.30	0.3	1	0.3
4	2	0.25	0.5	4	1
5	3	0.20	0.6	9	1.8
			$EX=\sum_{i=1}^{5} x_i p_i = 1.3$ $DX = EX^2-(EX)^2 = 1.51$		$EX^2=\sum_{i=1}^{5} x_i^2 p_i = 3.2$

დავადგინოთ დისპერსიის თვისებები, რომლებიც მუდმივად გამოიყენება გადაწყვეტილებების მიღების ალბათურ-სტატისტიკურ მეთოდებში.

I. მუდმივის დისპერსია ნულის ტოლია – $Dc=0$. მართლაც,

$$Dc=E(c-Ec)^2=E(c-c)^2=E0=0;$$

II. $D(aX+b)=a^2DX$ მართლაც, მათემატიკური ლოდინის ცნობილი თვისებების გამოყენებით გვაქვს:

$$D(aX+b)=E[(aX+b)-E(aX+b)]^2=E[aX+b-aEX-b)]^2=E[a(X-EX)]^2=$$
$$=E[a^2(X-EX)^2]=a^2E(X-EX)^2=a^2DX.$$

როგორც ვხედავთ, შემთხვევით სიდიდეზე მუდმივის დამატება მის დისპერსიას არ ცვლის (თუ ავიღებთ $a=1$, მაშინ მივიღებთ $D(X+b)=DX$), ხოლო მუდმივი მამრავლი დისპერსიის ნიშნის გარეთ გადის კვადრატში ახარისხებული (თუ ავიღებთ $b=0$, მივიღებთ $D(aX)=a^2DX$). კერძოდ, ეს ფორმულა გვიჩვენებს როგორ იცვლება დაკვირვების შედეგების დისპერსია ათვლის საწყისი წერტილისა და გაზომვის ერთეულის ცვლილებისას. ის გვაძლევს გამოსათვლელი ფორმულების გარდაქმნის წესს მასშტაბისა და გადატანის სხვა პარამეტრებზე გადასვლის დროს.

თეორემა 1. თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, მაშინ მათი ჯამის დისპერსია თითოეულის დისპერსიების ჯამია

$$D(X+Y)=DX+DY$$

დამტკიცება. ვისარგებლოთ იგივეობით

$$[(X+Y)-(EX+EY)]^2=[(X-EX)+(Y-EY)]^2=$$
$$=(X-EX)^2+2(X-EX)(Y-EY)+(Y-EY)^2.$$

მაშინ მათემატიკური ლოდინის თვისებების ძალით, ვღებულობთ:

$$D(X+Y)=E[(X+Y)-E(X+Y)]^2=$$
$$=E(X-EX)^2+2E[(X-EX)(Y-EY)]+E(Y-EY)^2=$$
$$=DX+DY+2E[(X-EX)(Y-EY)].$$

როგორც ცნობილია, თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, მაშინ აგრეთვე დამოუკიდებელია $X+a$ და $Y+b$ და $E(XY)=EX\cdot EY$ გარდა ამისა, $E(X-EX)=0$. ამიტომ საბოლოოდ გვაქვს:

$$D(X+Y)=DX+DY+2E(X-EX)\cdot E(Y-EY)]=DX+DY \qquad ■$$

დავალება. შეამოწმეთ, რომ თუ X და Y დამოუკიდებელი შემთხვევითი სიდიდეებია, მაშინ მათი სხვაობის დისპერსია თითოეულის დისპერსიების ჯამია

$$D(X-Y)=DX+DY.$$

თეორემა 2. თუ $X_1, X_2, \dots, X_n$ — წყვილ-წყვილად დამოუკიდებელი შემთხვევითი სიდიდეებია (ე. ი. X_i და X_j დამოუკიდებელია, თუ $i \neq j$). მაშინ ჯამის დისპერსია ტოლია დისპერსიების ჯამის

$$D(X_1+X_2+\dots+X_n)=DX_1+DX_2+\dots+DX_n.$$

დამტკიცება. შემოვიღოთ აღნიშვნები $a_i=X_i-EX_i$ $(i=1,2,\dots,n)$. მაშინ თუ ვისარგებლებთ აჯამვის სიმბოლოს შემდეგი თვისებით:

$$\left(\sum_{i=1}^{n}a_i\right)^2=\left(\sum_{i=1}^{n}a_i\right)\cdot\left(\sum_{j=1}^{n}a_j\right)=\sum_{i=1}^{n}a_i^2+\sum_{i\neq j}a_ia_j.$$

მივიღებთ, რომ

$$(X_1+X_2+\dots+X_n-EX_1-EX_2-\dots-EX_n)^2=$$
$$=\sum_{i=1}^{n}(X_i-EX_i)^2+\sum_{i\neq j}(X_i-EX_i)(X_j-EX_j).$$

თუ ახლა გავითვალისწინებთ, რომ ჯამის მათემატიკური ლოდინი მათემატიკური ლოდინების ჯამის ტოლია, ხოლო დამოუკიდებელი შემთხვევითი სიდიდეების ნამრავლის მათემატიკური ლოდინი მათემატიკური ლოდინების ნამრავლის ტოლია, მივიღებთ შესამოწმებელ თანაფარდობას:

$$D(X_1+X_2+\dots+X_n)=E[X_1+X_2+\dots+X_n-E(X_1+X_2+\dots+X_n)]^2=$$
$$=E\sum_{i=1}^{n}(X_i-EX_i)^2+E\sum_{i\neq j}(X_i-EX_i)(X_j-EX_j)=$$
$$=\sum_{i=1}^{n}E(X_i-EX_i)^2+\sum_{i\neq j}E(X_i-EX_i)\cdot E(X_j-EX_j)=\sum_{i=1}^{n}DX_i,$$

სადაც ბოლო ეტაპზე ჩვენ ვისარგებლეთ იგივეობით: $E(X-EX)=0$. ∎

შენიშვნა. რაც შეეხება ჯამის მათემატიკურ ლოდინს ის ყოველთვის შესაკრებთა მათემატიკური ლოდინების ჯამის ტოლია, მიუხედავად იმისა დამოუკიდებელია თუ დამოკიდებული შემთხვევითი სიდიდეები. ორი შემთხვევითი სიდიდის შემთხვევაში ჩვენ ეს უკვე შევამოწმეთ. მისი განზოგადოება ადვილად შეიძლება მათემატიკური ინდუქციის მეთოდის გამოყენებით.

დავალება. დაამტკიცეთ, რომ ნებისმიერი $X_1, X_2, \dots, X_n$ შემთხვევითი სიდიდეებისათვის ჯამის მათემატიკური ლოდინი მათემატიკური ლოდინების ჯამის ტოლია

$$E(X_1+X_2+\dots+X_n)=EX_1+EX_2+\dots+EX_n.$$

ეს ორი თანაფარდობა არსებით როლს თამაშობს მათემატიკურ სტატისტიკაში მონაცემთა შერჩევითი მახასიათებლების შესწავლის დროს, ვინაიდან შერჩევაში მონაწილე დაკვირვებებისა და გაზომვების შედეგები, მათემატიკურ სტატისტიკაში, გადაწყვეტილებების მიღების თეორიაში და ეკონომეტრიკაში, როგორც წესი განიხილება როგორც დამოუკიდებელი შემთხვევითი სიდიდეების რეალიზაციები.

მაგალითი 2. განვიხილოთ რაიმე A ხდომილება და X შემთხვევითი სიდიდე, ისეთი, რომ $X(\omega)=1$, თუ $\omega\in A$ და $X(\omega)=0$, თუ $\omega\notin A$

(ასეთ შემთხვევით სიდიდეს A ხდომილების **მახასიათებელი ფუნქცია** ეწოდება). ვაჩვენოთ, რომ

$$EX = P(A), \quad DX = P(A)\cdot(1-P(A)).$$

გასაგებია, რომ ამ X შემთხვევითი სიდიდის განაწილების კანონი იქნება

x_i	1	0
p_i	$P(A)$	$1-P(A)$

ასეთი კანონით განაწილებულ შემთხვევით სიდიდეს **ბერნულის შემთხვევით სიდიდეს** უწოდებენ.

მათემატიკური ლოდინის განმარტების თანახმად გვექნება

$$EX = 1\cdot P(A) + 0\cdot(1-P(A)) = P(A).$$

ანალოგიურად, $Y = (X-EX)^2 = (X-P(A))^2$ შემთხვევითი სიდიდის განაწილების კანონი იქნება:

y_i	$(1-P(A))^2$	$(P(A))^2$
p_i	$P(A)$	$1-P(A)$

შესაბამისად,

$$DX = EY = E(X-EX)^2 = (1-P(A))^2\cdot P(A) + (P(A))\cdot(1-P(A)) =$$
$$= P(A)\cdot(1-P(A))\cdot[1-P(A)+P(A)] = P(A)\cdot(1-P(A)). \quad \blacksquare$$

მაგალითი 3. განვიხილოთ n დამოუკიდებელი ექსპერიმენტი (ცდა), რომელთაგან თითოეულში გარკვეული A ხდომილება შეიძლება მოხდეს ან არ მოხდეს. შემოვიღოთ შემთხვევითი სიდიდეები $X_1, X_2, \dots, X_n$ შემდეგნაირად: $X_i(\omega)=1$, თუ i-ურ ექსპერიმენტში მოხდა A ხდომილება, და $X_i(\omega)=0$ – წინააღმდეგ შემთხვევაში. მაშინ როგორც უკვე ზემოთ ავღნიშნეთ $X_1, X_2, \dots, X_n$ შემთხვევითი სიდიდეები წყვილ-წყვილად დამოუკიდებელია. ავღნიშნოთ $p = P(A)$ (ზოგჯერ p-ს უწოდებენ "*წარმატების ალბათობას*" – თუ A ხდომილების მოხდენა განიხილება როგორც "*წარმატება*"). მაშინ წინა მაგალითის თანახმად $EX_i = p$ და $DX_i = p(1-p)$.

ბინომიალური განაწილება. $B = X_1 + X_2 + \dots + X_n$ შემთხვევით სიდიდეს, სადაც $X_1, X_2, \dots, X_n$ წინა მაგალითიდანაა, **ბინომიალური** ეწოდება. ცხადია, რომ მისი მნიშვნელობები მერყეობს 0-დან n-მდე ცდების ნებისმიერი შედეგების დროს, $0 \le B \le n$. იმისათვის რომ ვიპოვოთ მისი განაწილება (ე. ი. $P\{B=k\}$ ალბათობები, როცა $k = 0,1,\dots,n$), საკმარისია ვიცოდეთ p – ცალ-კეულ ცდაში განსახილველი A ხდომილების მოხდენის ალბათობა.

მართლაც, ხდომილება $\{B=k\}$ ხდება მაშინ და მხოლოდ მაშინ როცა A ხდომილება ხდება ზუსტად k ექსპერიმენტში (შესაბამისად, A ხდომილება არ ხდება ზუსტად $n-k$ ექსპერიმენტში, ამასთანავე მნიშვნელეობა არა აქვს რომელ k ექსპერიმენტში მოხდება A ხდომილება). ამდენად, $\{B=k\}$ ხდომილების ალბათობა ტოლია n დამოუკიდებელი ხდომილების ერთდროულად მოხდენის (ნამრავლის) ალბათობის, და რადაგანაც დამოუკიდებელ ხდომილებათა ნამრაველის ალბათობა ტოლია ალბათობების ნამრავლის, შესაბამისად, ალბათობა

იმისა, რომ n დამოუკიდებელ ცდაში A ხდომილება მოხდა ზუსტად k-ჯერ, ხოლო მისი საწინააღმდეგო $\overline{A}$ – კი $(n-k)$-ჯერ, ტოლია $\underbrace{p\cdot p\cdots p}_{k\text{-ჯერ}}\cdot\underbrace{(1-p)\cdot(1-p)\cdots(1-p)}_{(n-k)\text{-ჯერ}}=p^k(1-p)^{n-k}$.

რამდენ სხვადასხვანაირად შეიძლება k ადგილის შერჩევა n ადგილიდან? ეს შეიძლება განხორციელდეს C_n^k-ჯერ. ე. ი. ხდომილება $\{B=k\}$ წარმოიდგინება C_n^k ცალი უთავსებადი ხდომილების გაერთიანების სახით, რომელთაგან თითოეულის ალბათობაა $p^k(1-p)^{n-k}$ ამიტომ ალბათობათა შეკრების წესის თანახმად გვაქვს:

$$P\{B=k\}=C_n^k p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

სახელწოდება "*ბინომიალური განაწილება*" მოდის იქიდან, რომ $P\{B=k\}$ ალბათობა წარმოადგენს $(k+1)$-ე წევრს ნიუტონის ბინომის გაშლაში:

$$(a+b)^n=\sum_{k=0}^{n}C_n^k a^k b^{n-k},$$

თუ დავუშვებთ, რომ $a=p$ და $b=1-p$.

ვინაიდან ბინომიალური შემთხვევითი სიდიდე წარმოიდგინება n დამოუკიდებელი ბერნულის შემთხვევითი სიდიდის ჯამის სახით, ამიტომ თეორემა 2-სა და მაგალითი 2-ის თანახმად გვექნება:

$$EB=np \text{ და } DB=np(1-p).$$

ჰიპერგეომეტრიული განაწილების შემთხვევაში

$$\left(P\{X=m\}=\frac{C_M^m\cdot C_{N-M}^{n-m}}{C_N^n},\ m=0,1,\ldots,\min(M,n)\right):$$

$$EX=\frac{n\cdot M}{N} \text{ და } DX=\frac{n\cdot M\cdot(N-M)\cdot(N-n)}{N^2\cdot(N-1)}.$$

გამოვთვალოთ **პუასონის განაწილების**

$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\ldots$$

დისპერსია. როგორც ჩვენ უკვე ვნახეთ პუასონის განაწილების მათემატიკური ლოდინი ემთხვევა მის λ პარამეტრს. ვიპოვოთ ახლა პუასონის შემთხვევითი სიდიდის კვადრატის მათემატიკური ლოდინი. ამ მიზნით ვისარგებლოთ წარმოდგენით:

$$X^2=X(X-1)+X,$$

მაშინ

$$EX^2=E[X(X-1)]+EX=E[X(X-1)]+\lambda.$$

ამიტომ, თუ ვისარგებლებთ შემთხვევითი სიდიდის ფუნქციიდან მათემატიკური ლოდინის გამოთვლის წესითა და e^x ფუნქციის ხარისხოვან მწკრივად გაშლით, მივიღებთ, რომ:

$$E[X(X-1)]=\sum_{k=0}^{\infty}k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=2}^{\infty}k(k-1)\frac{\lambda^k}{k!}=$$

$$=\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}=\lambda^2e^{-\lambda}e^{\lambda}=\lambda^2.$$

საბოლოოდ, გვაქვს:

ე. ი. პუასონის განაწილების როგორც მათემატიკური ლოდინი, ისე დისპერსია ტოლია ამ განაწილების λ პარამეტრის.

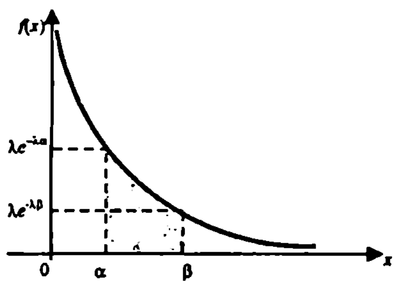
გეომეტრიული განაწილების შემთხვევაში ($P\{X=m\}=p(1-p)^{m-1}$, $m=1,2,\ldots$):

$$EX=1/p \text{ და } DX=(1-p)/p^2.$$

ექსპონენციალური განაწილება. X შემთხვევით სიდიდეს ეწოდება ექსპონენციალურად განაწილებული პარამეტრით λ ($\lambda>0$), თუ მის განა-წილების სიმკვრივეს აქვს სახე:

$$f(x)=\begin{cases}0, & x<0,\\ \lambda e^{-\lambda x}, & x\ge 0.\end{cases}$$

ექსპონენციელური განაწილების სიმკვრივის ფუნქციის გრაფიკს აქვს შემდეგი სახე:



აქ დაშტრიხული არის ფართობი ტოლია ექსპონენციალურად განაწილებული X შემთხვევით სიდიდის (α,β) ინტერვალში მოხვედრის ალბათობის.

ექსპონენციალური განაწილების სიმკვრივიდან შეგვიძლია აღვადგინოთ ექსპონენციალური განაწილების ფუნქცია. ადვილი დასანახია, რომ

$$F(x)=\begin{cases}0, & \text{როცა } x<0,\\ 1-e^{-\lambda x}, & \text{როცა } x\ge 0.\end{cases}$$

ვნახოთ რა ალბათური შინაარსი გააჩნია λ პარამეტრს. ამ მიზნით, გამოვთვალოთ ექსპონენციალური განაწილების მათემატიკური ლოდინი. უწყვეტი შემთხვევითი სიდიდისათვის მათემატიკური ლოდინის განმარტების თანახმად გვაქვს:

$$EX=\int_{-\infty}^{+\infty}xf(x)dx=\int_{0}^{+\infty}\lambda xe^{-\lambda x}dx=\frac{1}{\lambda}.$$

ექსპონენციალური განაწილების λ პარამეტრი წარმოადგენს განაწილების მათემატიკური ლოდინის შებრუნებულ სიდიდეს. მეორეს მხრივ, თუ გამოვითვლით დისპერსიას, დავინახავთ, რომ:

$$DX=1/\lambda^2,$$

შესაბამისად, საშუალო კვადრატული გადახრა იქნება $1/\lambda$.

გარდა ამისა, ვინაიდან განაწილების სიმკვრივე კლებადია, ამიტომ ის თავის მაქსიმალურ მნიშვნელობას აღწევს $x=0$ წერტილში. შესაბამისად, ექსპონენციალური განაწილების მოდა იქნება $M_0=0$.

შენიშვნა. ჩვეულებრივი უპირობო დისპერსიის ანალოგიურად განიმარტება პირობითი დისპერსია რაიმე B ხდომილების მიმართ:

$$D(X \mid B) = E\{[X - X(X \mid B)]^2 \mid B\} = E(X^2 \mid B) - [E(X \mid B)]^2.$$

სტანდარტული გადახრა. მომენტები.

განმარტება 1. X შემთხვევითი სიდიდის **საშუალო კვადრატული გადახრა** ეწოდება არითმეტიკულ კვადრატულ ფესვს ამ შემთხვევითი სიდიდის დისპერსიიდან და აღინიშნება σ_x სიმბოლოთი:

$$\sigma_x = +\sqrt{DX}$$

σ_x-ს ხშირად **სტანდარტულ გადახრასაც** უწოდებენ. გადახრის ამ მახასიათებლის შემოღება განპირობებულია იმით, რომ, განსხვავებით დისპერსიისაგან, იგი ზომის იგივე ერთეულებში გამოისახება, რაც X შემთხვევითი სიდიდე.

შემთხვევითი სიდიდის საშუალო კვადრატული გადახრა დაახლოებით მიუთითებს იმაზე, თუ რამდენად განსხვავდება შემთხვევითი სიდიდის დაკვირვებული მნიშვნელობა მათემატიკური ლოდინისაგან. კომერციული მოღვაწეობის ხშირ შემთხვევაში სტანდარტული გადახრა არის რისკის მახასიათებელი, მიუთიუთებს რა, თუ რამდენად განუსაზღვრელია სიტუაცია.

შემთხვევითი სიდიდის სტანდარტიზაცია. დავუშვათ, რომ X შემთხვევითი სიდიდის მათემატიკური ლოდინია EX, ხოლო საშუალო კვადრატული გადახრაა – σ_x. განვიხილოთ ახალი შემთხვევითი სიდიდე

$$Y = \frac{X - EX}{\sigma_x} \tag{1}$$

და გამოვთვალოთ მისი მათემატიკური ლოდინი და დისპერსია.

მათემატიკური ლოდინისა და დისპერსიის თვისებებიდან გამომდინარე ადვილი დასანახია, რომ $EY = 0$ და $DY = 1$. მართლაც, გვაქვს:

$$EY = E\left(\frac{X - EX}{\sigma_x}\right) = E\left[\frac{1}{\sigma_x} \cdot (X - EX)\right] = \frac{1}{\sigma_x} \cdot E(X - EX)$$

და

$$DY = D\left(\frac{X - EX}{\sigma_x}\right) = D\left[\frac{1}{\sigma_x} \cdot (X - EX)\right] = \frac{1}{\sigma_x^{\,2}} \cdot D(X - EX) = \frac{1}{DX} \cdot DX = 1.$$

(1) გარდაქმნას ეწოდება X შემთხვევითი სიდიდის **ცენტრირება** (მათემატიკური ლოდინის გამოკლება) და **ნორმირება** (საშუალო კვავდრატულ გადახრაზე გაყოფა) ან უფრო მოკლედ – X შემთხვევითი სიდიდის **სტანდარტიზაცია**.

სხვა სიტყვებით, რომ ვთქვათ სტანდარტიზაცია არის შემთხვევითი სიდიდის ისეთი წრფივი გარდაქმნა, რომელსაც გარკვეული მათემატიკური ლოდინისა და დისპერსიის მქონე შემთხვევითი სიდიდე დაყავს ნოლოვანი მათემატიკური ლოდინისა და ერთეულოვანი დისპერსიის მქონე (ანუ სტანდარტულ) შემთხვევით სიდიდეზე.

მომენტები, ასიმეტრია და ექსცესი. X შემთხვევითი სიდიდის n **რიგის საწყისი მომენტი** ეწოდება სიდიდეს $\mu_n := EX^n$ შესაბამისად,

პირველი რიგის მომენტი წარმოადგენს მათემატიკურ ლოდინს $\mu := \mu_1 = EX$.

X შემთხვევითი სიდიდის n **რიგის ცენტრალური მომენტი** ეწოდება სიდიდეს $\nu_n := E(X-\mu)^n$. ამ აღნიშვნებში გასაგებია, რომ მეორე რიგის ცენტრალური მომენტი წარმოადგენს დისპერსიას $\sigma^2 := \nu_2 = DX$. შესაბამისად, $\sigma = \sigma_x$. ადვილი დასანახია, რომ საწყის და ცენტრალურ მომენტებს შორის არსებობს შემდეგი კავშირი:

$$\nu_2 = \mu_2 - \mu^2,$$

$$\nu_3 = \mu_3 - 3\mu_2 \cdot \mu + 2\mu^2,$$

$$\nu_4 = \mu_4 - 4\mu_3 \cdot \mu + 6\mu_2 \cdot \mu^2 - 3\mu^4, \text{ და ა. შ.}$$

ცენტრალური მომენტების საშუალებით განიმარტება შემთხვევითი სიდიდის შემდეგი მნიშვნელოვანი რიცხვითი მახასიათებლები, კერძოდ, ასიმეტრიისა და ექსცესის კოეფიციენტები.

ექსცესის კოეფიციენტი ეწოდება სიდიდეს

$$e = \frac{\nu_4}{\sigma^4} - 3 = \frac{E(X-EX)^4}{[+\sqrt{E(X-EX)^2}\,]^4} - 3.$$

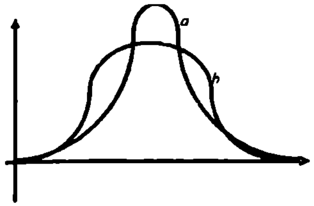
ექსცესის კოეფიციენტი ახასიათებს განაწილების კონცენტრაციის ხარისხს საშუალო მნიშვნელობის (μ-ს) ირგვლივ. რაც უფრო დიდია e, მით მეტადაა კონცენტრირებული განაწილება საშუალოს ირგვლივ, ანუ სიმკვრივეს μ წერტილში აქვს მაღალი პიკი, და პირიქით (იხ. ნახ. 1: შეს-აბამისად, a და b წირები).

ასიმეტრიის კოეფიციენტი ეწოდება სიდიდეს

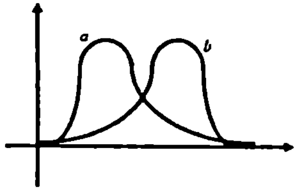
$$\alpha = \frac{\nu_3}{\sigma^3} = \frac{E(X-EX)^3}{[+\sqrt{E(X-EX)^2}\,]^3}.$$

ასიმეტრიის ზომას საფუძვლად უდევს საშუალო კუბური გადახრა, რომელიც საშუალებას იძლევა უფრო სრულად გავითვალისწინოთ შემთხვევითი სიდიდის დიდი გადახრები. განაწილების ასიმეტრიის შემთხვევაში განაწილების მრუდის ერთი მხარე იძლევა უფრო დიდ კუბურ გადახრას მეორე მხარესთან შედარებით და რადგან კუბური გადახრის დროს გადახრის ნიშანი ნარჩუნდება, ამიტომ კუბურ გადახრებს შორის განსხვავება აჩვენებს დადებით ან უარყოფით ასიმეტრიას.

თუ X შემთხვევითი სიდიდის განაწილება სიმეტრიულია თავისი საშუალო მნიშვნელობის ($\mu = EX$ მათემატიკური ლოდინის) მიმართ, მაშინ X შემთხვევითი სიდიდის ყველა კენტი რიგის ცენტრალური მომენტი ნულის ტოლია ($\nu_{2n-1} = 0$). შესაბამისად, ამ შემთხვევაში ასიმეტრიის კოეფიციენტიც ნულის ტოლი იქნება. თუ $\alpha < 0$, მაშინ განაწილება მარცხნივ ასიმეტრიულია, ხოლო თუ $\alpha > 0$, მაშინ განაწილება მარჯვნივ ასიმეტრიულია.(იხ. ნახ. 2: შესაბამისად, a და b წირები).



ნახ. 1



ნახ. 2

კოვარიაცია. კორელაციის კოეფიციენტი.

ორ ξ და η შემთხვევით სიდიდეს შორის კავშირის (დამოკიდებულების) მახასიათებელს წარმოადგენს ξ და η შემთხვევით სიდიდეების თავიანთი განაწილებების ცენტრებიდან (განაწილების ცენტრი ეწოდება შემთხვევითი სიდიდის მათემატიკურ ლოდინს) გადახრების ნამრავლის მათემატიკური ლოდინი, რომელსაც **კოვარიაციის კოეფიციენტი** ან უბრალოდ **კოვარიაცია** ეწოდება და აღინიშნება $cov(\xi;\eta)$ სიმბოლოთი:

$$cov(\xi;\eta) = E((\xi-E\xi)(\eta-E\eta)).$$

დავუშვათ, რომ ξ შემთხვევით სიდიდის შესაძლო მნიშვნელობებია $\{x_1, x_2, x_3,\ldots, x_n\}$, ხოლო η შემთხვევით სიდიდის შესაძლო მნიშვნელობებია $\{y_1, y_2, y_3,\ldots,y_k\}$. მაშინ მათემატიკური ლოდინის თვისებებიდან გამომდინარე გასაგებია, რომ:

$$cov(\xi;\eta)=\sum_{i=1}^{n}\sum_{j=1}^{k}(x_i-E\xi)\cdot(y_j-E\eta)\cdot P((\xi=x_i)\cap(\eta=y_j)). \qquad (1)$$

(1) ფორმულას შეიძლება მიეცეს ასეთი ინტერპრეტაცია: თუ ξ-ს დიდი მნიშვნელობებისათვის უფრო ალბათურია η-ს დიდი მნიშვნელობები, ხოლო ξ-ს მცირე მნიშვნელობებისათვის უფრო ალბათურია η-ს მცირე მნიშვნელობები, მაშინ (1) ფორმულის მარჯვენა მხარეში დომინირებენ დადებითი შესაკრებები, და კოვარიაცია ღებულობს დადებით მნიშვნელობას.

თუ კი უფრო ალბათურია ისეთი ნამრავლები $(x_i-E\xi)(y_j-E\eta)$, რომ-ლებიც შედგებიან სხვადასხვა ნიშნიანი თანამამრავლებისაგან, ანუ შემთხვევითი ექსპერიმენტის იმ შედეგებს, რომლებსაც მივყავართ ξ-ს დიდი მნიშვნელობებისაკენ, ძირითადად მივყავართ η-ს მცირე მნიშვნელობებთან. და პირიქით, მაშინ კოვარიაცია ღებულობს მოდულით დიდ უარყოფით მნიშვნელობებს.

პირველ შემთხვევაში მიღებულია ვილაპარაკოთ პირდაპირ კავშირზე: ξ-ს ზრდასთან ერთად η შემთხვევით სიდიდეს გააჩნია მატების (ზრდის) ტენდენცია. მეორე შემთხვევაში კი ლაპარაკობენ შებრუნებულ კავშირზე: ξ-ს ზრდასთან ერთად η შემთხვევით სიდიდეს გააჩნია შემცირების ანუ დაცემის ტენდენცია.

თუ კი ჯამში დაახლოებით ერთი და იგივე წელილი შეაქვთ დადებით და უარყოფით ნამრავლებს $(x_i-E\xi)(y_j-E\eta)p_{ij}$, მაშინ შეიძლება ითქვას, რომ ჯამში ისინი "აქრობენ" ერთმანეთს და კოვარიაცია ახლოს იქნება ნულთან. ამ შემთხვევაში ერთი შემთხვევითი სიდიდის მეორეზე დამოკიდებულება არ იკვეთება.

ადვილი საჩვენებელია, რომ თუ

$$P[(\xi = x_i)\cap(\eta = y_j)] = P(\xi = x_i)\cdot P(\eta = y_j),\quad i = 1,\dots n;\ j = 1,\dots,k \qquad (2)$$

მაშინ cov(ξ; η)= 0.

მართლაც, (1) ფორმულის თანახმად გვაქვს:

$$\mathrm{cov}(\xi;\eta) = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_i - E\xi)\cdot(y_j - E\eta)\cdot P(\xi = x_i)\cdot P(\eta = y_j) =$$

$$= \sum_{i=1}^{n}(x_i - E\xi)P(\xi = x_i)\cdot\sum_{j=1}^{k}(y_j - E\eta)P(\eta = y_j) =$$

$$= E(\xi - E\xi)\cdot E(\eta - E\eta) = 0\cdot 0 = 0.$$

აქ ჩვენ გამოვიყენეთ მათემატიკური ლოდინის შემდეგი მნიშვნელოვანი თვისება: *შემთხვევითი სიდიდის თავისი მათემატიკური ლოდინიდან გადახრის მათემატიკური ლოდინი ნულის ტოლია.*

მაგალითად, დისკრეტულ შემთხვევაში მართლაც გვაქვს:

$$E(\xi - E\xi) = \sum_{i=1}^{n}(x_i - E\xi)P(x_i) = \sum_{i=1}^{n}x_iP(x_i) - E\xi\sum_{i=1}^{n}P(x_i) = E\xi - E\xi = 0.$$

კოვარიაცია შეიძლება გადაიწეროს შემდეგი მოხერხებული ფორმით:

$$\mathrm{cov}(\xi;\eta) = E(\xi\eta - \xi E\eta - \eta E\xi + E\xi E\eta) = E(\xi\eta) - E(\xi E\eta) -$$
$$- E(\eta E\xi) + E(E\xi E\eta) =$$
$$E(\xi\eta) - E\eta E\xi - E\xi E\eta + E\xi E\eta = E(\xi\eta) - E\xi E\eta.$$

ე. ი. ორი შემთხვევითი სიდიდის კოვარიაცია ტოლია მათი ნამრავლის მათემატიკურ ლოდინს გამოკლებული მათემატიკური ლოდინების ნამრავლი.

ცნობილია, რომ თუ შემთხვევითი სიდიდეები დამოუკიდებელია (დისკრეტულ შემთხვევაში ეს იმას ნიშნავს, რომ შესრულებულია (2) თანაფარდობები), მაშინ ნამრავლის მათემატიკური ლოდინი მათემატიკური ლოდინების ნამრავლის ტოლია. ამიტომ, უკანასკნელი თანფარდობის თანახმად (ისევე როგორც ეს ზემოთ იყო აღნიშნული) დამოუკიდებელი ξ და η შემთხვევითი სიდიდეებისათვის cov(ξ; η)= 0. შევნიშნავთ, რომ საზოგადოდ შებრუნებული დებულება სწორი არ არის.

ამოცანა 1. მონეტას აგდებენ 5-ჯერ. ξ შემთხვევითი სიდიდე არის მოსულ გერბთა რიცხვი, ხოლო η შემთხვევითი სიდიდე კი ბოლო ორ აგდებაში მოსულ გერბთა რიცხვი. ავაგოთ ამ შემთხვევითი სიდიდეების ერთობლივი განაწილების კანონი და ვიპოვოთ კოვარიაცია.

ამოცანა 2. 32 კარტიდან შემთხვევით იღებენ 2-ს. ξ შემთხვევითი სიდიდე არის ამოღებული ტუზების რიცხვი, ხოლო η შემთხვევითი სიდიდე კი ამოღებული მეფეების რიცხვი. ავაგოთ ავაგოთ ამ შემთხვევითი სიდიდეების ერთობლივი განაწილების კანონი და ვიპოვოთ კოვარიაცია.

ადვილის დასანახია, რომ გარდა ზემოთმოყვანილი თვისებებისა, კოვარიაციის ფუნქციას გააჩნია შემდეგი თვისებები:

1). $\mathrm{cov}(\xi,\xi) = D\xi$;

2). $\mathrm{cov}(\xi,\eta) = \mathrm{cov}(\eta,\xi)$;

3). $\mathrm{cov}(c\cdot\xi,\eta) = c\cdot\mathrm{cov}(\xi,\eta)$;

4). $\mathrm{cov}(\xi\pm\eta,\varsigma) = \mathrm{cov}(\xi,\varsigma)\pm\mathrm{cov}(\eta,\varsigma)$;

5). $D(\xi \pm \eta) = D\xi \pm 2\operatorname{cov}(\xi,\eta) + D\eta$, კერძოდ, თუ ξ და η დამოუკიდებელი შემთხვევითი სიდიდეებია, მაშინ $D(\xi \pm \eta) = D\xi + D\eta$ (ვინაიდან ამ შემთხვევაში $\operatorname{cov}(\xi,\eta) = 0$);

6). $|\operatorname{cov}(\xi,\eta)| \le \sqrt{D\xi} \cdot \sqrt{D\eta}$.

გასაგებია, რომ $\operatorname{cov}(\xi;\eta)$ დამოკიდებულია იმ ზომის ერთეულებზე, რომლებშიც გამოსახულია ξ და η შემთხვევითი სიდიდეები (მაგალითად, ვთქვათ, ξ და η – გარკვეული დეტალის წრფივი სიგრძეებია. თუ ზომის ერთეულად ავიღებთ 1 სმ-ს, მაშინ $\operatorname{cov}(\xi;\eta)$ მიიღებს ერთ მნიშვნელობას, ხოლო თუ კი ზომის ერთეულად ავიღებთ 1მმ-ს, მაშინ $\operatorname{cov}(\xi;\eta)$ მიიღებს სხვა, უფრო მეტ მნიშვნელობას, თუ რა თქმა უნდა $\operatorname{cov}(\xi;\eta) \neq 0$). ამდენად, $\operatorname{cov}(\xi;\eta)$ -ს გამოყენება კავშირის მაჩვენებლად მოუხერხებელია.

იმისათვის, რომ საქმე გვქონდეს ზომის ერთეულისაგან დამოუკიდებ-ელ მაჩვენებელთან, განვიხილოთ შემთხვევითი სიდიდეები:

$$\xi^* = \frac{\xi - E\xi}{\sqrt{D\xi}} = \frac{\xi - E\xi}{\sigma_\xi}; \quad \eta^* = \frac{\eta - E\eta}{\sqrt{D\eta}} = \frac{\eta - E\eta}{\sigma_\eta}$$

ასეთ შემთხვევით სიდიდეებს ეწოდებათ ξ და η შემთხვევითი სიდიდეების **ნორმირებული გადახრები** ან **სტანდარტიზაცია**. თითოეულ მათგანს ცენტრად აქვს ნული, ხოლო დისპერსია კი ერთის ტოლია. მართლაც, გვაქვს:

$$E\xi^* = E\left(\frac{\xi - E\xi}{\sigma_\xi}\right) = \frac{1}{\sigma_\xi}\left(E\xi - E(E\xi)\right) = \frac{1}{\sigma_\xi}(E\xi - E\xi) = 0;$$

$$D\xi^* = D\left(\frac{\eta - E\eta}{\sigma_\xi}\right) = \frac{1}{\sigma_\xi^2} D(\eta - E\eta) = \frac{D\eta}{\sigma_\xi^2} = 1.$$

ξ^* და η^* შემთხვევითი სიდიდეების კოვარიაციას ეწოდება ξ და η შემთხვევითი სიდიდეების **კორელაციის კოეფიციენტი** და აღინიშნება $\rho(\xi;\eta)$ სიმბოლოთი:

$$\operatorname{cov}(\xi^*,\eta^*) = \rho(\xi;\eta) = E\left(\frac{\xi - E\xi}{\sigma_\xi} \cdot \frac{\eta - E\eta}{\sigma_\eta}\right) = \frac{E\left((\xi - E\xi)(\eta - E\eta)\right)}{\sigma_\xi \sigma_\eta} =$$

$$= \frac{\operatorname{cov}(\xi;\eta)}{\sigma_\xi \sigma_\eta} = \frac{E(\xi\eta) - E\xi E\eta}{\sigma_\xi \sigma_\eta}; \quad \sigma_\xi = \sqrt{D\xi}; \ \sigma_\eta = \sqrt{D\eta}.$$

ξ და η დამოუკიდებელი შემთხვევითი სიდიდეებია, მაშინ $\rho(\xi;\eta)=0$, ვინაიდან ამ შემთხვევაში $\operatorname{cov}(\xi;\eta)=0$. პირიქით, საზოგადოდ სწორი არ არის. შემთხვევითი სიდიდეები შეიძლება ფუნქციონალურადაც კი იყვნენ დამოკიდებულები (ერთი შემთხვევითი სიდიდის ნებისმიერ მნიშვნელობას შეესაბამება ერთადერთი მნიშვნელობა მეორე შემთხვევითი სიდიდის), მაგრამ მათი კორელაციის კოეფიციენტი ნულის ტოლი იყოს.

მაგალითი 1. დავუშვათ, რომ ξ შემთხვევითი სიდიდე სიმეტრიულა-დაა განაწილებული ნულის ირგვლივ. მაშინ $E\xi=0$. ვთქვათ, $\eta=\xi^2$. მაშინ $E(\xi\eta)=E(\xi^3)=0$, ვინაიდან ξ^3 აგრეთვე, სიამეტრიულადაა განაწილებული ნულის ირგვლივ. მეორეს მხრივ, $E\xi E\eta=0$, ვინაიდან $E\xi=0$. ამიტომ:

$$\rho(\xi;\eta)=\frac{E(\xi\eta)-E\xi E\eta}{\sigma_\xi\sigma_\eta}=0.$$

მაგალითი 2. დავუშვათ, რომ ξ და η შემთხვევითი სიდიდეების ერთობლივი განაწილების კანონი მოცემულია შემდეგი ცხრილით:

ξ \ η	1	2	
1	1/5	0	1/5
2	0	3/5	3/5
3	1/5	0	1/5
	2/5	3/5	

ცხადია, რომ:

$$E\xi=1\cdot 1/5+2\cdot 3/5+3\cdot 1/5=2;\quad E\eta=1\cdot 2/5+2\cdot 3/5=8/5;$$

$$E(\xi\eta)=1\cdot 1\cdot 1/5+2\cdot 2\cdot 3/5+3\cdot 1\cdot 1/5=16/5;\quad E(\xi\eta)-E\xi E\eta=0.$$

აქედან გამომდინარე, კორელაციის კოეფიციენტი ნულია, მაშინ როდესაც ნათელია, რომ ადგილი აქვს η შემთხვევითი სიდიდის ფუნქციონალურ დამოკიდებულებას ξ შემთხვევით სიდიდეზე.

კორელაციის კოეფიციენტი არ შეიცვლება, თუ ξ შემთხვევითი სიდიდის ნაცვლად განვიხილავთ $\xi_1=\xi+a$ ან $\xi_2=k\xi$ შემთხვევით სიდიდეს (სადაც a და k—მუდმივებია, $k>0$), ვინაიდან კოორდინატთა სათავის შეცვლისას ან მასშტაბის ცვლილებისას ნორმირებული გადახრა არ იცვლება. იგივე შეიძლება ითქვას η შემთხვევით სიდიდეზეც.

კორელაციის კოეფიციენტის თვისებები:

1. $-1\le\rho(\xi;\eta)\le 1$.
2. თუ $\rho(\xi;\eta)=1$, მაშინ $\eta=k\xi+b$, სადაც k და b— მუდმივებია, $k>0$.
3. თუ $\rho(\xi;\eta)=-1$, მაშინ $\eta=k\xi+b$, სადაც k და b— მუდმივებია, $k<0$.
4. თუ $\eta=k\xi+b$, $(k\neq 0)$ ან $\xi=k_1\eta+b_1$ $(k_1\neq 0)$, მაშინ
   $\rho(\xi;\eta)=1$ როცა $k_i>0$; $\rho(\xi;\eta)=-1$ როცა $k_i<0$ $(i=1,2)$.

კორელაციის კოეფიციენტი $\rho(\xi;\eta)$ აღწევს თავის სასაზღვრო მნიშვნელობებს -1-სა და 1-ს მაშინ და მხოლოდ მაშინ, როცა ξ და η შემთხვევითი სიდიდეების ყველა მნიშვნელობა კონცენტრირებულია (თავმოყრილია) ξ; η სიბრტყის გარკვეულ წრფეზე, ანუ ξ და η შემთხვევით სიდიდეებს შორის არსებობს წრფივი კავშირი.

თუ $|\rho(\xi;\eta)|<1$, მაშინ ასეთი წრფივი კავშირი არ არსებობს. თუმცა, $|\rho(\xi;\eta)|$-ს ერთთან მიახლოებასთან ერთად ξ და η –ს ერთობლივ განაწილებას გააჩნია გარკვეული წრფის ირგვლივ კონცენტრირების ტენდენცია და $|\rho(\xi;\eta)|$ სიდიდე შეიძლება ჩაითვალოს ξ და η სიდიდეებს შორის სრული წრფივი დამოკიდებულების საზომად.

მაგალითი. ξ და η შემთხვევითი სიდიდეების ქვემოთ მოყვანილი ერთობლივი განაწილების კანონის მიხედვით გამოვთვალოთ კორელაციის კოეფიციენტი $\rho(\xi;\eta)$.

η / ξ	1	2	3	
10	1/36	0	0	1/36
20	2/36	1/36	0	3/36
30	2/36	2/36	2/36	6/36
40	1/36	9/36	16/36	26/36
	6/36	12/36	18/36	

$$E\xi = 10\cdot 1/36 + 20\cdot 3/36 + 30\cdot 6/36 + 40\cdot 26/36 \cong 35,83$$

$$E\eta = 1\cdot 6/36 + 2\cdot 12/36 + 3\cdot 18/36 \cong 2,3$$

$$D\xi = (10-35,83)^2\cdot 1/36 + (20-35,83)^2\cdot 3/36 + (30-35,83)^2\cdot 6/36 +$$

$$+(40-35,83)^2\cdot 26/36 \cong 57,64$$

$$\sigma_\xi \cong 7,6$$

$$D\eta = (1-2,3)^2\cdot 6/36 + (2-2,3)^2\cdot 12/36 + (3-2,3)1^2\cdot 18/36 \cong 0,556$$

$$\sigma_\eta \cong 0,746$$

$$E(\xi\eta) = 10\cdot 1\cdot 1/36 + 20\cdot 1\cdot 2/36 + 20\cdot 2\cdot 1/36 + 30\cdot 1\cdot 2/36 + 30\cdot 2\cdot 2/36 +$$

$$+30\cdot 3\cdot 2/36 + 40\cdot 1\cdot 1/36 + 40\cdot 2\cdot 9/36 + 40\cdot 3\cdot 16/36 = 86,94$$

$$\rho(\xi;\eta) = (6,94 - 2,3\cdot 35.83)/(7,6\cdot 0,746) \cong 0,8.$$

დავუშვათ, რომ მოცემულია ξ და η ორი შემთხვევითი სიდიდის ერთობლივი განაწილების კანონი, და ξ შემთხვევითი სიდიდის პირობითი მათემატიკური ლოდინი იცვლება η შემთხვევითი სიდიდის მნიშვნელობის მიხედვით. მაშინ ლაპარაკობენ ξ შემთხვევითი სიდიდის **კორელაციურ დამოკიდებულებაზე** η შემთხვევით სიდიდეზე. თუ ξ-ს პირობითი მათემატიკური ლოდინი არის η-ს წრფივი ფუნქცია, მაშინ ξ და η შემთხვევით სიდიდეებს შორის არსებობს წრფივი კორელაციური კავშირი ანუ დამოკიდებულება.

როგორც წესი, კორელაციურ დამოკიდებულებაზე საუბრისას, მხედველობაში აქვთ წრფივი კორელაციური დამოკიდებულება. არაწრფივი კორელაციური დამოკიდებულების არსებობისას, მას სპეციალურად აღნიშნავენ.

ξ და η შემთხვევით სიდიდეებს შორის კორელაციური დამოკიდებულება შეიძლება განიმარტოს როგორც კავშირი ξ და η-ს ზრდის ტენდენციებს შორის. მაგალითად, ξ და η-ს შორის არსებობს პირდაპირი კორელაციური დამოკიდებულება, თუ ξ-ს ზრდასთან ერთად η შემთხვევით სიდიდეს გააჩნია ზრდის ტენდენცია (ეს ნიშნავს, რომ ξ-ს დიდი მნიშვნელობების შემთხვევაში დიდი ალბათობით შეგვხვდება η-ს დიდი მნიშვნელობებიც). თუ ξ-ს დიდ მნიშვნელობებს დიდი ალბათობით შეესაბამება η-ს მცირე მნიშვნელობები, ანუ ξ-ს ზრდასთან ერთად η შემთხვევით სიდიდეს გააჩნია კლების ტენდენცია, მაშინ ამბობენ, რომ ξ და η-ს შორის არსებობს შებრუნებული კორელაციური დამოკიდებულება.

კორელაციური დამოკიდებულების სიღრმე (ანუ სიმჭიდროვე) ხასიათდება კორელაციის კოეფიციენტით ρ(ξ;η). რაც უფრო ახლოსაა |ρ(ξ;η)| ერთთან. მით უფრო მჭიდროა კორელაციური დამოკიდებულება.

რაც უფრო ახლოსაა ξ შემთხვევითი სიდიდის პირობითი მათემატიკური ლოდინი η-ს მიმართ წრფივ დამოკიდებულებასთან და რაც უფრო მჭიდროდ ლაგდებიან ξ-ს მნიშვნელობები პირობითი მათემატიკური ლოდინის ირგვლივ, მით უფრო ღრმაა (მჭიდროა) კორელაციური დამოკიდებულება.

ჩვენ შეგვიძლია ვისაუბროთ ორი უწყვეტი შემთხვევითი სიდიდის ერთობლივ განაწილების კანონზე. უმეტეს შემთხვევაში შესაძლებელია უწყვეტი შემთხვევითი სიდიდეებიდან დისკრეტული შემთხვევითი სიდიდეების ერთობლივ განაწილებაზე გადასვლა შემდეგნაირად: ξ შემთხვევითი სიდიდის ცვლილების შუალედი $[a; b]$ უნდა გაიყოს ტოლი სიგრძის მონაკვეთებად $[c_0=a; c_1]$; $[c_1; c_2]$; $[c_2; c_3]$,…,$[c_{n-1}; c_n=b]$. ξ შემთხვევითი სიდიდის მნიშვნელობებად მივიღოთ თითოეული მონაკვეთის შუაწერტილი. ანალოგიურად, უნდა მოვიქცეთ η შემთხვევითი სიდიდის შემთხვევაში. მისი მნიშვნელობათა არე $[a; b]$ უნდა გაიყოს ტოლი სიგრძის შუალედებად $[c_0=a; c_1]$; $[c_1; c_2]$; $[c_2; c_3]$,…,$[c_{n-1}; c_n=b]$, და η-ს მნიშვნელობებად გამოვაცხადოთ $[g_{i-1}; g_i]$ შუალედების შუაწერტილები. ასეთნაირად, ჩვენ მივიღებთ $\xi^*=\{x_1; x_2; \dots x_n\}$ და $\eta^*=\{y_1; y_2; \dots y_k\}$ დისკრეტულ შემთხვევით სიდიდეებს, ამასთ-ანავე ყოველ წყვილს უსაბამებენ ალბათობას

$$P_{ij} = P((\xi\in[c_{i-1}; c_i])\cap(\eta\in[g_{j-1}; g_j])).$$

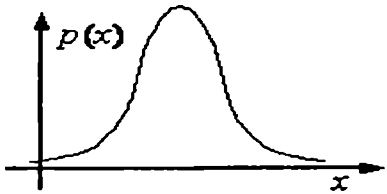
თავი V
ნორმალური და მასთან დაკავშირებული განაწილებები

თუ ξ შემთხვევითი სიდიდის განაწილების სიმკვრივე განისაზღვრება ფორმულით

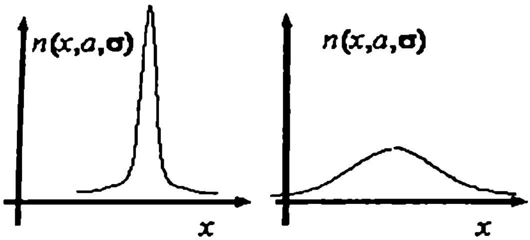
$$p(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

სადაც μ– ნებისმიერი რიცხვია, ხოლო σ– დადებითი რიცხვია, მაშინ ამბობენ რომ ξ განაწილებულია ნორმალური კანონით ანუ ξ “ნორმალური” შემთხვევითი სიდიდეა.

μ-სა და σ-ს მნიშვნელობები სრულად განსაზღვრავენ $p(x)$ ფუნქციას. ამ ფუნქციისათვის სარგებლობენ აღნიშვნით: $p(x):=n(x;\mu;\sigma)$. ნორმალური შემთხვევითი სიდიდის განაწილების სიმკვრივის გრაფიკს μ-სა და σ-ს გარკვეული მნიშვნელობებისათვის აქვს შემდეგი სახე:



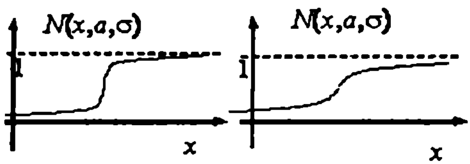
გრაფიკი სიმეტრიულია $x=\mu$ წრფის მიმართ, და სრულდება პირობა $p(x)\to 0$, როცა $x\to\pm\infty$. თუ დავიწყებთ μ-ს გაზრდას, ისე რომ σ-ს დავტოვებთ უცვლელად, მაშინ გრაფიკი დაიწყებს გადაადგილებას მარჯვნივ, ხოლო μ-ს შემცირებისას კი – მარცხნივ, ისე რომ არ შეიცვლის ფორმას. მეორეს მხრივ, თუ μ-ს მნიშვნელობა უცვლელია, მაშინ შედარებით მცირე σ-ს შეესაბამება $p(x)$-ის გრაფიკი აშკარად გამოხატული პიკით (როგორც ეს გამოსახულია ქვემოთ მოყვანილი ნახაზებიდან მარჯვენა ნახაზზე), ხოლო შედარებით დიდი σ-ს შემთხვევაში $p(x)$-ის გრაფიკი გაწოლილი წირია (როგორც ეს გამოსახულია ქვემოთ მოყვანილი ნახაზებიდან მარცხენა ნახაზზე).



ნორმალურად განაწილებული ξ შემთხვევითი სიდიდის განაწილების $F(x)$ ფუნქციის აღსანიშნავად ხმარობენ სიმბოლოს $N(x;\mu;\sigma)$. ის მიიღება განაწილების სიმკვრივის ინტეგრირებით:

$$F(x)=N(x;\mu;\sigma)=\int_{-\infty}^{x} n(t;\mu;\sigma)dt .$$

ქვემოთ მოყვანილია ნორმალურად განაწილებული შემთხვევითი სიდიდეების განაწილების $F(x)$ ფუნქციების გრაფიკები σ-ს შესაბამისად შედარებით მცირე (მარცხენა გრაფიკი) და შედარებით დიდი (მარჯვენა გრაფიკი) მნიშვნელობებისათვის:



ნორმალურად განაწილებული ξ შემთხვევითი სიდიდის განაწილების სიმკვრივის $p(x)$ ფუნქციის გრაფიკის $x=\mu$ წრფის მიმართ სიმეტრიულობიდან გამომდინარეობს, რომ $E\xi=\mu$

თუ გამოვითვლით ნორმალურად განაწილებული შემთხვევითი სიდიდის $D\xi$ დისპერსიას, აღმოჩნდება, რომ ის σ^2–ის ტოლია.

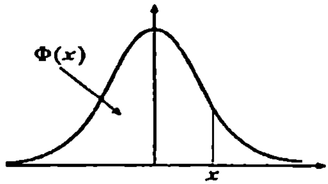
ამრიგად, μ და σ პარამეტრებს ნორმალურად განაწილებული შემთხვევითი სიდიდის განაწილების სიმკვრივის ფორმულაში, გააჩნიათ შემდეგი ალბათური შინაარსი: μ – არის მათემატიკური ლოდინი, ხოლო σ^2 - დისპერსია.

ალბათობა იმისა, რომ ნორმალურად განაწილებული ξ შემთხვევითი სიდიდე მიიღებს მნიშვნელობას (x_1, x_2) შუალედიდან, გამოითვლება შემდეგი ფორმულით:

$$P\left(x_1<\xi<x_2\right)=F(x_2)-F(x_1)=\Phi(\beta)-\Phi(\alpha).$$

აქ $\Phi(x)$ – ლაპლასის ინტეგრალური ფუნქციაა – $\Phi(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$;

$\beta=\frac{x_2-\mu}{\sigma}$; $\alpha=\frac{x_1-\mu}{\sigma}$.

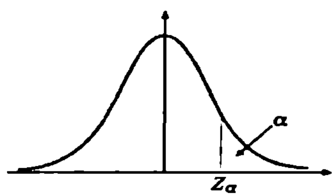


თუ ξ შემთხვევითი სიდიდის განაწილების სიმკვრივეა $n(x;0;1)$, მაშინ ის არის ნორმალურად განაწილებული შემთხვევითი სიდიდე მათემატიკური ლოდინით ნული და დისპერსიით ერთი. მას **სტანდარტული ნორმალური** შემთხვევითი სიდიდე ეწოდება. ასეთი შემთხვევითი სიდიდის განაწილების სიმკვრივის გრაფიკი სიმეტრიულია ორდინატთა ღერძის მიმართ, და მისთვის გვაქვს:

$$P(x_1<\xi<x_2)=\Phi(x_2)-\Phi(x_1).$$

სტანდარტული ნორმალური განაწილების **ზედა α-კრიტიკული წერტილი** ეწოდება ისეთ წერტილს აბსცისთა ღერძზე (მას აღნიშნავენ z_α სიმბოლოთი), რომლის მარჯვენივ სტანდარტული ნორმალური განაწილების სიმკვრივის წირის ქვეშ მოთავსებული არის ფართობი ტოლია α-სი:

$$P\{\xi>z_\alpha\}=\alpha.$$



დავუშვათ, რომ ξ და η – დამოუკიდებელი და ნორმალურად განაწილებული შემთხვევითი სიდიდეებია, ისეთი რომ $E\xi=a_1$, $D\xi=\sigma_1^2$, $E\eta=a_2$, $D\eta=\sigma_2^2$. მაშინ შემთხვევითი სიდიდე $\psi=c_1\xi+c_2\eta$ (სადაც c_1 და c_2 – ნებისმიერი მუდმივებია), აგრეთვე განაწილებულია ნორმალური კანონით. მისი მათემატიკური ლოდინი და დისპერსია გამოითვლება ფორმულებით:

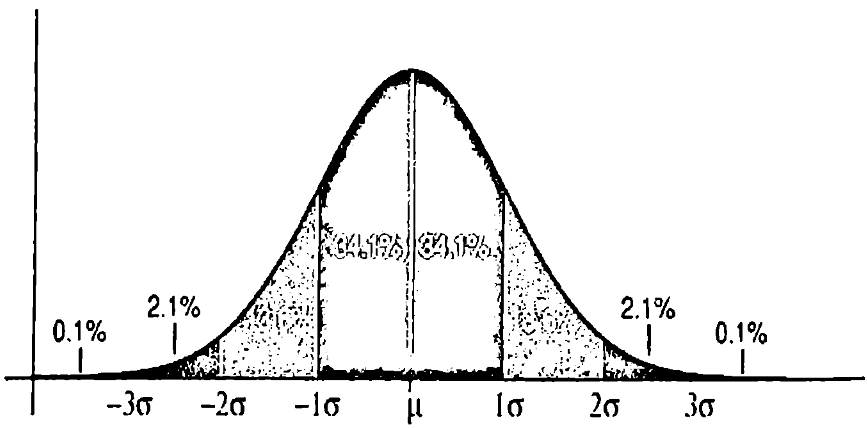
$E\psi=c_1a_1+c_2a_2$, $D\psi=c_1^2\sigma_1^2+c_2^2\sigma_2^2$.

ამოცანა. 12 ბოთლიანი ყუთის წონა – ნორმალურად განაწილებული შემთხვევითი სიდიდეა მათემატიკური ლოდინით 2კგ და საშუალოკვადრატული გადახრით 0.01კგ. ბოთლის მასა ლიმონათით – აგრეთვე ნორმალურად განაწილებული შემთხვევითი სიდიდეა მათემატიკური ლოდინით 0.8კგ და საშუალოკვადრატული გადახრით 0.04კგ. ვიპოვოთ ალბათობა იმისა, რომ ყუთის წონა 12 ბოთლი ლიმონათით მოთავსებული იქნება საზღვრებში: 11 კგ-დან 11.5კგ-მდე.

სამი σ–ს („სიგმას") წესი. დავუშვათ, მოცემულია ნორმალური კანონით განაწილებული ξ შემთხვევითი სიდიდე მათემატიკური ლოდინით μ და დისპერსიით σ^2. განვსაზღვროთ ξ შემთხვევითი სიდიდის ($\mu - 3\sigma$; $\mu + 3\sigma$) ინტერვალში მოხვედრის ალბათობა, ანუ ალბათობა იმისა, რომ ξ მიიღებს მნიშვნელობას, რომელიც მათემატიკური ლოდინიდან განსხვავდება არაუმეტეს სამი საშუალოკვადრატული გადახრით. ცხადია, რომ

$$P(\mu - 3\sigma < \xi < \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 2\Phi(3).$$

სტანდარტული ნორმალური განაწილების ფუნქციის (ლაპლასის ინტეგრალური ფუნქციის) ცხრილიდან ვპოულობთ, რომ $\Phi(3)=0,49865$, აქედან გამომდინარეობს, რომ $2\Phi(3)$ პრაქტიკულად ერთის ტოლია. ამრიგად, შეიძლება გაკეთდეს დასკვნა: ნორმალური შემთხვევითი სიდიდე ღებულობს მნიშვნელობებს, რომლებიც მისი მათემატიკური ლოდინიდან გადაიხრება არაუმეტეს 3σ-თი. ქვემოთ მოყვანილია ამ ფაქტის საილუსტრაციო ნახაზი, რომელზეც მითითებულია რომელ შუალედში რა ალბათობებით (პროცენტებში გამოსახული) ხვდება ნორმალური შემთხვევითი სიდიდე.



ხი კვადრატ განაწილება.

დავუშვათ, რომ მოცემულია n დამოუკიდებელი, ნორმალური კანონით განაწილებული $\xi_1, \xi_2, \ldots, \xi_n$ შემთხვევითი სიდიდე მათემატიკური ლოდინით ნული და დისპერსიით ერთი. მაშინ შემთხვევითი სიდიდე

$$\chi^2 = \sum_{i=1}^{n} \xi_i^2$$

განაწილებულია კანონით, რომელსაც ეწოდება „χ^2 განაწილება" ანუ „პირსონის განაწილება" თავისუფლების n ხარისხით. თუ შესაკრებები დაკავშირებულია რაიმე თანაფარდობით (მაგალითად, $\sum X_i = n\bar{X}$), მაშინ თავისუფლების ხარისხია $k = n - 1$.

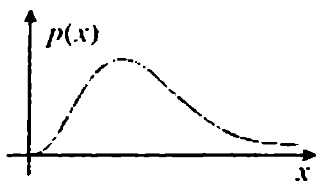
ამ განაწილების სიმკვრივეა

$$f(x)=\begin{cases} 0, & x\le 0; \\ \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\dfrac{k}{2}\right)}e^{-\frac{x}{2}}x^{\frac{k}{2}-1}, & x>0. \end{cases}$$

სადაც $\Gamma(x)=\int_0^{\infty} t^{x-1}e^{-t}dt$ – გამა ფუნქციაა, $\Gamma(n+1)=n!$.

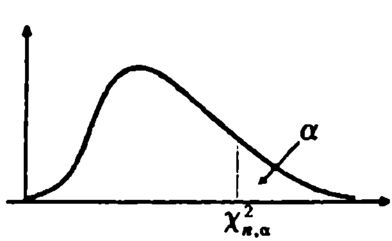
შესაბამისად, ხი კვადრატ განაწილება განისაზღვრება ერთი პარამეტრით, კერძოდ, თავისუფლების ხარისხის რიცხვით. აღსანიშნავია, რომ თავისუფლების ხარისხის რიცხვის ზრდასთან ერთად, ხი კვადრატ განაწილება თანდათანობით უახლოვდება ნორმალურ განაწილებას.

ცხადია, რომ χ^2 შემთხვევითი სიდიდე ღებულობს მხოლოდ არაუარყოფით მნიშვნელობებს. $n>1$ შემთხვევაში χ^2 შემთხვევითი სიდიდის განაწილების სიმკვრივის გრაფიკი გამოსახულია ქვემოთ მოყვანილ ნახაზზე:



იმისათვის, რომ განვსაზღვროთ χ^2 შემთხვევითი სიდიდის დადებით რიცხვთა სიმრავლიდან აღებულ რაიმე ინტერვალში მოხვედრის ალბათობა, გამოიყენება χ^2 განაწილების ცხრილი. ჩვეულებრივ, ასეთი ცხრილი საშუალებას იძლევა α ალბათობისა და თავისუფლების n ხარისხის მიხედვით განისაზღვროს ხი კვადრატ განაწილების **ზედა α-კრიტიკული წერტილი** χ^2_α, რომელიც განიმარტება შემდეგი თანაფარდობით:

$$P(\chi^2>\chi^2_\alpha)=\alpha.$$



ეს ფორმულა ნიშნავს შემდეგს: ალბათობა იმისა, რომ χ^2 შემთხვევითი სიდიდე მიიღებს მნიშვნელობას, რომელიც მეტია ვიდრე გარკვეული χ^2_α მნიშვნელობა, α-ს ტოლია.

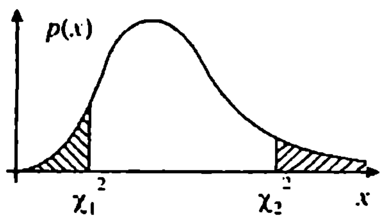
ქვემოთ მოყვანილია χ^2 განაწილების ცხრილის ერთი ფრაგმენტი. ის გვიჩვენებს, მაგალითად, რომ χ^2 შემთხვევითი სიდიდე თავისუფლების 10-ის ტოლი ხარისხით ალბათობით $\alpha = 0.95$ ღებულობს მნიშვნელობას მეტს, ვიდრე 3.94, ხოლო იგივე შემთხვევითი სიდიდე 1-ის ტოლი თავისუფლების ხარისხით ალბათობით $\alpha = 0.975$ არ აღემატება 0.00098-ს.

α / n	0.99	0.975	0.95		0.1	0.05	0.01
1	$0.0^3 15$	$0.0^3 98$	$0.0^2 39$	...	2.71	3.84	6.63
...	...	...	...	...	...	...	...
10	2.56	3.25	3.94	...	16.0	18.3	23.2
...	...	...	...	...	...	...	...

(აქ $0.0^3 15$ აღნიშნავს 0.00015, $0.0^2 39$=0.0039).

ამოცანა. ვიპოვოთ ისეთი ინტერვალი (χ_1^2, χ_2^2), რომელშიც χ^2 შემთხვევითი სიდიდე თავისუფლების 10-ის ტოლი ხარისხით მოხვდება ალბათობით 0.9.

ამოხსნა. ქვემოთ სქემატურად მოყვანილია თავისუფლების 10-ის ტოლი ხარისხის მქონე χ^2 შემთხვევითი სიდიდის განაწილების სიმკვრივის გრაფიკი.



ჩავთვალოთ, რომ დაშტრიხული არეების ფართობები (აქ მარჯვენა არე არ არის შემოსაზღვრული) ერთმანეთის ტოლია. თუ χ_1^2-სა და χ_2^2-ს შევარჩევთ პირობიდან

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_2^2) = (1 - 0.9)/2 = 0.05, \qquad (1)$$

მაშინ შესრულდება პირობა $P(\chi_1^2 < \chi^2 < \chi_2^2) = 0.9$.

(1) ტოლებები საშუალებას გვაძლევს χ^2 განაწილების ცხრილიდან განვსაზღვროთ: $\chi_2^2 = 18.3$. საძებნი ინტერვალის მარცხენა საზღვრის დასადგენად ვისარგებლოთ ტოლობით $P(\chi^2 > \chi_1^2) = 1\text{-}0.05=0.95$. მაშინ ცხრილიდან ვპოულობთ, რომ $\chi_1^2 = 3{,}94$, და ამიტომ ამოცანის პასუხი იქნება: χ^2 შემთხვევითი სიდიდის მნიშვნელობები ალბათობით 0.9 ეკუთვნის ინტერვალს (3.94, 18.3).

სტიუდენტის განაწილება.

სტატისტიკის ბევრ ამოცანას მივყავართ შემდეგი სახის შემთხვევით სიდიდემდე

$$t = \xi\sqrt{k}/\sqrt{\eta},$$

სადაც ξ და η დამოუკიდებელი შემთხვევითი სიდიდეებია, ამასთანავე ξ – ნორმალურად განაწილებული შემთხვევითი სიდიდეა პარამეტრებით $E\xi = 0$ და $D\xi = 1$, ხოლო η განაწილებულია χ^2 განაწილების კანონით თავისუფლების ხარისხით k. t შემთხვევითი სიდიდის განაწილების კანონს ეწოდება **სტიუდენტის განაწილება** თავისუფლების ხარისხით k.

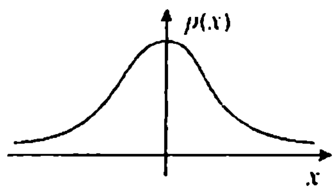
სტიუდენტის განაწილების სიმკვრივეს აქვს შემდეგი სახე:

$$s(t,n) = B_n\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}$$

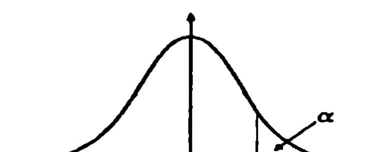
სადაც

$$B_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi(n-1)}\,\Gamma\left(\frac{n-1}{2}\right)}.$$

სტიუდენტის განაწილების სიმკვრივის გრაფიკი სქემატურად გამოსახულია ქვემოთ მოყვანილ ნახაზზე:



განაწილების სიმკვრივის წირი მსგავსია ნორმალური განაწილების ანალოგიური წირის.



სტიუდენტის განაწილების **ზედა α-კრიტიკული წერტილი** აღინიშნება $t_{n,\alpha}$ სიმბოლოთი. ეს ისეთი წერტილია, რომლის მარცხნივ, სტიუდენტის განაწილების სიმკვრივის ქვეშ, მოქცეული არის ფართობის ტოლია α-სი.

სტიუდენტის განაწილების ცხრილები საშუალებას იძლევა თავისუფლების ხარისხის მოცემული n რიცხვისათვის α ალბათობის მიხედვით განისაზღვროს ისეთი მნიშვნელობა t_α, რომლისთვისაც სრულდება თანა-

ფარდობა $P(|t|>t_\alpha)=\alpha$. ამ ცხრილის ერთი ფრაგმენტი მოყვანილია ქვემოთ:

α / n	0.1	0.05		0.01	0.005	
1	6.314	12.71		63.57	318	
12	1.782	2.179		3.055	3.428	

ამოცანა 1. ვიპოვოთ სიმეტრიული ინტერვალი, რომელშიც სტიუდენტის კანონით გათვალისწინებული შემთხვევითი სიდიდე თავისუფლების ხარისხით 12, მოხვდება ალბათობით 0.9.

ამოხსნა. ცხადია, რომ

$$P(-x<t<x)=P(|t|<x)=1-P(|t|\geq x)=0.9.$$

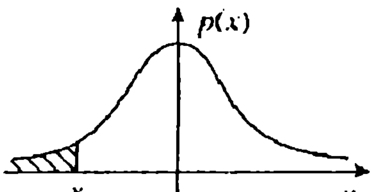
უკანასკნელი ტოლობიდან გამომდინარეობს, რომ:

$$P(|t|\geq x)=0.1\ ,\ (n=12).$$

არ უნდა გაგვიკვირდეს, რომ აქ გვაქვს არამკაცრი უტოლობა. რამდენადაც საქმე გვაქვს უწყვეტ შემთხვევით სიდიდესთან, ის კონკრეტულ მნიშვნელობას ღებულობს ნულოვანი ალბათობით. ამიტომ არამკაცრი უტოლობა იცვლება მისი ექვივალენტური მკაცრი უტოლობით. ცხრილიდან ვადგენთ, რომ: $x=1.782$.

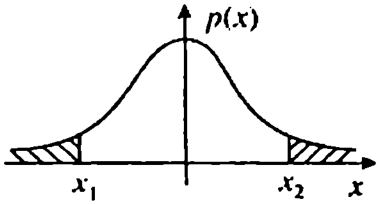
ამოცანა 2. ვიპოვოთ მნიშვნელობა x პირობიდან $P(t>x)=0.995$, სადაც t — სტიუდენტის კანონით განაწილებული შემთხვევითი სიდიდეა 12-ის ტოლი თავისუფლების ხარისხით.

ამოხსნა. ქვემოთ მოყვანილია თავისუფლების 12 ხარისხის მქონე სტიუდენტის განაწილების სიმკვრივის გრაფიკი:



ალბათობა იმისა, რომ შემთხვევითი სიდიდე მიიღებს მნიშვნელობას x_1 წერტილის მარჯვნივ მდებარე არიდან ტოლია 0.995-ის, შესაბამისად, ამ წერტილის მარცხნივ მდებარე არეში შემთხვევითი სიდიდე მოხვდება 0.005-ის ტოლი ალბათობით. იმისათვის, რომ ვიპოვოთ x_1, განვიხილოთ

ორი სიმეტრიული არე, რომლებიც გამოსახულია ქვემოთ მოყვანილ ნახაზზე:



დავუშვათ, რომ თითოეულ ამ შუალედში, შემთხვევითი სიდიდის მნიშვნელობა აღმოჩნდება ალბათობით 0.005. მაშინ მივიღებთ: $x_1=-x$, $x_2=x$, ამასთანავე x განისაზღვრება პირობიდან $P(|t|>x)=0.01$. ცხრილიდან გვაქვს, რომ: $x=3.055$. ამიტომ ამოცანის პასუხი იქნება:

$$P(t>-3.055)=0.995.$$

ფიშერის განაწილება.

სტატისტიკაში მნიშვნელოვანი გამოყენებები გააჩნია შემთხვევით სიდიდეს

$$F=\left(\frac{\xi}{n}\right)/\left(\frac{\eta}{m}\right)=\frac{m\xi}{n\eta},$$

სადაც ξ – შემთხვევითი სიდიდე განაწილებულია χ^2 განაწილების კანონით თავისუფლების ხარისხით n, ხოლო η– შემთხვევითი სიდიდე განაწილებულია χ^2 განაწილების ხარისხით თავისუფლების ხარისხით m, ამასთანავე ξ და η შემთხვევითი სიდიდეები დამოუკიდებელია.

F შემთხვევითი სიდიდე განაწილებულია კანონით, რომელსაც ეწოდება **ფიშერის განაწილება** თავისუფლების ხარისხებით n და m. მისი განაწილების სიმკვრივეს აქვს შემდეგი სახე:

$$f(x)=\begin{cases} 0, & x\le 0; \\ C_0\dfrac{x^{\frac{n-2}{2}}}{(m+nx)^{\frac{n+m}{2}}}, & x>0, \end{cases}$$

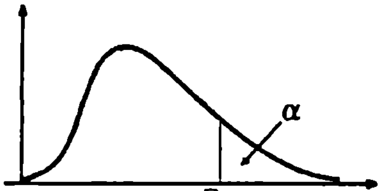
სადაც

$$C_0=\frac{\Gamma\left(\frac{n+m}{2}\right)n^{\frac{n}{2}}m^{\frac{m}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}.$$

ამრიგად, ფიშერის განაწილება განისაზღვრება ორი პარამეტრით, კერძოდ, თავისუფლების ხარისხების რიცხვებით.

მოცემული n და m რიცხვებისათვის, და მოცემული α ალბათობით ფიშერის განაწილების ცხრილიდან განისაზღვრება ისეთი მნიშვნელობა F_α, რომ

$$P(F > F_\alpha) = \alpha$$



ასეთი მნიშვნელობას ფიშერის განაწილების ზედა α-კრიტიკული წერტილი ეწოდება და აღინიშნება $F_{n,m,\alpha}$ სიმბოლოთი.

როგორც წესი, ცხრილები დგება α-ს მიმნიშვნელობებისათვის, რომელიც ტოლია 0.05-ის ან 0.01-ის, ხოლო ზოგჯერ ორივე მნიშვნელობისათვის. ამ ცხრილის ერთი ფრაგმენტი მოყვანილია ქვემოთ:

n m	1		10		20	
1	161.4 647.8		241.9 6056		248 6209	
...	...	...	...	...	...	...
10	4.96 10.04		2.97 4.85		2.77 4.41	
...	...	...	...	...	...	...

ამ ცხრილში ყოველი უჯრის ზედა ნაწილში მოცემულია F_α-ს მნიშვნელობა, როცა $\alpha = 0.05$, ხოლო ქვედა ნაწილში კი როცა $\alpha = 0.01$.

თავი VI
ალბათობის თეორიის ზღვარითი თეორემები

ჩებიშევის უტოლობა.

ჩებიშევის უტოლობა აფასებს შემთხვევითი სიდიდის გადახრას თავისი მათემატიკური ლოდინიდან. თუ X რაიმე შემთხვევითი სიდიდეა, მაშინ ნებისმიერი $\varepsilon > 0$ რიცხვისათვის სამართლიანია უტოლობა:

$$p(|X - E(X)| < \varepsilon) \ge D(X)/\varepsilon^2. \quad 1-$$

ამ უტოლობას **ჩებიშევის უტოლობას** უწოდებენ. იგი სამართლიანია როგორც დისკრეტული, ისე უწყვეტი შემთხვევითი სიდიდეებისათვის.

შევამოწმოთ ჩებიშევის უტოლობა დისკრეტულ შემთხვევაში. დავუშვათ, რომ შემთხვევითი სიდიდე მოცემულია განაწილების კანონით:

X	x_1	x_2	...	x_n
p	p_1	p_2	...	p_n

ვინაიდან $|X - E(X)| < \varepsilon$ და $|X - E(X)| \ge \varepsilon$ საწინააღმდეგო ხდომილებებია, ამიტომ $p(|X - E(X)| < \varepsilon) + p(|X - E(X)| \ge \varepsilon) = 1$, შესაბამისად,

$$p(|X - E(X)| < \varepsilon) = 1 - p(|X - E(X)| \ge \varepsilon).$$

ვიპოვოთ $p(|X - E(X)| \ge \varepsilon)$. დისპერსიის განმარტების თანახმად გვაქვს:

$$D(X) = (x_1 - E(X))^2 p_1 + (x_2 - E(X))^2 p_2 + \ldots + (x_n - E(X))^2 p_n.$$

გადავაგდოთ ამ ჯამიდან ის შესაკრებები, რომელთათვისაც $|X - M(X)| < \varepsilon$. ამის შედეგად ჯამი მხოლოდ შემცირდება, ვინაიდან ყველა მასში შემავალი შესაკრები არაუარყოფითია. გარკვეულობისათვის ვიგულისხმოთ, რომ გადაგდებულია პირველი k შესაკრები. მაშინ

$$D(X) \ge (x_{k+1} - E(X))^2 p_{k+1} + (x_{k+2} - E(X))^2 p_{k+2} + \cdots + (x_n - E(X))^2 p_n \ge$$
$$\le \varepsilon^2 (p_{k+1} + p_{k+2} + \cdots + p_n).$$

შევნიშნოთ, რომ $p_{k+1} + p_{k+2} + \ldots + p_n$ არის ალბათობა იმისა, რომ $|X - E(X)| \ge \varepsilon$, ვინაიდან ეს არის ჯამი X შემთხვევითი სიდიდის ყველა შესაძლო მნიშვნელობის, რომელთათვისაც აღნიშნული უტოლობა სამართლიანია. შესაბამისად, $D(X) \ge \varepsilon^2 p(|X - E(X)| \ge \varepsilon)$, ანუ $p(|X - E(X)| \ge \varepsilon) \le D(X)/\varepsilon^2$. მაშინ საწინააღმდეგო ხდომილების ალბათობა იქნება

$$p(|X - E(X)| < \varepsilon) \ge D(X)/\varepsilon^2. \quad ■$$
1-

დიდ რიცხვთა კანონი.

ჩებიშევის უტოლობა საშუალებას იძლევა დავამტკიცოთ ფუნდამენტური შედეგი, რომელიც საფუძვლად უდევს მათემატიკურ სტატისტიკას – ე. წ. *დიდ რიცხვთა კანონი.* ამ შედეგის თანახმად შერჩევითი მახასიათებლები ცდების (ექსპერიმენტების) რიცხვთა ზრდისას უახლოვდება თეორიულ მახასიათებლებს, რაც საშუალებას იძლევა ამა თუ იმ რეალური მოვლენის ალბათური მოდელების პარამეტრები შევაფასოთ ცდების მიერ მიღებული შედეგების გამოყენებით. დიდ რიცხვთა კანონის გარეშე არ გვექნებოდა გამოყენებითი მათემატიკური სტატისტიკის მნიშვნელოვანი ნაწილი.

სტატისტიკური კანონზომიერებების შესწავლა საშუალებას იძლევა დავადგინოთ, რომ გარკვეულ პირობებში შემთხვევით სიდიდეთა დიდი რაოდენობის ჯამური ქცევა (ეფექტი) თითქმის კარგავს შემთხვევით ხასიათს და ხდება კანონზომიერი (სხვა სიტყვებით რომ ვთქვათ, ურთიერთ ჩაიხშობა შემთხვევითი გადახრები გარკვეული საშუალო ქცევიდან). კერძოდ, თუ ცალკეული შესაკრებების გავლენა ჯამზე თანაბრად მცირეა, მაშინ ჯამის განაწილების კანონი უახლოვდება ნორმალურს. ამ მტკიცებულების მათემატიკური ფორმულირება ატარებს სწორედ დიდ რიცხვთა კანონის სახელს. მოვიყვანოთ ერთ-ერთი ამ ტიპის მტკიცებულება.

ჩებიშევის თეორემა. ვთქვათ, შემთხვევითი სიდიდეები $X_1, X_2, ..., X_n$ წყვილ-წყვილად დამოუკიდებელია და არსებობს ისეთი რიცხვი C, რომ $DX_i \le C$, $i = 1, 2, ..., n$. მაშინ ნებისმიერი დადებითი ε რიცხვისათვის სრულდება უტოლობა:

$$P\{|\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{EX_1 + EX_2 + \cdots + EX_n}{n}| \ge \varepsilon\} \le \frac{C}{n\varepsilon^2}. \qquad (1)$$

დამტკიცება. განვიხილოთ შემთხვევითი სიდიდეები $Y_n = X_1 + X_2 + \cdots + X_n$ და $Z_n = Y_n / n$. მათემატიკური ლოდინისა და დისპერსიის თვისებების ძალით გვექნება შემდეგი თანაფარდობები:

$$EY_n = EX_1 + EX_2 + \cdots + EX_n, \quad DY_n = DX_1 + DX_2 + \cdots + DX_n.$$

გარდა ამისა,

$$EZ_n = EY_n / n \text{ და } DZ_n = DY_n / n^2$$

შესაბამისად,

$$EY_n = [EX_1 + EX_2 + \cdots + EX_n]/n \text{ და } DY_n = [DX_1 + DX_2 + \cdots + DX_n]/n^2.$$

ამიტომ თეორემის პირობებში გვაქვს:

$$DY_n = Cn/n^2 = C/n.$$

თუ ახლა Z_n შემთხვევითი სიდიდისათვის გამოვიყენებთ ჩებიშევის უტოლობას, მაშინ (1) თანაფარდობის მარცხენა მხარისათვის მივიღებთ შეფასებას:

$$P\{|Z_n - EZ_n| \ge \varepsilon\} \le \frac{DZ_n}{\varepsilon^2} \le \frac{C}{n\varepsilon^2}. \qquad \blacksquare$$

ეს თეორემა, ისევე როგორც, საკუთრივ ჩებიშევის უტოლობები მიღებულ იქნა პ. ჩებიშევის მიერ 1867 წელს გამოქვეყნებულ ნაშრომში: "საშუალო მნიშვნელობების შესახებ".

მაგალითი 1. ვთქვათ, $C = 1$ და $\varepsilon = 0{,}1$. n-ის რომელი მნიშვნელობისათვის არ აღემატება (1) უტოლობის მარჯვენა მხარე 0.1-ს?, 0.05-ს?, 0.00001-ს?

განსახილველ შემთხვევაში (1) უტოლობის მარჯვენა მხარე ტოლია $100/n$ -ის. შესაბამისად, ის არ აღემატება 0.1-ს, თუ n არაა ნაკლები 1000-ზე, არ აღემატება 0.05-ს, თუ n არაა ნაკლები 2000-ზე და არ აღემატება 0.00001-ს, თუ n არაა ნაკლები 10 000 000-ზე.

(1) უტოლობის მარჯვენა მხარე, და მასთან ერთად მარცხენაც, n-ის ზრდასთან ერთად, ფიქსირებული C და ε-ის შემთხვევაში, უახლოვდება

ნულს. შესაბამისად, ალბათობა იმისა, რომ დამოუკიდებელი შემთხვევითი სიდიდეების საშუალო არითმეტიკული თავისი მათემატიკური ლოდინისაგან განსხვავდება ε-ზე ნაკლებით, უახლოვდება 1-ს შემთხვევით სიდიდეთა რიცხვის ზრდასთან ერთად, ნებისმიერი ε-ის შემთხვევაში. ამ მტკიცებულებას უწოდებენ **დიდ რიცხვთა კანონს.**

გადაწყვეტილებების მიღების ალბათურ-სტატისტიკური მეთოდებისათვის (და მთლიანად მათემატიკური სტატისტიკისათვის) განსაკუთრებით მნიშვნელოვანია ის შემთხვევა, როცა ყველა X_i შემთხვევით სიდიდეს ($i=1,2,...$) გააჩნია ერთი და იგივე მათემატიკური ლოდინი EX_1 და ერთი და იგივე დისპერსია $\sigma^2=DX_1$. მკვლევარისათვის უცნობი მათემატიკური ლოდინის ნაცვლად (შეფასებად) იყენებენ შერჩევით საშუალო არითმეტიკულს:

$$\bar{X}=\frac{X_1+X_2+\cdots+X_n}{n}.$$

დიდ რიცხვთა კანონიდან გამომდინარეობს, რომ ცდების (ექსპერიმენტების, გაზომვების) რიცხვის ზრდასთან ერთად $\bar{X}$ რაგინდ ახლოს უახლოვდება EX_1-ს, რაც მოკლედ ასე ჩაიწერება:

$$\bar{X}\xrightarrow{P}EX_1,\ \text{როცა}\ n\to\infty.$$

სიმბოლო $\xrightarrow{P}$ აღნიშნავს "ალბათობით კრებადობას". საჭიროა აღინიშნოს, რომ "ალბათობით კრებადობის" ცნება განსხვავდება მათემატიკურ ანალიზში მიღებული "ზღვარზე გადასვლის" ცნებისაგან. გავიხსენოთ, რომ a_n რიცხვით მიმდევრობას აქვს ზღვარი a, როცა $n\to\infty$, თუ ნებისმიერი, რაგინდ მცირე, $\delta>0$ რიცხვისათვის, არსებობს ისეთი $n(\delta)$ რიცხვი, რომ ყოველი $n>n(\delta)$ ნომრისათვის სრულდება თანაფარდობა: $a_n\in(a-\delta,a+\delta)$. "ალბათობით კრებადობის" ცნების გამოყენებისას მიმდევრობის წევრები Y_n შემთხვევითი სიდიდეებია, ვიხილავთ რაგინდ მცირე $\varepsilon>0$ რიცხვს და თანაფარდობა $Y_n\in(a-\varepsilon,a+\varepsilon)$ იგულისხმება რომ სრულდება არა გარანტირებულად, არამედ არანაკლებ $(1-\varepsilon)$-ის ტოლი ალბათობით.

სიხშირეების კრებადობა ალბათობებისაკენ. როგორც აღნიშნული იყო რაიმე A ხდომილების ალბათობა – ეს არის ის რიცხვი, რომელსაც უახლოვდება A ხდომილების მოხდენათა რიცხვის შეფარდება ექსპერიმენტების საერთო რიცხვთან, როცა ექსპერიმენტების რიცხვი უსასრულოდ იზრდება. ეს დებულება, მათემატიკური მოდელის ჩარჩოებში, მე-17 საუკუნის მიწურულს პირველად დაამტკიცა ცნობილმა მათემატიკოსმა იაკობ ბერნულმა, მაგრამ დამტკიცება გამოქვეყნებულ იქნა ი. ბერნულის სიკვდილის შემდეგ 1713 წელს (ი. ბერნული ცხოვრობდა შვეიცარიის ქალაქ ბაზელში 1654-1705 წლეებში). ბერნულის თეორემის თანამედროვე ფორმულირება შემდეგია:

ბერნულის თეორემა. დავუშვათ, m არის n დამოუკიდებელ ექსპერიმენტში A ხდომილების მოხდენათა რიცხვი, ხოლო p არის A ხდომილ-

ების მოხდენის ალბათობა ცალკეულ ექსპერიმენტში. მაშინ ნებისმიერი $\varepsilon > 0$ რიცხვისათვის სამართლიანია უტოლობა

$$P\{|\frac{m}{n} - p| \geq \varepsilon\} \leq \frac{p(1-p)}{n\varepsilon^2}. \qquad (2)$$

დამტკიცება. როგორც ჩვენ უკვე ვნახეთ m შემთხვევით სიდიდეს აქვს ბინომიალური განაწილება წარმატების ალბათობით p და იგი წარმოიდგინება n დამოუკიდებელი X_i, $i=1,2,\dots,n$ შემთხვევითი სიდიდის ჯამის სახით, რომელთაგან თითოეული ბერნულის შემთხვევითი სიდიდეა: X_i ტოლია 1-ის ალბათობით p და ტოლია 0-ის ალბათობით $1-p$, ანუ $m = X_1 + X_2 + \cdots + X_n$. თუ ახლა გამოვიყენებთ ჩებიშევის თეორემას $X_1, X_2, \dots, X_n$ შემთხვევითი სიდიდეებისათვის, სადაც $C = p(1-p)$, ადვილად დავრწმუნდებით (2) უტოლობის სამართლიანობაში. ■

შენიშვნა. ვინაიდან $1/4 - p(1-p) = (p - 1/2)^2 \geq 0$, შესაბამისად, $p(1-p) \leq 1/4$, ამიტომ ჩებიშევის თეორემაში ჩვენ შეგვეძლო აგვეღო $C = 1/4$. მაშინ ნებისმიერი p-სა და ფიქსირებული $\varepsilon > 0$ რიცხვისათვის (2) უტოლობის მარჯვენა მხარე n-ის ზრდასთან ერთად უახლოვდება ნულს. ეს კი თავის მხრივ გვიჩვენებს, რომ ალბათობის მათემატიკური განმარტება (მაგალითად, ა. ნ. კოლმოგოროვის მიხედვით) სრულ თანხვედრაშია ბუნებისმეტყველთა (მაგალითად, პ. მიზესის (1883-1953)) მიერ მოყვანილ განმარტებასთან, რომლის თანახმად ალბათობა არის სიხშირეების ზღვარი ექსპერიმენტების უსასრულო მიმდევრობაში.

რაც შეეხება პირდაპირ ექსპერიმენტალურ დადასტურებას იმისა, რომ გარკვეული ხდომილებების განხორციელების სიხშირეები ახლოსაა ალბათობებთან, რომლებიც განიმარტება თეორიული მოსაზრებებით, ეს ჩვენ ადრე უკვე მოვიყვანეთ შესავალ ნაწილში სხვადასხვა დროს სხვადასხვა მეცნიერის მიერ სიმეტრიული მონეტის აგდების მაგალითზე. ასე მაგალითად, მე-18 საუკუნეში ფრანგი მეცნიერის ბიუფონის მიერ მონეტის 4040-ჯერ აგდებისას გერბის მისვლის ფარდობითი სიხშირე იყო 0.507; ინგლისელი სტატისტიკოსის კ. პირსონის მიერ მონეტის 12000-ჯერ აგდებისას შესაბამისი სიხშირე აღმოჩნდა 0.5016, ხოლო მის მიერვე მონეტის 24000-ჯერ აგდებისას კი – 0.5005. ყველა შემთხვევაში სიხშირეები მხოლოდ უმნიშვნელოდ განსხვავდებოდნენ თეორიული ალბათობისაგან, რომელიც 0.5-ის ტოლია (ვინაიდან სიმეტრიული მონეტის შემთხვევაში გერბისა და საფასურის მოსვლა თანაბრადშესაძლებელია).

სტატისტიკური ჰიპოთეზების შემოწმების შესახებ. (2) უტოლობის გამოყენებით ჩვენ შეგვიძლია გარკვეული დასკვნები გამოვიტანოთ პროდუქციის ხარისხის წინასწარ მოცემულ მოთხოვნებთან შესაბამისობასთან დაკავშირებით.

დავუშვათ, რომ პროდუქციის 100 000 ერთეულიდან 30 000 აღმოჩნდა დეფექტური. ეთანხმება თუ არა ეს ჰიპოთეზას იმის შესახებ, რომ პროდუქციის დეფექტურობის ალბათობა ტოლია 0,23-ის? როგორი ალბათური მოდელის გამოყენებაა მიზანშეწონილი? ჩავთვალოთ, რომ ტარდება რთული ცდა, რომელიც შედგება 100 000 ექსპერიმენტისაგან, რომელთაგან თითო-

ული გულისხმობს პროდუქციის 100 000 ერთეულიდან ცალკეულის შემოწმებას გამოსადეგიანობაზე. ითვლება, რომ ექსპერიმენტები წყვილ-წყვილად დამოუკიდებელია და ყოველ ექსპერიმენტში ალბათობა იმისა, რომ პროდუქციის ერთეული დეფექტურია ტოლია p-სი.

რეალურ ცდაში მიღებულია, რომ ხდომილება "პროდუქციის ერთეული დეფექტურია" განხორციელდა 30 000-ჯერ 100 000 ექსპერიმენტში. ეთანხმება თუ არა ეს ჰიპოთეზას იმის შესახებ, რომ პროდუქციის დეფექტურობის ალბათობაა 0,23?

ვისარგებლოთ (2) უტოლობით. განსახილველ შემთხვევაში $n=100000$, $m=30000$, $m/n=0.3$, $p=0.23$, $m/n-p=0.07$. ჰიპოთეზის შესამოწმებლად იქცევიან შემდეგნაირად. შევაფასოთ ალბათობა იმისა, რომ m/n განსხვავდება p-სა-გან ისევე როგორც განსახილველ შემთხვევაში, ან უფრო მეტით, ე. ი. შევაფასოთ $|m/n-p|\geq 0.07$ უტოლობის შესრულების ალბათობა. ვიგულისხმოთ, რომ (2) უტოლობაში $p=0.23$ და $\varepsilon=0.07$. მაშინ ბერნულის თეორემის თანახმად

$$P\{|\frac{m}{n}-0.23|\geq 0.07\}\leq\frac{0.23\cdot 0.77}{0.0049n}\approx\frac{36.11}{n}. \qquad (3)$$

როცა $n=100000$ (3) უტოლობის მარჯვენა მხარე ნაკლებია 1/2500. ამიტომ ალბათობა იმისა, რომ გადახრა იქნება არანაკლები დაკვირვებულზე, ერთობ მცირეა. შესაბამისად, თუ საწყისი ჰიპოთეზა სამართლიანია, მაშინ განსახილველ ცდაში განხორციელდა ხდომილება, რომლის ალბათობა ნაკლებია 1/2500. ვინაიდან, 1/2500 – ძალიან პატარა რიცხვია, ამიტომ საწყისი ჰიპოთეზა უნდა უკუვაგდოთ (უარვყოთ).

ცენტრალური ზღვარითი თეორემა.

დიდ რიცხვთა კანონი არ იკვლევს შემთხვევით სიდიდეთა ჯამის განაწილების კანონის სახეს. ეს საკითხი შეისწავლება თეორემების ჯგუფში, რომლებსაც **ცენტრალური ზღვარითი თეორემა** ეწოდება. ეს თეორემა ამტკიცებს, რომ შემთხვევით სიდიდეთა ჯამის განაწილების კანონი, რომელთაგან ცალკეულ შესაკრებს შეიძლება ჰქონდეს განსხვავებული განაწილება, უახლოვდება ნორმალურს შესაკრებთა საკმაოდ დიდი რიცხვის შემთხვევაში. ამით აიხსნება ნორმალური განაწილების კანონის უაღრესად დიდი მნიშვნელობა პრაქტიკულ გამოყენებებში.

ცენტრალური ზღვარითი თეორემა ერთნაირად განაწილებული შემთხვევითი სიდიდეებისათვის ასე ჩამოყალიბდება.

თეორემა 1. თუ $X_1, X_2,\ldots, X_n,\ldots$ – დამოუკიდებელი შემთხვევითი სიდიდეების მიმდევრობაა, ერთი და იგივე განაწილების კანონით, მათემატიკური ლოდინით m და დისპერსიით σ^2, მაშინ n –ის უსასრულოდ ზრდისას $Y_n=\sum_{k=1}^{n}X_k$ ჯამის განაწილების კანონი უახლოვდება ნორმალურ განაწილების კანონს:

$$\lim_{n\to\infty}P\{\frac{Y_n-m}{\sigma\sqrt{n}}\leq x\}=N(x;0;1).$$

ა. ლიაპუნოვმა დაამტკიცა ცენტრალური ზღვარითი თეორემა უფრო ზოგად შემთხვევაში.

თეორემა 2 (ლიაპუნოვის თეორემა). თუ X შემთხვევითი სიდიდე წარმოადგენს დამოუკიდებელ შემთხვევით სიდიდეთა ძალიან დიდი რიცხვის ჯამს, რომელთათვისაც შესრულებულია პირობა:

$$\lim_{n\to\infty}\left(\sum_{k=1}^{n} b_k\right)/\left[\left(\sum_{k=1}^{n} D_k\right)^{\frac{3}{2}}\right]=0,$$

სადაც b_k – მესამე რიგის აბსოლუტური ცენტრალური მომენტია X_k შემთხვევითი სიდიდის, ხოლო D_k – მისი დისპერსია, მაშინ X შემთხვევით სიდიდეს გააჩნია განაწილება, რომელიც ახლოსაა ნორმალურ განაწილებასთან.

შევნიშნავთ, რომ ლიაპუნოვის თეორემის პირობა ნიშნავს იმას, რომ ცალკეული შესაკრების გავლენა ჯამზე მიზერულია.

აღსანიშნავია, რომ ცენტრალური ზღვარითი თეორემა პრაქტიკულად შესაძლებელია გამოყენებულ იქნეს შემთხვევით სიდიდეთა საკმაოდ არა დიდი რიცხვის შემთხვევაში. გამოცდილება აჩვენებს, რომ თუნდაც 10 ან უფრო ნაკლები შესაკრებების რაოდენობის შემთხვევაშიც ჯამის განაწილება შესაძლებელია შეცვლილ იქნეს ნორმალურით.

მუავრ-ლაპლასის თეორემა. ცენტრალური ზღვარითი თეორემის კერძო შემთხვევას დისკრეტული შემთხვევითი სიდიდეების შემთხვევაში წარმოადგენს მუავრ-ლაპლასის თეორემა.

თეორემა 3 (მუავრ-ლაპლასის თეორემა). თუ ტარდება n დამოუკიდებელი ცდა, რომელთაგან თითოეულში A ხდომილება ხდება ალბათობით p, $np>15$, მაშინ სამართლიანია თანაფარდობა:

$$p\left(\alpha<\frac{Y-np}{\sqrt{npq}}<\beta\right)\approx\Phi(\beta)-\Phi(\alpha),$$

სადაც Y – A ხდომილების მოხდენათა რიცხვია n ცდაში, $q=1-p$, ხოლო

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

(ამ ფუნქციის მნიშვნელობები მოყვანილია სპეციალურ ცხრილებში, ამასთანავე $\Phi(-x)=1-\Phi(x)$).

დამტკიცება. შეგვიძლია ჩავთვალოთ, რომ $Y=\sum_{i=1}^{n}X_i$, სადაც X_i – A ხდომილების მოხდენათა რიცხვია i –ურ ცდაში (ანუ ბერნულის შემთხვევითი სიდიდეებია მნიშვნელობებით 0 ან 1). მაშინ თეორემა 1-ის თანახმად $Z=\dfrac{Y-m_y}{\sigma_y}$ შემთხვევითი სიდიდე შეიძლება ჩაითვალოს ნორმალური კანონით განაწილებულ, ნორმირებულ (სტანდარტიზებულ) შემთხვევით სიდიდედ. შესაბამისად, მისი (α, β) ინტერვალში მოხვედრის ალბათობა გამოითვლება ფორმულით

$$p(\alpha<Z<\beta)=\Phi(\beta)-\Phi(\alpha).$$

ვინაიდან Y შემთხვევით სიდიდეს აქვს ბინომიალური განაწილება,

$$m_y = np,\ D_y = npq.\ \sigma_y = \sqrt{npq}.$$

ამიტომ $Z = \dfrac{Y - np}{\sqrt{npq}}$. თუ ჩავსვამთ ამ გამოსახულებას წინა ფორმულაში, მივიღებთ დასამტკიცებელ თანაფარდობას.

ეს თეორემა ლიტერატურაში აგრეთვე ცნობილია მუავრ-ლაპლასის ინტეგრალური თეორემის სახელწოდებით.

შედეგი (მუავრ-ლაპლასის ლოკალური თეორემა). მუავრ-ლაპლასის თეორემის პირობებში $p_n(k)$ — ალბათობა იმისა, რომ A ხდომილება n ცდაში მოხდება ზუსტად k-ჯერ, ცდათა დიდი რიცხვის შემთხვევაში, თუ $np > 15$, შეიძლება გამოითვალოს შემდეგი ფორმულით:

$$p_n(k) \approx \frac{1}{\sqrt{npq}} \cdot \varphi(x),$$

სადაც $x = \dfrac{k - np}{\sqrt{npq}}$, ხოლო $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (ამ ფუნქციის მნიშვნელობები მოყვანილია სპეციალურ ცხრილებში, ამასთანავე $\varphi(-x) = \varphi(x)$).

შემოვიღოთ განაწილების ფუნქციებისათვის შემდეგი აღნიშვნები:

ჰიპერგეომეტრიული განაწილების ფუნქცია –

$$H(x; l, s, n) = \sum_{k \le x} \frac{C_l^k \cdot C_{n-l}^{s-k}}{C_n^s};$$

ბინომური განაწილების ფუნქცია –

$$Bi(x; p, n) = \sum_{k \le x} C_n^k p^k q^{n-k};$$

პუასონის განაწილების ფუნქცია –

$$\Pi(x; \lambda) = \sum_{0 \le k \le x} \frac{\lambda^k}{k!} e^{-\lambda};$$

სტანდარტული ნორმალური განაწილების ფუნქცია –

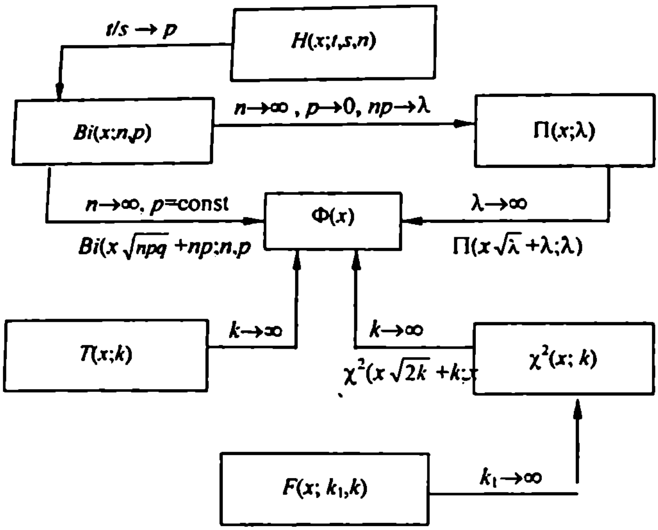
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt;$$

ხი კვადრატ განაწილების ფუნქცია (იხ. §. 5) – $\chi^2(x; k)$;

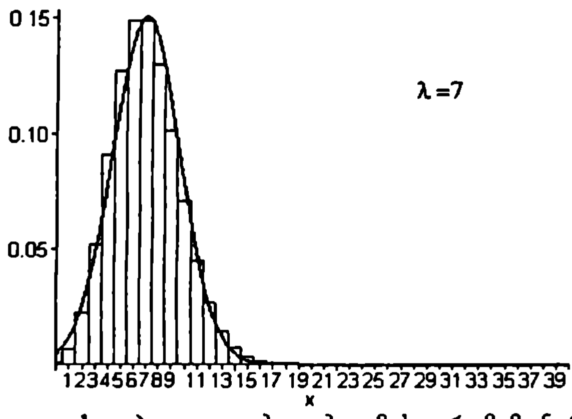
სტიუდენტის განაწილების ფუნქცია (იხ. §. 5) – $T(x; k)$;

ფიშერის განაწილების ფუნქცია (იხ. §. 5) – $F(x; k_1, k_2)$.

აღსანიშნავია, რომ ზემოთ მოყვანილი განაწილებების ფუნქციებს შორის ადგილი აქვს ქვემოთ მოყვანილ სქემაზე გამოსახულ ზღვრულ თანაფარდობებს:



ქვემოთ მოყვანილ ნახაზზე შედარებულია ნორმალური და პუასონის განაწილებები $\lambda = 7$-ის შემთხვევაში:



მაგალითი 1. ვიპოვოთ ალბათობა იმისა, რომ მონეტის 100-ჯერ აგდებისას გერბთა მოსვლის რიცხვი აღმოჩნდება საზღვრებში 40-დან 60-მდე.

ვისარგებლოთ მუავრ-ლაპლასის ინტეგრალური თეორემით. ჩვენს შემთხვევაში $p = 0.5$, $np = 100 \cdot 0.5 = 50$. ამიტომ თუ $40 < Y < 60$, მაშინ

$$-2 < \frac{Y-50}{5} < 2.$$

შესაბამისად, გვაქვს:

$$p(40 < Y < 60) = p\left(-2 < \frac{Y-50}{5} < 2\right) = \Phi(2) - \Phi(-2) = 0,9772 - 0,0228 = 0,9544.$$

მაგალითი 2. წინა მაგალითის პირობებში ვიპოვოთ ალბათობა იმისა, რომ გერბი მოვა 45-ჯერ.

ამ შემთხვევაში $x = \frac{k-np}{\sqrt{npq}} = \frac{45-50}{5} = -1$, ამიტომ

$$p_{100}(45) \approx \frac{1}{5} \cdot \varphi(-1) = \frac{1}{5} \cdot \varphi(1) = \frac{1}{5} \cdot 0.2420 = 0.0484.$$

სტატისტიკური დასკვნების თეორია

თავი I
მათემატიკური სტატისტიკის ძირითადი ცნებები

მათემატიკური სტატისტიკის ძირითადი მიზანია ისეთი მეთოდების დამუშავება, რომლებიც დაკვირვებებისა და ექსპერიმენტების შედეგებზე დაყრდნობით, მასობრივ მოვლენებზე და პროცესებზე მეცნიერულად დასაბუთებული დასკვნების მიღების საშუალებას იძლევა. ეს დასკვნები შეეხება არა ცალკეულ ექსპერიმენტებს, რომელთა განმეორებითაც ყალიბდება მოცემული მასობრივი მოვლენა, არამედ წარმოადგენს მტკიცებულებებს მოცემული პროცესის ზოგადი ალბათური მახასიათებლების შესახებ (ანუ ალბათობებზე, განაწილებების კანონებზე, მათემატიკურ ლოდინებზე, დისპერსიებზე და ა. შ.).

დავუშვათ, რომ ჩვენ გაგვაჩნია მონაცემები, მაგალითად, გარკვეულ პირობებში დამზადებულ პროდუქციაში დეფექტური ნაწარმის რიცხვის შესახებ ან ნაწარმის გამძლეობაზე შემოწმების ექსპერიმენტის შედეგების შესახებ და ა. შ. ჩვენს მიერ შეგროვილი მონაცემები შესაძლებელია წარმოადგენდეს უშუალო ინტერესის საგანს პროდუქციის ამა თუ იმ პარტიის ხარისხზე ინფორმაციის თვალსაზრისით. სტატისტიკური ამოცანა კი იწყება მაშინ, როდესაც ჩვენ იმავე ინფორმაციაზე დაყრდნობით ვიწყებთ დასკვნების გაკეთებას მოვლენათა უფრო ფართო წრის შესახებ. ასე მაგალითად, ჩვენ შეიძლება გვაინტერესებდეს ტექნოლოგიური პროცესის ხარისხი, რისთვისაც ჩვენ ვაფასებთ ამ პროცესში დეფექტური ნაწარმის მიღების ალბათობას ან ნაწარმის საშუალო სიცოცხლის ხანგრძლივობას. ამ შემთხვევაში, შეგროვილ მასალას ჩვენ ვიხილავთ როგორც გარკვეულ საცდელ ჯგუფს ან შერჩევას, რომელიც წარმოადგენს მხოლოდ სერიას შესაძლო შედეგებიდან, რომლებიც შესაძლებელია შეგვხვდეს მოცემულ პირობებში მასობრივ პროცესზე დაკვირვებების გაგრძელების შემთხვევაში. დაკვირვებების შედეგების საფუძველზე გაკეთებული დასკვნები და შეფასებები ასახავენ საცდელი ჯგუფის შემთხვევით შემადგენლობას და ამიტომ ითვლება, რომ ისინი ალბათური ხასიათის მიახლოებითი შეფასებებია. თეორია გვიჩვენებს, თუ როგორ უნდა გამოვიყენოთ არსებული ინფორმაცია იმისათვის, რომ მივიღოთ რაც შეიძლება ზუსტი და საიმედო მახასიათებლები და ამასთანავე მივუთითოთ მონაცემთა მარაგის შეზღუდულობით გამოწვეული დასკვნების საიმედოობის ხარისხი.

მათემატიკურ სტატისტიკაში იხილავს ამოცანათა ორი ძირითადი კატეგორია: შეფასება და ჰიპოთეზათა სტატისტიკური შემოწმება. პირველი ამოცანა, თავის მხრივ, იყოფა განაწილების პარამეტრების წერტილოვან და ინტერვალურ შეფასებებად. მაგალითად, შესაძლებელია დაკვირვებების საფუძველზე წარმოიშვას მათემატიკური ლოდინისა და დისპერსიის წერტილოვანი შეფასების აუცილებლობა. თუ კი ჩვენ გვინდა მივიღოთ რაიმე ინტერვალი, რომელიც ამა თუ იმ საიმედოობის ხარისხით მოიცავს პარამეტრის ჭეშმარიტ მნიშვნელობას, მაშინ ეს არის ინტერვალური შეფასების ამოცანა.

მეორე ამოცანა – ჰიპოთეზათა შემოწმება – მდგომარეობს იმაში, რომ ჩვენ ვაკეთებთ დაშვებას შემთხვევითი სიდიდის ალბათური განაწილების შესახებ (მაგალითად, განაწილების ფუნქციის სახის შესახებ, ან განაწილების ფუნქციის ერთი ან რამოდენიმე პარამეტრის მნიშვნელობის შესახებ) და ვადგენთ არის თუ არა განაწილების სახე ან პარამეტრების მნიშვნელობები შესაბამისობაში (გარკვეული აზრით) დაკვირვებების მიღებულ შედეგებთან.

შერჩევითი მეთოდი. დავუშვათ, რომ ჩვენი მიზანია შევისწავლოთ საქონლის გარკვეული პარტიის რაოდენობრივი ნიშანი. პარტიის შემოწმება (კონტროლი) შესაძლებელია მოხდეს ორი გზით:

1. ჩავატაროთ მთელი პარტიის შემოწმება;
2. ჩავატაროთ პარტიის მხოლოდ ნაწილის შემოწმება.

პირველი გზა ყოველთვის არაა განხორციელებადი, მაგალითად, პარტიაში საქონლის დიდი რიცხვის გამო, შემოწმების ოპერაციის ჩატარების სიძვირის გამო ან იმის გამო, რომ შემოწმება იწვევს საქონლის განადგურებას (ელექტრო ნათურის შემოწმებისას მუშაობის ხანგრძლივობაზე).

მეორე შემთხვევაში, შემთხვევითი გზით შერჩეული ობიექტების სიმრავლეს **შერჩევითი ერთობლიობა ან შერჩევა** ეწოდება. ობიექტთა მთლიან ერთობლიობას, საიდანაც ხდება შერჩევა, **გენერალური ერთობლიობა** ეწოდება. შერჩევაში ელემენტთა რაოდენობას **შერჩევის მოცულობა** ეწოდება. როგორც წესი, ითვლება, რომ გენერალური ერთობლიობის მოცულობა უსასრულოა.

შერჩევა შეიძლება იყოს **განმეორებითი** (დაბრუნებით) და **განმეორების გარეშე** (დაბრუნების გარეშე).

ჩვეულებრივ, ხორციელდება განმეორების გარეშე შერჩევები, თუმცა გენერალური ერთობლიობის მოცულობის სიდიდის (უსასრულობის) გამო, მხოლოდ განმეორებითი შერჩევების დროს სამართლიანი გათვლები წარმოებს და დასკვნები კეთდება.

შერჩევა საკმარისად სრულად უნდა ასახავდეს გენერალური ერთობლიობის ყველა ობიექტის განსაკუთრებულობებს, ანუ სხვა სიტყვებით, რომ ვთქვათ, შერჩევა უნდა იყოს **რეპრეზენტატული** (წარმომადგენლობითი).

ჩვენი ცოდნა, მსჯელობა და ქცევები მნიშვნელოვანწილად დამყარებულია შერჩევით მონაცემებზე. ეს მტკიცებულება ერთნაირად სამართლიანია როგორც ყოველდღიურ საქმიანობაში, ისე სამეცნიერო გამოკვლევებში. შთაბეჭდილება იმ დაწესებულებაზე, რომელშიც ყოველდღიურად კეთდება ათასობით სხვადასხვა ოპერაცია (ქმედება), ხშირ შემთხვევაში იქმნება მრავალი წლის განმავლობაში ამ დაწესებულების მხოლოდ ერთჯერ ან ორჯერ მონახულების საფუძველზე. მოგზაური, რომელმაც გაატარა 10 დღე უცხო ქვეყანაში, ფიქრობს დაწეროს წიგნი სადაც აპირებს ურჩიოს ამ ქვეყნის მოსახლეობას თუ როგორ აღორძინონ წარმოება, გარდაქმნან პოლიტიკური სისტემა, გაზარდონ ქვეყნის ბიუჯეტი, აამაღლონ სწავლების დონე ზოგად საგანმანათლებო დასესებულებებში, განამტკიცონ ვალუტის კურსი, აამაღლონ ქვეყნის თავდაცვისუნარიანობა და ა. შ. ცხადია ეს ადამიანი არ გამოიყურება სერიოზულად. მაგრამ, სინამდვილეში იგი იმ მეცნ-

იერ-საზოგადოებათმცოდნისაგან, რომელმაც ამ ქვეყანაში იცხოვრა 20 წელი და ამ პერიოდის განმავლობაში სწავლობდა მას, განსხვავდება მხოლოდ იმით, რომ თავის დასკვნებს აკეთებს დაკვირვებათა გაცილებით მცირე რიცხვის საფუძველზე და ამასთანავე, ალბათ ნაკლებად გარკვეულია თავისი უცოდინრობის ხარისხში. როგორც მეცნიერებაში, ისე საყოფაცხოვრებო საქმიანობაში ჩვენთვის ხელმისაწვდომია მხოლოდ ფრაგმენტი იმ ზოგადი სურათის, რამაც უნდა გააფართოოს ჩვენი ცოდნა.

სტატისტიკოსისათვის უაღრესად მნიშვნელოვანია თუ როგორ უნდა აგოს შერჩევა და მიღებული შერჩევიდან როგორ უნდა მივიღოს დასაბუთებული დასკვნები. თუ კი სიმრავლე საიდანაც ხდება შერჩევა (პოპულაცია) ერთგვაროვანია, მაშინ შერჩევის მიღების მეთოდს დიდი მნიშვნელობა არ ექნება და ნებისმიერი შერჩევა მოგვცემს დაახლოებით ერთი და იგივე შედეგს. მაგალითად, დასკვნა ადამიანის სისხლის შემადგენლობის შესახებ კეთდება მისი სისხლის რამოდენიმე წვეთისაგან გაკეთებული ლაბორატორიული გამოკვლევით. ეს მეთოდი ემყარება იმ დაშვებას, რომ ცირკულაციის შედეგად სისხლი კარგადაა შერეული და მისი ყოველი წვეთი ერთი და იგივე ინფორმაციის მატარებელია. მაგრამ, იმ შემთხვევაში როცა ჩვენ ვიკვლევთ არაერთგვაროვან პოპულაციას, როგორც ეს ხშირად ხდება პრაქტიკაში; შერჩევის მიღების მეთოდი იძენს გადამწყვეტ მნიშვნელობას, ხოლო იმ მეთოდების შესწავლა, რომელიც საშუალებას იძლევა მივიღოთ ამომწურავი მონაცემები, ხდება ძალზე მნიშვნელოვანი.

ობიექტების ამორჩევის ხერხების მიხედვით განასხვავებენ შემდეგი ტიპის შერჩევებს:

1. *მარტივი შემთხვევითი ამორჩევა.*

გენერალური ერთობლიობის ყველა ელემენტი გადაინომრება და შემთხვევით რიცხვთა ცხრილიდან იღებენ, მაგალითად, ნებისმიერ ერთმანეთის მომდევნო 50 რიცხვის მიმდევრობას და შერჩევაში შეყავთ ამოსული ნომრების მქონე ობიექტები.

2. *ტიპიური ამორჩევა.*

ასეთი ამორჩევა წარმოებს იმ შემთხვევაში, თუ გენერალური ერთობლიობა შესაძლებელია წარმოდგეს ისეთ ქვესიმრავლეთა გაერთიანებად, რომელთა ელემენტები ერთგვაროვანია რაიმე ნიშნის მიხედვით, თუმცა მთელ ერთობლიობას ასეთი ერთგვაროვნება არ გააჩნია (საქონლის პარტია შედგება რამოდენიმე ჯგუფისაგან, რომლებიც წარმოებულია სხვადასხვა საწარმოს მიერ). მაშინ, თითოეულ ქვესიმრავლეში ატარებენ მარტივ შემთხვევით შერჩევას, და შერჩევაში აერთიანებენ ყველა მიღებულ ობიექტს.

3. *მექანიკური ამორჩევა.*

გენერალური ერთობლიობიდან იღებენ ყოველ მეცხრე (ორმოცდამეათე) ობიექტს.

4. *სერიული ამორჩევა.*

შერჩევაში აერთიანებენ იმ ობიექტებს, რომლებიც წარმოებულია რაიმე წარმოების სფეროში დროის გარკვეულ ინტერვალში.

შემდგომში, გენერალური ერთობლიობის ქვეშ, ჩვენ ვიგულისხმებთ არა თვითონ ობიექტთა სიმრავლეს, არამედ იმ შემთხვევითი სიდიდის მნი-

შენელობათა სიმრავლეს, რომელიც ღებულობს რიცხვით მნიშვნელობებს თითოეულ ობიექტზე. სინამდვილეში, გენერალური ერთობლიობა, როგორც ობიექტთა სიმრავლე შეიძლება არც არსებობდეს. მაგალითად, აზრი აქვს ვილაპარაკოთ იმ დეტალების სიმრავლეზე, რომლებიც შეიძლება წარმოებულ იქნეს, თუ გამოვიყენებთ მოცემულ ტექნოლოგიურ პროცესს. გამოვიყენებთ რა ამ პროცესის ჩვენთვის ცნობილ მახასიათებლებს, ჩვენ შეგვიძლია შევაფასოთ დეტალების ამ არ არსებული სიმრავლის პარამეტრები. დეტალის ზომა – ეს შემთხვევითი სიდიდეა, რომლის მნიშვნელობა განისაზღვრება ტექნოლოგიური პროცესის შემადგენელი მრავალი ფაქტორის ზემოქმედებით. ჩვენ, მაგალითად, შეიძლება გვაინტერესებდეს თუ რა ალბათობით ღებულობს ეს შემთხვევითი სიდიდე მნიშვნელობას გარკვეული ინტერვალიდან. ამ კითხვაზე პასუხის გაცემა შესაძლებელია თუ ჩვენ გვეცოდინება ამ შემთხვევითი სიდიდის განაწილების კანონი და მისი ისეთი პარამეტრები, როგორიცაა ლოდინი და დისპერსია.

ამრიგად, გენერალური ერთობლიობის, როგორც ობიექტთა სიმრავლის ცნებიდან, რომლებიც ხასიათდებიან გარკვეული ნიშნით (თვისებით), ჩვენ გადავდივართ გენერალურ ერთობლიობაზე, როგორც შემთხვევით სიდიდეზე, რომლის განაწილების კანონი და პარამეტრები განისაზღვრება შერჩევითი მეთოდის საშუალებით.

განვიხილოთ n მოცულობის შერჩევა, რომელიც წარმოადგენს მოცემულ გენერალურ ერთობლიობას. პირველ შერჩევით მნიშვნელობას x_1-ს განვიხილავთ როგორც რეალიზაციას, როგორც ერთ-ერთ შესაძლებელ მნიშვნელობას ξ_1 შემთხვევითი სიდიდის, რომელსაც გააჩნია იგივე განაწილების კანონი რაც ξ შემთხვევით სიდიდეს. მეორე შერჩევითი მნიშვნელობა x_2-ს განვიხილავთ როგორც ერთ-ერთ შესაძლებელ მნიშვნელობას ξ_2 შემთხვევითი სიდიდის, რომელსაც გააჩნია იგივე განაწილების კანონი რაც ξ შემთხვევით სიდიდეს და ა. შ. იგივე შეიძლება ითქვას $x_3, x_4, \ldots, x_n$ მნიშვნელობებზე.

ამრიგად, შერჩევას ჩვენ ვუყურებთ როგორც ერთობლიობას დამოუკიდებელი $\xi_1, \xi_2, \ldots, \xi_n$ შემთხვევითი სიდიდეების, რომლებიც იმავე კანონით არიან განაწილებული როგორც ξ შემთხვევითი სიდიდე, რომელიც წარმოადგენს გენერალურ ერთობლიობას. შერჩევითი მნიშვნელობები $x_1, x_2, \ldots, x_n$ – ეს ის მნიშვნელობებია, რაც მიიღეს შემთხვევითმა სიდიდეებმა პირველი, მეორე, და ა. შ. მე-n ექსპერიმენტის შედეგად.

შერჩევის მოცულობის განსაზღვრა. შერჩევითი გამოკვლევის დაგეგმვისას ყოველთვის დგება მომენტი, როცა უნდა გადავწყვიტოთ როგორი უნდა იყოს შერჩევის მოცულობა. ამ გადაწყვეტილებას არსებითი მნიშვნელობა აქვს. ძალიან დიდი შერჩევა მოითხოვს ძალიან დიდ დანახარჯებს, ხოლო ძალიან პატარა კი ამცირებს მიღებული შედეგების სარგებლიანობას (ვარგისიანობას). მიღებული გადაწყვეტილება ყოველთვის არაა დამაკმაყოფილებელი, ვინაიდან ხშირად ინფორმაციის ნაკლებობის გამო ჩვენ არ შეიძლება ვიყოთ დარწმუნებული, რომ დავადგინეთ შერჩევის საუკეთესო

მოცულობა. შერჩევითი მეთოდის თეორია შესაძლებლობას იძლევა მეცნიერულად დავასაბუთოთ ასეთი გადაწყვეტილება.

გადაწყვეტილების მიღების ეტაპების დემონსტრირების მიზნით განვიხილოთ ჰიპოთეტური მაგალითი: ანთროპოლოგი აპირებს გამოიკვლიოს გარკვეული კუნძულის მოსახლეობა. სხვა ნიშნებთან ერთად, ის გეგმავს შეაფასოს მოსახლეობის იმ ნაწილის პროცენტი, რომელთაც გააჩნიათ I ჯგუფის სისხლი. პროცედურა ისეა ორგანიზებული, რომ შესაძლებელია მიიღოს მარტივი შემთხვევითი შერჩევა. როგორი უნდა იყოს ამ შერჩევის მოცულობა? ამ კითხვაზე პასუხის გაცემა შეუძლებელია, თუ წინასწარ არ მივიღებთ პასუხს სხვა კითხვაზე, კერძოდ, რამდენად დამაჯერებლად სურს ანთროპოლოგს იცოდეს I ჯგუფის მქონე ადამიანების პროცენტი? შეიძლება მან გვიპასუხოს, რომ კმაყოფილი იქნებოდა თუ ეს პროცენტი აღმოჩნდებოდა საზღვრებში ±5% იმ აზრით, რომ თუ შერჩევის მონაცემების მიხედვით I ჯგუფის სისხლი ექნებოდა ადამიანების 43%-ს, მაშინ კუნძულის ყველა მოსახლისათვის ეს პროცენტი სინამდვილეში აღმოჩნდებოდა 43%-5%=38%-სა და 43%+5%=48%-ს შორის.

იმისათვის რომ თავიდან ავიცილოთ გაუგებრობა, ჩვენ უნდა ავუხსნათ ანთროპოლოგს, რომ ცალკეული გამოკვლევის მიხედვით არ შეიძლება აბსოლუტური (მყარი) გარანტიების მიცემა, რომ შედეგი იქნება ±5%-ის ფარგლებში, თუ კი არ ჩავატარებთ მოსახლეობის მთლიან გამოკვლევას. რაოდენ დიდიც არ უნდა იყოს შერჩევის მოცულობა n ყოველთვის არსებობს შანსი იმისა, რომ მივიღოთ იმდენად არასახარბიელო შერჩევა, რომ მისი შეცდომა გადააჭარბებს სასურველ 5%-ს. ანთროპოლოგი აცხადებს, რომ მან იცის ამის შესახებ, მაგრამ ის თანახმაა გარისკოს იმ შემთხვევაში, თუ კი შანსი იმისა, რომ მიღებულ იქნას არასახარბიელო შეცდომა შეადგენს 1-ს 20-ის წინააღმდეგ.

ამის შემდეგ ჩვენ შეგვიძლია მივიღოთ შერჩევის მოცულობის უხეში შეფასება. საქმის გამარტივების მიზნით იგულისხმება, რომ შერჩევითი პროცენტი (პროპორცია) p განაწილებულია ნორმალურად. რამდენად შესაძლებელია ეს დაშვება შესაძლებელი იქნება შევამოწმოთ მას შემდეგ რაც ვიპოვით შერჩევის მოცულობის პირველ მოახლოებას.

მათემატიკურად p უნდა მოთავსებული იყოს ინტერვალში $(P-5, P+5)$ ალბათობით 19/20. ვინაიდან იგულისხმება, რომ p განაწილებულია ნორმალურად მათემატიკური ლოდინით P, ამიტომ ალბათობით 19/20 ის მოთავსებულია ინტერვალში $(P-2\sigma, P+2\sigma)$. გარდა ამისა, $\sigma \approx \sqrt{PQ/n}$ $(Q=1-P)$, ამიტომ ჩვენ შეგვიძლია დავუშვათ, რომ

$$2\sqrt{PQ/n}=5 \text{ ანუ } n=4PQ/25.$$

აქ წარმოიქმნება სიძნელეები, რომელის ახასიათებს შერჩევის მოცულობის მოძებნის ყველა ამოცანას. მართალია n-ის მოსაძებნი ფორმულა მიღებულია, მაგრამ ის დამოკიდებულია გარკვეულ მახასიათებელზე, რომელიც თვითონ ექვემდებარება შერჩევით გამოკვლევას. ჩვენს მაგალითში ასეთი მახასიათებელია P, რომელიც ჩვენ ჯერ კიდევ შესაფასებელი გვაქვს. ამიტომ ჩვენ ვეკითხებით ანთროპოლოგს, ხომ არ წარმოუდგენია მას როგორი შეიძლება იყოს P-ს შესაძლო მნიშვნელობა. მან შეიძლება

გეიპასუხოს, რომ გამომდინარე სხვა ეთნიკური ჯგუფების წინა გამოკვლევებიდან და მისი დაშვებებით კუნძულის მოსახლეობის რასობრივი წარმომავლობის შესახებ, ნაკლებად სავარაუდოა , რომ P გავიდეს (30%, 60%) ინტერვალის საზღვრებიდან.

ეს მონაცემები საკმარისია დამაკმაყოფილებელი პასუხის მისაღებად. P-ს ნებისმიერი მნიშვნელობისათვის 30-დან 60-მდე PQ ნამრავლი იცვლება 2100-დან 2500-მდე (მაქსიმალური მნიშვნელობა 2500 მიიღება როცა $P=50$). n-ის შესაბამისი მნიშვნელობა მოთავსებულია 336-სა და 400-ს შორის. სრული გარანტიისათვის n-ის პირველ მიახლოებად ავიღოთ $n=400$ მნიშვნელობა.

პოპულაციის სასრულობის შესწორება. ცნობილია, რომ n მოცულობის მქონე შემთხვევითი შერჩევისათვის უსასრულო პოპულაციიდან საშუალოს დისპერსია არის σ^2/n. თუ ერთობლიობა სასრულია, მაშინ უნდა შემოვიღოთ მამრავლი $(N-n)/N$. $(N-n)/N$ მამრავლს დისპერსიისათვის, ხოლო $\sqrt{(N-n)/N}$ მამრავლს სტანდარტული გადახრისათვს უწოდებენ **პოპულაციის სასრულობის შესწორებას (პსშ)**. ზოგიერთი ავტორი გამყოფად იღებს ($N-1$)-ს N-ის ნაცვლად. თუ შერჩევის წილი n/N საკმაოდ მცირეა, მაშინ ეს მამრავლები ახლოსაა ერთთან და პოპულაციის მოცულობა თავისთავად უშუალო გავლენას არ ახდენს შერჩევითი საშუალოს სტანდარტულ შეცდომაზე. მაგალითად, თუ ორი პოპულაციისათვის s ერთნაირია, მაშინ 200000-იანი პოპულაციიდან აღებული შერჩევა, რომლის მოცულობაა 500 უზრუნველყოფს ერთობლიობის საშუალოს შეფასების თითქმის იგივე სიზუსტეს, რასაც 10000-იანი შერჩევიდან აღებული 500-ის ტოლი მოცულობის მქონე შერჩევა. პრაქტიკულ ამოცანებში პსშ შეიძლება არ გაეითვალისწინოთ, თუ შერჩევის მოცულობა არ აღემატება 5%-ს, ხოლო ზოგიერთი მიზნებისათვის 10%-საც კი.)

ახლა გავაანალიზოთ ჩვენს მიერ გაკეთებული დაშვებები. როცა $n=400$ და P მოთავსებულია 30-სა და 60-ს შორის p-ს განაწილება ახლოსაა ნორმალურთან. საჭიროა თუ არა პსშ, დამოკიდებულია კუნძულის მოსახლეობის რაოდენობაზე. თუ პოპულაციის ელემენტთა რიცხვი მეტია 8000-ზე, მაშინ შერჩევის მოცულობა იქნება 5%-ზე ნაკლები და n-ის შესწორება პსშ-ს საჭირო არ იქნება.

შერჩევის მოცულობის დადგენა წილების შეფასების დროს. ზოგჯერ ჩვენ გვინდა შევაფასოთ საერთო რიცხვი, წილი ან პროცენტი პოპულაციის იმ ერთეულების, რომელთაც გააჩნიათ გარკვეული ნიშანი ან თვისება ან ეკუთვნიან ერთეულების გარკვეულ კლასს. როგორც წესი ამ სახით ქვეყნდება მოსახლეობის აღწერის შედეგების დიდი ნაწილი, მაგალითად, უმუშევრების რიცხვი, პროცენტი მოსახლეობის იმ ნაწილის, რომელიც ლაპარაკობს ლათინურ ენაზე და ა. შ. ასეთი კლასიფიკაცია შეიძლება შეტანილ იქნეს უშუალოდ გამოსაკითხ ფურცელში ისეთი კითხვების დასმით, რომელიც მოითხეს მარტივ პასუხებს "კი" ან "არა". სხვა შემთხვევებში საწყისი დაკვირვების შედეგებს გააჩნიათ მეტ ნაკლებად უწყვეტი ხასიათი და კლასიფიკაცია წარმოებს გამოკვლევის მონაცემების დახარისხებისა და დაჯგუფების დროს. ასეთნაირად ჩვენ შეგვიძლია დავარეგისტრირ-

ოთ გამოკითხულების ასაკი ერთი წლის სიზუსტით, მაგრამ გამოვქავეყნოთ 60 წლისა და უფრო მეტი ასაკის მოსახლეობის პროცენტი.

ვიგულისხმოთ, რომ პოპულაციის ყველა ერთეული მიეკუთვნება ერთ-ერთს ორი კლასიდან: C^1 და C^2 C^1 კლასის ერთეულების რაოდენობა პოპულაციაში იყოს A, ხოლო შერჩევაში კი a. მაშინ C^1 კლასის ერთეულების წილი პოპულაციაში იქნება $P = A/N$, ხოლო შერჩევაში კი - $p = a/n$. P-ს შერჩევით შეფასებას წარმოადგენს p, ხოლო A-ს შერჩევით შეფასებას წარმოადგენს Np ან Na/p. სტატისტიკურ პრაქტიკაში a და p ტიპის შეფასებების განხილვისას ხშირად გამოიყენება *ბინომიალური* განაწილება. როგორც აღმოჩნდა, სასრული პოპულაციების შემთხვევაში საჭიროა გამოვიყენოთ *ჰიპერგეომეტრიული* განაწილება, თუმცა ბინომიალური განაწილება როგორც წესი იძლევა დამაკმაყოფილებელ სიზუსტეს.

ყოველი ერთეულისათვის შერჩევაში ან პოპულაციაში დავუშვათ, რომ $y_i = 1$, თუ ერთეული ეკუთვნის C^1 კლასს, და $y_i = 0$, თუ ერთეული ეკუთვნის C^2 კლასს. მაშინ ცხადია, რომ

$$Y = \sum_{i=1}^{N} y_i = A \text{ და } \overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{A}{N} = P.$$

ანალოგიურად შერჩევისათვის გვექნება:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{a}{n} = p.$$

შესაბამისად, A-სა და P-ს შეფასების ამოცანა შეიძლება განვიხილოთ როგორც ისეთი პოპულაციის ჯამური და საშუალო მნიშვნელობების შეფასების ამოცანა, რომელშიც ნებისმიერი წევრი ან 1-ია ან 0. პოპულაციისა და შერჩევის შესწორებული დისპერსიები S^2 და s^2 გამოვსახოთ P-სა და p-ს საშუალებით. ვინაიდან, $y_i^2 = y_i$, შესაბამისად,

$$\sum_{i=1}^{N} y_i^2 = A = NP \text{ და } \sum_{i=1}^{n} y_i^2 = a = np.$$

ამიტომ

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{Y})^2}{N-1} = \frac{\sum_{i=1}^{N} y_i^2 - N\overline{Y}^2}{N-1} = \frac{1}{N-1}(NP - NP^2) = \frac{N}{N-1}PQ,$$

სადაც $Q = 1 - P$. ანალოგიურად,

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} = \frac{n}{n-1}pq.$$

ცნობილია რომ: ა). შერჩევითი წილი $p = a/n$ წარმოადგენს პოპულაციის $P = A/N$ წილის ჩაუნაცვლებელ შეფასებას;

ბ). $p = a/n$-ის დისპერსია ტოლია:

$$Dp = E(p-P)^2 = \frac{S^2}{n} \cdot (\frac{N-n}{N}) = \frac{PQ}{n} \cdot (\frac{N-n}{N-1});$$

გ). დისპერსიის გამოსახულება ძალაში რჩება მაშინაც, როცა p და P არის შესაბამისად შერჩევისა და პოპულაციისათვის იმ ერთეულების *პროცენტი*, რომლებიც მიეკუთვნებიან C^1 კლასს;

დ). C^1 კლასის ერთეულების საერთო რიცხვის შეფასების $\hat{A} = Np$ დისპერსია ტოლია:

$$D\hat{A} = D(Np) = \frac{N^2 PQ}{n} \cdot (\frac{N-n}{N-1});$$

ე). ჩაუნაცვლებელი შეფასება p-ს დისპერსიის შერჩევის მონაცემების მიხედვით არის:

$$s_p^2 = \frac{N-n}{(n-1)N} pq;$$

ვ). C^1 კლასის ერთეულების საერთო რიცხვის შეფასების $\hat{A} = Np$-ს დისპერსიის ჩაუნაცვლებელი შეფასებაა

$$s_{Np}^2 = \frac{N(N-n)}{n-1} pq.$$

მაგალითი. პოპულაცია შედგება 3042 გვარისა და მისამართების სიისაგან. 200-ის ტოლი მოცულობის მქონე მარტივმა შემთხვევითმა შერჩევამ აჩვენა რომ 38 მისამართი არ არის სწორი. შევაფასოთ ამ სიის იმ მისამართების საერთო რიცხვი, რომლებიც საჭიროებენ შესწორებას და ვიპოვოთ ამ შეფასების სტანდარტული გადახრა.

ამოხსნა. გვაქვს:

$$N = 3042, \ n = 200, \ a = 38, \ p = a/n = 38/200 = 0.19.$$

არასწორი მისამართების საერთო რიცხვის შეფასება ტოლია:

$$\hat{A} = Np = 3042 \cdot 0.19 = 578,$$

$$s_{\hat{A}} = \sqrt{\frac{N(N-n)}{n-1} pq} = \sqrt{\frac{3042 \cdot 2842}{119} \cdot 0{,}19 \cdot 0.81} = \sqrt{6685} = 81.8.$$

ვინაიდან შერჩევის წილი შეადგენს 7%-ზე ნაკლებს, პსშ-ს გავლენა უმნიშვნელოა. იმისათვის, რომ ის გამოირიცხოს, ჩავსვათ $(N-n)$-ის ნაცვლად ჩავსვათ N. თუ გარდა ამისა, $(n-1)$-ის ადგილას ჩავსვამთ n-ს, მივიღებთ უფრო მარტივ ფორმულას

$$s_{Np} = N\sqrt{pq/n} = 3042 \cdot \sqrt{0.19 \cdot 0.81/200} = 84.4$$

რაც საკმაოდ კარგად ეთანხმება წინა შედეგს – 81.8-ს)

ვიგულისხმოთ, რომ პოპულაციის ყველა ერთეული მიეკუთვნება ერთ-ერთს ორი კლასიდან: C^1 და C^2 ჩვენ ვთანხმდებით C^1 კლასის წილის p შეფასების შეცდომის გარკვეული ზღვრული მნიშვნელობის შესახებ d და გვინდა, რომ მხოლოდ მცირე α რისკით ფაქტიური შეცდომა იყოს d-ზე მეტი, ე. ი. რომ:

$$\mathrm{P}(|p-P| \ge d) = \alpha.$$

იგულისხმება, რომ წარმოებს მარტივი შემთხვევითი შერჩევა და p-ს აქვს ნორმალური განაწილება. ბ).-ს თანახმად

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{PQ}{n}}.$$

შესაბამისად ფორმულას, რომელიც აკავშირებს შერჩევის n მოცულობას სიზუსტის სასურველ ხარისხთან, ექნება სახე

$$d = z_{\alpha/2} \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{PQ}{n}}.$$

თუ ამ განტოლებას ამოეხსნით n-ის მიმართ, მივიღებთ, რომ

$$n = \frac{\dfrac{z_{\alpha/2}^2 PQ}{d^2}}{1 + \dfrac{1}{N}(\dfrac{z_{\alpha/2}^2 PQ}{d^2} - 1)} = \frac{z_{\alpha/2}^2 PQN}{Nd^2 + z_{\alpha/2}^2 PQ - d^2}.$$

პრაქტიკული გამოყენების მიზნით ამ ფორმულაში P-ს ადგილას სვამენ გარკვეულ წინასწარ შეფასებას, p. თუ N დიდია, მაშინ პირველ მიახლოებად იღებენ რიცხვს

$$n_0 = \frac{z_{\alpha/2}^2 pq}{d^2} = \frac{pq}{d_1},$$

სადაც $d_1 = \dfrac{d^2}{z_{\alpha/2}^2}$ – შერჩევითი წილის სასურველი დისპერსიაა.

პრაქტიკაში თავიდან ითვლიან n_0-ს. თუ n_0/N პრაქტიკულად უგულებელყოფადია, მაშინ n_0 იქნება დამაკმაყოფილებელი მიახლოება n-სათვის. წინააღმდეგ შემთხვევაში n-ისა და n_0-ის ფორმულების შედარებით შეიძლება დავასკვნათ, რომ n გამოითვლება ფორმულით:

$$n = \frac{n_0}{1+(n_0-1)/N} \approx \frac{n_0}{1+(n_0/N)}.$$

მაგალითი. ჰიპოთეტურ მაგალითში სისხლის ჯგუფთან დაკავშირებით

$$d = 0.05, \quad p = 0.5, \quad \alpha = 0.05, \quad z_{\alpha/2} = 2.$$

ამიტომ

$$n_0 = \frac{pq}{d_1} = \frac{0.5 \cdot 0.5}{\dfrac{0.0025}{4}} = 400.$$

ვიგულისხმოთ, რომ კუნძულზე ცხოვრობს დაახლოებით 3200 ადამიანი. საჭიროა გავითვალისწინოთ პსშ და მივიღებთ

$$n = \frac{n_0}{1+(n_0-1)/N} = \frac{400}{1+\dfrac{399}{3200}} = 356.$$

ფორმულა n_0-სათვის სამართლიანია მაშინაც, როცა სიდიდეები d, p და q გამოსახულია არა წილადებში, არამედ პროცენტებში. ვინაიდან

pq ნამრაველი იზრდება, როცა p უახლოვდება 1/2-ს (ანუ 50%-ს), ამიტომ n-ის უფრო ზუსტი შეფასება მიიღება, თუ იმ ინტერვალში, რომელშიც იგულისხმება, რომ მოთავსებულია p, ავიღებთ p-ს როლში 1/2-თან (შესაბამისად, 5-%-თან) უახლოეს რიცხვს. მაგალითად, თუ p მოთავსებულია 5%-სა და 9%-ს შორის, მაშინ n-ის განსაზღვრისას ჩვენ ვიღებთ მნიშვნელობას $p = 9\%$.

შერჩევის მოცულობის ფორმულა უწყვეტი სიდიდეების შემთხვევაში. დავუშვათ, რომ $\bar{y}$ – დაკვირვებების საშუალო მნიშვნელობაა მარტივი შემთხვევითი შერჩევის დროს, და ჩვენ გვინდა, რომ

$$P(|\bar{y} - \bar{Y}| \geq d) = \alpha,$$

სადაც d – არჩეული შეცდომის ზღვრული მნიშვნელობაა, ხოლო α – გარკვეული მცირე ალბათობაა. დავუშვათ, რომ $\bar{y}$ განაწილებულია ნორმალურად, მაშინ მისი სტანდარტული შეცდომა ტოლია:

$$\sigma_{\bar{y}} = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}.$$

ამიტომ

$$d = z_{\alpha/2} \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}.$$

საიდანაც:

$$n = \frac{\left(\frac{z_{\alpha/2}S}{d}\right)^2}{1 + \frac{1}{N} \cdot \left(\frac{z_{\alpha/2}S}{d}\right)^2} = \frac{(z_{\alpha/2}S)^2 N}{Nd^2 + (z_{\alpha/2}S)^2}.$$

ისევე როგორც ზემოთ, დიდი N-ებისათვის შერჩევის ზომის პირველ მიახლოებად ვიღებთ

$$n_0 = \left(\frac{z_{\alpha/2}S}{d}\right)^2 = \frac{S^2}{d_1}$$

(სადაც $d_1 = \frac{d^2}{z_{\alpha/2}^2}$). ეს ფორმულა გამოსადეგია მხოლოდ მაშინ, როცა n_0/N მცირეა. წინააღმდეგ შემთხვევაში n უნდა გამოვთვალოთ ფორმულიდან

$$n = \frac{n_0}{1 + n_0/N}$$

თუ ფასდება პოპულაციის ჯამური მნიშვნელობა Y, ზღვრული შეცდომით d, მაშინ შერჩევის მოცულობის პირველადი მიახლოების როლში, როცა n_0/N მცირეა, უნდა ავიღოთ

$$n_0 = \left(\frac{z_{\alpha/2}SN}{d}\right)^2 = \frac{(SN)^2}{d_1},$$

ხოლო წინააღმდეგ შემთხვევაში კი ისევ

$$n = \frac{n_0}{1 + n_0 / N}.$$

მაგალითი. საობიექტებში, სადაც გასაყიდად გამოყავთ ხეების ნერგები, ზამთრის ბოლოსა და გაზაფხულის დასაწყისში მიზანშეწონილია შეფასდეს გამოსადეგი ნერგების რიცხვი, რადგანაც მასზეა დამოკიდებული მოთხოვნების დაკმაყოფილება და შეკვეთების მიღება. ნერგების საერთო რიცხვის შეფასების შერჩევითი მეთოდები შესწავლილი იყო ჯონსონის (Johnson, 1943) მიერ. ქვემოთ მოყვანილი მონაცემები მიღებული იყო ვერცხლისფერი ნეკერჩხლის ნერგების მწკრივის შესწავლისას, რომლის სიგანე იყო 1 ფუტი, ხოლო სიგრძე 430 ფუტი. შერჩევის ერთეულად ითვლებოდა მწკრივის სიგრძის 1 ფუტი. ასე, რომ პოპულაციის მოცულობა $N = 430$. ნერგების მწკრივის მთლიანი შესწავლის შედეგად აღმოჩნდა რომ ერთობლიობისათვის $\overline{Y}$-ისა და S^2-ის რეალური მნიშვნელობებია: $\overline{Y} = 19$ და $S^2 = 85.6$.

დავსვათ ასეთი კითხვა: რამდენი ერთეულის აღებაა საჭირო მარტივი შემთხვევითი შერჩევის შემთხვევაში, რომ შევაფასოთ $\overline{\overline{Y}}$ 19/20-ის ტოლი ალბათობით შეცდომით, რომელიც მოთავსებულია 10%-იან საზღვრებში? ვინაიდან, ამ შემთხვევაში $\alpha = 1 - 19/20 = 1 - 0.95 = 0.05$, $z_{\alpha/2} = z_{0.025} = 2$, $S^2 = 85.6$, $d = 19 \cdot 10/100 = 1.9$, ამიტომ უწყვეტი სიდიდის შემთხვევაში შერჩევის მოცულობის ფორმულიდან გვაქვს:

$$n_0 = \frac{(z_{\alpha/2})^2 \cdot S^2}{d^2} = \frac{2^2 \cdot 85.6}{1.9^2} \approx 95.$$

რადგანაც $n_0 / N = 95/430 \approx 0.22$-ს უგულებელყოფა არ შეიძლება, ამიტომ ვღებულობთ, რომ

$$n = \frac{n_0}{1 + n_0 / N} = \frac{95}{1 + \dfrac{95}{430}} = 78.$$

ე. ი. იმისათვის, რომ მივიღოთ სასურველი სიზუსტე, აუცილებელია რომ შევისწავლოთ ნერგები მწკრივების თითქმის 20%-ზე.

ვარიაციული მწკრივი. დავუშვათ, რომ გენერალური ერთობლიობის ობიექტებისათვის განსაზღრულია გარკვეული ნიშანი ან რიცხვითი მახასიათებელი, რომლის გაზომვა შესაძლებელია (დეტალის სიგრძე, ნიტრატების ფარდობითი შემცველობა საზამთროში, ძრავის მუშაობის ხმაური). ეს მახასიათებელი არის შემთხვევითი სიდიდე ξ, რომელიც ყოველ ობიექტზე ღებულობს გარკვეულ რიცხვით მნიშვნელობას. n მოცულობის შერჩევიდან ვღებულობთ ამ შემთხვევითი სიდიდის მნიშვნელობებს n რიცხვისგან შედგენილი მწკრივის სახით:

$$x_1, x_2, \ldots, x_n. \tag{1}$$

ამ რიცხვებს ნიშნის მნიშვნელობებს უწოდებენ.

(1) მწკრივის რიცხვებს შორის შესაძლებელია იყოს ერთი და იგივე რიცხვები. თუ ნიშნის მნიშვნელობებს დავალაგებთ, ანუ რიცხვებს განვალაგებთ ზრდადობის ან კლებადობის მიხედვით, ამასთანავე ყოველ მნიშ-

ნელობას დავწერთ მხოლოდ ერთჯერ, ხოლო შემდეგ ყოველი x_i მნიშვნელობის ქვეშ დავწერთ m_i რიცხვს, რომელიც გვიჩვენებს თუ რამდენჯერ შეგვხვდა x_i მნიშვნელობა (1) მწკრივში, მივიღებთ ცხრილს, რომელსაც **დისკრეტული ვარიაციული მწკრივი** ეწოდება:

x_1	x_2	x_3	...	x_k
m_1	m_2	m_3	...	m_k

m_i რიცხვს ნიშნის i-ური მნიშვნელობის სიხშირე ეწოდება.

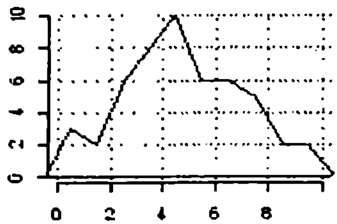
ცხადია, რომ (1) მწკრივის x_i შეიძლება არ ემთხვეოდეს x_i-ს ვარიაციული მწკრივიდან. ნათელია აგრეთვე, რომ

$$\sum_{i=1}^{k} m_i = n.$$

თუ შერჩევის მინიმალურ და მაქსიმალურ მნიშვნელობებს შორის ინტერვალს გავყოფთ ერთი და იგივე სიგრძის რამოდენიმე ინტერვალად, და ყოველ ინტერვალს შევუსაბამებთ ამ ინტერვალში მიხვედრილი ნიშნის შერჩევითი მნიშვნელობების რიცხვს, მაშინ მივიღებთ **ინტერვალურ ვარიაციულ მწკრივს**. თუ ნიშანს შეუძლია მიიღოს ნებისმიერი მნიშვნელობა გარკვეული ინტერვალიდან, ე. ი. წარმოადგენს უწყვეტ შემთხვევით სიდიდეს, მაშინ შერჩევა სწორედ ასეთი მწკრივის სახით უნდა წარმოვადგინოთ. თუ ინტერვალურ ვარიაციულ მწკრივში ყოველ $[a_i, a_{i+1})$ ინტერვალს შევცვლით ამ ინტერვალის შუაში მდებარე რიცხვით – $(a_i + a_{i+1})/2$, მაშინ მივიღებთ დისკრეტულ ვარიაციულ მწკრივს. ასეთი შეცვლა სრულიად ბუნებრივია, ვინაიდან, მაგალითად, დეტალის სიგრძის გაზომვისას ერთი მილიმეტრის სიზუსტით, ყველა სიგრძეს [49.5,50.5) ინტერვალიდან შეესაბამება ერთი რიცხვი, რომელიც ტოლია 50-ის.

ემპირიული განაწილების ფუნქცია.

გამოსაკვლევი შემთხვევითი სიდიდის თვალსაჩინო წარმოსახვისათვის შერჩევის მიხედვით შესაძლებელია აიგოს სხვადასხვა გრაფიკები. ერთ-ერთიასეთი გრაფიკია – **სიხშირეთა პოლიგონი**: ტეხილი, რომლის მონაკვეთები აერთებენ წერტილებს კოორდინატებით $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$, სადაც x_i გადაიზომება აბსცისთა ღერძზე, ხოლო n_i – ორდინატთა ღერძზე. თუ ორდინატთა ღერძზე გადავზომავთ არ აბსოლუტურ (n_i), არამედ ფარდობით (w_i) სიხშირეებს, მაშინ მივიღებთ **ფარდობით სიხშირეთა პოლიგონს**:



შემთხვევითი სიდიდის განაწილების ფუნქციის ანალოგიით, შეიძლება განსაზღვრულ იქნეს გარკვეული ფუნქცია, კერძოდ, $X \le x$ ხდომილების ფარდობითი სიხშირე.

განმარტება. შერჩევით (ემპირიულ) განაწილების ფუნქციას უწოდებენ ფუნქციას $F^*(x)$, რომელიც x –ის ნებისმიერი მნიშვნელობისათვის განსაზღვრავს $X \le x$ ხდომილების ფარდობით სიხშირეს. ამრიგად,

$$F^*(x) = \frac{n_x}{n},$$

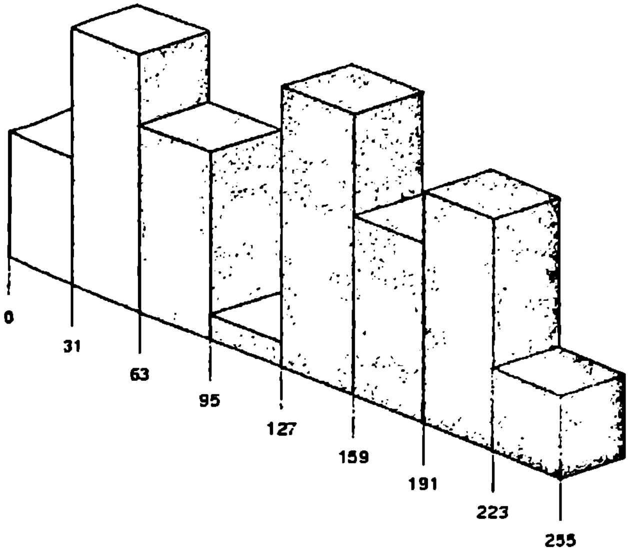
სადაც n_x – ვარიანტების რიცხვია, რომლებიც არ არემატება x –ს, ხოლო n – შერჩევის მოცულობა.

განსხვავებით ემპირიული განაწილების ფუნქციისაგან, რომელიც იგება შერჩევის მიხედვით, გენერალური ერთობლიობის $F(x)$ განაწილების ფუნქციას *თეორიული განაწილების ფუნქციას* უწოდებენ. იგი განსაზღვრავს $X \le x$ ხდომილების ალბათობას, ხოლო $F^*(x)$ კი -- მის ფარდობით სიხშირეს. საკმაოდ დიდი n–ებისათვის, როგორც ამას ამტკიცებს დიდ რიცხვთა კანონი, $F^*(x)$ ფუნქცია კრებადია ალბათობით $F(x)$ ფუნქციისაკენ.

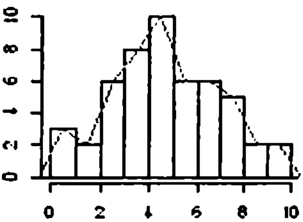
ემპირიული განაწილების ფუნქციის განმარტებიდან ადვილი დასანახია, რომ მისი თვისებები ემთხვევა $F(x)$ ფუნქციის თვისებებს, კერძოდ:

1. $0 \le F^*(x) \le 1$.
2. $F^*(x)$ – არაკლებადი ფუნქციაა.
3. თუ x_1– უმცირესი ვარიანტია, მაშინ $F^*(x) = 0$, როცა $x < x_1$; თუ x_k – უდიდესი ვარიანტია, მაშინ $F^*(x) = 1$, როცა $x \ge x_k$.
4. $F^*(x)$ – მარჯვნიდან უწყვეტი ფუნქციაა.

უწყვეტი მონაცემების შემთხვევაში გრაფიკულ ილუსტრაციას წარმოადგენს ე. წ. **ჰისტოგრამა**, ე. ი. საფეხურა ფიგურა, რომელიც შედგება მართკუთხედებისაგან, რომელთა ფუძეებია h სიგრძის ინტერვალები, ხოლო სიმაღლეები – მონაკვეთები სიგრძით n_i/h (სიხშირეების ჰისტოგრამა) ან w_i/h (ფარდობითი სიხშირეების ჰისტოგრამა). პირველ შემთხვევაში ჰისტოგრამის ფართობი ტოლია შერჩევის მოცულობის, ხოლო მეორე შემთხვევაში – ერთის.



ჰისტოგრამა წარმოდგენას გვაძლევს გენერალური ერთობლიობის განაწილების სიმკვრივეზე. შერჩევის დიდი მოცულობის შემთხვევაში ის ახლოსაა თეორიულ სიმკვრივესთან. ქვემოთ, ერთ ნახაზზე, მოყვანილია პოლიგონი და ჰისტოგრამა.

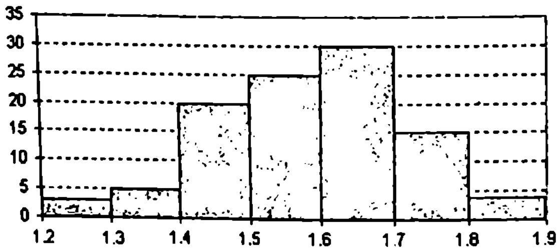


თავი II
ნორმალური განაწილება პირველი წაკითხვისათვის

რა არის ნორმალური? სამედიცინო მკვლევარები განსაზღვრავენ ე. წ. ნორმალურ ინტერვალს ადამიანის სისხლის წნევისათვის, ქოლესტერინის შემცველობის დონისათვის, აზოტის შემცველობის დონისათვის და ა. შ. მაგალითად, ადამიანის სისხლის წნევის ნორმალური დიაპაზონია 110-დან 140-მდე. ამ სიდიდეების გაზომვით მედიკოსს შეუძლია დაადგინოს არის თუ არა პაციენტის ჯანმრთელობის მდგომარეობის ესა თუ ის მაჩვენებელი ნორმალურ ინტერვალში, თუ საჭიროა გარკვეული ძალისხმევა რათა მისი შესაბამისი მაჩვენებელი მოექცეს ნორმის ფარგლებში და მომავალში აცილებულ იქნეს ავადმყოფობა. შესაბამისად იბადება კითხვა: "როგორ უნდა განისაზღვროს ე. წ. ნორმალური ინტერვალი?"

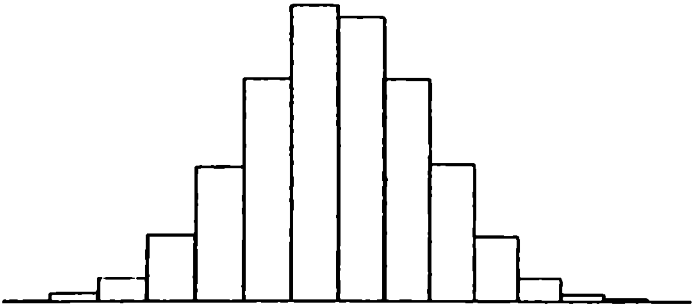
ჩვენ ქვემოთ შევისწავლით როგორ უნდა განსაზღვროს მკვლევარმა კონკრეტული სამედიცინო ტესტისათვის ნორმალური ინტერვალი ნორმალური განაწილების გამოყენებით. ჩვენ დავინახავთ აგრეთვე, რომ იგივე მეთოდების გამოყენებით განისაზღვრება ელემენტების სიცოცხლის ხანგრძლივობა, თოკის წინააღმდეგობის ძალა და მრავალი სხვადასხვა სახის მახასიათებელი.

შემთხვევითი სიდიდე შეიძლება იყოს როგორც დისკრეტული, ისე უწყვეტი ტიპის. გავიხსენოთ, რომ დისკრეტულ სიდიდეს არ შეუძლია მიიღოს ნებისმიერი მნიშვნელობა თავის ნებისმიერ ორ მნიშვნელობას შორის. მეორეს მხრივ, უწყვეტი სიდიდე ღებულობს ყველა მნიშვნელობას თავის ნებისმიერ ორ მნიშვნელობას შორის. უწყვეტი სიდიდის მაგალითებია სრულწლოვანი ადამიანის სიმაღლე, თაგვის სხეულის ტემპერატურა, ქოლესტერინის დონე ადამიანის ორგანიზმში. ბევრ უწყვეტ სიდიდეს, მსგავსად აღნიშნული მაგალითებისა, გააჩნია ზარის ფორმის განაწილება და მათ *დაახლოებით ნორმალურად განაწილებულ სიდიდეებს* უწოდებენ. მაგალითად, თუ მკვლევარი აირჩევს 100 ქალისაგან შემდგარ შემთხვევით შერჩევას, გაზომავს მათ სიმაღლეებს და ააგებს ჰისტოგრამას, მაშინ მკვლევარი მიიღებს გრაფიკს, რომელიც ანალოგიური იქნება ნახაზი 1-ის.

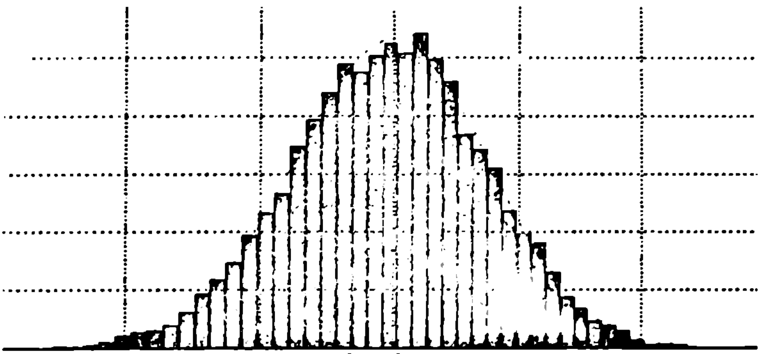


ნახაზი 1

შერჩევის მოცულობის გაზრდისა და კლასების სიგანის შემცირების შედეგად ჰისტოგრამა ჯერ მიიღებს ნახაზი 2-ის, ხოლო შემდგომ ნახაზი 3-ის ანალოგიურ ფორმას.

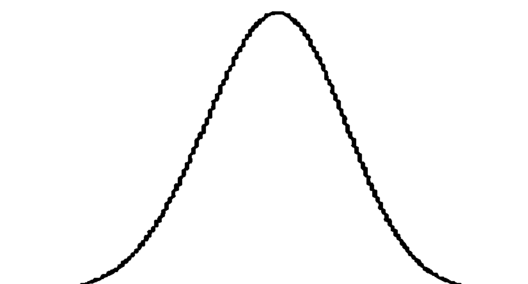


ნახაზი 2



ნახაზი 3

საბოლოოდ, თუ დავუშვებთ, რომ ყველა სრულწლოვანი ქალის სიმაღლის გაზომვა შესაძლებელია ზუსტად, მაშინ ჰისტოგრამა სულ უფრო და უფრო მიუახლოვდება ე. წ. ***ნორმალურ განაწილებას***, რომლის გრაფიკი მოყვანილია მე-4 ნახაზზე. ეს განაწილება აგრეთვე ცნობილია ***გაუსის განაწილების*** სახელწოდებით, გერმანელი მათემატიკოსის კარლ ფრიდრიხ გაუსის საპატივცემლოდ (1777-1855).

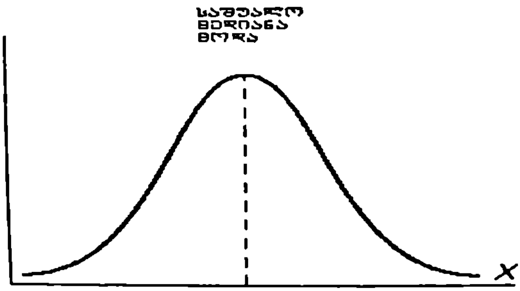


ნახაზი 4

არ არსებობს სიდიდე, რომლის განაწილება ზუსტად ემთხვევა ნორმალურ განაწილებას, ვინაიდან ნორმალური განაწილება თეორიული განაწილებაა. მიუხედავად ამისა, ნორმალური განაწილების გამოყენება შესაძლებელია მრავალი სიდიდის აღსაწერად, ვინაიდან მათი გადახრა ნორმალური განაწილებიდან არის ძალიან მცირე.

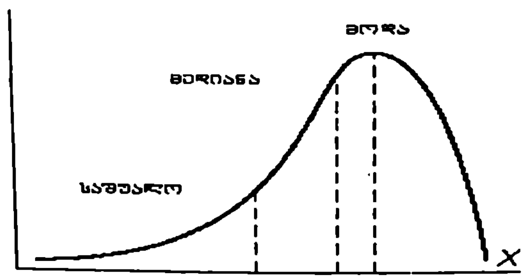
იმ შემთხვევაში, როდესაც სიდიდის მნიშენელობები თანაბრადაა განაწილებული საშუალოს ირგვლივ, განაწილებას ეწოდება **სიმეტრიული**.

სიმეტრიული

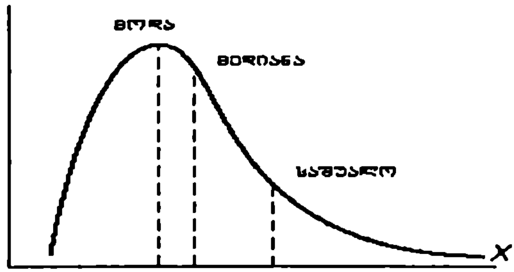


როდესაც სიდიდის მნიშენელობების უმრავლესობა ვარდება საშუალოს მარჯვნივ ან მარცხნივ, მაშინ განაწილებას ეწოდება *ასიმეტრიული*. თუ სიდიდის მნიშენელობების უმრავლესობა მოთავსებულია საშუალოს მარჯვნივ, მაშინ განაწილებას ეწოდება **უარყოფითად** ან **მარცხნივ ასიმეტრიული**. ამ შემთხვევაში საშუალო არის მედიანის მარცხნივ, და საშუალო და მედიანა არიან მოდის მარცხნივ.

უარყოფითად ასიმეტრიული

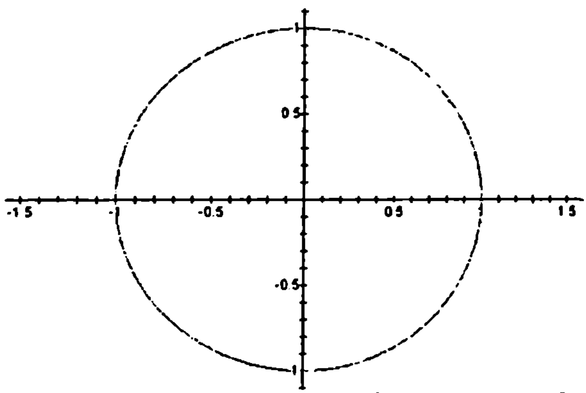


თუ სიდიდის მნიშენელობების უმრავლესობა მოთავსებულია საშუალოს მარცხნივ, მაშინ განაწილებას ეწოდება **დადებითად** ან **მარჯვნივ ასიმეტრიული**. ამ შემთხვევაში საშუალო არის მედიანის მარჯვნივ, ხოლო საშუალო და მედიანა ორივე არის მოდის მარჯვნივ.



წირის "კუდი" მიუთითებს ასიმეტრიის მიმართულებაზე (მარჯვენა არის დადებითი, ხოლო მარცხენა – უარყოფითი).

მათემატიკაში წირები წარმოიდგინება განტოლების სახით. მაგალითად, წრეწირის განტოლებას აქვს სახე: $x^2+y^2=r^2$, სადაც r – წრეწირის რადიუსია.



წრეწირი შეიძლება გამოყენებულ იქნეს მრავალი ფიზიკური ობიექტის წარმოსადგენად, რომელთაც გააჩნიათ წრიული ფორმა (ბორბალი, წრიული მექანიზმები). მიუხედავად იმისა, რომ შეუძლებელია დამზადდეს ბორბალი, რომელიც იქნებოდა სრულიად (იდეალურად) მრგვალი, წრეწირის განტოლება და მისი თვისებები შეიძლება გამოყენებულ იქნეს ბორბლის მრავალი ასპექტის (როგორიცაა ფართობი, სიჩქარე, აჩქარება) შესწავლის დროს. ანალოგიურად, თეორიული წირი, რომელსაც უწოდებენ *ნორმალური განაწილების წირს,* შეიძლება გამოყენებულ იქნეს მრავალი სიდიდის შესწავლის დროს, რომელიც არ არის ზუსტად ნორმალურად განაწილებული, მაგრამ მიუხედავად ამისა არის დაახლოებით ნორმალური.

ნორმალური განაწილების მათემატიკურ განტოლებას აქვს სახე:

$$y=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

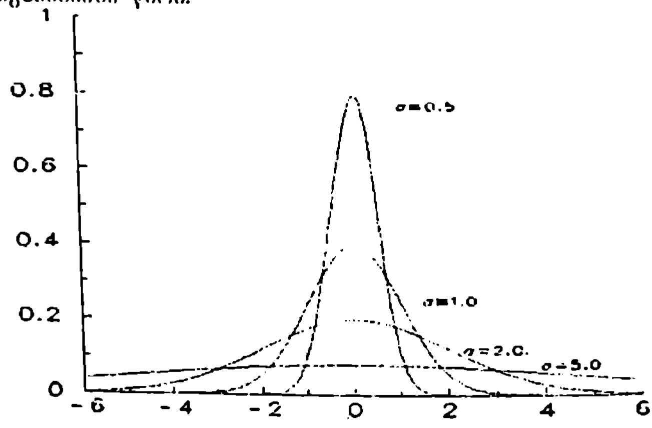
სადაც e — ნეპერის რიცხვია $e \approx 2.718$, $\pi \approx 3.14$, μ -- პოპულაციის საშუალოა და σ — პოპულაციის სტანდარტული გადახრაა.

შეიძლება ეს განტოლება ძნელი აღსაქმელი იყოს, მაგრამ გამოყენებით სტატისტიკაში, კონკრეტული ამოცანების ამოხსნის დროს, ამ განტოლების ნაცვლად გამოიყენება ნორმალური განაწილების ცხრილები.

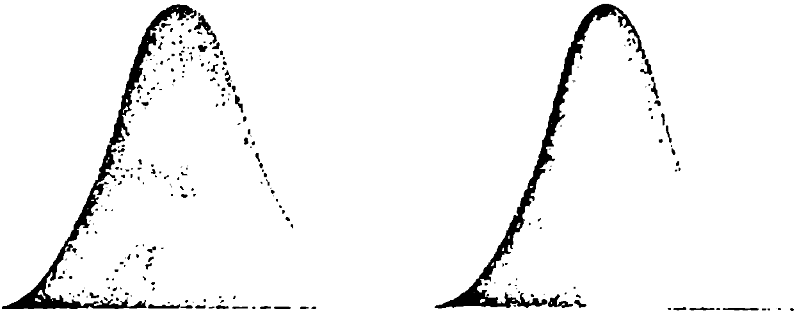
გამოყენებითი სტატისტიკის მეორე მნიშვნელოვანი ასპექტი იმაში მდგომარეობს, რომ ნორმალური განაწილების წირის ქვეშ მოქცეული ფართობი უფრო მნიშვნელოვანია, ვიდრე სიხშირეები. ამიტომ, როცა ხატავენ ნორმალურ განაწილებას, y ღერძი, რომელზეც გამოსახავენ სიხშირეებს, შეიძლება გამოტოვებული იქნეს.

ისევე როგორც წრეწირი შეიძლება იყოს სხვადასხვა ზომის მისი რადიუსის მიხედვით და შეიძლება გამოყენებულ იქნეს სხვადასხვა ზომის ბორბლის წარმოსადგენად, ანალოგიურად, ნორმალური განაწილების წირს შეიძლება ჰქონდეს სხვადასხვა ფორმა და შეიძლება გამოყენებულ იქნეს სხვადასხვა სიდიდეების წარმოსადგენად.

ნორმალური განაწილების ფორმა და მდებარეობა დამოკიდებულია ორ პარამეტრზე: *საშუალოსა* და *სტანდარტულ გადახრაზე*. ნებისმიერ ნორმალურად განაწილებულ სიდიდეს აქვს თავისი ნორმალური განაწილების წირი, რომელიც დამოკიდებულია სიდიდის საშუალოსა და სტანდარტული გადახრის მნიშვნელობაზე. ქვემოთ გამოსახულია ნორმალური განაწილების წირები ერთი და იგივე საშუალოთი და სხვადასხვა სტანდარტული გადახრებით. რაც უფრო დიდია სტანდარტული გადახრა, მით უფრო გაწოლილია შესაბამისი წირი:



ქვემოთ მოყვანილია ორი ნორმალური განაწილება ერთი და იგივე სტანდარტული გადახრითა და განსხვავებული საშუალოთი. ამ წირებს აქვთ ერთი და იგივე ფორმა, მაგრამ განლაგებული არიან x ღერძის სხვადასხვა პოზიციებზე:

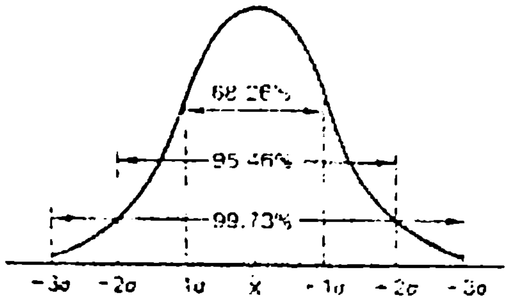


ისტორიული შენიშვნა. ნორმალური განაწილების განტოლების აღმოჩენა უკავშირდება სამი მათემატიკოსის სახელს. 1733 წელს ფრანგმა მათემატიკოსმა აბრაამ დე მუავრმა (Abraham DeMoivre) გამოიყვანა ნორმალური განაწილების განტოლება, როცა ის იკვლევდა მონეტის დიდ რიცხვჯერ აგდებისას გერბის მოსვლის ვარიაციას, მაგრამ მან თავისი ფორმულა აჩვენა მხოლოდ რამოდენიმე ამხანაგს. 100 წლის შემდეგ ორმა მათემატიკოსმა პიერ ლაპლასმა (Pierre Laplace) საფრანგეთში და კარლ გაუსმა (Carl Gauss) გერმანიაში მიიღო ნორმალური განაწილების წირის განტოლება ერთმანეთისაგან დამოუკიდებლად. 1924 წელს კარლ პირსონმა (Karl Pearson) აღმოაჩინა, რომ მუავრს მიღებული ჰქონდა აღნიშნული ფორმულა ლაპლასზე და გაუსზე უფრო ადრე.

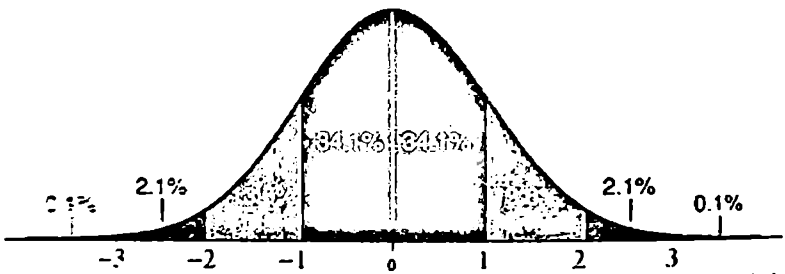
თეორიული ნორმალური განაწილების თვისებები:

1. ნორმალური განაწილების წირი ზარის ფორმისაა.
2. საშუალო, მედიანა და მოდა ტოლია და მოთავსებულია განაწილების ცენტრში.
3. ნორმალური განაწილების წირი უნიმოდალურია (ე. ი. აქვს მხოლოდ ერთი მოდა).
4. წირი სიმეტრიულია საშუალოს მიმართ, რაც ექვივალენტურია იმისა, რომ გვეთქვა: მისი ფორმა ერთი და იგივეა ცენტრზე გამავალი ვერტიკალური წრფის მიმართ.
5. წირი უწყვეტია – ე. ი. მას არა აქვს გამოტოვებული ადგილი. X-ის ნებისმიერ მნიშვნელობას შეესაბამება Y-ს მნიშვნელობა.
6. წირი არსად არ ეხება x ღერძს. თეორიულად, მიუხედავად იმისა თუ რამდენად შორს გავაგრძელებთ წირს ორივე მხარეს, ის არასოდეს შეეხება x ღერძს, მაგრამ უახლოვდება მას რაგინდ მცირე მანძილზე.
7. ნორმალური განაწილების წირის ქვეშ მოქცეული მთლიანი არის ფართობია 1 ანუ 100%. ეს ფაქტი შეიძლება უჩვეულოდ გამოიყურებოდეს, ვინაიდან წირი არსად არ ეხება x ღერძს, მაგრამ მისი დამტკიცება მათემატიკურად შესაძლებელია.
8. ნორმალური განაწილების წირის ქვეშ საშუალოდან ერთი სტანდარტული გადახრის მანძილზე მოთავსებული არის ფართობი დაახლოებით 0.68 ანუ 68%-ია, საშუალოდან ორი სტანდარტული

გადახრის მანძილზე დაახლოებით 0.95 ანუ 95%, და სამი სტანდარტული გადახრის მანძილზე – დაახლოებით 0.997 ანუ 99.7%:



რადგანაც ყოველ ნორმალურად განაწილებულ სიდიდეს აქვს საკუთარი საშუალო და სტანდარტული გადახრა, ამიტომ მათი წირების ფორმა და მდებარეობა განსხვავებულია. პრაქტიკულ ამოცანებში მოუხერხებელია სხვადასხვა ცხრილების გამოყენება. ამიტომ სიტუაციის გამარტივებისათვის სტატისტიკოსები იყენებენ ე. წ. *სტანდარტულ ნორმალურ განაწილებას*. ნორმალურ განაწილებას საშუალოთი 0 და სტანდარტული გადახრით 1 ეწოდება **სტანდარტული ნორმალური განაწილება**. ქვემოთ მოყვანილია სტანდარტული ნორმალური განაწილების წირის ქვეშ მოქცეული თითოეული ნაწილის შესაბამისი პროცენტი (მაგალითად, იმ არის ფართობი, რომელიც მოთავსებულია საშუალოსა და ერთ სტანდარტულ გადახრას შორის საშუალის მარჯვენივ ან მარცხნივ ტოლია 0.3413-ის ანუ 34.13%-ის):



სტანდარტული ნორმალური განაწილების განტოლებას აქვს სახე:

$$y=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

ნორმალურად განაწილებული სიდიდე გარდაიქმნება სტანდარტულ ნორმალურად განაწილებულ სიდიდედ სტანდარტიზირების ფორმულის გამოყენებით:

$$z=\frac{\text{მნიშვნელობა - საშუალო}}{\text{სტანდარტული გადახრა}} \quad \text{ანუ} \quad z=\frac{X-\mu}{\sigma}.$$

როგორც აღნიშნული იყო ნორმალური განაწილების წირის ქვეშ

მოქცეული არე გამოიყენება პრაქტიკული ამოცანების ამოხსნის დროს. მაგალითად, იმის გასაგებად, თუ სრულწლოვანი ქალების რამდენი პროცენტის სიმაღლეა მოთავსებული 155სმ-სა და 165სმ-ს შორის, ან იმის ალბათობის გასაგებად, რომ ელემენტი იმუშავებს 4 წელზე უფრო დიდხანს. შესაბამისად, ქვემოთ ჩვენ ძირითად ყურადღებას დავუთმობთ z-ის ნებისმიერი მნიშვნელობისათვის ნორმალური განაწილების წირის ქვეშ მოქცეული არის ფართობის პოვნის პროცედურის აღწერას. პირველ რიგში X-ს მნიშვნელობას გარდავქმნით ზემოთ მოყვანილი ფორმულის გამოყენებით, და მიღებულ მნიშვნელობას ვუწოდებთ z **მნიშვნელობას.** z მნიშვნელობა სინამდვილეში არის სტანდარტული გადახრის რაოდენობა X-ის მნიშვნელობის დაშორებაში საშუალოდან. ნორმალური განაწილების ცხრილში მოყვანილია სტანდარტული ნორმალური განაწილების წირის ქვეშ z მნიშვნელობამდე მოთავსებული არის ფართობის მნიშვნელობები.

შენიშვნა. სახელმძღვანელოებში მოყვანილია როგორც $\Phi_0(z)$, ისე $\Phi(z)$ ფუნქციების ცხრილები და ორივეს უწოდებენ ნორმალური განაწილების ცხრილებს. ამასთანავე, $\Phi_0(z)$ გვიჩვენებს სტანდარტული ნორმალური განაწილების წირის ქვეშ 0-დან დადებით z მნიშვნელობამდე მოთავსებული არის ფართობს და ბუნებრივია მისი ცხრილებით ვისარგებლებთ სწორედ აღნიშნული არის ფართობის გამოთვლისას. მეორეს მხრივ, $\Phi(z)$ გვიჩვენებს სტანდარტული ნორმალური განაწილების წირის ქვეშ $-\infty$-დან დადებით z მნიშვნელობამდე მოთავსებული არის ფართობს (განმარტების თანახმად სწორედ ის არის განაწილების ფუნქცია) და მისი ცხრილებით ბუნებრივია ვისარგებლოთ დადებითი z მნიშვნელობის მარცხნივ მდებარე უსასრულო არის ფართობის გამოთვლის შემთხვევაში. ამ ფუნქციებს შორის არსებობს მარტივი კავშირი $\Phi(z)=\Phi_0(z)+0.5000$, რაც აიხსნება იმით, რომ 0-ის მარცხნივ მდებარე უსასრულო არის ფართობი ტოლია 0.5000-ის (მთლიანად ნორმალური წირის ქვეშ მდებარე არის ფართობი ტოლია 1-ის). ორივე აღნიშნული ცხრილი მოყვანილია მხოლოდ დადებითი z მნიშვნელობებისათვის. მაგრამ, ნორმალური წირის საშუალოს მიმართ სიმეტრიულობის ძალით, ცხადია, რომ $\Phi_0(-z)=\Phi_0(z)$, ხოლო $\Phi(-z)=1-\Phi(z)$.

$$\Phi(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-t^2/2}dt \qquad \Phi_0(z)=\frac{1}{\sqrt{2\pi}}\int_{0}^{z}e^{-t^2/2}dt$$

სტანდარტული ნორმალური განაწილების წირის ქვეშ მოთავსებული არის ფართობის გამოსათვლელად საჭიროა ოთხი ნაბიჯი:

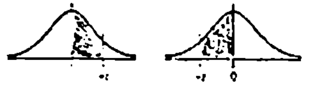
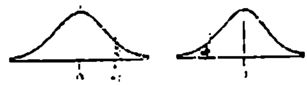
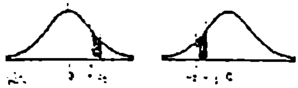
ნაბიჯი 1. დავხაზოთ ნახაზი.

ნაბიჯი 2. მოვნიშნოთ სასურველი არე.

ნაბიჯი 3. ვიპოვოთ შესაბამისი ნახაზი პროცედურულ ცხრილში.

ნაბიჯი 4. პროცედურული ცხრილის შესაბამის ბლოკში მითითებული პროცედურით ვიპოვოთ სასურველი ფართობი.

სტანდარტული ნორმალური განაწილების წირის ქვეშ მდებარე არის ფართობის მოსაძებნი პროცედურული ცხრილი

1. 0-სა და ნებისმიერ z-ს შორის: ნორმალური განაწილების ცხრილიდან ვპოულობთ $\Phi_0(z)$-ს. 	2. ნებისმიერი კუდის შემთხვევა: ა. ვნახოთ $\Phi_0(z)$-ის მნიშვნელობა ცხრილში. ბ. 0.5000-ს გამოვაკლოთ $\Phi_0(z)$. 
3. z-ის ორ მნიშვნელობას შორის საშუალოს ერთ მხარეს: ა. ვიპოვოთ $\Phi_0(z_1)$ და $\Phi_0(z_2)$. ბ. დიდს გამოვაკლოთ პატარა. 	4. z-ის ორ მნიშვნელობას შორის საშუალოს სხვადასხვა მხარეს: ა. ვიპოვოთ $\Phi_0(z_1)$ და $\Phi_0(z_2)$. ბ. შევკრიბოთ ისინი. 
5. საშუალოზე მეტი z-ის მარცხნივ: ა. ცხრილში ვპოულობთ $\Phi_0(z)$-ს. ბ. ვუმატებთ მას 0.5000-ს. 	6. საშუალოზე ნაკლები $-z$-ის მარჯვნივ: ა. ცხრილში ვპოულობთ $\Phi_0(z)$-ს. ბ. ვუმატებთ მას 0.5000-ს. 
7. ორი კუდის შემთხვევა: ა. ვიპოვოთ $\Phi_0(z_1)$ და $\Phi_0(z_2)$. ბ. 0.5000-ს გამოვაკლოთ თითოეული. გ. შევკრიბოთ სხვაობები. 	პროცედურა 1. დავხაზოთ ნახაზი. 2. მოვნიშნოთ სასურველი არე. 3. ვიპოვოთ შესაბამისი ბლოკი. 4. ვიმოქმედოთ ბლოკის სქემით. შენიშვნა: $\Phi(z)=\Phi_0(z)+0.5000$, როცა $z>0$.

სიტუაცია 1. ვიპოვოთ ნორმალური წირის ქვეშ 0-სა და z-ს შორის მოქცეული არის ფართობი.

მაგალითი 1. ვიპოვოთ ნორმალური წირის ქვეშ $z=0$-სა და $z=2.34$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის პირველი ბლოკის სქემით. ნორმალური განაწილების ცხრილის მარცხენა სვეტში ვიპოვოთ რიცხვი 2.3, ხოლო ზედა სტრიქონში კი – რიცხვი .04. 2.3-ის შესაბამისი სტრიქონისა და .04-ის შესაბამისი სვეტის გადაკვეთაზე მდგომი რიცხვი 0.4904 ანუ 49.04% იქნება პასუხი.

მაგალითი 2. ვიპოვოთ ნორმალური წირის ქვეშ $z=0$-სა და $z=1.8$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის პირველი ბლოკის სქემით. ნორმალური განაწილების ცხრილის მარცხენა სვეტში ვიპოვოთ რიცხვი 1.8, ხოლო ზედა სტრიქონში კი – რიცხვი .00. 1.8-ის შესაბამისი სტრიქონისა და .00-ის შესაბამისი სვეტის გადაკვეთაზე მდგომი რიცხვი 0.4641 იქნება პასუხი.

მაგალითი 3. ვიპოვოთ ნორმალური წირის ქვეშ $z=-1.75$-სა და $z=0$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. როგორც ვიცით, ნორმალური განაწილების ცხრილში არ მონაწილეობს z-ის უარყოფითი მნიშვნელობები. მაგრამ, ვინაიდან ნორმალური განაწილება სიმეტრიულია საშუალოს (ამ შემთხვევაში 0-ის) მიმართ, ამიტომ ნულის მარცხნივ მოთავსებული არის ფართობი ემთხვევა ნულის მარჯვნივ მოთავსებული მისი სიმეტრიული არის ფართობს. შესაბამისად, საძიებელი ფართობი ტოლი იქნება $z=0$-სა და $z=1.8$-ს შორის მოთავსებული არის ფართობის. ამიტომ პასუხი იქნება 1.7-ის შესაბამისი სტრიქონისა და .05-ის სესაბამისი სვეტის გადაკვეთაზე მდებარე რიცხვი: 0.4599 ანუ 45.99%.

სიტუაცია 2. ვიპოვოთ ნორმალური წირის კუდის ქვეშ მოქცეული არის ფართობი.

მაგალითი 4. ვიპოვოთ ნორმალური წირის ქვეშ $z=1.11$-ის მარჯვნივ მოთავსებული არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მეორე ბლოკის სქემით. ვინაიდან ნორმალური განაწილების ცხრილი გვაძლევს $z=0$-სა და $z=1.11$-ს შორის მოქცეული არის ფართობს, ამიტომ პასუხს მივიღებთ თუ ნახევარი არის ფართობს (0.5000-ს) გამოვაკლებთ 1.1-ის შესაბამისი სტრიქონისა და .01-ის შესაბამისი სვეტის გადაკვეთაზე მდგომ რიცხვს – 0.3665-ს. პასუხი იქნება 0.5000-0.3665=0.1335 ანუ 13.35%.

მაგალითი 5. ვიპოვოთ ნორმალური წირის ქვეშ $z=-1.93$-ის მარცხნივ მოთავსებული არის ფართობი.

ამოხსნა. ნორმალური განაწილების სიმეტრიულობის თვისებიდან გამომდინარე საძიებელი ფართობი ტოლია $z=1.93$-ის მარჯვნივ მოთავსებული არის ფართობის. ამის შემდეგ ვიქცევით როგორც წინა მაგალითში. 0.5000-ს ვაკლებთ 0-სა და 1.93-ს შორის მოთავსებული არის ფართობს – 0.4732-ს და ვღებულობთ რომ საძიებელი ფართობია 0.5000-0.4732=0.0268 ანუ 2.68%.

სიტუაცია 3. ვიპოვოთ ნორმალური წირის ქვეშ საშუალოს ერთი და იგივე მხარეს ორ მნიშვნელობას შორის მოქცეული არის ფართობი.

მაგალითი 6. ვიპოვოთ ნორმალური წირის ქვეშ $z=2.00$-სა და $z=2.47$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მესამე ბლოკის სქემით. ვიპოვოთ $z=0$-სა და $z=2.47$-ს შორის მოქცეული არის ფართობი (ის ტოლია 0.4932-ის), $z=0$-სა და $z=2$-ს შორის მოქცეული არის ფართო-

ბი (ის ტოლია 0.4772-ის) და პირველ ფართობს გამოვაკლოთ მეორე ფართობი. გვაქვს: 0.4932-0.4772=0.0160 ანუ 1.60%.

მაგალითი 7. ვიპოვოთ ნორმალური წირის ქვეშ $z=-2.48$-სა და $z=-0.83$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. ნორმალური განაწილების სიმეტრიულობის გამო საძიებელი ფართობი ტოლია ფართობის $z=0.83$-სა და $z=2.48$-ს შორის. ამიტომ წინა მაგალითის ანალოგიით პასუხი იქნება 0.4934-0.2967=0.1967 ანუ 19.67%.

სიტუაცია 4. ვიპოვოთ ნორმალური წირის ქვეშ საშუალოს სხვადასხვა მხარეს ორ მნიშვნელობას შორის მოქცეული არის ფართობი.

მაგალითი 8. ვიპოვოთ ნორმალური წირის ქვეშ $z=-1.37$-სა და $z=1.68$-ს შორის მოქცეული არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მეოთხე ბლოკის სქემით. ვინაიდან აღნიშნული არე წარმოადგენს საშუალოს მარცხნივ და მარჯვენივ მდებარე ორი არის გაერთიანებას, ამიტომ ჩვენ უნდა ვიპოვოთ თითოეული არის ფართობი და შევკრიბოთ. $z=-1.37$-სა და $z=0$-ს შორის მოქცეული არის ფართობი ემთხვევა $z=0$-სა და $z=1.37$-ს შორის მოქცეული არის ფართობს და ტოლია 0.4147-ის, ხოლო $z=0$-სა და $z=1.68$-ს შორის მოქცეული არის ფართობი ტოლია 0.4535-ის. ამიტომ საძიებელი ფართობი იქნება 0.4147+0.4535=0.8682 ანუ 86.82%.

სიტუაცია 5. ვიპოვოთ ნორმალური წირის ქვეშ საშუალოზე დიდი მნიშვნელობის მარცხნივ მოთავსებულილი არის ფართობი.

მაგალითი 9. ვიპოვოთ ნორმალური წირის ქვეშ $z=1.99$-ის მარცხნივ მდებარე არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მეხუთე ბლოკის სქემით. ვინაიდან ნორმალური განაწილების ცხრილი გვაძლევს $z=0$-სა და $z=1.99$-ს შორის მდებარე არის ფართობს (რომელიც ტოლია 0.4767-ს), მას უნდა მივუმატოთ $z=0$-ის მარცხნივ მდებარე მთელი არის ფართობი (რომელიც ტოლია 0.5000-ის). ამიტომ საძიებელი ფართობი იქნება

0.4767+0.5000=0.9767 ანუ 97.67%.

სიტუაცია 6. ვიპოვოთ ნორმალური წირის ქვეშ საშუალოზე ნაკლები მნიშვნელობის მარჯვნივ მოთავსებულილი არის ფართობი.

მაგალითი 10. ვიპოვოთ ნორმალური წირის ქვეშ $z=-1.16$-ის მარჯვნივ მდებარე არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მეექვსე ბლოკის სქემით. საძიებელი ფართობი ტოლია $z=-1.16$-სა და $z=0$-ს შორის მდებარე არის ფართობისა (რომელიც ემთხვევა $z=0$-სა და $z=1.16$-ს შორის მდებარე არის ფართობს და ტოლია 0.3770-ის) და $z=0$-ის მარჯვენივ მდებარე მთელი არის ფართობის (რომელიც ტოლია 0.5000-ის) ჯამის. ამიტომ საძიებელი ფართობი იქნება: 0.3770+0.5000=0.8770 ანუ 87.7%.

სიტუაცია 7. ვიპოვოთ ნორმალური წირის ორი კუდის ერთობლივი ფართობი.

მაგალითი 11. ვიპოვოთ ნორმალური წირის ქვეშ $z=-3.01$-ის მარცხნივ და $z=2.43$-ის მარჯვენივ მდებარე არის ფართობი.

ამოხსნა. ვისარგებლოთ პროცედურული ცხრილის მეშვიდე ბლოკის სქემით. უნდა ვიპოვოთ თითოეული კუდის ფართობი და შევკრიბოთ ისინი. $z=-3.01$-ის მარცხნივ მდებარე არის ფართობი ტოლია (იხ. მაგალითი 5) 0.5000-0.4987=0.0013. $z=2.43$-ის მარჯვნივ მდებარე არის ფართობი ტოლია (იხ. მაგალითი 4) 0.5000-0.4925=0.0075. ამიტომ საძიებელი ფართობი იქნება 0.0013+0.0075=0.0088 ანუ 0.88%.

ნორმალური განაწილების წირი როგორც ალბათური განაწილების წირი.

ნორმალური განაწილების წირი შეიძლება განვიხილოთ როგორც ნორმალურად განაწილებული შემთხვევითი სიდიდის ალბათური განაწილების წირი. გავიხსენოთ, რომ ნორმალური განაწილება არის *უწყვეტი განაწილება*. მისი უწყვეტობა ნიშნავს, რომ წირს არა აქვს წყვეტა (გამოტოვებული ადგილი), სხვა სიტყვებით, ნებისმიერ z მნიშვნელობას x ღერძზე შეესაბამება გარკვეული მნიშვნელობა y ღერძზე.

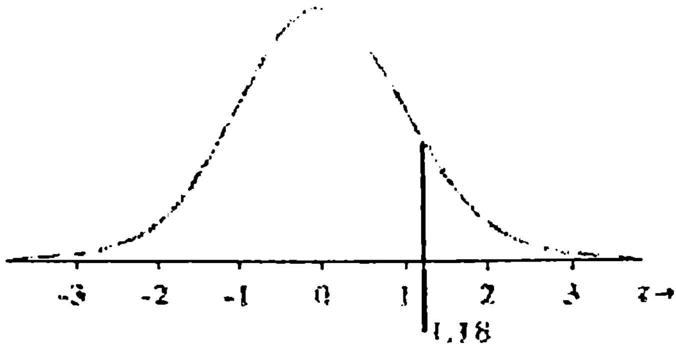
როგორც ზემოთ იყო აღნიშნული, ნორმალური წირის ქვეშ არის ფართობი მნიშვნელოვანია ვიდრე სიხშირეები. *ეს ფართობები შეესაბამება ალბათობებს.* თუ შესაძლებელია, რომ z-ის ნებისმიერი მნიშვნელობა შეირჩეს შემთხვევით, მაშინ ალბათობა იმისა, რომ ის არჩეული იქნება 0-სა და 2.00-ს შორის იქნება იგივე რაც ნორმალური წირის ქვეშ 0-სა და 2.00-ს შორის მოთავსებული არის ფართობი. ამიტომ ალბათობა იმისა, რომ z-ის შერჩეული მნიშვნელობა აღმოჩნდება 0-სა და 2.00-ს შორის ტოლია 0.4772-ის. შესაბამისად, რაიმე შუალედში სიდიდის მნიშვნელობის მოხვედრის ალბათობის გამოთვლის ამოცანა იხსნება იმავე წესით, როგორც წინა მაგალითებში იხსნებოდა შესაბამისი არის ფართობის პოვნის ამოცანა. მაგალითად, თუ ამოცანა მდგომარეობს შერჩეული z-ის მნიშვნელობის 2.25-სა და 2.94-ს შორის მოხვედრის ალბათობის გამოთვლაში, ის უნდა ამოიხსნას პროცედურული ცხრილის მესამე ბლოკში მითითებული სქემის მიხედვით. ამასთანავე, ალბათობების შემთხვევაში სხვა აღნიშვნები გამოიყენება. კერძოდ, წინა მაგალითში ალბათობა ასე იწერება:

$$P(2.25<z<2.94).$$

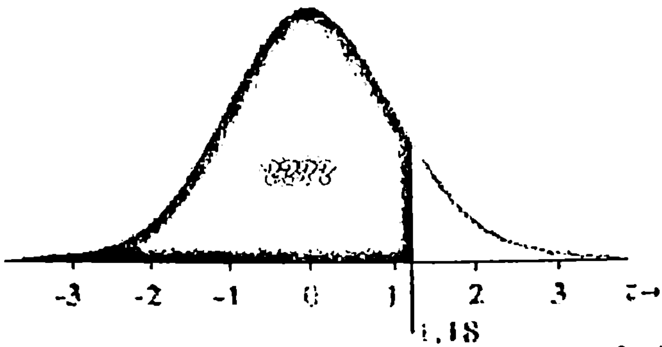
შენიშვნა 1. ვინაიდან მონაკვეთის ფართობი ნულის ტოლია, ამიტომ ალბათობების გამოთვლის დროს მნიშვნელობა არა აქვს უტოლობა მკაცრი იქნება თუ არამკაცრი. მაგალითად, $P(Z<1.79)=P(Z\le 1.79)$.

მაგალითი 1. ვიპოვოთ ალბათობა იმისა, რომ ნორმალური სიდიდე ნაკლებია 1.18-ზე, $P(Z<1.18)$?

ამოხსნა. დავხაზოთ სტანდარტული ნორმალური განაწილების სიმკვრივის წირი და აბსცისთა ღერძზე ავღნიშნოთ 1.18-ის შესაბამისი წერტილი (ამ შემთხვევაში $z=1.18$).



საძიებელი ალბათობა ემთხვევა ნორმალური წირის ქვეშ 1.18-ის მარცხნივ მოთავსებული არის ფართობს, შესაბამისად უნდა ვისარგებლოთ პროცედურული ცხრილის მეხუთე ბლოკის სქემით.

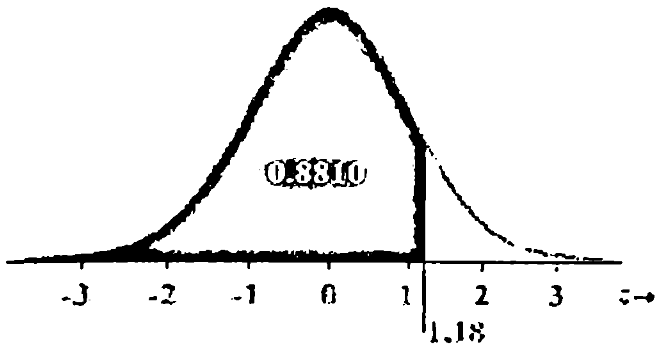


ნორმალური განაწილების ცხრილის პირველ სვეტში მოვძებნოთ რიცხვი 1.1, ხოლო პირველ სტრიქონში კი – რიცხვი .08.

z	.00	.01		.08	.09
0.0	.5000	.5040		.5319	.5359
0.1	.5398	.5438		.5714	.5753
...	...	...		...	...
1.0	.8413	.8438		.8599	.8621
1.1	.8643	.8665		**.8810**	8830
1.2	.8849	.8869		.8997	.9015
...	...	...		...	...

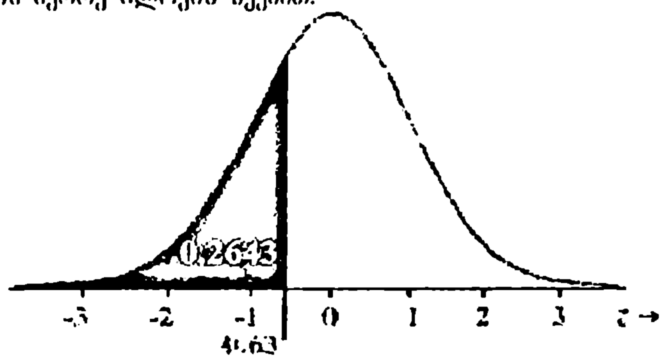
1.1-ის შესაბამისი სტრიქონისა და .08-ის შესაბამისი სვეტის გადაკვეთაზე ვპოულობთ რიცხვს – 0.8810. ამიტომ საძიებელი ალბათობა იქნება

$P(Z<1.18)=0.8810$ ანუ 88.10%.



მაგალითი 2. ვიპოვოთ ალბათობა იმისა, რომ ნორმალური სიდიდე ნაკლებია -0.63-ზე, $P(Z<-0.63)$?

ამოხსნა. ამ შემთხვევაში $z=-0.63$ და ვსარგებლობთ პროცედურული ცხრილის მეორე ბლოკის სქემით.



საძიებელი ალბათობა იქნება $P(Z<-0.63)=0.5000-0.2357=0.2643$ ანუ 26.43%.

შენიშვნა 2. იგივე ალბათობა ჩვენ შეგვეძლია ვიპოვოთ სტანდარტული ნორმალური განაწილების ფუნქციის განმარტების გამოყენებით. როგორც ცნობილია, თუ Z სტანდარტული ნორმალური შემთხვევითი სიდიდეა, მაშინ:

$$\Phi(z)=P(Z\le z)=P(Z<z).$$

გარდა ამისა, ვინაიდან ნორმალური განაწილების წირის ქვეშ მოთავსებული მთელი არის ფართობი ტოლია 1-ის და ნორმალური განაწილების წირი სიმეტრიულია საშუალოს მიმართ, ამიტომ ცხადია, რომ ნებისმიერი უარყოფითი $-z$ მნიშვნელობისათვის:

$$P(Z<-z)=P(Z>z).$$

მეორეს მხრივ, საწინააღმდეგო ხდომილების ალბათობის გამოთვლის წესის თანახმად, შეგვეძლია დავწეროთ, რომ:

$$P(Z>z)=1-P(Z\le z).$$

ყოველივე ზემოთ თქმულის გაერთიანებით ვრწმუნდებით, რომ:

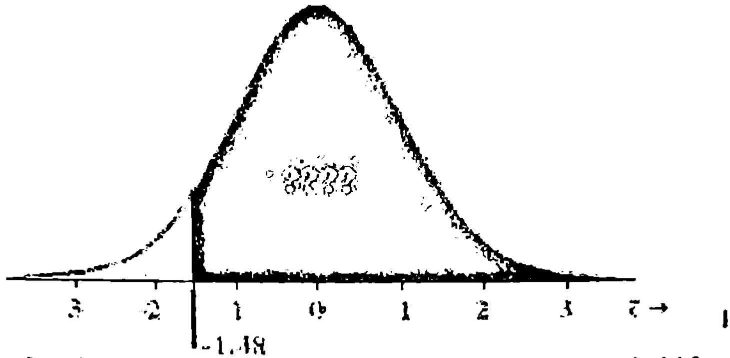
$$P(Z<-z)=P(Z>z)=1-P(Z\le z)=1-\Phi(z).$$

შესაბამისად, სტანდარტული ნორმალური განაწილების ცხრილებიდან ვასკვნით, რომ საძიებელი ალბათობა ტოლია

$$P(Z<-0.63)=1-\Phi(0.63)=1-0.7357=0.2643.$$

მაგალითი 3. ვიპოვოთ ალბათობა იმისა, რომ ნორმალური სიდიდე მეტია -1.48-ზე, $P(Z>-1.48)$?

ამოხსნა. I გზა. ამ შემთხვევაში $z=-1.48$ და გამოსათვლელია ნორმალური განაწილების წირის ქვეშ -1.48-ის მარჯვნივ მოთავსებული არის ფართობი.



ვისარგებლობთ პროცედურული ცხრილის მექვსე ბლოკის სქემით. $z=-1.48$-სა და $z=0$-ს შორის მდებარე არის ფართობი ემთხვევა $z=0$-სა და $z=1.48$-ს შორის მოთავსებული არის ფართობს და ტოლია 0.4306-ის. ამიტომ საძიებელი ალბათობა იქნება:

$$P(Z>-1.48)=0.5000+0.4306=0.9306 \quad \text{ანუ} \quad 93.06\%.$$

II გზა. იგივე შედეგს მივიღებთ, თუ ვისარგებლებთ ნორმალური განაწილების სიმეტრიულობითა და $\Phi(z)$ ფუნქციის ცხრილებით:

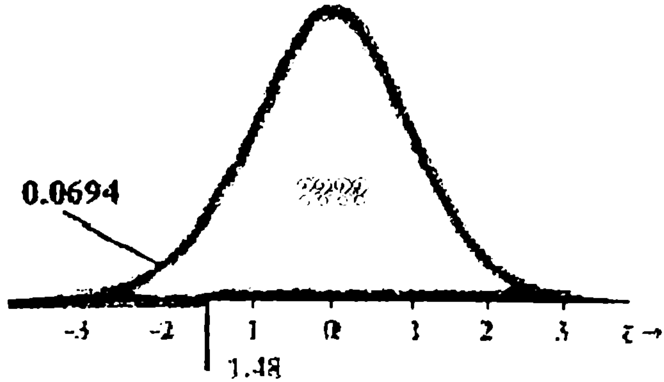
$$P(Z>-1.48)=P(Z<1.48)=\Phi(1.48)=0.9306.$$

III გზა. შეგვიძლია ვისარგებლოთ საწინააღმდეგო ხდომილების ალბათობის გამოსათვლელი ფორმულით და მაშინ მაგალითი დაიყვანება წინა მაგალითზე:

$$P(Z>-1.48)=1-P(Z\le-1.48)=1-P(Z<-1.48).$$

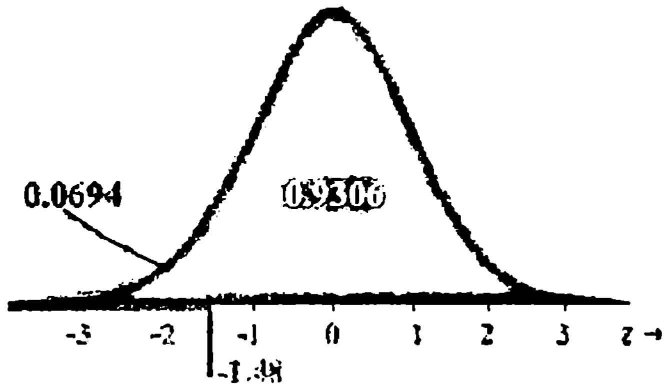
წინა მაგალითის ანალოგიურად გვაქვს:

$$P(Z<-1.48)=0.5000-0.4306=0.0694.$$



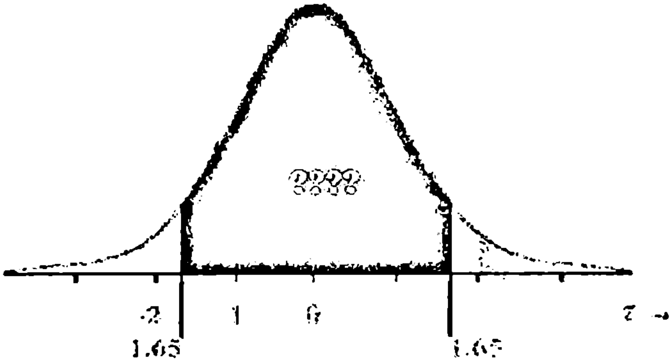
საბოლოოდ კი ვღებულობთ, რომ საძიებელი ალბათობა ტოლია:

$$P(Z>-1.48)=1-P(Z<-1.48)=1-0.0694=0.9306.$$



მაგალითი 4. ვიპოვოთ ალბათობა იმისა, რომ ნორმალური სიდიდე მოთავსებულია -1.65-სა და 1.65-ს შორის, $P(-1.65<Z<1.65)$?

ამოხსნა. I გზა. ამ შემთხვევაში საძიებელი ალბათობა იგივე რაც საშუალოს სხვადასხვა მხარეს მოთავსებულ ორ მნიშვნელობას $-z=-1,65$ -სა და $z=1.65$-ს შორის მოთავსებული არის ფართობი.



ჩვენ შეგვიძლია ვისარგებლოთ პროცედურული ცხრილის მეოთხე ბლოკის სქემით. შესაბამისად, ჩვენ უნდა ვიპოვოთ როგორც $-z=-1,65$-სა და $z=0$-ს შორის მდებარე, ისე $z=0$-სა და $z=1.65$-ს შორის მდებარე არეების ფართობები და შემდეგ შევკრიბოთ ისინი. ამიტომ საძიებელი ალბათობა იქნება:

$$P(-1.65<Z<1.65)=0.4505+0.4505=0.9010 \quad \text{ანუ} \quad 90.10\%.$$

II გზა. ნორმალური განაწილების სიმეტრიულობის გამო საძიებელი ალბათობა იქნება $z=0$-სა და $z=1.65$-ს შორის მდებარე არის გაორკეცებული ფართობი. ამიტომ პროცედურული ცხრილის პირველი ბლოკის სქემის მიხედვით გვექნება, რომ:

$$P(-1.65<Z<1.65)=2\times 0.4505=0.9010.$$

III გზა. ვისარგებლოთ სხვაობის ალბათობის ფორმულით (თუ $A\subset B$, მაშინ $P(B\setminus A)=P(B)-P(A)$). ცხადია, რომ თუ $z_1<z_2$, მაშინ იმ მნიშვნელობების სიმრავლე, რომელიც z_1-ზე ნაკლებია, მითუმეტეს ნაკლები იქნება z_2-ზე, ანუ $(Z\le z_1)\subset(Z<z_2)$. მეორეს მხრივ, გასაგებია, რომ

$$(z_1<Z<z_2)=(Z<z_2)\setminus(Z\le z_1).$$

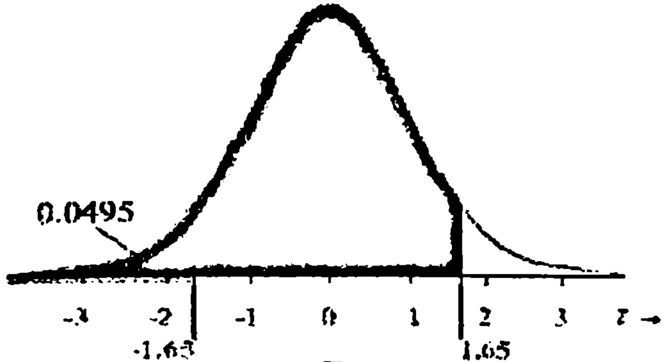
ამიტომ სხვაობის ალბათობის ფორმულის თანახმად გვაქვს:

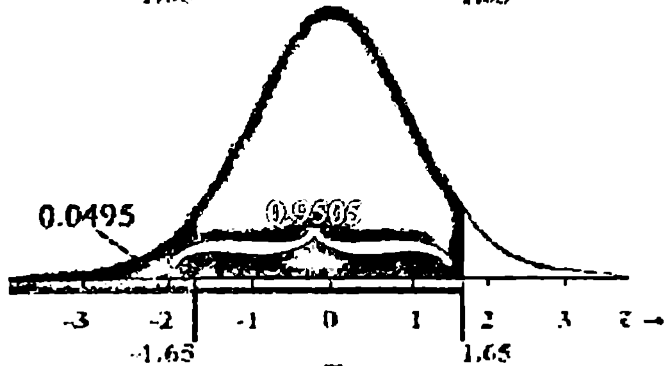
$$P(z_1<Z<z_2)=P[(Z<z_2)\setminus(Z\le z_1)]=P(Z<z_2)-P(Z\le z_1)=\Phi(z_2)-\Phi(z_1).$$

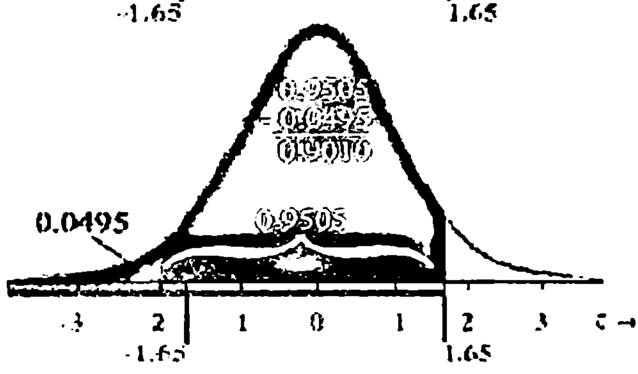
ამ შემთხვევაში საძიებელი ალბათობა იქნება;

$$P(-1.65<Z<1.65)=\Phi(1.65)-\Phi(-1.65)=$$
$$=\Phi(1.65)-[1-\Phi(1.65)]=2\Phi(1.65)-1=2\times 0.9505-1=0.9010.$$

შევნიშნავთ, რომ მესამე გზის პროცედურები სქემატურად გამოსახულია ქვემოთ მოყვანილ სამ ნახაზზე:

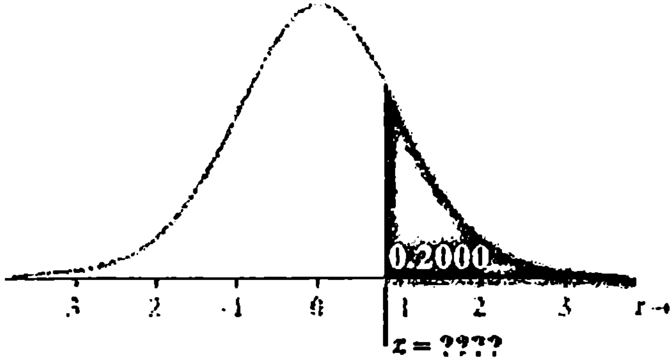




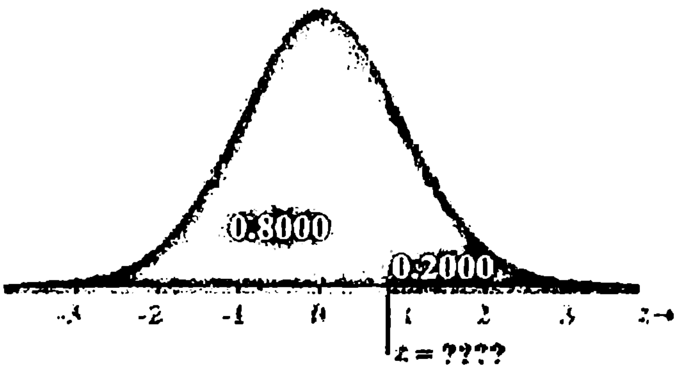


აღსანიშნავია, რომ ზოგჯერ საჭიროა გარკვეული აზრით შებრუნებული ამოცანების ამოხსნა: საჭიროა მოიძებნოს z-ის ისეთი მნიშვნელობა (შესაბამისად, მნიშვნელობები), რომ მის მიერ (შესაბამისად, მათ მიერ) განსაზღვრული არის ფართობი ტოლი იყოს წინასწარ მოცემული რიცხვის.

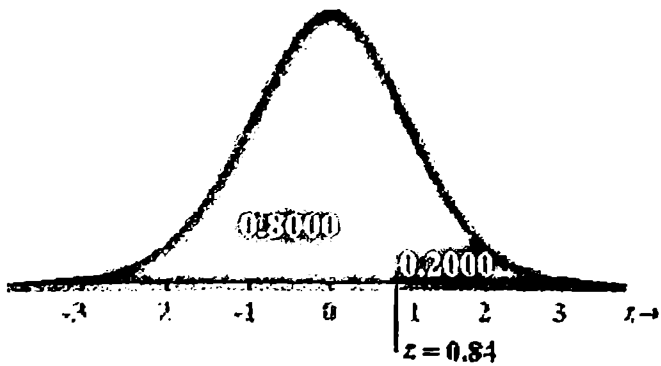
მაგალითი 5. ვიპოვოთ z-ის ისეთი მნიშვნელობა, რომლის მარჯვნივ ნორმალური განაწილების წირის ქვეშ მოთავსებული არის ფართობი ტოლია 0.2000-ის?



ამოხსნა. ცხადია, რომ ეს ამოცანა ტოლფასია z-ის ისეთი მნიშვნელობის მოძებნის, რომლის მარცხნივ ნორმალური განაწილების წირის ქვეშ მოთავსებული არის ფართობი ტოლია 1-0.2000=0.8000-ის.



ნორმალური განაწილების ცხრილში ვპოულობთ 0.8000-სთან ყველაზე ახლოს მდგომ რიცხვს -- 0.7995-ს. ეს რიცხვი დგას 0.8-ის შესაბამისი სტრიქონისა და .04-ის შესაბამისი სვეტის გადაკვეთაზე. ამიტომ გასაგებია, რომ z-ის საძიებელი მნიშვნელობაა $z = 0.84$.



შენიშვნა 3. მაგალითი 5 ექვივალენტურია შემდეგი მაგალითის: ვიპოვოთ z-ის ისეთი მნიშვნელობა, რომლისთვისაც ალბათობა იმისა, რომ სტანდარტული ნორმალური კანონით განაწილებული შემთხვევითი სიდიდე მიიღებს მნიშვნელობას z-ის მარჯვნივ ტოლია 0.2000-ის,

$$P(Z > z) = 0.2000\,?$$

ვიგულისხმოთ, რომ Z – სტანდარტული ნორმალური კანონით განაწილებული შემთხვევითი სიდიდე, ხოლო $0 < \alpha < 1$ – მოცემული რიცხვია.

განმარტება 1. z-ის იმ მნიშვნელობას, რომლისთვისაც

$$P(Z > z) = \alpha$$

სტანდარტული ნორმალური განაწილების **ზედა α-კრიტიკული წერტილი** ეწოდება და აღინიშნება z_α სიმბოლოთი.

მაგალითი 5-ის თანახმად: $z_{0.2000} = 0.84$.

განმარტება 2. x-ის იმ მნიშვნელობას, რომლისთვისაც

$$P(Z \le x) = \alpha$$

სტანდარტული ნორმალური განაწილების **α-დონის კვანტილი** (ან უბრალოდ **α-კვანტილი**) ეწოდება და აღინიშნება x_α სიმბოლოთი.

ცხადია, რომ $x_\alpha = z_{1-\alpha}$ და $z_\alpha = x_{1-\alpha}$, ანუ, თუ რაიმე წერტილის მარჯვნივ მოთავსებული არის ფართობი ტოლია α-სი, მაშინ მის მარცხნივ მოთავსებული არის ფართობი იქნება $1-\alpha$ და პირიქით.

ნორმალური განაწილების გამოყენება.

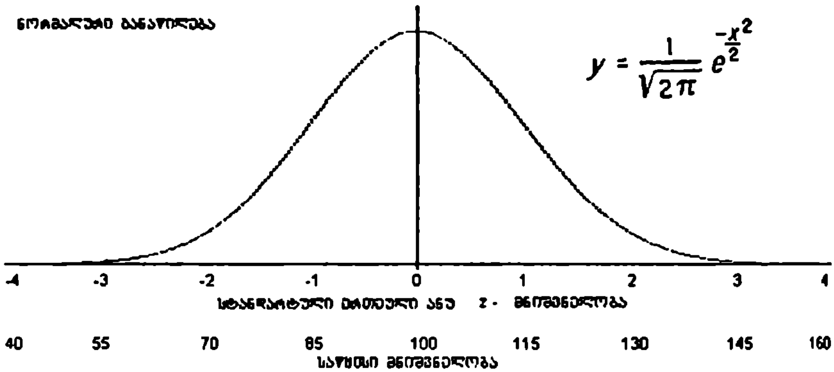
სტანდარტული ნორმალური განაწილების წირი შეიძლება გამოყენებულ იქნეს მრავალი განსხვავებული პრაქტიკული ამოცანის გადაწყვეტისას. ერთადერთი მოთხოვნა მდგომარეობს იმაში, რომ სიდიდე იყოს ნორმალური ან დაახლოებით ნორმალურად განაწილებული. არსებობს მრავალი მათემატიკური ტესტი იმის დასადგენად არის თუ არა სიდიდე ნორმალურად განაწილებული. ყველგან ქვემოთ ჩვენ ვიგულისხმებთ, რომ სიდიდე ნორმალურია ან დაახლოებით ნორმალურადაა განაწილებული.

იმისათვის, რომ ამოცანები გადავწყვიტოთ სტანდარტული ნორმალური განაწილების გამოყენებით, გარდავქმნათ საწყისი სიდიდე სტანდარტ-

ულ ნორმალურად განაწილებულ სიდიდეს შემდეგი ფორმულის გამოყენებით:

$$z = \frac{\text{მნიშვნელობა - საშუალო}}{\text{სტანდარტული გადახრა}} \quad \text{ანუ} \quad z = \frac{X - \mu}{\sigma}.$$

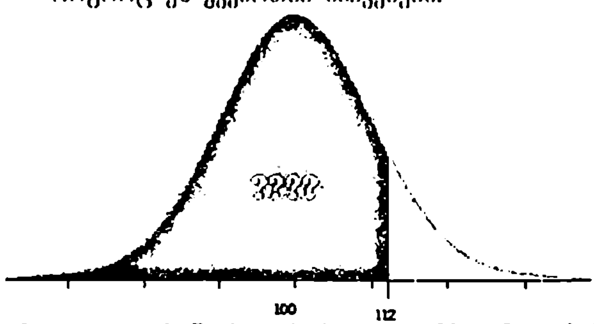
მას შემდეგ რაც სიდიდე გარდაქმნილია, ჩვენ შეგვიძლია ვისარგებლოთ არეთა ფართობის მოძებნის პროცედურული ცხრილითა და სტანდარტული ნორმალური განაწილების ცხრილით. მაგალითად, საწყისი მნიშვნელობები ნორმალურადაა განაწილებული საშუალოთი 100 და სტანდარტული გადახრით 15. როცა სიდიდე გარდაქმნილია z სიდიდედ, მაშინ ორი განაწილება ერთმანეთს ემთხვევა ისე როგორც ეს ნახაზზეა ნაჩვენები:



მაგალითი 1. სიდიდე განაწილებულია ნორმალურად საშუალოთი 100 და სტანდარტული გადახრით 15. ვიპოვოთ იმ მნიშვნელობების პროცენტი, რომლებიც მოთავსებული არიან 112-ის ქვემოთ.

ამოხსნა.

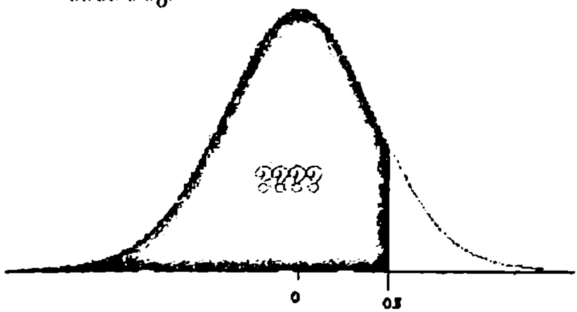
ნაბიჯი 1. დავხაზოთ გრაფიკი და მოვნიშნოთ საძიებელი არე ისე როგორც ეს ქვემოთაა ნაჩვენები:



ნაბიჯი 2. ვიპოვოთ საწყისი 112-ის ტოლი მნიშვნელობის z მნიშვნელობა:

$$z = \frac{X-\mu}{\sigma} = \frac{112-100}{15} = \frac{12}{15} = 0.8.$$

შესაბამისად, 112 არის 0.8 სტანდარტული გადახრით ზევით ვიდრე 100, როგორც ეს ნაჩვენებია ქვემოთ მოყვანილ ნახაზზე:



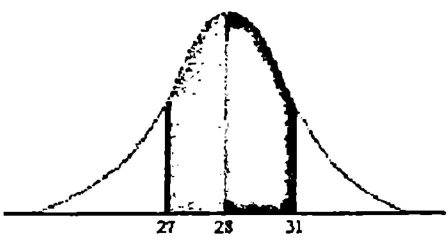
ნაბიჯი 3. ვისარგებლოთ პროცედურული ცხრილის მეხუთე ბლოკის სქემით. $z=0$-სა და $z=0.8$-ს შორის მდებარე არის ფართობია 0.2881. ამიტომ სასურველი არის ($z=0.8$-ის მარცხნივ მოთავსებული არის) ფართობია 0.5000+0.2881=0.7881. შესაბამისად, 78.81% მნიშვნელობების ვარდება 112-ს ქვემოთ.

ისტორიული შენიშვნა. მე-18. საუკუნეში ასტრონომები ნორმალურ განაწილებაზე დაფუძნებულ პრინციპებს იყენებდნენ პლანეტების განლაგების დიაგრამაში დაშვებული შეცდომების კორექტირების დროს.

მაგალითი 2. თვის განმავლობაში პენსიონერი ფეხით გადის საშუალოდ 28 კმ-ს. სტანდარტული გადახრა ტოლია 2-ის. ვიპოვოთ ალბათობა იმისა, რომ შემთხვევით შერჩეული პენსიონერი თვეში გადის არანაკლებ 27კმ-ს და არაუმეტეს 31კმ-ს, $P(27 \le X \le 31)$?

ამოხსნა.

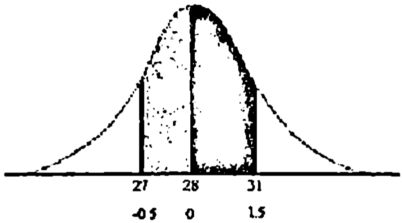
ნაბიჯი 1. დავხაზოთ გრაფიკი და მოვნიშნოთ საძიებელი არე ისე როგორც ეს ქვემოთაა ნაჩვენები:



ნაბიჯი 2. ვიპოვოთ z მნიშვნელობები:

$$z_1 = \frac{X-\mu}{\sigma} = \frac{27-28}{2} = -\frac{1}{2} = -0.5,$$

$$z_2 = \frac{X-\mu}{\sigma} = \frac{31-28}{2} = \frac{3}{2} = 1.5\,.$$



ნაბიჯი 3 ა. ვისარგებლოთ პროცედურული ცხრილის მეოთხე ბლოკის სქემით. $z=-0.5$-სა და $z=0$-ს შორის ფართობია 0.1915, ხოლო $z=0$-სა და $z=1.5$-ს შორის კი – 0.4332. ამიტომ საძიებელი ალბათობა იქნება: $P(27 \le X \le 31) =$

$= P(-0.5 \le Z \le 1.5) = 0.1915 + 0.4332 = 0.6247$ ანუ 62.47%.

ნაბიჯი 3 ბ. საძიებელი ალბათობა გამოვთვალოთ წინა პარაგრაფში მიღებული ფორმულით $P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$:

$$P(-0.5 \le Z \le 1.5) = \Phi(1.5) - \Phi(-0.5) = \Phi(1.5) + \Phi(0.5) - 1.$$

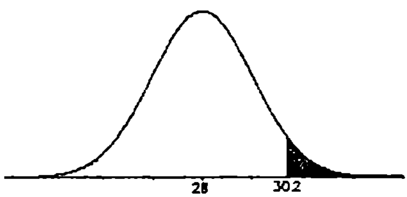
მაშინ თუ ვისარგებლებთ ნორმალური განაწილების ცხრილით მივიღებთ, რომ:

$$P(-0.5 \le Z \le 1.5) = 0.9332 + 0.6915 - 1 = 0.6247$$ ანუ 62.47%.

მაგალითი 3. წინა მაგალითის პირობებში ვიპოვოთ ალბათობა იმისა, რომ შემთხვევით შერჩეული პენსიონერი თვეში გაივლის 30.2კმ-ზე მეტს, $P(X > 30.2)$?

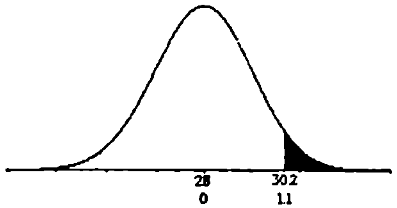
ამოხსნა.

ნაბიჯი 1. დავხაზოთ გრაფიკი და მოვნიშნოთ საძიებელი არე ისე როგორც ეს ქვემოთაა ნაჩვენები:



ნაბიჯი 2. ვიპოვოთ 30.2-ის z მნიშვნელობა:

$$z = \frac{X-\mu}{\sigma} = \frac{30.2-28}{2} = \frac{2.2}{2} = 1.1\,.$$



ნაბიჯი 3 ა. ვისარგებლოთ პროცედურული ცხრილის მეორე ბლოკის სქემით. $z = 0$-სა და $z = 1.1$-ს შორის ფართობია 0.3643. ამიტომ საძიებელი ალბათობა იქნება:

$$P(X > 30.2) = P(Z > 1.1) = 0.5000 - 0.3643 = 0.1357 \text{ ანუ } 13.57\%.$$

ნაბიჯი 3 ბ. საძიებელი ალბათობა გამოვთვალოთ საწინააღმდეგო ხდომილების ალბათობის გამოსათვლელი ფორმულითა და სტანდარტული ნორმალური განაწილების ცხრილის საშუალებით:

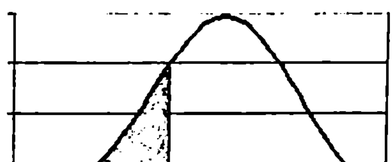
$$P(X > 30.2) = P(Z > 1.1) = 1 - P(Z \le 1.1) =$$
$$= 1 - \Phi(1.1) = 1 - 0.8643 = 0.1357 \text{ ანუ } 13.57\%.$$

ნორმალური განაწილება შეიძლება აგრეთვე გამოყენებულ იქნეს პასუხის გასაცემად კითხვაზე: “რამდენი?”

მაგალითი 4. დავუშვათ, რომ ავტომობილის დაზიანების შემთხვევაში ავარიულ გამოძახებაზე რეაგირების საშუალო დრო არის 25 წუთი. ჩავთვალოთ, რომ ეს სიდიდე განაწილებულია დაახლოებით ნორმალურად და მისი სტანდარტული გადახრა ტოლია 4.5 წუთის. შემთხვევით შერჩეულ იქნა 80 გამოძახება. დაახლოებით რამდენ მათგანზე მოხდება რეგირება 15 წუთზე ნაკლებ დროში?

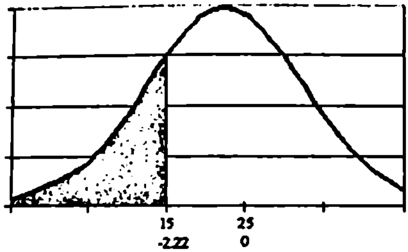
ამოხსნა. ამ ამოცანის ამოსახსნელად ვიპოვოთ ნორმალური წირის ქვეშ 15-ის მარცხნივ მდებარე არის ფართობი.

ნაბიჯი 1. დავხაზოთ გრაფიკი და მოვნიშნოთ საძიებელი არე ისე როგორც ეს ქვემოთაა ნაჩვენები:



ნაბიჯი 2. ვიპოვოთ 15-ის z მნიშვნელობა:

$$z = \frac{X - \mu}{\sigma} = \frac{15 - 25}{4.5} = -\frac{10}{4.5} = -2.22.$$



ნაბიჯი 3. ვიპოვოთ $z=-2.22$-სა და $z=0$-ს შორის მდებარე არის ფართობი. ის ტოლია 0.4868-ის.

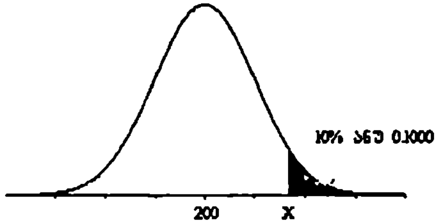
ნაბიჯი 4. გამოვაკლოთ 0.5000-ს 0.4868. მივიღებთ 0.0132-ს.

ნაბიჯი 5. იმისათვის, რომ გავიგოთ რანდენ გამოძახებაზე მოხდა რეაგირება 15 წუთზე ნაკლებ დროში, გავამრავლოთ შერჩევის მოცულობა (80) მიღებულ ფართობზე (0.0132). მაშინ მივიღებთ 1.056-ს. მაშასადამე, 1.056 ანუ დაახლოებით ერთ გამოძახებაზე მოხდება რეაგირება 15 წუთზე ნაკლებ დროში.

ნორმალური განაწილება შეიძლება აგრეთვე გამოყენებულ იქნეს მოცემული პროცენტის მიხედვით კონკრეტული მნიშვნელობის მოსაძებნად.

მაგალითი 5. ცნობილია, რომ აბიტურიენტების მხოლოდ 10% შეიძლება გახდეს სტუდენტი. ჩავთვალოთ, რომ აბიტურიენტების მიერ მოგროვილი ქულების მნიშვნელობები ნორმალურადაა განაწილებული საშუალოთი 200 და სტანდარტული გადახრით 20. ვიპოვოთ ის მინიმალური ქულა, რომელიც საჭიროა რათა აბიტურიენტი გახდეს სტუდენტი.

ამოხსნა. ვინაიდან აბიტურიენტების მიერ მოგროვილი ქულების მნიშვნელობები ნორმალურადაა განაწილებული, ამიტომ იმ ქულის მნიშვნელობა (X), რომლის ზევითაც აბიტურიენტი გახდება სტუდენტი, არის ისეთი რიცხვი, რომლის მარჯვნივ ნორმალური წირის ქვეშ მოთავსებული არის ფართობი ტოლია 10%-ის ანუ 0.1000-ის:



ნაბიჯი 1. იმისათვის, რომ ვიპოვოთ ნორმალური წირის ქვეშ 200-სა და X-ს შორის მდებარე არის ფართობი 0.5000-ს გამოვაკლოთ 0.1000, მივიღებთ 0.4000-ს.

ნაბიჯი 2. ვიპოვოთ z მნიშვნელობა, რომელიც შეესაბამება ნორმა-

ლური განაწილების ცხრილში 0.4000-ს. იმ შემთხვევაში, როცა ცხრილში არ იძებნება ზუსტად ეს მნიშვნელობა ვიღებთ მასთან ყველაზე ახლოს მყოფს, ამ შემთხვევაში 0.3997-ს. შესაბამისი $z=1.28$.

ნაბიჯი 3. შევიტანოთ 1.28 z მნიშვნელობის გამოსათვლელ ფორმულაში $z=(X-\mu)/\sigma$ და ამოეხსნათ X.

$$1.28=\frac{X-\mu}{\sigma}$$
$$1.28\times 20+200=X$$
$$X=25.60+200=225.60$$
$$X=226.$$

ე.ი. იმისათვის, რომ აბიტურიენტი გახდეს სტუდენტი, მან უნდა მოაგროვოს 226 ქულა.

შენიშვნა. იმ შემთხვევაში, როცა ფართობის მნიშვნელობა ვარდება ცხრილის ორი მნიშვნელობის ზუსტად შუაში, მაშინ ვიღებთ მათი შესაბამისი z მნიშვნელობებიდან უფრო დიდს. მაგალითად, თუ ფართობის მნიშვნელობაა 0.4500, ის იმყოფება 0.4495-ისა და 0.4505-ის შუაში და z მნიშვნელობად ვიღებთ 1.65-ს და არა 1.64-ს.

გამოვიყვანოთ X მნიშვნელობის გამოსათვლელი ფორმულა z მნიშვნელობის გამოსათვლელი ფორმულიდან

$$z=\frac{X-\mu}{\sigma}.$$

გავამრავლოთ ტოლობის ორივე მხარე σ-ზე

$$z\cdot\sigma=X-\mu.$$

ტოლობის ორივე მხარეს მივუმატოთ μ

$$z\cdot\sigma+\mu=X.$$

შევცვალოთ ტოლობის მარჯვენა და მარცხენა მხარეები

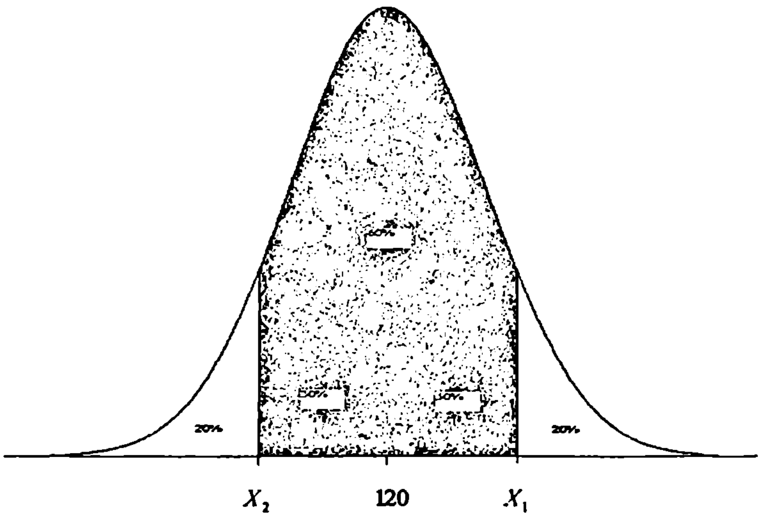
$$X=z\cdot\sigma+\mu.$$

საწყისი მნიშვნელობის გამოთვლა z მნიშვნელობის მიხედვით:

$$X=z\cdot\sigma+\mu.$$

მაგალითი 6. სამედიცინო გამოკვლევის მიზნით მკვლევარს სურს შეარჩიოს არტერიული წნევის საფუძველზე შედგენილი პოპულაციის შუაში მდგომი ადამიანების 60%. ვიგულისხმოთ, რომ არტერიული წნევა განაწილებულია ნორმალურად საშუალოთი 120 და სტანდარტული გადახრით 8. დაადგინოთ შესარჩევი ადამიანების არტერიული წნევის ზედა და ქვედა საზღვარი.

ამოხსნა. გასაგებია, რომ შესარჩევი ადამიანების არტერიული წნევის ზედა და ქვედა საზღვარი იქნება ის ორი მოპირდაპირე რიცხვი, რომელთა გარეთ ნორმალური წირის ქვეშ მოქცეული თითოეული კუდის ფართობია 20%.



ცხადია, რომ z მნიშვნელობა, რომელიც შეესაბამება ნორმალური განაწილების ცხრილში 0.5000-0.2000=0.3000 ფართობს ტოლია 0.84-ის. შევიტანოთ იგი $X = z\cdot\sigma+\mu$ ფორმულაში. მივიღებთ, რომ

$$X_1 = z\cdot\sigma+\mu = 0.84\times 8+120 = 126.72\,.$$

მეორეს მხრივ, სტანდარტული ნორმალური განაწილების საშუალოს მიმართ სიმეტრიულობის გამო, X_2-ის გამოსათვლელად უნდა ავიღოთ $z=-0.84$. შესაბამისად გვექნება:

$$X_2 = z\cdot\sigma+\mu = -0.84\times 8+120 = 113.28\,.$$

ე. ი. მკვლევარმა უნდა შეარჩიოს ისეთი ადამიანები, რომელთა არტერიული წნევა მოთავსებულია 113.8-სა და 126.72-ს შორის,

$$113.28 < X < 126.72\,.$$

თავი III
შეფასების თეორია პირველი წაკითხვისათვის

წერტილოვანი შეფასებები.

მასობრივი ინფორმაციის საშუალებიდან თქვენ შესაძლებელია მიიღოთ შემდეგი ტიპის ინფორმაცია: ამ თუ იმ ორგანიზაციის მიერ ჩატარებულმა გამოკითხვამ აჩვენა, რომ სატელევიზიო პროგრამით უკმაყოფილო ადამიანების 45% ტელევიზორს გადართავს სხვა არხზე, მაშინ როდესაც ადამიანების 15% საერთოდ თიშავს თავის ტელევიზორს. აქვე თქვენ შეიტყობთ, რომ გამოკითხული იყო 4000 სრულწლოვანი ადამიანი და ცდომილება შეადგენს 3%-ს. ასეთ შემთხვევაში თქვენ ბუნებრივად გებადებათ შემდეგი ტიპის კითხვები:

1. როგორ თანადობაშია ეს შეფასება ადამიანების რეალურ პოპულაციაში შესაბამის რეალურ მაჩვენებელთან?

2. რა იგულისხმება სიტყვების ქვეშ: ცდომილება შეადგენს 3%-ს?

3. 4000 ადამიანის გამოკითხვა საკმარისია თუ არა იმისათვის, რომ სრულად წარმოვაჩინოთ ის სრულწლოვანი მოსახლეობა, რომელიც ტელევიზორს უყურებს ქვეყნის მასშტაბით?

მას შემდეგ რაც ახსნილი იქნება თუ როგორ იყენებს სტატისტიკოსი სტატისტიკურ მონაცემებს პარამეტრების შესაფასებლად, ჩვენ შეგვეძლება პასუხი გავცეთ ამ ტიპის კითხვებს.

სტატისტიკური დასკვნების თეორიის ერთ-ერთ მნიშვნელოვან მიმართულებას წარმოადგენს **შეფასების თეორია**, რომელიც გულისხმობს შერჩევიდან მიღებული ინფორმაციის საფუძველზე პარამეტრის უცნობი მნიშვნელობის შეფასებას. იმის გათვალისწინებით, რომ პოპულაცია, საიდანაც მიღებულ უნდა იქნეს ეს მნიშვნელობები, დიდია – ეს მნიშვნელობები მხოლოდ *შეფასებებია* პარამეტრის ჭეშმარიტი მნიშვნელობის და მიღებულია იმ მონაცემებიდან, რომლებიც თავმოყრილია შერჩევაში. ქვემოთ მოყვანილი და ახსნილი იქნება ის სტატისტიკური პროცედურები, რომლებიც გამოიყენება პოპულაციის საშუალოს, პროპორციის, დისპერსიისა და საშუალო კვადრატული გადახრის (სტანდარტული გადახრის) შესაფასებლად.

პირველი მნიშვნელოვანი საკითხი, რომელიც წარმოიშობა ამა თუ იმ პარამეტრის შეფასებისას არის შერჩევის მოცულობის სიდიდე. რამდენად დიდი უნდა იყოს შერჩევის მოცულობა იმისათვის, რომ გავაკეთოთ სწორი შეფასება? ამ კითხვაზე პასუხის გაცემა არ არის მარტივი, რადგანაც შერჩევის მოცულობა დამოკიდებულია მრავალ ფაქტორზე, როგორიცაა მაგალითად, სასურველი სიზუსტე და ალბათობა, რომელიც უზრუნველყოფს შეფასების სიზუსტეს. ქვემოთ ჩვენ განვიხილავთ შერჩევის მოცულობის დადგენის წესს.

დავუშვათ, რომ კოლეჯის დირექტორს სურს შეაფასოს იმ სტუდენტების საშუალო ასაკი, რომლებიც ლექციებზე დადიან მიმდინარე სემესტრში. დირექტორს შეუძლია აირჩიოს შემთხვევითი შერჩევა, რომელიც შედგება 100 სტუდენტისაგან და იპოვოს ამორჩეული სტუდენტების საშუალო ასაკი. დავუშვათ, რომ ის აღმოჩნდა 22.3 წელი. შერჩევის ამ საშუალოს მიხედვით დირექტორს შეუძლია დაასკვნას, რომ კოლეჯის ყველა სტუდე-

ნტის საშუალო ასაკია 22.3 წელი. ამ ტიპის შეფასებას *წერტილოვან შეფასებას* უწოდებენ.

წერტილოვანი შეფასება (a point estimate) წარმოადგენს შესაფასებელი პარამეტრის სპეციფიკურ რიცხობრივ მნიშვნელობას. პოპულაციის μ საშუალოს საუკეთესო წერტილოვან შეფასებას წარმოადგენს შერჩევითი საშუალო $\overline{X}$.

ბუნებრივად ჩნდება კითხვა: ცენტრალური ტენდენციის სხვა საზომები, როგორიცაა მედიანა და მოდა, რატომ არ გამოიყენება პოპულაციის საშუალოს შესაფასებლად. ეს იმით აიხსნება, რომ შერჩევითი საშუალო ნაკლებად იცვლება, ვიდრე სხვა სტატისტიკები (როგორიცაა მედიანა და მოდა), როცა ერთი და იგივე პოპულაციიდან არჩეულია რამოდენიმე სხვადასხვა შერჩევა. ამიტომ შერჩევითი საშუალო პოპულაციის საშუალოს საუკეთესო შეფასებაა.

შერჩევითი საზომები (ანუ სტატისტიკები) გამოიყენება პოპულაციის საზომების (ანუ პარამეტრების) შესაფასებლად. ამ სტატისტიკებს აგრეთვე **შეფასებები** ეწოდება (ინგლისურად იხმარება სიტყვა: "estimator", რომელიც ითარგმნება როგორც შემფასებელი, მაგრამ ქართულ ტერმინოლოგიაში გამოიყენება სიტყვა შეფასება). საუკეთესო შეფასება უნდა აკმაყოფილებდეს ქვემოთ მოყვანილ შემდეგ სამ თვისებას:

1. შეფასება უნდა იყოს **ჩაუნაცვლებელი**. რაც იმას ნიშნავს, რომ მოცემული მოცულობის შერჩევიდან მიღებული შეფასების მათემატიკური ლოდინი ანუ საშუალო ტოლი უნდა იყოს შესაფასებელი პარამეტრის ჭეშმარიტი მნიშვნელობის.
2. შეფასება უნდა იყოს **ძალმოსილი**. რაც იმას გულისხმობს, რომ როცა შერჩევის მოცულობა უსასრულოდ იზრდება შეფასების მნიშვნელობა უახლოვდება შესაფასებელი პარამეტრის ჭეშმარიტ მნიშვნელობას.
3. შეფასება უნდა იყოს **ფარდობითად ეფექტური**. ეს ნიშნავს, რომ ყველა იმ სტატისტიკას შორის, რომელიც გამოიყენება პარამეტრის შესაფასებლად, ფარდობითად ეფექტურ სტატისტიკას აქვს მინიმალური ვარიაცია (დისპერსია).

ინტერვალური შეფასებები. ნდობის ინტერვალი.

ძალიან ხშირ შემთხვევაში შესაძლებელია შერჩევითი საშუალო გარკვეულ წილად განსხვავდებოდეს პოპულაციის საშუალოსაგან (ანუ ადგილი ქონდეს შეფასების ცდომილებას). შესაბამისად, იბადება შემდეგი ბუნებრივი კითხვა: რამდენად კარგია წერტილოვანი შერფასება? ამ კითხვაზე პასუხის გასაცემად უნდა გაირკვეს რამდენად ახლოსაა წერტილოვანი შეფასება პოპულაციის საშუალოსთან. ეს პასუხი ხასიათდება წერტილოვანი შეფასების სიზუსტის გარკვეულ საიმედოობით. ამ მიზნით სტატისტიკოსები მიმართავენ განსხვავებული ტიპის შეფასებას, რომელიც *ინტერვალური შეფასების* სახელწოდებითაა ცნობილი.

პარამეტრის **ინტერვალური შეფასება (confidence interval)** წარმოადგენს ინტერვალს ან მნიშვნელობების რანგს (დიაპაზონს), რომელიც გამოიყენება პარამეტრის შესაფასებლად. ეს შეფასება შეიძლება მოიცავდეს ან არ მოიცავდეს შესაფასებელი პარამეტრის მნიშვნელობას.

ინტერვალური შეფასების შემთხვევაში პარამეტრი განისაზღვრება როგორც ორ მნიშვნელობას შორის მოქცეული (მდებარე). მაგალითად, კოლეჯის ყველა სტუდენტის საშუალო ასაკის ინტერვალური შეფასება შეიძლება იყოს $26.9 < \mu < 27.7$, ანუ 27.3 ± 0.4 წელი.

ნებისმიერი ინტერვალი შეიცავს ან არ შეიცავს პარამეტრს. საიმედოობის ხარისხი (როგორც წესი, პროცენტებში გამოსახული) მოიცემა წინასწარ, მანამ სანამ გაკეთდება ინტერვალური შეფასება. მაგალითად, ჩვენ შეიძლება გვსურდეს, რომ ინტერვალი 95%-იანი საიმედოობით მოიცავდეს პოპულაციის საშუალოს ჭეშმარიტ მნიშვნელობას. თუმცა, მაშინვე შეიძლება წარმოიშვას სხვა კითხვა: რატომ 95%? რატომ არა 99% ან 99.5%?

თუ ვინმეს სურს უფრო მეტი საიმედოობა, როგორიცაა 99% ან 99.5%, მაშინ ინტერვალი უნდა იყოს უფრო განიერი. მაგალითად, კოლეჯის სტუდენტების ასაკის საშუალოს 99%-იანი ნდობის ინტერვალი შეიძლება იყოს $26.7 < \mu < 27.9$, ანუ 27.3 ± 0.6.

ისტორიული შენიშვნა. წერტილოვანი და ინტერვალური შეფასებები ცნობილი იყო კიდევ 1700 წელს. მაგრამ, მხოლოდ 1937 წელს მოხდა მათი მათემატიკური ფორმალიზება და პრაქტიკული გამოყენებების ფორმულირება ჯ. ნეიმანის მიერ.

პარამეტრის ინტერვალური შეფასების **საიმედოობის დონე** ეწოდება იმ ალბათობას, რომლითაც ინტერვალური შეფასება მოიცავს (ფარავს) პარამეტრს.

ნდობის ინტერვალი წარმოადგენს პარამეტრის შეფასების სპეციალურ ინტერვალს, რომელიც განისაზღვრება შერჩევიდან მიღებული მონაცემების გამოყენებით და შეფასების სპეციალური საიმედოობის დონის მიხედვით.

ამ გზით აგებულ ინტერვალებს უწოდებენ *ნდობის ინტერვალებს.* პრაქტიკაში გამოიყენება სამი ტიპის ნდობის ინტერვალი: 90%-იანი, 95%-იანი და 99%-იანი ნდობის ინტერვალები. საშუალოსათვის ნდობის ინტერვალის განმსაზღვრელი ფორმულის ალგებრული გამოყვანა მოყვანილი იქნება მოგვიანებით. წინასწარ მოვიყვანოთ მოკლე ინტუიციური ახსნა.

ცენტრალური ზღვარითი თეორემის საფუძველზე შესაძლებელია დავასკვნათ, რომ როცა შერჩევის მოცულობა საკმაოდ დიდია, შერჩევათა საშუალოების დაახლოებით 95% ვარდება (თავსდება) პოპულაციის საშუალოსაგან ±1.96 სტანდარტული გადახრის ფარგლებში, ანუ

$$\mu \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

თუ ახლა შერჩევის სპეციალური საშუალო შერჩეულია, ვთქვათ $\overline{X}$, მაშინ 95%-იანი ალბათობით ის მოთავსებული იქნება დიაპაზონში $\mu - 1.96(\sigma/\sqrt{n})$ -დან $\mu + 1.96(\sigma/\sqrt{n})$-მდე. ანალოგიურად, 95%-იანი ალბათობით ინტერვალი მოცემული საზღვრებით

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

მოიცავს μ-ს, როგორც ეს ნაჩვენები იქნება ქვემოთ. სხვა სიტყვებით:

$$\overline{X}-1.96\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+1.96\frac{\sigma}{\sqrt{n}}.$$

შესაბამისად, 95%-იანი საიმედოობით პოპულაციის საშუალო მოთავსებული იქნება ზემოთ მოყვანილ ინტერვალში, როცა მონაცემები აღებულია ნორმალურად განაწილებული პოპულაციიდან.

95%-იანი ნდობის ინტერვალისათვის გამოყენებული მნიშვნელობა 1.96 მიღებულია სტანდარტული ნორმალური განაწილების ცხრილიდან. 99%-იანი ნდობის ინტერვალისათვის 1.96-ის ნაცვლად უნდა გამოვიყენოთ მნიშვნელობა 2.58, რომელიც აგრეთვე მიიღება ნორმალური განაწილების ცხრილიდან. ვინაიდან, სტატისტიკაში გამოიყენება სხვა ნდობის ინტერვალებიც, საზოგადოდ ნდობის ინტერვალის ფორმულაში ხმარობენ სიმბოლოს $z_{\alpha/2}$ (იკითხება — ზეტ ინდექსით ალფა ნახევარი). შევნიშნავთ, რომ $z_{\alpha/2}$ ისეთი რიცხვია, რომ $\Phi(z_{\alpha/2})=1-\alpha/2$ ანუ $\Phi_0(z_{\alpha/2})=1/2-\alpha/2$ და მას სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$ **კრიტიკულ წერტილი** ეწოდება.

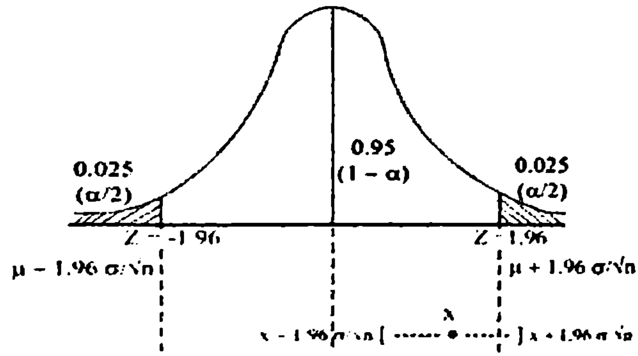
კავშირი α-სა და საიმედოობის დონეს შემდეგნაირია: საიმედოობის დონე არის ის პროცენტი, რომელიც ექვივალენტურია $1-\alpha$ სიდიდის ათწილადური მნიშვნელობის და პირიქით. როცა საქმე გვაქვს 95%-იან ნდობის ინტერვალთან, მაშინ $\alpha=0.05$, ვინაიდან $1-\alpha=0.95$ ანუ 95%. როცა $\alpha=0.01$, მაშინ $1-\alpha=1-0.01=0.99$, და მაშასადამე, უნდა აიგოს 99%-იანი ნდობის ინტერვალი.

საშუალოს ნდობის ინტერვალის ფორმულა (როცა σ ცნობილია)

$$\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

სადაც $\overline{X}$ — არის შერჩევითი საშუალო, $z_{\alpha/2}$ — არის სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$ კრიტიკული წერტილი, σ — არის პოპულაციის სტანდარტული გადახრა, ხოლო n — შერჩევის მოცულობა.

სიდიდეს — $z_{\alpha/2}(\sigma/\sqrt{n})$ უწოდებენ *შეფასების მაქსიმალურ შეცდომას*. კონკრეტული მნიშვნელობისათვის, ვთქვათ $\alpha=0.05$, როგორც უკვე აღნიშნული იყო, შერჩევათა საშუალოების 95% შეიძლება აღმოჩნდეს გადახრილი პოპულაციის საშუალოს ნებისმიერ მხარეს არა უმეტეს ამ სიდიდისა. იხილეთ ნახაზი:



შეფასების მაქსიმალური შეცდომა წარმოადგენს მაქსიმალურ განსხვავებას (სხვაობას) პარამეტრის წერტილოვან შეფასებასა და ამ პარამეტრის ჭეშმარიტ (რეალურ) მნიშვნელობას შორის.

დამრგვალების წესი საშუალოსათვის ნდობის ინტერვალის აგებისას. *ნედლი მონაცემების* საშუალებით პოპულაციის საშუალოსათვის ნდობის ინტერვალის აგებისას დამრგვალება უნდა მოვახდინოთ ერთით მეტი ათობითი ნიშნით, ვიდრე ამას ადგილი აქვს საწყის მონაცემებში. იმ შემთხვევაში კი, როცა პოპულაციის საშუალოსათვის ნდობის ინტერვალს ვაგებთ შერჩევითი საშუალოსა და სტანდარტული გადახრის გამოყენებით, მაშინ დამრგვალებას ვახდენთ იმდენივე ათობითი ნიშნით, რამდენი ათობითი ნიშანიც აქვს საშუალოს.

მაგალითი 1. დიდი უნივერსიტეტის პრეზიდენტს სურს შეაფასოს ამჟამად ჩარიცხული სტუდენტების საშუალო ასაკი. წარსული გამოკვლევებიდან ცნობილია, რომ სტანდარტული გადახრა არის 2 წელი. შერჩეულია 50 სტუდენტისაგან შემდგარი შერჩევა და მისთვის გამოთვლილი საშუალო ტოლია 23.2 წლის. იპოვეთ 95%-იანი ნდობის ინტერვალი პოპულაციის საშუალოსათვის.

ამოხსნა. ვინაიდან 95%-იანი ნდობის ინტერვალს შეესაბამება $\alpha = 0.05$, ამიტომ ნორმალური განაწილების ცხრილებიდან ვპოულობთ $z_{\alpha/2} = z_{0.025} = 1.96$. შევიტანოთ ნდობის ინტერვალის

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ფორმულაში მონაცემები: $\overline{X} = 23.2$, $z_{0.025} = 1.96$, $\sigma = 2$ და $n = 50$. გვაქვს:

$$23.2 - 1.96 \cdot \frac{2}{\sqrt{50}} < \mu < 23.2 + 1.96 \cdot \frac{2}{\sqrt{50}},$$

$$23.2 - 0.6 < \mu < 23.2 + 0.6,$$

$$22.6 < \mu < 23.8,$$

ანუ 23.2 ± 0.6. შესაბამისად, 50 სტუდენტის ასაკიდან გამომდინარე, 95%-იანი საიმედოობით, უნივერსიტეტის პრეზიდენტს შეუძლია თქვას, რომ

სტუდენტების საშუალო ასაკი მოთავსებულია 22.6 წელსა და 23.8 წელს შორის.

მაგალითი 2. ცნობილია, რომ გარკვეული მედიკამენტის გამოყენების შემდეგად პულსის რიცხვი მატულობს. ცნობილია, რომ პულსის რიცხვის სტანდარტული გადახრა არის 5 დარტყმა წუთში. შერჩეულ იქნა ამ მედიკამენტის 30 მომხმარებელი და მათთვის პულსის რიცხვის საშუალო აღმოჩნდა 104 დარტყმა წუთში. ვიპოვოთ 99%-იანი ნდობის ინტერვალი პულსის რიცხვის ჭეშმარიტი საშუალოსათვის.

ამოხსნა. ვინაიდან 99%-იანი ნდობის ინტერვალს შეესაბამება $\alpha = 0.01$, ამიტომ ნორმალური განაწილების ცხრილებიდან ვპოულობთ $z_{\alpha/2} = z_{0.005} = 2.58$. შევიტანოთ ნდობის ინტერვალის

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ფორმულაში მონაცემები: $\overline{X} = 104$, $z_{0.005} = 2.58$, $\sigma = 5$ და $n = 30$. გვექნება:

$$104 - 2.58\cdot\frac{5}{\sqrt{30}} < \mu < 104 + 2.58\cdot\frac{5}{\sqrt{30}},$$

$$104 - 2.4 < \mu < 104 + 2.4, \text{ ანუ } 101.6 < \mu < 106.6.$$

დავამრგვალოთ ერთეულებამდე სიზუსტით:

$$102 < \mu < 106 \text{ ანუ } 104 \pm 2.$$

მაშასადამე, 30 მომხმარებლიანი შერჩევის საფუძველზე, 99%-იანი საიმედოობით, ჩვენ შეგვიძლია დავასკვნათ, რომ ამ მედიკამენტის ყველა მომხმარებლის პულსის რიცხვის საშუალო მოთავსებულია (მერყეობს) 102-სა და 106-ს შორის.

არანორმალური პოპულაციის შემთხვევა.

შენიშვნა. იმ შემთხვევაში, როდესაც საწყისი სიდიდე განაწილებულია ნორმალურად და α ცნობილია ნდობის ინტერვალის ასაგებად ყოველთვის შეიძლება გამოვიყენოთ სტანდარტული ნორმალური განაწილება მიუხედავად იმისა თუ რა მოცულობისაა შერჩევა. როდესაც $n \geq 30$ საშუალოს განაწილება იქნება დაახლოებით ნორმალური იმ შემთხვევაშიც კი თუ საწყისი სიდიდის განაწილება განსხვავებულია ნორმალურისაგან. გარდა ამისა, თუ $n \geq 30$ (ზოგიერთი ავტორი იყენებს $n > 30$), ნდობის ინტერვალის ფორმულაში σ შეიძლება შეიცვალოს s-ით და საშუალოსათვის ნდობის ინტერვალის ასაგებად შეიძლება ისევ გამოვიყენოთ სტანდარტული ნორმალური განაწილება, ისე როგორც ეს ნაჩვენები იქნება შემდეგ მაგალითში.

მაგალითი 3. ქვემოთ მოყვანილია სამხრეთ-აღმოსავლეთ პენსილვანიის 30 საკრედიტო კავშირის აქტივების (მილიონ დოლარებში) შერჩევა. ააგეთ 90%-იანი ნდობის ინტერვალი საშუალოსათვის.

12.23	8.74	16.56	3.17	4.39	18.13
2.89	7.92	1.24	4.78	2.17	16.85
13.19	40.22	9.16	2.42	1.42	21.58
73.25	5.01	1.91	1.47	14.64	12.24

11.59 2.27 6.69 12.77 1.06 2.76

ამოხსნა.

ნაბიჯი 1. გამოვთვალოთ წარმოდგენილი მონაცემების საშუალო და სტანდარტული გადახრა, რისთვისაც ვისარგებლოთ ფორმულებით:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

$$s = \sqrt{\frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]}.$$

ამ შემთხვევაში $n = 30$ და გამოთვლების ჩატარების შედეგად ვღებულობთ, რომ საშუალო – $\overline{X} = 11.091$, ხოლო სტანდარტული გადახრა – $s = 14.4.5$.

ნაბიჯი 2. ვიპოვოთ $\alpha/2$. ვინაიდან მოითხოვება 90%-იანი ნდობის ინტერვალის აგება, ამიტომ $\alpha = 1 - 0.90 = 0.10$, და შესაბამისად $\alpha/2 = 0.05$.

ნაბიჯი 3. ვიპოვოთ $z_{\alpha/2}$. ვისარგებლოთ სტანდარტული ნორმალური განაწილების ცხრილით (ე. ი. $\Phi(x)$ ფუნქციის ცხრილით), გამოვაკლოთ 1-ს 0.05, 1-0.05=0.95 და ცხრილში მოვძებნოთ ისეთი რიცხვი, რომელშიც სტანდარტული ნორმალური განაწილების ფუნქციის მნიშვნელობა არის 0.95. ასეთი რიცხვია 1.65, $\Phi(1.65) = 0.95$, ე. ი. $z_{\alpha/2} = 1.65$ (იგივე შედეგს მივიღებთ $\Phi_0(x)$ ფუნქციის ცხრილებიდან: ამ შემთხვევაში 0.5-ს ვაკლებთ 0.05-ს, ვღებულობთ 0.45 და $\Phi_0(x)$-ს ცხრილიდან ვხედავთ, რომ $\Phi_0(1.65) = 0.45$, ანუ ისევ $z_{\alpha/2} = 1.65$).

ნაბიჯი 4. შევიტანოთ შესაბამისი მონაცემები ფორმულაში

$$\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

(ვინაიდან $n \geq 30$ უცნობი σ-ს ადგილას ვწერთ s-ს), მივიღებთ, რომ:

$$11.091 - 1.65 \cdot \frac{14.405}{\sqrt{30}} < \mu < 11.091 + 1.65 \cdot \frac{14.405}{\sqrt{30}},$$

$$11.091 - 4.339 < \mu < 11.091 + 4.339,$$

$$6.752 < \mu < 15.430.$$

ამრიგად, 30 საკრედიტო კავშირის შესაბამისი შერჩევიდან, 90%-იანი საიმედოობით ჩვენ შეგვიძლია დავასკვნათ, რომ ყველა საკრედიტო კავშირის აქტივების პოპულაციის საშუალო მოთავსებულია 6.752 მილიონ დოლარსა და 15.430 მილიონ დოლარს შორის.

როგორც აღნიშნული იყო შერჩევითი საშუალოს განაწილება მიახლოებით ნორმალურია, როცა პოპულაციიდან აღებული შერჩევის მოცულობა დიდია (კერძოდ, როცა $n \geq 30$). შესაბამისად, ამ შემთხვევაში სტანდარტიზებული შერჩევითი საშუალო

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

მიახლოებით სტანდარტული ნორმალური კანონითაა განაწილებული. ამიტომ ალბათობა იმისა, რომ z მოთავსებული იქნება $-z_{\alpha/2}$-სა და $+z_{\alpha/2}$-ს შორის ტოლია $1-\alpha$. შესაბამისად,

$$-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}.$$

უტოლობის სამივე ნაწილი გავამრავლოთ მნიშვნელზე:

$$-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

უტოლობის სამივე ნაწილს გამოვაკლოთ $\overline{X}$:

$$-\overline{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

უტოლობის სამივე ნაწილი გავამრავლოთ -1-ზე:

$$\overline{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} > \mu > \overline{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

თუ ახლა მოვაბრუნებთ მიღებულ უტოლობას, ჩვენ მივიღებთ საშუალოსათვის ნდობის ინტერვალის ფორმულას:

$$\overline{X}-z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

შერჩევის მოცულობა.

შერჩევის მოცულობის განსაზღვრა მჭიდრო კავშირში იმყოფება სტატისტიკურ შეფასებასთან. საკმაოდ ხშირად და ბუნებრივად ისმის კითხვა: "რამდენად დიდი მოცულობის შერჩევაა საჭირო, რომ მივიღოთ კარგი შეფასება?" პასუხი ამ კითხვაზე არ არის მარტივი, ვინაიდან ის დამოკიდებულია სამ სიდიდეზე: შეფასების მაქსიმალურ შეცდომაზე, პოპულაციის სტანდარტულ გადახრაზე და საიმედოობის ხარისხზე (დონეზე). მაგალითად, რამდენად ახლოს უნდა იყოს შეფასება ჭეშმარიტ საშუალოსთან (2 ერთეული, 5 ერთეული, და ა. შ.), და როგორი საიმედოობა გვსურს (90%-იანი, 95%-იანი, 99%-იანი და ა. შ.). ქვემოთ ჩვენ ვიგულისხმებთ, რომ პოპულაციის სტანდარტული გადახრა ცნობილია ან მისი შეფასება შესაძლებელია საწყის ეტაპზე.

შერჩევის მოცულობის ფორმულა მიიღება შეფასების მაქსიმალური შეცდომის ფორმულიდან

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

რომელიც უნდა ამოიხსნას n-ის მიმართ შემდეგნაირად:

$$E\sqrt{n} = z_{\alpha/2}\cdot\sigma,$$

$$\sqrt{n} = \frac{z_{\alpha/2}\cdot\sigma}{E},$$

$$n = \left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2.$$

პოპულაციის საშუალოს ინტერვალური შეფასებისათვის

$$n=\left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2,$$

სადაც $z_{\alpha/2}$ – არის სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$ კრიტიკული წერტილი, σ – პოპულაციის სტანდარტული გადახრა, ხოლო E – შეფასების მაქსიმალური შეცდომა. ამასთანავე, მიღებული n რიცხვი უნდა დავამრგვალოთ მეტობით, რათა მივიღოთ ნატურალური რიცხვი. კერძოდ, თუ მიღებული რიცხვი შეიცავს წილად ან ათწილად ნაწილს, მასში შერჩევის მინიმალური მოცულობა იქნება მასზე დიდი უახლოესი ნატურალური რიცხვი.

მაგალითი 1. კოლეჯის პრეზიდენტმა დაავალა სტატისტიკის მასწავლებელს შეაფასოს კოლეჯის სტუდენტების საშუალო ასაკი. რამდენად დიდია აუცილებელი შერჩევა? სტატისტიკის მასწავლებელი თვლის, რომ საიმედოობა უნდა იყოს 99%, რათა შეფასება იყოს სწორი ერთი წლის სიზუსტით. წინა გამოკვლევებიდან ცნობილია, რომ ასაკთა სტანდარტული გადახრა არის 3 წელი.

ამოხსნა. ამ შემთხვევაში $E=1$; $\sigma=3$; $\alpha=1-0.99=0.01$, $\alpha/2=0.005$, $1-\alpha/2=1-0.005=0.995$, $\Phi(2.58)=0.995$. ამიტომ $z_{\alpha/2}=z_{0.995}=2.58$.

შევიტანოთ ეს სიდიდეები შერჩევის მოცულობის ფორმულაში, მივიღებთ, რომ

$$n=\left(\frac{z_{\alpha/2}\cdot\sigma}{E}\right)^2=\left(\frac{2.58\cdot3}{1}\right)^2=59.9.$$

ეს რიცხვი უნდა დავამრგვალოთ 60-მდე. ამრიგად, იმისათვის, რომ მასწავლებელმა შეაფასოს სტუდენტთა ასაკის საშუალოს ჭეშმარიტი მნიშვნელობა ერთი წლის სიზუსტით 99%-იანი საიმედოობით, მას სჭირდება სულ ცოტა 60 სტუდენტისაგან შემდგარი შერჩევა.

როგორც ვხედავთ შერჩევის მოცულობის დასადგენი ფორმულა მოითხოვს პოპულაციის სტანდარტული გადახრის ცოდნას. რა ხდება იმ შემთხვევაში, როცა σ უცნობია? ამ შემთხვევაში უნდა შევეცადოთ σ-ს შეფასებას. σ-ს შეფასების ერთ-ერთი გზაა s სტანდარტული გადახრის გამოყენება, რომელიც გამოითვლება შერჩევის მონაცემების მიხედვით. სტანდარტული გადახრა ასევე შეიძლება შეფასდეს 4-ზე გაყოფილი გაბნევის დიაპაზონით.

ზოგჯერ ინტერვალურ შეფასებებს უპირატესობა ენიჭება წერტილოვან შეფასებებთან შედარებით. მაგალითად, თქვენ შეიძლება წაიკითხოთ ასეთი მტკიცებულება: "200 ოჯახიანი შერჩევის საფუძველზე ჩატარებულმა გამოკითხვამ აჩვენა რომ ორ წევრიანი ამერიკული ოჯახი პროდუქტებზე კვირაში საშუალოდ ხარჯავს 84 დოლარს. 95%-იანი საიმედოობით ამ შეფასების სიზუსტე ჭეშმარიტი საშუალოს მიმართ შეადგენს 3 დოლარს". ეს მტკიცებულება ნიშნავს, რომ საშუალოს ჭეშმარიტი მნიშვნელობის 95%-იანი ნდობის ინტერვალია:

$$84\$-3\$<\mu<84\$+3\$,$$

$$81\$<\mu<87\$.$$

ნდობის ინტერვალი საშუალოსათვის (σ უცნობია და $n < 30$).

როგორც ზემოთ იყო აღნიშნული, იმ შემთხვევაში როცა σ ცნობილია და სიდიდე ნორმალურადაა განაწილებული ან σ უცნობია და $n \geq 30$, მაშინ საშუალოსათვის ნდობის ინტერვალის ასაგებად იყენებენ სტანდარტულ ნორმალურ განაწილებას. მაგრამ, ძალიან ბევრ პრაქტიკულ სიტუაციაში, პოპულაციის სტანდარტული გადახრა უცნობია და შერჩევის მოცულობა ნაკლებია 30-ზე. ასეთ სიტუაციებში ნდობის ინტერვალის აგებისას პოპულაციის უცნობი სტანდარტული გადახრის ადგილას შეიძლება გამოყენებულ იქნას შერჩევის სტანდარტული გადახრა. იმ შემთხვევაში როცა შერჩევის მოცულობა ნაკლებია 30-ზე და სიდიდე ნორმალურად ან დაახლოებით ნორმალურად არის განაწილებული, მაშინ უნდა გამოვიყენოთ გარკვეულწილად განსხვავებული განაწილება, ე. წ. t განაწილება.

ისტორიული შენიშვნა. t განაწილება პირველად მიღებულ იქნა 1908 წელს ირლანდიელი ლუდის მწარმოებლის უ. ს. გოსეტის (W. S. Gossel) მიერ. გოსეტი იკვლევდა ლუდის წარმოების ახალ მეთოდებს. ვინაიდან ლუდის მწარმოებლებს აკრძალული ჰქონდათ სამეცნიერო შედეგების გამოქვეყნება, გოსეტმა თავისი მიგნება გამოაქვეყნა სტიუდენტის (Student) ფსევდონიმით. შესაბამისად, t განაწილებას ზოგჯერ სტიუდენტის განაწილებას უწოდებენ.

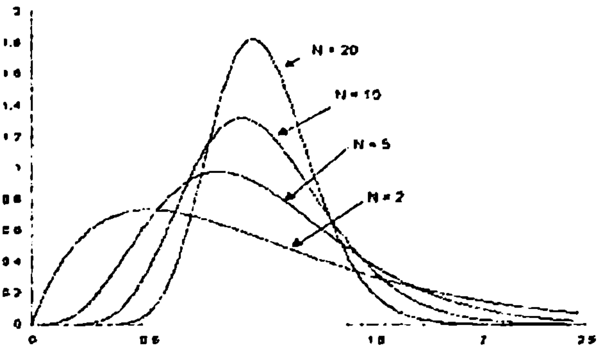
t განაწილების მახასიათებლები

t განაწილებას გააჩნია ნორმალური განაწილების ზოგიერთი თვისება და მისგან განსხვავებული თვისებებიც. t განაწილება ანალოგიურია სტანდარტული ნორმალური განაწილების შემდეგი მიმართულებებით:

1. ის ზარის ფორმისაა.
2. ის სიმეტრიულია საშუალოს მიმართ.
3. საშუალო, მედიანა და მოდა ტოლია 0-ის და განლაგებული არიან განაწილების ცენტრში.
4. წირი (გრაფიკი) არსად არ ეხება აბსცისთა ღერძს.

t განაწილება განსხვავდება სტანდარტული ნორმალური განაწილებისაგან შემდეგი მიმართულებებით:

1. დისპერსია მეტია 1-ზე.
2. t განაწილება ფაქტიურად არის წირების ოჯახი დაფუძნებული *თავისუფლების ხარისხის* ცნებაზე, რომელიც დაკავშირებულია შერჩევის მოცულობასთან.
3. შერჩევის მოცულობის ზრდასთან ერთად t განაწილება უახლოვდება სტანდარტულ ნორმალურ განაწილებას. იხ. ნახაზი (N – გვიჩვენებს თავისუფლების ხარისხს):



თავისუფლების ხარისხის ცნებას იყენებს ბევრი სტატისტიკური განაწილება და სხვადასხვა სტატისტიკური კრიტერიუმებისათვის თავისუფლების ხარისხის გამოსათვლელი ფორმულა განსხვავებულია. **თავისუფლების ხარისხი (degrees of freedom)** არის იმ მონაცემების რაოდენობა, რომლებიც არიან თავისუფალი ცვლილებისათვის მას შემდეგ რაც შერჩევის სტატისტიკა გამოთვლილია, და ის კარნახობს მკვლევარს თუ რომელი კონკრეტული წირი უნდა გამოიყენოს, როცა განაწილება შედგება წირების ოჯახისაგან.

მაგალითად, თუ 5 მონაცემის საშუალო არის 10 (ე. ი. ამ ხუთი მონაცემის ჯამია $10\times5=50$), მაშინ 4 ამ 5 მონაცემიდან თავისუფალია ცვლილებების მიმართ. მაგრამ, მას შემდეგ რაც 4 მონაცემი შერჩეულია, მე-5 მონაცემი უკვე ცალსახად განისაზღვრება (კონკრეტული რიცხვი იქნება) პირობიდან, რომ ხუთივეს ჯამი იყოს 50, რათა საშუალო გამოვიდეს 10 ($50:5=10$). შესაბამისად, ამ შემთხვევაში თავისუფლების ხარისხი არის $5-1=4$, და ეს რიცხვი მიუთითებს მკვლევარს თუ რომელი t წირი უნდა გამოიყენოს.

თავისუფლების ხარისხის აღსანიშნავად გამოვიყენებთ სიმბოლოს – **თ.ხ. (d.f.)**. საშუალოსათვის ნდობის ინტერვალის თავისუფლების ხარისხი გამოითვლება შერჩევის მოცულობიდან 1-ის გამოკლებით, ანუ თ.ხ.$=n-1$. არ უნდა გვეგონოს, რომ ანალოგიურად გამოითვლება სხვა სტატისტიკური კრიტერიუმების თავისუფლების ხარისხი.

საშუალოსათვის ნდობის ინტერვალის მოსაძებნ ფორმულას, რომელიც იყენებს t განაწილებას აქვს შემდეგი სახე.

საშუალოსათვის ნდობის ინტერვალის ფორმულა

(როცა σ უცნობია და $n<30$)

$$\overline{X}-t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}}<\mu<\overline{X}+t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}},$$

სადაც $\overline{X}$ – არის შერჩევითი საშუალო, $t_{n-1,\alpha/2}$ – არის $n-1$ თავისუფლების ხარისხის მქონე t განაწილების ზედა $\alpha/2$ კრიტიკული წერტილი, s – არის შერჩევის სტანდარტული გადახრა, ხოლო n – შერჩევის მოცულობა.

$t_{n-1,\alpha/2}$-ს მნიშვნელობის პოვნა ხდება t განაწილების ზედა კრიტიკული წერტილების ცხრილიდან. ამ ცხრილის პირველ სვეტში მითითებულია თავისუფლების ხარისხი, ხოლო პირველ სტრიქონში კი α -ს მნიშვნელობები.

მაგალითი 1. ვიპოვოთ $t_{n-1,\alpha/2}$-ის მნიშვნელობა 95%-იანი ნდობის ინტერვალისათვის, როცა შერჩევის მოცულობა არის 22.

ამოხსნა. $თ.ხ.= n-1 = 22-1 = 21$. $\alpha = 1-0.95 = 0.05$, შესაბამისად, $\alpha/2 = 0.05/2 = 0.025$. ამიტომ t განაწილების ზედა კრიტიკული წერტილების ცხრილში 21-ე სტრიქონისა და 0.025-ის შესაბამისი სვეტის გადაკვეთაზე ვპოულობთ $t_{n-1,\alpha/2}$-ის მნიშვნელობას:

$$t_{n-1,\alpha/2} = t_{21,0.025} = 2.080\,.$$

მაგალითი 2. გააჩერეს შემთხვევით შერჩეული 10 ავტომობილი და გაუზომეს მარჯვენა წინა საბურავის პროტექტორის სიღრმე. საშუალო აღმოჩნდა 0.32 დიუმი (1 დიუმი არის 2.54 სმ.), ხოლო სტანდარტული გადახრა — 0.08 დიუმი. იპოვეთ 95%-იანი საიმედოობის ნდობის ინტერვალი პროტექტორის სიღრმის საშუალოსათვის. იგულისხმეთ, რომ სიდიდე მიახლოებით ნორმალურადაა განაწილებული.

ამოხსნა. ამ შემთხვევაში $n = 10$, $\overline{X} = 0.32$, $s = 0.08$. ვინაიდან პოპულაციის სტანდარტული გადახრა σ უცნობია, ის უნდა შევცვალოთ შერჩევის სტანდარტული გადახრით $s = 0.08$.

$თ.ხ.= n-1 = 10-1 = 9$. $\alpha = 1-0.95 = 0.05$, $\alpha/2 = 0.05/2 = 0.025$. შესაბამისად t განაწილების ზედა კრიტიკული წერტილების ცხრილში მე-9 სტრიქონისა და 0.025-ის შესაბამისი სვეტის გადაკვეთაზე ვპოულობთ $t_{n-1,\alpha/2}$-ის მნიშვნელობას: $t_{n-1,\alpha/2} = t_{9,0.025} = 2.262$.

შევიტანოთ ეს მონაცემები ნდობის ინტერვალის ფორმულაში

$$\overline{X} - t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}},$$

$$0.32 - 2.262\cdot\frac{0.08}{\sqrt{10}} < \mu < 0.32 + 2.262\cdot\frac{0.08}{\sqrt{10}},$$

$$0.32 - 0.057 < \mu < 0.32 + 0.057,$$

$$0.26 < \mu < 0.38\,.$$

ამრიგად, 10 მოცულობის მქონე შერჩევის საფუძველზე შეიძლება დავასკვნათ, რომ ყველა წინა მარჯვენა საბურავების პოპულაციის პროტექტორების სიღრმის საშუალო 95%-იანი საიმედოობის დონით მოთავსებულია 0.26 დიუმსა და 0.38 დიუმს შორის.

მაგალითი 3. ქვემოთ მოყვანილია ელექტროობით გაჩენილი სახლის ხანძრების რიცხვი უკანასკნელი 7 წლის განმავლობაში. იპოვეთ 99%-იანი საიმედოობის ნდობის ინტერვალი ყოველწლიურად ელექტროობით გაჩენილი სახლის ხანძრების რიცხვის საშუალოსათვის.

5460 5900 6090 6310 7160 8440 9930

ამოხსნა.

ნაბიჯი 1. გამოვთვალოთ ამ მონაცემების საშუალო და სტანდარტული გადახრა, რისთვისაც გამოვიყენოთ ფორმულები

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n},$$

$$s=\sqrt{\frac{1}{n-1}[(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2]}.$$

გვაქვს: $\overline{X}=7041.4$, $s=1610.3$.

ნაბიჯი 2. ვიპოვოთ $t_{n-1,\alpha/2}$. ამ შემთხვევაში $n=7$; თ.ხ.$=n-1=7-1=6$; $\alpha=1-0.99=0.01$, $\alpha/2=0.01/2=0.005$. ამიტომ t განაწილების ზედა კრიტიკული წერტილების ცხრილში მე-6 სტრიქონისა და 0.005-ის შესაბამისი სვეტის გადაკვეთაზე ვპოულობთ $t_{n-1,\alpha/2}$-ის მნიშვნელობას: $t_{n-1,\alpha/2}=t_{6,0.005}=3.707$.

ნაბიჯი 3. შევიტანოთ ეს მონაცემები ნდობის ინტერვალის ფორმულაში

$$\overline{X}-t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}}<\mu<\overline{X}+t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}},$$

$$7041.4-3.707\cdot\frac{1610.3}{\sqrt{7}}<\mu<7041.4+3.707\cdot\frac{1610.3}{\sqrt{7}},$$

$$7041.4-2256.2<\mu<7041.4+2256.2,$$

$$4785.2<\mu<9297.6.$$

ამრიგად, უკანასკნელი 7 წლის განმავლობაში ელექტროობით გაჩენილი სახლის ხანძრების მონაცემების საფუძველზე შეგვიძლია დავასკვნათ, რომ 99%-იანი საიმედოობის დონით ყოველწლიურად ელექტროობით გაჩენილი სახლის ხანძრების პოპულაციის საშუალო მერყეობს 4785.2-დან 9297.6-მდე.

საშუალოსათვის ნდობის ინტერვალის აგების დროს სტუდენტებს ზოგჯერ უჭირთ გადაწყვეტა თუ როდის უნდა გამოიყენონ $z_{\alpha/2}$ ან $t_{n-1,\alpha/2}$. როგორც აღნიშნული იყო, როცა σ ცნობილია $z_{\alpha/2}$-ის გამოყენება შეიძლება მიუხედავად იმისა თუ როგორია შერჩევის მოცულობა ნორმალურად განაწილებული სიდიდის შემთხვევაში ან თუ $n\geq30$ ნებისმიერი სიდიდის შემთხვევაში. როცა σ უცნობია და $n\geq30$ ნდობის ინტერვალის ფორმულაში σ-ს ადგილას შეიძლება დაიწეროს s და გამოიყენება $z_{\alpha/2}$. და ბოლოს, როცა σ უცნობია და $n<30$, ფორმულაში გამოიყენება s და $t_{n-1,\alpha/2}$, თუ კი სიდიდე დაახლოებით ნორმალურად განაწილებულია. ეს წესი სქემატურად ასე გამოიყურება:

როდის გამოიყენება z ან t განაწილება

ცნობილია თუ არა σ? $\xrightarrow{\text{დიახ}}$ გამოიყენე $z_{\alpha/2}$ შერჩევის მოცულობის მიუხედავად*

$\downarrow$ არა

არის თუ არა $n \geq 30$? $\xrightarrow{\text{დიახ}}$ გამოიყენე $z_{\alpha/2}$ და ფორმულაში σ შეცვალე s-ით

$\downarrow$ არა

გამოიყენე $t_{n-1,\alpha/2}$ და s ფორმულაში**

* სიდიდე უნდა იყოს ნორმალურად განაწილებული როცა $n<30$.
** სიდიდე უნდა იყოს მიახლოებით ნორმალური.

ნდობის ინტერვალი და შერჩევის მოცულობა პროპორციისათვის.

მტკიცებულებაში: "საფრანგეთის მოქალაქეების 10% ლაპარაკობს არაბულ ენაზე" – პარამეტრს 10% ეწოდება **პროპორცია**. ეს ნიშნავს, რომ მთელი საფრანგეთის მასშტაბით ყოველი 100 მოქალაქიდან 10 ლაპარაკობს არაბულ ენაზე. პროპორცია წარმოადგენს მთელის ნაწილს. ის შეიძლება წარმოდგენილ იქნეს წილადის, ათწილადის ან პროცენტის სახით. ამ შემთხვევაში, $10\% = 0.10 = \dfrac{10}{100}$ ან $\dfrac{1}{10}$. პროპორცია წარმოადგენს აგრეთვე ალბათობას. ამ შემთხვევაში, თუ საფრანგეთის მოქალაქე შერჩეულია შემთხვევით, მაშინ ალბათობა იმისა, რომ იგი ლაპარაკობს არაბულად არის 0.10.

პროპორციები შეიძლება მივიღოთ როგორც შერჩევიდან, ისე პოპულაციიდან.

პროპორციის ცნებაში გამოყენებული სიმბოლოები

p – პოპულაციის პროპორციის სიმბოლო

$\hat{p}$ (იკითხება "p ქუდიანი") – შერჩევის პროპორციის სიმბოლო

შერჩევის პროპორციისათვის:

$$\hat{p}=\frac{X}{n} \text{ და } \hat{q}=\frac{n-X}{n} \text{ ან } \hat{q}=1-\hat{p},$$

სადაც X – შერჩევის იმ მონაცემების რაოდენობა, რომელთაც ახასიათებთ საინტერესო თვისება და n – შერჩევის მოცულობა.

მაგალითად, გამოკვლევისას 200 ადამიანს დაუსვეს შეკითხვა: არიან თუ არა ისინი კმაყოფილი თავიანთი სამუშაოთი ან პროფესიით? 162 ადამიანმა უპასუხა, რომ ის კმაყოფილია თავისი სამუშაოთი ან პროფესიით. ამ შემთხვევაში, $n=200$, $X=162$, და შესაბამისად,

$$\hat{p}=\frac{X}{n}=\frac{162}{200}=0.81.$$

შეიძლება ითქვას, რომ ამ შერჩევისათვის გამოკითხული ადამიანების 0.81 ან 81% კმაყოფილია თავისი სამსახურით ან პროფესიით. შერჩევის პროპორცია არის $\hat{p}=0.81$.

იმ ადამიანების პროპორცია, რომლებიც არ არიან კმაყოფილი არც სამსახურით და არც პროფესიით არის $\hat{q}$, სადაც $\hat{q}=(n-X)/n$. ამ გამოკითხვის შემთხვევაში

$$\hat{q}=\frac{200-162}{200}=\frac{38}{200} \text{ ან } 0.19 \text{ ან } 19\%.$$

იმ შემთხვევაში, როცა $\hat{p}$ და $\hat{q}$ მოცემულია წილადებში ან ათწილადებში, მაშინ $\hat{p}+\hat{q}=1$, ხოლო როცა $\hat{p}$ და $\hat{q}$ მოცემულია პროცენტებში, მაშინ $\hat{p}+\hat{q}=100\%$. აქედან გამომდინარე, $\hat{q}=1-\hat{p}$ ან $\hat{p}=1-\hat{q}$, როცა $\hat{p}$ და $\hat{q}$ მოცემულია წილადებში ან ათწილადებში. შესაბამისად, ამ გამოკითხვის დროს $\hat{q}$ შეგვეძლია ვიპოვოთ $\hat{q}=1-\hat{p}$ ფორმულის გამოყენებით და რა თქმა უნდა მივიღებთ იგივე შედეგს: $\hat{q}=1-\hat{p}=1-0.81=0.19$.

ანალოგიური თანაფარდობები სამართლიანია პოპულაციის პროპორციებისთვისაც. კერძოდ, $p=1-q$, $q=1-p$ და $p+q=1$, როცა p და q გამოსახულია წილადებით ან ათწილადებით. როცა p და q გამოსახულია პროცენტების სახით, მაშინ $p+q=100\%$, $p=100\%-q$ და $q=100\%-p$.

მაგალითი 1. გამოკითხული 150 ოჯახიდან 54-ს გააჩნია ჰაერის ცენტრალური კონდიცირების სისტემა. იპოვეთ $\hat{p}$ და $\hat{q}$, სადაც $\hat{p}$ – წარმოადგენს იმ ოჯახების პროპორციას, რომელთაც გააჩნია ჰაერის ცენტრალური კონდიცირების სისტემა.

ამოხსნა. ვინაიდან $X=54$ და $n=150$, ამიტომ

$$\hat{p}=\frac{X}{n}=\frac{54}{150}=0.36 \text{ ან } 36\%,$$

$$\hat{q}=\frac{n-X}{n}=\frac{150-54}{150}=\frac{96}{150}=0.64 \text{ ან } 64\%.$$

ისევე როგორც საშუალოს შემთხვევაში, სტატისტიკოსი მოცემული შერჩევითი პროპორციის მიხედვით ცდილობს შეაფასოს პოპულაციის საშუალო. პოპულაციის პროპორციის როგორც წერტილოვანი, ისე ინტერვალური შეფასება შესაძლებელია აიგოს შერჩევითი პროპორციის გამოყენებით. პოპულაციის პროპორციის p წერტილოვან შეფასებად გამოიყენება შერჩევითი პროპორცია $\hat{p}$. კარგი შემფასებლის სამი თვისების თანახმად, $\hat{p}$ არის ჩაუნაცვლებელი, ძალმოსილი და ფარდობითად ეფექტური შეფას-

ეპა. მაგრამ, ისევე როგორც საშუალოს შემთხვევაში, აქაც ჩვენ არ შეგვიძლია ვთქვათ თუ რამდენად ზუსტია წერტილოვანი შეფასება p. ამიტომ, სტატისტიკოსი აგრეთვე იყენებს პროპორციის ინტერვალურ შეფასებას და განსაზღვრავს ალბათობას, რომლითაც ინტერვალი მოიცავს პოპულაციის პროპორციას.

პროპორციისათვის ნდობის ინტერვალის ასაგებად უნდა გამოყენებულ იქნას შეფასების მაქსიმალური შეცდომა, რომელიც ტოლია

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

პროპორციისათვის ნდობის ინტერვალის ფორმულა

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}},$$

როცა $np \geq 5$ და $nq \geq 5$.

პროპორციისათვის ნდობის ინტერვალის დამრგვალების წესი: მძიმის შემდეგ უნდა დავტოვოთ სამი ათწილადი ნიშანი.

მაგალითი 2. მოხუცების 500 მომელელისაგან შემდგარ შერჩევაში 60 მამაკაცია. იპოვეთ 90%-იანი საიმედოობის ნდობის ინტერვალი მოხუცების მოვლის პროგრამაში მონაწილე მამაკაცთა ჭეშმარიტი პროპორციისათვის.

ამოხსნა. ვინაიდან $\alpha = 1 - 0.90 = 0.10$, ამიტომ $z_{\alpha/2} = z_{0.05} = 1.65$. გარდა ამისა, $X = 60$, $n = 500$. შესაბამისად,

$$\hat{p} = \frac{X}{n} = \frac{60}{500} = 0.12 \text{ და } \hat{q} = 1 - \hat{p} = 1 - 0.12 = 0.88$$

შევიტანოთ ეს მნიშვნელობები ნდობის ინტერვალის ფორმულაში

$$\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}},$$

$$0.12 - 1.65 \cdot \sqrt{\frac{0.12 \times 0.88}{500}} < p < 0.12 + 1.65 \cdot \sqrt{\frac{0.12 \times 0.88}{500}},$$

$$0.12 - 0.024 < p < 0.12 + 0.024,$$

$$0.096 < p < 0.144 \text{ ან } 9.6\% < p < 14.4\%.$$

ამრიგად, მოხუცების მოვლის პროგრამაში მონაწილე მამაკაცების რეალური პროპორცია 90%-იანი საიმედოობით მოთავსებულია 9.6%-სა და 14.4%-ს შორის.

მაგალითი 3. 200000 გამოკითხული ძაღლის მეპატრონედან 12%-მა განაცხადა, რომ მისი ძაღლის საყვარელი სახელია *ბინგო.* იპოვეთ 95%-იანი საიმედოობის ნდობის ინტერვალი იმ ადამიანთა რეალური პროპორციის, რომელთა ძაღლის საყვარელი სახელია *ბინგო.*

ამოხსნა. ამ შემთხვევაში $\hat{p} = 0.12$; $n = 200000$; $\alpha = 1 - 0.95 = 0.05$; $z_{\alpha/2} = z_{0.025} = 1.96$. შევიტანოთ ეს რიცხვები ნდობის ინტერვალის ფორმულაში:

$$0.12-1.96\cdot\sqrt{\frac{0.12\times0.88}{200000}}<p<0.12+1.96\cdot\sqrt{\frac{0.12\times0.88}{200000}},$$
$$0.119<p<0.121.$$

შესაბამისად, 95%-იანი საიმედოობით შეგვიძლია დავასკვნათ, რომ ძაღლის მეპატრონეებიდან იმ ადამიანების რეალური პროპორცია ეისი ძაღლის საყვარელი სახელია *ბინგო* მერყეობს 11.9%-დან 12.1%-მდე.

შერჩევის მოცულობა. პოპულაციის პროპორციის ინტერვალური შეფასებისათვის საჭირო შერჩევის მინიმალური მოცულობის გამოსათვლელი ფორმულა ადვილად მიიღება თუ შეფასების მაქსიმალური შეცდომის გამოსათვლელი გამოსახულებიდან

$$E=z_{\alpha/2}\cdot\sqrt{\hat{p}\hat{q}/n}$$

ამოვხსნით n-ს.

პოპულაციის პროპორციის ინტერვალური შეფასებისათვის საჭირო შერჩევის მინიმალური მოცულობის ფორმულა

$$n=\hat{p}\hat{q}\cdot\left(\frac{z_{\alpha/2}}{E}\right)^2$$

ის მეტობით უნდა დამრგვალდეს უახლოეს ნატურალურ რიცხვამდე.

შესაძლებელია ორი განსხვავებული სიტუაცია: პირველი, როცა $\hat{p}$-ს გარკვეული აპროქსიმაცია ცნობილია (ისევე როგორც წინა მაგალითებში), და სწორედ ეს მნიშვნელობა გამოიყენება ფორმულაში.

მეორე, როცა $\hat{p}$-ს აპროქციმაცია უცნობია, მაშინ უნდა გამოვიყენოთ $\hat{p}=0.5$. ეს მნიშვნელობა იძლევა შერჩევის საკმაოდ დიდ მოცულობას იმისათვის, რომ გარანტირებული იყოს ინტერვალური შეფასების მოცემული საიმედოობის დონე და შეფასების შეცდომა. ეს აიხსნება იმით, რომ როცა თითოეული $\hat{p}$ და $\hat{q}$ ტოლია 0.5-ის, მაშინ ნამრავლი $\hat{p}\cdot\hat{q}$ მაქსიმალურია, რაც კარგად ჩანს შემდეგი ცხრილიდან:

$\hat{p}$	$\hat{q}$	$\hat{p}\cdot\hat{q}$
0.1	0.9	0.09
0.2	0.8	0.16
0.3	0.7	0.21
0.4	0.6	0.24
0.5	**0.5**	**0.25**
0.6	0.4	0.24
0.7	0.3	0.21
0.8	0.2	0.16
0.9	0.1	0.09

მაგალითი 4. მკვლევარს სურს 95%-იანი საიმედოობით შეაფასოს იმ ადამიანების პროპორცია, რომლებსაც სახლში აქვთ საკუთარი კომპიუტე-

რი. წინა გამოკვლევამ აჩვენა, რომ გამოკითხულთა 40%-ს სახლში აქვს საკუთარი კომპიუტერი. მკვლევარს სურს 2%-იანი სიზუსტე ჭეშმარიტი პროპორციისათვის. ვიპოვოთ შერჩევის მოცულობის აუცილებელი მინიმუმი.

ამოხსნა. რადგანაც $z_{\alpha/2} = z_{(1-0.95)/2} = z_{0.025} = 1.96$, $E = 0.02$, $\hat{p} = 0.40$ და $\hat{q} = 0.60$, ამიტომ შერჩევის მნიმიმალური მოცულობა იქნება:

$$n = \hat{p}\hat{q}\cdot(\frac{z_{\alpha/2}}{E})^2 = 0.40\times0.60\times(\frac{1.96}{0.02})^2 = 2304.96,$$

რომელიც მეტობით დამრგვალების შემდეგ გვაძლევს, რომ უნდა გამოიკითხოს 2305 ადამიანი.

მაგალითი 5. მკვლევარს სურს 90%-იანი საიმედოობითა და 5%-იანი სიზუსტით შეაფასოს ალმასრულებელი ხელისუფლების იმ წარმომადგენელთა პროპორცია, რომლებსაც აქვთ მანქანის ტელეფონი.

ამოხსნა. ვინაიდან უცნობია $\hat{p}$-ის შეფასება, სტატისტიკოსი ირჩევს $\hat{p} = 0.5$ და $\hat{q} = 0.5$ მნიშვნელობებს. ამ მნიშვნელობებით მიღებული შერჩევის მოცულობა იქნება საკმარისად დიდი იმისათვის, რომ უზრუნველყოთ სასურველი საიმედოობა და სიზუსტე. გარდა ამისა, ამ შემთხვევაში $E = 0.05$, $\alpha = 1 - 0.90 = 0.1$, $z_{\alpha/2} = z_{0.05} = 1.65$. შესაბამისად,

$$n = \hat{p}\hat{q}\cdot(\frac{z_{\alpha/2}}{E})^2 = 0.5\times0.5\times(\frac{1.65}{0.05})^2 = 272.25,$$

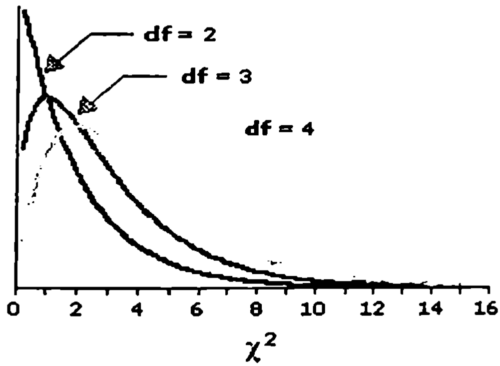
რომელიც მეტობით დამრგვალების შედეგად გვაძლევს, რომ უნდა გამოიკითხოს ალმასრულებელი ხელისუფლების 273 წარმომადგენელი.

ნდობის ინტერვალი დისპერსიისა და სტანდარტული გადახრისათვის.

ჩვენ უკვე ვიცით როგორ გამოვთვალოთ ნდობის ინტერვალი საშუალოსა და პროპორციისათვის. ახლა ჩვენ ავხსნით თუ როგორ ავაგოთ ნდობის ინტერვალი დისპერსიისა და სტანდარტული გადახრისათვის. სტატისტიკაში სიდიდის დისპერსია და სტანდარტული გადახრა ისევე მნიშვნელოვანია როგორც საშუალო. მაგალითად, როცა ორი ერთმანეთთან დასაკავშირებელი დეტალი მზადდება უაღრესად მნიშვნელოვანია რომ მათი დიამეტრების დისპერსია (variance) იყოს რაც შეიძლება მცირე, რათა ისინი ერთმანეთს მოერგოს, წინააღმდეგ შემთხვევაში ისინი იქცევიან ჯართის გროვად. ფარმაცევტულ წარმოებაში მედიკამენტის პილიულების (ტაბლეტების) დისპერსია და სტანდარტული გადახრა უაღრესად მნიშვნელოვან როლს თამაშობს იმაში, რომ პაციენტმა მიიღოს პრეპარატის საჭირო დოზა. ამ აზრით, ნდობის ინტერვალის ცოდნა დისპერსიისა და სტანდარტული გადახრისათვის აუცილებელია.

იმისათვის რომ ავაგოთ ნდობის ინტერვალი დისპერსიისა და სტანდარტული გადახრისათვის საჭიროა ახალი სტატისტიკური განაწილება. მას ეწოდება **ხი-კვადრატ განაწილება (chi-square distribution)**.

ხი-კვადრატ სიდიდე ანალოგიურია t სიდიდის იმ აზრით, რომ მისი განაწილება აგრეთვე არის წირების ოჯახი დაფუძნებული თავისუფლების ხარისხზე. ხი-კვადრატის აღსანიშნავად გამოიყენება სიმბოლო χ^2 (χ – ბერძნული "ხი"). რამოდენიმე განაწილება ნაჩვენებია ქვემოთ მოყვანილ ნახაზზე შესაბამისი თავისუფლების ხარისხის მითითებით. აღსანიშნავია, რომ თავისუფლების ხარისხის რიცხვის ზრდასთან ერთად ხი-კვადრატ განაწილების წირი ემსგავსება ნორმალური განაწილების წირს. სინამდვილეში ხი-კვადრატ განაწილება მიიღება $(n-1)s^2/\sigma^2$ გამოსახულების მნიშვნელობებით, როცა შემთხვევითი შერჩევა აღებულია ნორმალურად განაწილებული პოპულაციიდან, რომლის დისპერსია არის σ^2 (ცენტრალური ზღვარითი თეორემის თანახმად ნორმირებული სიდიდის $(\chi_n^2-n)/\sqrt{2n}$ განაწილება უახლოვდება სტანდარტულ ნორმალურ განაწილებას, როცა n უსასრულოდ იზრდება):

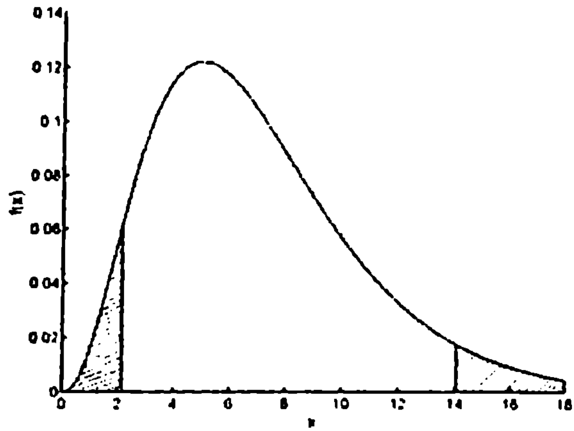


ისტორიული შენიშვნა. ხი-კვადრატ განაწილება თავისუფლების ხარისხით 2 ფორმულირებული იქნა ჰერშელის (Hershel) მიერ 1869 წელს, როცა იგი სწავლობდა სროლის სიზუსტეს. ამის შემდეგ ბევრმა მათემატიკოსმა შეიტანა თავისი წვლილი ხი-კვადრატ განაწილების შესწავლაში. 1903 წელს კ. პირსონის (K. Pearson) მიერ შემოთავაზებულ იქნა სტატისტიკური ჰიპოთეზების შემოწმების ხი-კვადრატ კრიტერიუმი.

ხი-კვადრატ სიდიდე არ შეიძლება იყოს უარყოფითი და განაწილებას გააჩნია დადებითი ასიმეტრია. დაახლოებით 100-ის ტოლი თავისუფლების ხარისხის შემთხვევაში ხი-კვადრატ განაწილება ხდება სიმეტრიული. ხი-კვადრატ განაწილების ქვეშ მოქცეული არის ფართობია 1.

როგორც ჩვენ ქვემოთ დავინახავთ დისპერსიისა და სტანდარტული გადახრისათვის ნდობის ინტერვალში მონაწილეობს ხი-კვადრატ განაწილების ორი განსხვავებული მნიშვნელობა. ერთი უნდა მოვძებნოთ ხი-კვადრატ განაწილების ცხრილის მარცხენა, ხოლო მეორე კი მარჯვენა მხარეს. მაგალითად, იმისათვის რომ ცხრილიდან ვიპოვოთ 95%-იანი ნდობის ინტერვალის შესაბამისი მნიშვნელობები, საჭიროა 95% შევცვალოთ ათწილადით 0.95, 1-ს გამოვაკლოთ 0.95, 1-0.95=0.05, შედეგი გავყოთ 2-ზე ($\alpha/2=$

$=0.05/2=0.025$). 0.025-ის შესაბამისი სვეტი მოვძებნოთ ცხრილის მარჯვენა მხარეს და ამ სვეტისა და $n-1$ თავისუფლების ხარისხის შესაბამისი სტრიქონის გადაკვეთაზე მდგომი მნიშვნელობა აღვნიშნოთ $\chi^2_{\text{მარჯვ.}}(\alpha/2)$ სიმბოლოთი. მეორე (მარცხენა) მნიშვნელობის მისაღებად 1-ს გამოვაკლოთ $\alpha/2$ $(1-\alpha/2=1-0.025=0.975)$, 0.975-ის შესაბამისი სვეტი მოვძებნოთ ცხრილის მარცხენა მხარეს და ამ სვეტისა და $n-1$ თავისუფლების ხარისხის შესაბამისი სტრიქონის გადაკვეთაზე მდგომი მნიშვნელობა აღვნიშნოთ $\chi^2_{\text{მარცხ.}}(1-\alpha/2)$ სიმბოლოთი. სინამდვილეში ხი-კვადრატ განაწილების ქვეშ $\chi^2_{\text{მარჯვ.}}(\alpha/2)$-ის მარჯვენივ მოქცეული არის ფართობია $\alpha/2$, ხოლო $\chi^2_{\text{მარცხ.}}(1-\alpha/2)$-ის მარცხნივ მოქცეული არის ფართობია $1-\alpha/2$. ანალოგიური პროცედურა გამოიყენება 90%-იანი და 99%-იანი ნდობის ინტერვალებისათვის.



$\chi^2_{\text{მარცხ.}}(\alpha)$ $\chi^2_{\text{მარჯვ.}}(\alpha)$

მაგალითი 1. ვიპოვოთ $\chi^2_{\text{მარცხ.}}$-სა და $\chi^2_{\text{მარჯვ.}}$-ს მნიშვნელობები 90%-იანი საიმედოობის დონისათვის, როცა $n=25$.

ამოხსნა. ჯერ ვიპოვოთ $\chi^2_{\text{მარჯვ.}}$: 1-ს გამოვაკლოთ 0.90, 1-0.90=0.10 და შედეგი გავყოთ 2-ზე, მივიღებთ 0.05-ს. მაშინ $\chi^2_{\text{მარჯვ.}}$ იქნება ხი-კვადრატ განაწილების ცხრილში 0.05 სვეტისა და იმ სტრიქონის გადაკვეთაზე მდგომი მნიშვნელობა, რომელიც შეესაბამება $n-1=25-1=24$-ის ტოლ თავისუფლების ხარისხს: $\chi^2_{\text{მარჯვ.}}=36.415$.

ახლა ვიპოვოთ $\chi^2_{\text{მარცხ.}}$: 1-ს გამოვაკლოთ 0.05, მივიღებთ 0.95-ს. ამიტომ $\chi^2_{\text{მარცხ.}}$ იქნება ხი-კვადრატ განაწილების ცხრილში 0.95 სვეტისა და იმ სტრიქონის გადაკვეთაზე მდგომი მნიშვნელობა, რომელიც შეესაბამება $n-1=25-1=24$-ის ტოლ თავისუფლების ხარისხს: $\chi^2_{\text{მარცხ.}}=13.848$.

σ^2-ისა და σ-ს კარგი შეფასებებია შესაბამისად, s^2 და s.

იმისათვის, რომ ავაგოთ ნდობის ინტერვალი დისპერსიისა და სტანდარტული გადახრისათვის ჩვენ უნდა ვიგულისხმოთ, რომ სიდიდე ნორმალური კანონითაა განაწილებული.

ნდობის ინტერვალის ფორმულა დისპერსიისათვის

$$\frac{(n-1)s^2}{\chi^2_{\text{მარჯ.}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{მარცხ.}}},$$

$$\text{თ.ხ.} = n-1.$$

ნდობის ინტერვალის ფორმულა სტანდარტული გადახრისათვის

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{მარჯ.}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{მარცხ.}}}},$$

$$\text{თ.ხ.} = n-1.$$

დამრგვალების წესი. იმ შემთხვევაში, როცა ნდობის ინტერვალს დისპერსიისა ან სტანდარტული გადახრისათვის ვაგებთ ნედლი მონაცემების გამოყენებით, მაშინ დამრგვალებას ვახდენთ ერთი ათობითი ნიშნით მეტზე, ვიდრე ათობითი ნიშნები გვაქვს ნედლ მონაცემებში.

იმ შემთხვევაში, როცა ნდობის ინტერვალს დისპერსიისა ან სტანდარტული გადახრისათვის ვაგებთ შერჩევითი დისპერსიისა ან შერჩევითი სტანდარტული გადახრის საშუალებით, მაშინ დამრგვალებას ვახდენთ იმდენივე ათობით ნიშნამდე, რამდენი ათობითი ნიშანიც აქვს შერჩევითი დისპერსიას ან შერჩევით სტანდარტულ გადახრას.

მაგალითი 2. იპოვეთ თამბაქოში ნიკოტინის შემცველობის დისპერსიისა და სტანდარტული გადახრისათვის 95%-იანი ნდობის ინტერვალი, თუ ცნობილია, რომ 20 სიგარეტისაგან შედგენილ შერჩევაში ნიკოტინის შემცველობის სტანდარტული გადახრა 1.6 მილიგრამის ტოლია.

ამოხსნა. ვინაიდან ამ შემთხვევაში $\alpha = 1-0.95 = 0.05$, ამიტომ ხი-კვადრატ განაწილების ორი კრიტიკული მნიშვნელობა, რომელიც შეესაბამება $\alpha/2 = 0.025$-სა და $1-\alpha/2 = 0.975$-ს და თავისუფლების ხარისხის რიცხვს, რომელიც ტოლია $n-1 = 20-1 = 19$, შესაბამისად იქნება: $\chi^2_{\text{მარჯ.}} = 32.852$ და $\chi^2_{\text{მარცხ.}} = 8.907$. შევიტანოთ ეს მონაცემები ნდობის ინტერვალის ფორმულაში:

$$\frac{(n-1)s^2}{\chi^2_{\text{მარჯ.}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{მარცხ.}}},$$

$$\frac{(20-1)\cdot(1.6)^2}{32.852} < \sigma^2 < \frac{(20-1)\cdot(1.6)^2}{8.907},$$

$$1.4805 < \sigma^2 < 5.4608.$$

თუ ახლა ვისარგებლებთ დამრგვალების წესით, მივიღებთ:

$$1.5 < \sigma^2 < 5.5.$$

ამრიგად, ჩვენ შეგვიძლია 95%-იანი საიმედოობით ვიგულისხმოთ, რომ სიგარეტში ნიკოტინის შემცველობის რეალური დისპერსია მოთავსებულია 1.5 მილიგრამსა და 5.5 მილიგრამს შორის.

ნდობის ინტერვალი სტანდარტული გადახრისათვის იქნება

$$\sqrt{1.5} < \sigma < \sqrt{5.5}\,,$$

$$1.2247 < \sigma < 2.3452\,,$$

$$1.2 < \sigma < 2.3\,.$$

მაშასადამე, 20 სიგარეტიანი შერჩევის საფუძველზე ჩვენ ვასკვნით, რომ სიგარეტში ნიკოტინის შემცველობის რეალური სტანდარტული გადახრის 95%-იანი ნდობის ინტერვალია (1.2მგ, 2.3მგ).

მაგალითი 3. ვიპოვოთ სრულწლოვანი ადამინისათვის ერთი დღის განმავლობაში ზამთრის კურორტზე საბაგირო გზის ბილეთის ფასის დისპერსიისა და სტანდარტული გადახრისათვის 90%-იანი ნდობის ინტერვალი, ქვემოთ მოყვანილი შემთხვევით შერჩეული 10 ზამთრის კურორტისათვის აღნიშნული ბილეთის ფასის (დოლარებში) მონაცემების მიხედვით. ვიგულისხმოთ, რომ ეს სიდიდე განაწილებულია ნორმალურად.

59	54	53	52	51
39	49	46	49	48

ამოხსნა.

ნაბიჯი 1. ვიპოვოთ შერჩევის დისპერსია. ვისარგებლოთ ფორმულით

$$s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]\,.$$

მივიღებთ, რომ $s^2 = 28.2$.

ნაბიჯი 2. ხი-კვადრატ განაწილების ცხრილიდან ვიპოვოთ $\chi^2_{\text{მარცხ}}$ და $\chi^2_{\text{მარჯვ}}$. გვაქვს: $\alpha = 1 - 0.90 = 0.10$, $\alpha/2 = 0.05$, $1 - \alpha/2 = 0.95$, თ.ხ.$= n - 1 = 10 - 1 = 9$. ამიტომ $\chi^2_{\text{მარცხ}} = 3.325$ და $\chi^2_{\text{მარჯვ}} = 16.919$.

ნაბიჯი 3. მიღებული მნიშვნელობები შევიტანოთ დისპერსიისათვის ნდობის ინტერვალის ფორმულაში

$$\frac{(n-1)s^2}{\chi^2_{\text{მარჯვ.}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{მარცხ}}}\,,$$

$$\frac{(10-1)\cdot 28.2}{16.919} < \sigma^2 < \frac{(10-1)\cdot 28.2}{3.325}\,,$$

$$15.0008 < \sigma^2 < 76.3308\,,$$

$$15.0 < \sigma^2 < 76.3\,.$$

ნაბიჯი 4. ნდობის ინტერვალი სტანდარტული გადახრისათვის იქნება:

$$\sqrt{15.0} < \sigma < \sqrt{76.3}\,,$$

$$3.8729 < \sigma < 8.7349\,,$$

$$3.87 < \sigma < 8.73\,.$$

მაშასადამე, 10 შემთხვევით შერჩეული ზამთრის კურორტის მონაც მების მიხედვით ზამთრის კურორტებზე სრულწლოვანი ადამიანისათვი საბაგირო გზის ერთდღიანი ბილეთის ფასის პოპულაციის სტანდარტულ გადახრა 90%-იანი საიმედოობით მოთავსებულია 3.87$-სა და 8.73$-ს შორი (მძიმის შემდეგ ორი ათიბითი ნიშანი აღებულია არა დამრგვალების წეს ის გამო, არამედ იმიტომ, რომ ფასი გამოისახება დოლარებში და ცენტებ ში).

დასკვნა. სტატისტიკური დასკვნების თეორიის უმნიშვნელოვანეს ნაწილია შეფასების თეორია. პოპულაციის პარამეტრების შეფასება დაკავ შირებულია ამ პოპულაციიდან შემთხვევითი შერჩევის გაკეთებასთან და ისეთი სტატისტიკის არჩევასთან და გამოთვლასთან, რომელიც წარმოადგ ენს ამ თუ იმ პარამეტრის საუკეთესო შეფასებას. კარგი შემფასებელი უნ და იყოს ჩაუნაცვლებელი, ძალმოსილი და ფარდობითად ეფექტური. μ-სა და p-ს საუკეთესო შეფასებებია შესაბამისად $\overline{X}$ და $\hat{p}$, ხოლო σ^2-სა და σ-ს საუკეთესო შეფასებებია შესაბამისად s^2 და s.

არსებობს პარამეტრის ორი ტიპის შეფასება: წერტილოვანი შეფასება და ინტერვალური შეფასება. წერტილოვანი შეფასება წარმოადგენს კონკრეტულ მნიშვნელობას. მაგალითად, თუ მკვლევარს სურს შეაფასოს ზრდასრული თევზის საშუალო სიგრძე, ის აკეთებს ამ თევზების შერჩევას, ზომავს მათ სიგრძეებს და ითვლის შერჩევით საშუალოს – დავუშვათ, რომ ის ქლმოჩნდა 32მმ. შერჩევის ამ საშუალოს მიხედვით მკვლევარი აფასებს პოპულაციის საშუალოს 32 მილიმეტრით.

წერტილოვანი შეფასების ნაკლი იმაში მდგომარეობს, რომ ამ დროს შეუძლებელია განისაზღვროს შეფასების სიზუსტე. ამის გამო სტატისტიკოსები უპირატესობას ანიჭებენ ინტერვალურ სეფასებებს. მოცემული შერჩევისათვის ნდობის ინტერვალის აგებით სტატისტიკოსი 95%-იანი ან 99%-იანი (ან სხვა პროცენტიანი) საიმედოობით თვლის რომ მისი შეფასება მოიცავს პარამეტრის რეალურ მნიშვნელობას. საიმედოობის დონე განისაზღვრება მკვლევარის მიერ. საიმედოობის მაღალ დონეს უფრო ფართე ნდობის ინტერვალი შეესაბამება.

პოპულაციის საშუალოსათვის ნდობის ინტერვალის აგებად გამოიყენება ან z სტატისტიკის, ან t სტატისტიკის მნიშვნელობა, რაც დამოკიდებულია იმაზე ცნობილია თუ არა პოპულაციის სტანდარტული გადახრა და რა მოცულობისაა შერჩევა. თუ σ ცნობილია ან $n \geq 30$, მაშინ გამოიყენება z სტატისტიკის მნიშვნელობა. თუ σ უცნობია, მაშინ t სტატისტიკის მნიშვნელობა გამოიყენება, როცა შერჩევის მოცულობა ნაკლებია 30-ზე, და პოპულაცია ნორმალურადაა განაწილებული.

ნდობის ინტერვალის აგება დაკავშირებულია შერჩევის მოცულობის განსაზღვრასთან. შერჩევის აუცილებელი მინიმალური მოცულობის დასადგენად უნდა გაეითვალისწინოთ შემდეგი ინფორმაცია:

წინასწარ უნდა იყოს დადგენილი საიმედოობის დონე.

პოპულაციი სტანდარტული გადახრა უნდა იყოს ცნობილი ან შეიძლებოდეს მისი შეფასება.

შეფასების მაქსიმალური შეცდომა უნდა იყოს დადგენილი.

ნდობის ინტერვალი და შერჩევის მოცულობა აგრეთვე ითვლება პროპორციისათვის ნორმალური განაწილების გამოყენებით. გარდა ამისა, იგება ნდობის ინტერვალი დისპერსიისათვის და სტანდარტული გადახრისათვის ხი-კვადრატ განაწილების საშუალებით.

დავუბრუნდეთ ზემოთ განხილულ მაგალითს: გამოკითხვამ აჩვენა, რომ სატელევიზიო პროგრამით უკმაყოფილო ადამიანების 45% ტელევიზორს გადართავს სხვა არხზე, მაშინ როდესაც ადამიანების 15% საერთოდ თიშავს თავის ტელევიზორს. გამოკითხული იყო 4000 სრულწლოვანი ადამიანი და ცდომილება შეადგენს 3%-ს. ჩვენ უკვე შეგვიძლია პასუხი გავცეთ ამ მაგალითის მოყვანისას გაჩენილ კითხვებს.

აქ მოყვანილი შეფასება წარმოადგენს წერტილოვან შეფასებას. მაგრამ, ვინაიდან შეფასების მაქსიმალური შეცდომა შეადგენს 3%-ს, ამიტომ აქედან ადვილად მიიღება ინტერვალური შეფასებაც. რადგანაც ადამიანების 45% გადართავს ტელევიზორს სხვა არხზე, ამიტომ ნდობის ინტერვალი იმ ადამიანების რეალური პროცენტისათვის, რომლებიც ტელევიზორს გადართავენ სხვა არხზე იქნება $42\% < p < 48\%$. აღნიშნულ მაგალითს აკლია თუ რა საიმედოობის დონითაა აგებული ნდობის ინტერვალი.

თუ გამოვიყენებთ შერჩევის აუცილებელი მინიმალური მოცულობის გამოსათვლელ ფორმულას, მაშინ ადვილად დავრწმუნდებით, რომ 95%-იანი საიმედოობის ნდობის ინტერვალის ასაგებად საჭირო შერჩევის მოცულობის აუცილებელია მინიმუმია 1068. მართლაც, ვინაიდან $\hat{p}$-ის მნიშვნელობა უცნობია, გამოვიყენოთ $\hat{p}=\hat{q}=0.5$ და გამოვთვალოთ n:

$$n=\hat{p}\hat{q}\cdot\left(\frac{z_{\alpha/2}}{E}\right)^2=0.5\cdot 0.5\cdot\left(\frac{1.96}{0.03}\right)^2=1067.1.$$

შესაბამისად, $n=1068$.

თავი IV
პოპულაციის პარამეტრების წერტილოვანი შეფასებები

შერჩევითი საშუალო და დისპერსია.

ძალიან ბევრ შემთხვევაში ჩვენ გაგვაჩნია ინფორმაცია შემთხვევითი სიდიდის განაწილების კანონის სახის შესახებ (ნორმალური, ბერნულის, თანაბარი და ა. შ.), მაგრამ არ ვიცით ამ განაწილების ისეთი პარამეტრები, როგორიცაა $E\xi$ და $D\xi$. ამ პარამეტრების განსაზღვრისათვის გამოიყენება შერჩევითი მეთოდი.

დავუშვათ, რომ n მოცულობის შერჩევა წარმოდგენილია ვარიაციული მწკრივის სახით. **შერჩევითი საშუალო** ეწოდება სიდიდეს:

$$\bar{x} = \frac{x_1 m_1 + x_2 m_2 + \ldots + x_k m_k}{n} = x_1 \frac{m_1}{n} + x_2 \frac{m_2}{n} + \ldots + \frac{m_k}{n}.$$

სიდიდეს $\nu_i = m_i / n$ ნიშნის x_i მნიშვნელობის **ფარდობითი სიხშირე** ეწოდება. თუ შერჩევიდან მიღებულ ნიშნის მნიშვნელობებს არ დავაჯგუფებთ და არ წარმოვადგენთ ვარიაციული მწკრივის სახით, მაშინ შერჩევითი საშუალოს გამოსათვლელად უნდა ვისარგებლოთ შემდეგი ფორმულით:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

ბუნებრივია $\bar{x}$ სიდიდე ჩაითვალოს $E\xi$ პარამეტრის შერჩევით შეფასებად. პარამეტრის შერჩევით შეფასებას, რომელიც წარმოადგენს რიცხვს, **წერტილოვანი შეფასება** ეწოდება.

შერჩევითი დისპერსია ეწოდება სიდიდეს:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 \nu_i = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

ის შეიძლება ჩაითვალოს გენერალური ერთობლიობის $D\xi$ დისპერსიის წერტილოვან შეფასებად.

დავუშვათ, რომ გენერალური ერთობლიობის ყოველი ობიექტი ხასიათდება ორი რაოდენობრივი x და y ნიშნით. მაგალითად, დეტალს შეიძლება ჰქონდეს ორი ზომა – სიგრძე და სიგანე, შეიძლება სხვადასხვა რეგიონში გაიზომოს მავნე ნივთიერებების კონცენტრაცია და დაფიქსირდეს თვის განმავლობაში მოსახლეობაში ფილტვების დაავადებების რაოდენობა, შეიძლება დროის ტოლ შუალედებში შევადაროთ მოცემული კორპორაციის აქციების შემოსავლიანობა რაიმე ინდექსს, რომელიც ახასიათებს აქციების მთელი ბაზრის საშუალო შემოსავლიანობას. ასეთ შემთხვევაში, გენერალური ერთობლიობა წარმოადგენს ორგანზომილებიან შემთხვევით სიდიდეს ξ, η. ეს შემთხვევითი სიდიდე გენერალური ერთობლიობის ობიექტების სიმრავლეზე ღებულობს მნიშვნელობებს x, y. თუ ჩვენ არ ვიცით ξ და η შემთხვევითი სიდიდეების ერთობლივი განაწილების კანონი, ჩვენ არ შეგვიძლია ვილაპარაკოთ მათ შორის კორელაციური კავშირის არსებ-

ობაზე ან სიძლიერეზე, მაგრამ, მიუხედავად ამისა, შერჩევითი მეთოდის გამოყენებით შესაძლებელია ზოგიერთი დასკვნის გაკეთება.

ასეთ შემთხვევაში, n მოცულობის შერჩევა წარმოიდგინება ცხრილის სახით, სადაც i-ური ამორჩეული ობიექტი ($i=1,2,...,n$) წარმოდგენილია რიცხვთა წყვილით x_i, y_i:

x_1	x_2	...	x_n
y_1	y_2	...	y_n

შერჩევითი კორელაციის კოეფიციენტი ეწოდება სიდიდეს:

$$r_{xy} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y},$$

სადაც

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \,,\ \sigma_x = \sqrt{\sigma_x^{\,2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

$$\sigma_y = \sqrt{\sigma_y^{\,2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

შერჩევითი კორელაციის კოეფიციენტი შეიძლება განხილულ იქნეს როგორც წერტილოვანი შეფასება კორელაციის კოეფიციენტის $\rho_{\xi\eta}$, რომელიც ახასიათებს გენერალურ ერთობლიობას.

შერჩევითი პარამეტრები $\overline{x}$, σ^2, r_{η} ან ნებისმიერი სხვა დამოკიდებულია იმაზე, გენერალური ერთობლიობის რომელი ობიექტები მოხვდნენ შერჩევაში და განსხვავდებიან შერჩევიდან შერჩევამდე. ამიტომ ისინი თვითონ წარმოადგენენ შემთხვევით სიდიდეებს.

დავუშვათ, რომ შერჩევითი პარამეტრი δ განიხილება როგორც გენერალური ერთობლიობის Δ პარამეტრის შერჩევითი შეფასება. შერჩევით შეფასებას ეწოდება **გადაუადგილებადი** (ან **ჩაუნაცვლებელი**), თუ

$$E\delta = \Delta.$$

იმისათვის რომ დავამტკიცოთ ზოგიერთი წერტილოვანი შეფასების გადაუადგილებადობა, n მოცულობის შერჩევას განვიხილავთ როგორც სისტემას n დამოუკიდებელი $\xi_1, \xi_2, ..., \xi_n$ შემთხვევითი სიდიდის, რომელთაგან თითოეული გააჩნია იგივე განაწილების კანონი, იგივე პარამეტრებით, რაც ξ შემთხვევით სიდიდეს, რომელიც წარმოადგენს გენერალურ ერთობლიობას. ასეთი მიდგომის შემთხვევაში ცხადი ხდება თანაფარდობები:

$$Ex_i = E\xi_i = E\xi;\ Dx_i = D\xi_i = D\xi\ \ (\underline{k} = 1,2,...n).$$

ახლა ვაჩვენოთ, რომ შერჩევითი საშუალო $\overline{x}$ წარმოადგენს გენერალური ერთობლიობის საშუალოს გადაუადგილებად შეფასებას, ან რაც იგივეა ξ შემთხვევითი სიდიდის მათემატიკური ლოდინის გადაუადგილებად შეფასებას. მართლაც, გვაქვს:

$$E\bar{x} = E\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\left(E\xi_1 + E\xi_2 + \cdots + E\xi_n\right) = \frac{1}{n} n E\xi = E\xi .$$

გამოვთვალოთ შერჩევითი საშუალოს დისპერსია. ცხადია, რომ:

$$D\bar{x} = D\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n^2}\left(D\xi_1 + D\xi_2 + \cdots + D\xi_n\right) = \frac{1}{n^2} n D\xi = \frac{D\xi}{n} .$$

ვიპოვოთ ახლა რისი ტოლია შერჩევითი დისპერსიის მათემატიკური ლოდინი, რისთვისაც თავიდან σ^2 გარდავქმნათ შემდეგნაირად:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - E\xi + E\xi - \bar{x}\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - E\xi\right)^2 - 2\left(x_i - E\xi\right)\left(\bar{x} - E\xi\right) + \left(\bar{x} - E\xi\right)^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - E\xi\right)^2 - \left(\bar{x} - E\xi\right)^2$$

(შევნიშნავთ, რომ ჩვენ აქ გამოვიყენეთ გარდაქმნა:

$$\sum_{i=1}^{n} 2\left(x_i - E\xi\right)\left(\bar{x} - E\xi\right) = 2\left(\bar{x} - E\xi\right)\sum_{i=1}^{n}\left(x_i - E\xi\right) =$$

$$= 2\left(\bar{x} - E\xi\right)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} E\xi\right) = 2\left(\bar{x} - E\xi\right)\left(n\bar{x} - nE\xi\right) = 2n\left(\bar{x} - E\xi\right)^2).$$

ამიტომ შერჩევითი დისპერსიის მათემატიკური ლოდინი იქნება:

$$E\sigma^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - E\xi\right)^2 - \left(\bar{x} - E\xi\right)^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left(x_i - E\xi\right)^2 - E\left(\bar{x} - E\xi\right)^2 = \frac{1}{n} n\, D\xi - D\bar{x} =$$

$$= D\xi - \frac{D\xi}{n} = \frac{n-1}{n} D\xi .$$

როგორც ვხედავთ, $E\sigma^2 \neq D\xi$. ამიტომ *შერჩევითი დისპერსია არ წარმოადგენს გენერალური ერთობლიობის დისპერსიის გადაუადგილებად შეფასებას.*

იმისათვის რომ მივიღოთ გენერალური ერთობლიობის დისპერსიის გადაუადგილებადი შეფასება, საჭიროა შერჩევითი დისპერსია გავამრავლოთ მამრავლზე $n/(n-1)$. მიღებული სიდიდე აღინიშნება s^2-ით და მას **შესწორებული შერჩევითი დისპერსია** ეწოდება:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

(ზოგჯერ s^2 გამოიყენება შერჩევითი დისპერსიის აღსანიშნავად და, ამ შემთხვევაში, შესწორებულ შერჩევით დისპერსიას აღნიშნავენ s'^2 სიმბოლოთი).

თუ ჩვენ გვაქვს გენერალური ერთობლიობის ერთი და იგივე პარამეტრის რამოდენიმე ჩაუნაცვლებელი შეფასება, მაშინ იმ შეფასებას, რომელსაც გააჩნია უმცირესი დისპერსია, **ეფექტური** ეწოდება.

n მოცულობის შერჩევიდან მიღებულ გენერალური ერთობლიობის Δ პარამეტრის წერტილოვან δ_n შეფასებას ეწოდება **ძალმოსილი**, თუ ის ალბათობით კრებადია Δ-სკენ. ეს იმას ნიშნავს, რომ ნებისმიერი დადებითი ε და γ რიცხვებისათვის, მოიძებნება ისეთი რიცხვი $n_{\varepsilon\gamma}$, რომ ყველა n რიცხვისათვის, რომელიც აკმაყოფილებს უტოლობას $n > n_{\varepsilon\gamma}$, სრულდება პირობა

$$P\left(\left|\delta_n - \Delta\right| < \varepsilon\right) > 1 - \gamma .$$

შევნიშნავთ, რომ $\bar{x}$ და s^2 შესაბამისად წარმოადგენენ გენერალური ერთობლიობის მათემატიკური ლოდინისა და დისპერსიის ჩაუნაცვლებელ, ძალმოსილ და ეფექტურ შეფასებებს.

შერჩევითი პარამეტრების განაწილება ნორმალური პოპულაციისათვის.

დავუშვათ, რომ $\xi_1, \xi_2, \ldots, \xi_n$ წარმოადგენს შერჩევას ნორმალური გენერალური ერთობლიობიდან, $\xi_i = N(x;\mu;\sigma^2)$, $i = 1,2,\ldots,n$. რადაგანაც $\bar{x}$ წარმოადგენს დამოუკიდებელი ნორმალურად განაწილებული შემთხვევითი სიდიდეების წრფივ კომბინაციას, ამიტომ

$$\overline{X} = N(x;\mu;\sigma^2 / n).$$

გავარკვიოთ ახლა შერჩევითი დისპერსიის განაწილების კანონი. ჯერ განვიხილოთ ის შემთხვევა, როცა გენერალური ერთობლიობის საშუალო (მათემატიკური ლოდინი) ცნობილია. რადგანაც, $\xi_i = N(x;\mu;\sigma^2)$, ამიტომ

$$(\xi_i - \mu)/\sigma = N(x;0;1).$$

თუ გავითვალისწინებთ, რომ შერჩევით დისპერსიას აქვს სახე

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mu)^2 ,$$

ადვილად დავრწმუნდებით, რომ სამართლიანია თანაფარდობა:

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{\xi_i - \mu}{\sigma}\right)^2$$

ამრიგად, nS^2/σ^2 წარმოიდგინება $N(x;0;1)$ კანონით განაწილებული დამოუკიდებელი შემთხვევითი სიდიდეების კვადრატების ჯამის სახით. ე. ი. მას აქვს χ^2 განაწილება თავისუფლების ხარისხით n:

$$\frac{nS^2}{\sigma^2} = \chi^2(n).$$

უცნობი საშუალოს შემთხვევაში (განსხვავებით განხილული შემთხვევისაგან, სადაც $\xi_i - \mu$ $(i = 1,2,\ldots,n)$ დამოუკიდებელი შემთხვევითი სიდიდეებია), nS^2/σ^2 აღარ წარმოადგენს დამოუკიდებელი შემთხვევითი სიდიდ-

ეების კვადრატების ჯამს, $\xi_i - \overline{X}$ შემთხვევით სიდიდეებს გააჩნიათ ერთი „ბმა". კერძოდ, ვინაიდან $\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, ამიტომ $\sum_{i=1}^{n}(\xi_i - \overline{X}) = 0$. ამ შემთხვევაში ადგილი აქვს თანაფარდობას

$$\frac{nS^2}{\sigma^2} = \chi^2(n-1),$$

რაც ექვივალენტურია შემდეგი თანაფარდობის:

$$\frac{(n-1)S'^2}{\sigma^2} = \chi^2(n-1)$$

(ვინაიდან, $nS^2 = (n-1)S'^2$).

გარდა ამისა, მტკიცდება რომ $\overline{X}$ და S^2 დამოუკიდებელი შემთხვევითი სიდიდეებია.

სტატისტიკაში ხშირად გამოიყენება ე. წ. Z სტატისტიკა, და T სტატისტიკა:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{და} \quad T = \frac{\overline{X} - \mu}{S/\sqrt{n-1}} = \frac{\overline{X} - \mu}{S'/\sqrt{n}}.$$

ცხადია, რომ $Z = N(x;0;1)$. რაც შეეხება T სტატისტიკას, ის გადავწეროთ შემდეგი სახით:

$$T = \frac{Z}{\sqrt{\frac{nS^2}{\sigma^2}/(n-1)}} = \frac{Z}{\sqrt{\frac{(n-1)S'^2}{\sigma^2}/(n-1)}}.$$

ვინაიდან, $\frac{(n-1)S'^2}{\sigma^2} = \chi^2(n-1)$ და Z და $\frac{(n-1)S'^2}{\sigma^2}$ შემთხვევითი სიდიდეები დამოუკიდებელია, ამიტომ T სტატისტიკას აქვს სტიუდენტის განაწილება თავისუფლების ხარისხით $n-1$.

შეფასებათა აგების მეთოდები

მაქსიმალური დასაჯერობის მეთოდი.

დავუშვათ, რომ X – დისკრეტული შემთხვევითი სიდიდეა, რომელმაც ექსპერიმენტის შედეგად მიიღო მნიშვნელობები $x_1, x_2, \ldots, x_n$. დავუშვათ, რომ ჩვენთვის ცნობილია ამ შემთხვევითი სიდიდის განაწილების კანონი, რომელიც განისაზღვრება Θ პარამეტრით, მაგრამ უცნობია ამ პარამეტრის რიცხვითი მნიშვნელობა. ჩვენი მიზანია ვიპოვოთ ამ პარამეტრის წერტილოვანი შეფასება.

ვთქვათ, $p(x_i, \Theta)$ – არის ალბათობა იმისა, რომ ექსპერიმენტის შედეგად X შემთხვევითი სიდიდე მიიღებს x_i მნიშვნელობას. დისკრეტული X შემთხვევითი სიდიდის **მაქსიმალური დასაჯერობის ფუნქცია** ეწოდება Θ არგუმენტის ფუნქციას, რომელიც განისაზღვრება ფორმულით:

$$L(x_1, x_2, \ldots, x_n; \Theta) = p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta).$$

Θ პარამეტრის წერტილოვანი შეფასების როლში იღებენ მის ისეთ $\Theta^* = \Theta(x_1, x_2, \ldots, x_n)$ მნიშვნელობას, რომლის დროსაც მაქსიმალური დასაჯე-

რობის ფუნქცია აღწევს თავის მაქსიმუმს. Θ^* შეფასებას **მაქსიმალური დასაჯერობის შეფასებას** უწოდებენ.

რადგანაც ფუნქციები L და $\ln L$ მაქსიმუმს აღწევენ Θ-ს ერთი და იგივე მნიშვნელობისათვის, უფრო მოხერხებულია მოვძებნოთ $\ln L$ ფუნქციის მაქსიმუმი (ვინაიდან ნამრავლის ლოგარითმი ლოგარითმების ჯამია და ამდენად კრიტიკული წერტილების პოვნისას ნამრავლის გაწარმოების ნაცვლად მოგვიწევს ჯამის გაწარმოება, რაც გაცილებით მარტივია). ამ ფუნქციას **მაქსიმალური დასაჯერობის ლოგარითმული ფუნქცია** ეწოდება.

$\ln L$ ფუნქციის მაქსიმუმის მიმნიჭებელი წერტილის მოსაძებნად საჭიროა შემდეგი პროცედურების ჩატარება:

1). ვიპოვოთ წარმოებული $\dfrac{d\ln L}{d\Theta}$;

2). გავუტოლოთ წარმოებული ნულს (მივიღებთ ე. წ. **მაქსიმალური დასაჯერობის განტოლებას**) და ვიპოვოთ კრიტიკული წერტილები'

3). ვიპოვოთ მეორე წარმოებული $\dfrac{d^2\ln L}{d\Theta^2}$; თუ ის უარყოფითია კრიტიკულ წერტილში, მაშინ ეს წერტილი – მაქსიმუმის წერტილია.

აღსანიშნავია, რომ მაქსიმალური დასაჯერობის მეთოდით მიღებული შეფასებები ძალმოსილია (თუმცა შესაძლებელია არ იყოს ჩაუნაცვლებელი), განაწილებული არიან ასიმპტოტურად ნორმალურად შერჩევის დიდი მოცულობის შემთხვევაში და გააჩნიათ უმცირესი დისპერსია სხვა ასიმპტოტურად ნორმალურ შეფასებებთან შედარებით. თუ შესაფასებელი Θ პარამეტრისათვის არსებობს ეფექტური Θ^* შეფასება, მაშინ მაქსიმალური დასაჯერობის განტოლებას გააჩნია ერთადერთი ამონახსნი Θ^*. ეს მეთოდი ყველაზე სრულად იყენებს შერჩევის მონაცემებს და ამიტომ განსაკუთრებით სასარგებლოა მცირე შერჩევების დროს.

მაქსიმალური დასაჯერობის მეთოდის ნაკლად შეიძლება ჩაითვალოს გამოთვლების სირთულე.

უწყვეტი შემთხვევითი სიდიდის შემთხვევაში, რომლის $f(x)$ განაწილების სიმკვრივის სახე ცნობილია, მაგრამ იგი შეიცავს უცნობ Θ პარამეტრს, მაქსიმალური დასაჯერობის ფუნქციას აქვს შემდეგი სახე:

$$L(x_1,x_2,...,x_n;\Theta)=f(x_1,\Theta)\cdot f(x_2,\Theta)\cdots f(x_n,\Theta)\,.$$

უცნობი პარამეტრის მაქსიმალური დასაჯერობის შეფასების საპოვნელად უნდა ჩავატაროთ იგივე პროცედურები, რაც დისკრეტულ შემთხვევაში.

მომენტთა მეთოდი.

მომენტთა მეთოდი დაფუძნებულია იმ გარემოებაზე, რომ საწყისი და ცენტრალური ემპირიული მომენტები წარმოადგენენ შესაბამისი საწყისი და ცენტრალური თეორიული მომენტების ძალმოსილ შეფასებებს. ამიტომ ჩვენ შეგვიძლია თეორიული მომენტები გავუტოლოთ იმავე რიგის შესაბამის ემპირიულ მომენტებს. თუ მოცემულია განაწილების $f(x,\Theta)$ სიმკვრივის სახე, რომელიც განისაზღვრება ერთი უცნობი Θ პარამეტრით (დამოკიდებულია ერთ უცნობ პარამეტრზე), მაშინ ამ პარამეტრის შესაფასებლ-

ად საკმარისია გვქონდეს ერთი განტოლება. მაგალითად, შეგვიძლია გავუ-
ტოლოთ ერთმანეთს პირველი რიგის საწყისი მომენტები:

$$\overline{x}_B = E(X) = \int_{-\infty}^{\infty} xf(x;\Theta)dx = \varphi(\Theta),$$

და მივიღებთ განტოლებას Θ პარამეტრის საპოვნელად. მისი ამონახსნი Θ^* იქნება Θ პარამეტრის წერტილოვანი შეფასება, რომელიც წარმოადგენს შერჩევითი საშუალოს ფუნქციას და, შესაბამისად, შერჩევის ფუნქციას:

$$\Theta = \psi\,(x_1, x_2, \ldots, x_n).$$

თუ განაწილების სიმკვრივე განისაზღვრება ორი Θ_1 და Θ_2 პარამეტრით (დამოკიდებულია ორ პარამეტრზე), მაშინ მოითხოვება შევადგინოთ ორი განტოლება, მაგალითად, $\nu_1 = M_1$, $\mu_2 = m_2$.

აქედან ვღებულობთ ორი განტოლებისაგან შემდგარ სისტემას ორი Θ_1 და Θ_2 უცნობით: $\begin{cases} E(X) = \overline{x}_B \\ D(X) = D_B \end{cases}$

მისი ამოხსნები Θ_1^* და Θ_2^* იქნებიან Θ_1 და Θ_2 პარამეტრების წერტილოვანი შეფასებები დამოკიდებული შერჩევაზე:

$$\Theta_1 = \psi_1\,(x_1, x_2, \ldots, x_n),$$
$$\Theta_2 = \psi_2(x_1, x_2, \ldots, x_n).$$

თავი V
ინტერვალური შეფასებები. ნდობის ინტერვალი

გენერალური ერთობლიობის პარამეტრების წერტილოვანი შეფასებები შეიძლება მიღებულ იქნეს შერჩევითი მონაცემების დამუშავების საორიენტაციო, პირველად შედეგებად. მათი ნაკლი იმაში მდგომარეობს, რომ უცნობია რა სიზუსტით ფასდება პარამეტრი. დიდი მოცულობის შერჩევებისათვის სიზუსტე როგორც წესი საკმარისია (შეფასებების გადაუადგილებადობის, ძალმოსილებისა და ეფექტურობის პირობებში), მაშინ როდესაც მცირე მოცულობის შერჩევებისათვის შეფასების სიზუსტის საკითხი ძალიან მნიშვნელოვანია.

შემოვიღოთ გენერალური ერთობლიობის (ან ξ შემთხვევითი სიდიდის, რომელიც განმარტებულია ამ გენერალური ერთობლიობის ობიექტების სიმრავლეზე) უცნობი პარამეტრის ინტერვალური შეფასების ცნება. აღვნიშნოთ ეს პარამეტრი Δ-თი. მოცემული შერჩევიდან გარკვეული წესით იძებნება ისეთი რიცხვები Δ_1 და Δ_2, რომ სრულდებოდეს პირობა:

$$P(\Delta_1 < \Delta < \Delta_2) = P(\Delta \in (\Delta_1; \Delta_2)) = \gamma.$$

Δ_1 და Δ_2 რიცხვებს უწოდებენ **ნდობის საზღვრებს**, ხოლო (Δ_1, Δ_2) ინტერვალს – Δ პარამეტრის **ნდობის ინტერვალს**. γ რიცხვს ეწოდება **ნდობის ალბათობა** ან გაკეთებული შეფასების **საიმედოობა**.

თავიდან მოიცემა საიმედოობა. ჩვეულებრივ, მას ირჩევენ 0.95-ის, 0.99 -ის ან 0.999-ის ტოლს. მაშინ ალბათობა იმისა, რომ ჩვენთვის საინტერესო პარამეტრი მოხვდა (Δ_1, Δ_2) ინტერვალში საკმარისად მაღალია. რიცხვი $(\Delta_1+\Delta_2)/2$ – ნდობის ინტერვალის შუაწერტილი – იძლევა Δ პარამეტრის მნიშვნელობას $(\Delta_2-\Delta_1)/2$-ს ტოლი სიზუსტით, რომელიც წარმოადგენს ნდობის ინტერვალის სიგრძის ნახევარს.

საზღვრები Δ_1 და Δ_2 განისაზღვრება შერჩევითი მონაცემებიდან და წარმოადგენენ $X_1, X_2, \dots, X_n$ შემთხვევითი სიდიდეების ფუნქციებს. შესაბამისად, საზღვრები თვითონაც შემთხვევითი სიდიდეებია. აქედან გამომდინარე, ნდობის ინტერვალი (Δ_1, Δ_2) -- აგრეთვე შემთხვევითია. ის შეიძლება ფარავდეს ან არ ფარავდეს Δ პარამეტრს. სწორედ ასეთი აზრით უნდა გავიგოთ შემთხვევითი ხდომილება, რომელიც მდგომარეობს იმაში, რომ ნდობის ინტერვალი ფარავს Δ რიცხვს.

ნდობის ინტერვალი ნორმალური განაწილების მათემატიკური ლოდინისათვის ცნობილი დისპერსიის შემთხვევაში:

დავუშვათ, რომ შემთხვევითი სიდიდე ξ (შეიძლება ვილაპარაკოთ გენერალურ ერთობლიობაზე) განაწილებულია ნორმალური განაწილების კანონის მიხედვით, რომლის დისპერსია ცნობილია $D\xi = \sigma^2$ $(\sigma > 0)$. გენერალური ერთობლიობიდან (რომლის ობიექტების სიმრავლეზე განმარტებულია შემთხვევითი სიდიდე) კეთდება n მოცულობის შერჩევა. შერჩევა $X_1, X_2, \dots, X_n$ განიხილება როგორც ერთობლიობა n დამოუკიდებელი შემთ-

ხვევითი სიდიდის, რომლებიც იგივე კანონით არიან განაწილებული როგორც ξ. ამ შემთხვევაში, როგორც ჩვენ უკვე ვნახეთ:

$$EX_1 = EX_2 = \cdots = EX_n = E\xi, \quad DX_1 = DX_2 = \cdots = DX_n = D\xi,$$

$$E\overline{X} = E\xi, \quad D\overline{X} = D\xi / n.$$

ცნობილია, რომ მოცემულ შემთხვევაში შემთხვევითი სიდიდე $\overline{X}$ აგრეთვე განაწილებულია ნორმალური განაწილების კანონით. აღვნიშნოთ უცნობი მათემატიკური ლოდინი a-თი, $E\xi := a$ და მოცემული γ საიმედოობისათვის ($1-\gamma := \alpha$ სიდიდეს მნიშვნელოვნების დონეს უწოდებენ) შევარჩიოთ $d > 0$ რიცხვი ისე, რომ შესრულდეს პირობა:

$$P(|\bar{x} - a| < d) = \gamma \qquad ($$

ვინაიდან შემთხვევითი სიდიდე $\overline{X}$ განაწილებულია ნორმალურად მათემატიკური ლოდინით $E\overline{X} = E\xi = a$ და დისპერსიით $D\overline{X} = D\xi/n = \sigma^2/n$, ამიტომ გვაქვს:

$$P(|\overline{X} - a| < d) = P(a - d < \overline{X} < a + d) =$$

$$= \Phi\left((a + d - a)\sqrt{n}/\sigma\right) - \Phi\left((a - d - a)\sqrt{n}/\sigma\right) = 2\Phi\left(d\sqrt{n}/\sigma\right) - 1.$$

ახლა შევარჩიოთ $d > 0$ ისე, რომ შესრულდეს ტოლობა

$$2\Phi\left(d\sqrt{n}/\sigma\right) - 1 = \gamma \text{ ანუ } \Phi\left(d\sqrt{n}/\sigma\right) = (1+\gamma)/2.$$

ნებისმიერი $\gamma \in [0;1]$ რიცხვისათვის ნორმალური განაწილების ფუნქციის ცხრილიდან შეგვიძლია ვიპოვოთ ისეთი t რიცხვი, რომ

$$\Phi(t) = (1+\gamma)/2.$$

ამ t რიცხვს $(1+\gamma)/2$-კვანტილი ეწოდება. მას აღნიშნავენ აგრეთვე $x_{(1+\gamma)/2}$ სიმბოლოთი.

ტოლობიდან $d\sqrt{n}/\sigma = x_{(1+\gamma)/2}$ ვპოულობთ d-ს მნიშვნელობას: $d = \sigma x_{(1+\gamma)/2}/\sqrt{n}$. საბოლოო შედეგს მივიღებთ, თუ (I) ფორმულას წარმოვადგენთ შემდეგი სახით:

$$P\left(\overline{X} - \sigma x_{(1+\gamma)/2}/\sqrt{n} < a < \overline{X} + \sigma x_{(1+\gamma)/2}/\sqrt{n}\right) = \gamma.$$

უკანასკნელი ფორმულის აზრი მდგომარეობს შემდეგში: საიმედოობით γ ნდობის ინტერვალი

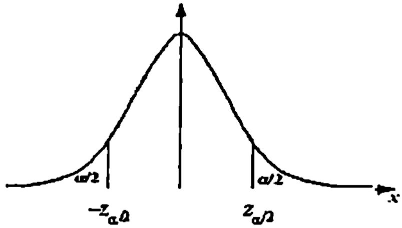
$$\left(\overline{X} - \sigma x_{(1+\gamma)/2}/\sqrt{n}; \overline{X} + \sigma x_{(1+\gamma)/2}/\sqrt{n}\right)$$

ფარავს (მოიცავს) გენერალური ერთობლიობის უცნობ პარამეტრს $a = E\xi$-ს. შეიძლება ითქვას სხვანაირად: წერტილოვანი შეფასება $\overline{X}$ განსაზღვრავს $E\xi$ პარამეტრის მნიშვნელობას $d = \sigma x_{(1+\gamma)/2}/\sqrt{n}$ სიზუსტითა და γ საიმედოობით.

გადავიდეთ $\alpha = 1-\gamma$ სიდიდეზე. მაშინ $(1+\gamma)/2 = 1-\alpha/2$ და შესაბამისად, $x_{(1+\gamma)/2} = x_{1-\alpha/2} = z_{\alpha/2}$, სადაც $z_{\alpha/2}$ – სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$-კრიტიკული წერტილია. ამიტომ

α მნიშვნელოვნების დონის მქონე ნდობის ინტერვალი ნორმალური განაწილების მათემატიკური ლოდინისათვის ცნობილი σ^2 დისპერსიის შემთხვევაში არის:

$$\left(\overline{X}-\sigma\cdot\frac{z_{\alpha/2}}{\sqrt{n}},\ \overline{X}+\sigma\cdot\frac{z_{\alpha/2}}{\sqrt{n}}\right).$$



ამოცანა. დავუშვათ გვაქვს გენერალური ერთობლიობა გარკვეული მახასიათებლით, რომელიც განაწილებულია ნორმალური კანონით, რომლის დისპერსია ტოლია 6.25-ის. ჩატარებულია $n=27$ მოცულობის შერჩევა და მიღებულია მახასიათებლის საშუალო შერჩევითი მნიშვნელობა $\overline{x}=12$. ვიპოვოთ ნდობის ინტერვალი, რომელიც ფარავს გენერალური ერთობლიობის გამოსაკვლევი მახასიათებლის უცნობ მათემატიკურ ლოდინს საიმედოობით $\gamma=0.99$.

ამოხსნა. პირველ რიგში, ლაპლასის ფუნქციის ცხრილებიდან ვიპოვოთ t-ს მნიშვნელობა ტოლობიდან $\phi(t)=(1+\gamma)/2=0.495$. მიღებული $t=2.58$ მნიშვნელობიდან განვსაზღვროთ შეფასების სიზუსტე (ანუ ნდობის ინტერვალის სიგრძის ნახევარი) d: $d=2.5\times2.58/\sqrt{27}\approx1.24$. აქედან ვღებულობთ საძებნ ნდობის ინტერვალს: (10.76, 13.24).

ნდობის დონის სიზუსტე და შერჩევის მოცულობის მოძებნა.

მოცემული γ ნდობის ალბათობისათვის დავადგინოთ შერჩევის ის მინიმალური n^* მოცულობა, რომელიც უზრუნველყოფს შეფასების წინასწარ ფიქსირებულ სიზუსტეს (შეფასება მით უფრო ზუსტია, რაც უფრო ნაკლებია ნდობის ინტერვალის სიგრძე). ცხადია, რომ რაც უფრო დიდია γ, მით უფრო დიდია $x_{(1+\gamma)/2}$, და შესაბამისად, განიერია ნდობის ინტერვალი და პირიქით. აქედან გამომდინარე, თუ შერჩევის მოცულობა ფიქსირებულია, ნდობის ინტერვალის სიგრძის (ანუ შეფასების სიზუსტის) შემცირება შესაძლებელია მხოლოდ ნდობის ალბათობის შემცირების ხარჯზე. ფიქსირებული ნდობის ალბათობის დროს ინტერვალის სიგრძე მით უფრო მცირეა, რაც უფრო დიდია შერჩევის მოცულობა. ყოველივე ზემოთ თქმულიდან ვასკვნით, რომ შეფასების ფიქსირებული სიზუსტე ნიშნავს ნდობის ინტერვალის ფიქსირებულ l სიგრძეს.

ვინაიდან, რომ γ ნდობის ინტერვალის სიგრძეა

$$2\frac{\sigma}{\sqrt{n}}x_{(1+\gamma)/2},$$

ამიტომ შერჩევის n^* მოცულობა უნდა შეირჩეს, როგორც შემდეგი განტოლების ამონახსნი

$$2\frac{\sigma}{\sqrt{n}}x_{(1+\gamma)/2}=l,$$

ე. ი.

$$n^*=(\frac{2\sigma}{l}x_{(1+\gamma)/2})^2=(\frac{2\sigma}{l}x_{1-\alpha/2})^2=(\frac{2\sigma}{l}z_{\alpha/2})^2.$$

ვინაიდან ასეთნაირად მოძებნილი n^* შეიძლება არ იყოს მთელი რიცხვი, ამიტომ n^*-ის როლში იღებენ მიღებული სიდიდის მთელ ნაწილს მიმატებულ ერთს:

$$n^*=[(\frac{2\sigma}{l}x_{(1+\gamma)/2})^2]+1=[(\frac{2\sigma}{l}z_{\alpha/2})^2]+1.$$

ნდობის ინტერვალი ნორმალური განაწილების მათემატიკური ლოდინისათვის უცნობი დისპერსიის შემთხვევაში.

დავუშვათ, რომ ξ – ნორმალური კანონით განაწილებული შემთხვევითი სიდიდეა უცნობი მათემატიკური ლოდინით $E\xi$, რომელიც აღვნიშნოთ a ასოთი. ჩავატაროთ n მოცულობის შერჩევა. განვსაზღვროთ შერჩევითი საშუალო $\overline{X}$ და შესწორებული შერჩევითი დისპერსია S'^2 ზემოთ მოყვანილი ფორმულების მიხედვით.

ცნობილია, რომ შემთხვევითი სიდიდე

$$t=\left(\overline{X}-a\right)\sqrt{n}/S'$$

განაწილებულია სტიუდენტის კანონის მიხედვით თავისუფლების $n-1$ ხარისხით. ამოცანა მდგომარეობს იმაში, რომ მოცემული γ საიმედოობისა და თავისუფლების $n-1$ ხარისხის მიხედვით, ვიპოვოთ ისეთი t^γ რიცხვი, რომ შესრულდეს ტოლობა:

$$P\left(\left|\left(\overline{X}-a\right)\sqrt{n}/S'\right|<t^\gamma\right)=\gamma,$$

(2)

ან მისი ექვივალენტური ტოლობა

$$P\left(\overline{X}-t^\gamma S'/\sqrt{n}<a<\overline{X}+t^\gamma S'/\sqrt{n}\right)=\gamma. \qquad (3)$$

აქ ფრჩხილებში წერია იმის პირობა, რომ უცნობი a პარამეტრის მნიშვნელობა ეკუთვნის გარკვეულ შუალედს, რომელიც არის სწორედ ნდობის ინტერვალი. მისი საზღვრები დამოკიდებულია γ საიმედოობაზე და აგრეთვე, შერჩევის $\overline{x}$ და s პარამეტრებზე.

იმისათვის, რომ γ სიდიდის მიხედვით ვიპოვოთ t_γ-ს მნიშვნელობა, (2) ტოლობა გადავწეროთ შემდეგი სახით:

$$P\left(\left|\left(\overline{X}-a\right)\sqrt{n}/S'\right|\geq t^\gamma\right)=1-\gamma.$$

აქედან, სტიუდენტის განაწილების სიმეტრიულობის ძალით, შეგვიძლია დავწეროთ, რომ:

$$P\{\frac{\overline{X}-a}{S'/\sqrt{n}} \geq t^{\gamma}\} = \frac{1-\gamma}{2} = \frac{\alpha}{2}.$$

აქედან ვასკვნით, რომ t^{γ} არის თავისუფლების $n-1$ ხარისხის მქონე სტიუდენტის განაწილების ზედა $\alpha/2$-კრიტიკული წერტილი:

$$t^{\gamma} = t_{n-1,(1-\gamma)/2} = t_{n-1,\alpha/2}$$

და α მნიშვნელოვნების დონის მქონე ნდობის ინტერვალს ნორმალური განაწილების მათემატიკური ლოდინისათვის უცნობი დისპერსიის შემთხვევაში აქვს შემდეგი სახე:

$$\overline{X} - t_{n-1,\alpha/2} \cdot \frac{S'}{\sqrt{n}} < a < \overline{X} + t_{n-1,\alpha/2} \cdot \frac{S'}{\sqrt{n}}.$$

ამოცანა. 20 ელექტრონათურის საკონტროლო შემოწმებისას მათი მუშაობის საშუალო ხანგრძლივობა აღმოჩნდა 2000 საათის ტოლი, ხოლო საშუალო კვადრატული გადახრა (გამოთვლილი როგორც კვადრატული ფესვი შესწორებული შემთხვევითი დისპერსიიდან) კი 11 საათის ტოლი. ცნობილია, რომ ნათურის მუშაობის ხანგრძლივობა წარმოადგენს ნორმალური კანონით განაწილებულ შემთხვევით სიდიდეს. განვსაზღვროთ ამ შემთხვევითი სიდიდის მათემატიკური ლოდინის ნდობის ინტერვალი საიმედოობით 0.95.

ამოხსნა. ამ შემთხვევაში $1-\gamma = 0.05$. სტიუდენტის განაწილების ცხრილიდან (თავისუფლების 19-ის ტოლი ხარისხით) ვპოულობთ, რომ $t^{\gamma} = 2.093$. გამოვთვალოთ შეფასების სიზუსტე: $2.093 \times 11/\sqrt{20} = 5.2$. ამიტომ ნდობის ინტერვალი იქნება (1994.8, 2005.2).

ნდობის ინტერვალი პოპულაციის საშუალოსათვის შერჩევის დიდი მოცულობის შემთხვევაში.

აქამდე ჩვენ ვგულისხმობდით, რომ პოპულაცია ნორმალურად იყო განაწილებული. ამასთანავე, შერჩევის მოცულობა ნებისმიერი იყო. ნორმალური პოპულაციის შემთხვევაში შერჩევითი საშუალოსა და დისპერსიის განაწილება ზუსტად არის ცნობილი, რასაც საზოგადოდ, ვერ ვიტყვით იმ პოპულაციებისათვის, რომლებიც არ არის ნორმალურად განაწილებული. მიუხედავად ამისა, ალბათობის თეორიის ზღვარითი თეორემები საშუალებას იძლევა ავაგოთ ნდობის ინტერვალი პოპულაციის უცნობი საშუალოსათვის ზოგად შემთხვევაშიც, როცა შერჩევის მოცულობა საკმაოდ დიდია.

იმ შემთხვევაში, როცა პოპულაციის დისპერსია, σ^2, ცნობილია, ცენტრალური ზღვარითი თეორემის თანახმად, თუ შერჩევის n მოცულობა საკმაოდ დიდია ($n \geq 30$)

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

სტატისტიკას მიახლოებით სტანდარტული ნორმალური განაწილება გააჩნია. ამიტომ, ზემოთ მოყვანილი მსჯელობების ანალოგიურად, შეგვიძლია დავასკვნათ, რომ: α მნიშვნელოვნების დონის მქონე ასიმპტოტური ნდობის ინტერვალი საშუალოსათვის იქნება

$$(\overline{X}_n - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}).$$

იმ შემთხვევაში, როცა პოპულაციის დისპერსიაც უცნობია, თუ შერჩევის n მოცულობა საკმაოდ დიდია ($n \geq 30$), მაშინ დაახლოებით ნორმალურია

$$T_n = \frac{\overline{X}_n - \mu}{S'_n / \sqrt{n}}$$

სტატისტიკა. ამ შემთხვევაში ინტერვალი

$$(\overline{X}_n - z_{\alpha/2}\cdot\frac{S'}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\cdot\frac{S'}{\sqrt{n}})$$

წარმოადგენს $1-\alpha$ საიმედოობის ასიმპტოტურ ნდობის ინტერვალს (მისი ნდობის დონე მიახლოებით ($1-\alpha$)-ს ტოლია).

ნდობის ინტერვალი ნორმალური განაწილების დისპერსიისათვის.

დავუშვათ, რომ ξ შემთხვევითი სიდიდე განაწილებულია ნორმალური განაწილების კანონით, რომლის დისპერსია $D\xi$ უცნობია. განიხილება ორი შემთხვევა: 1). პოპულაციის საშუალო μ ცნობილია და 2). პოპულაციის საშუალო უცნობია. კეთდება n მოცულობის შერჩევა. მისი საშუალებით განისაზღვრება შერჩევითი დისპერსია s^2 და შესწორებული შერჩევითი დისპერსია s'^2. ცნობილია რომ, შემთხვევითი სიდიდე $\chi^2 = nS^2 / D\xi$ (შესაბამისად, $\chi^2 = (n-1)S'^2 / D\xi$) განაწილებულია χ^2 განაწილების კანონით თავისუფლების ხარისხით n (შესაბამისად, $n-1$). მოცემული γ საიმედოობისათვის შეიძლება ვიპოვოთ ინტერვალების ისეთი საზღვრები χ_1^2 და χ_2^2, რომ

$$P\{\chi_1^2 < \chi^2 < \chi_2^2\} = \gamma \qquad (1)$$

ვიპოვოთ χ_1^2 და χ_2^2 შემდეგი პირობებიდან:

$$P\{\chi^2 \leq \chi_1^2\} = (1-\gamma)/2,$$

$$P\{\chi^2 \geq \chi_2^2\} = (1-\gamma)/2.$$

ნათელია, რომ ამ ორი უკანასკნელი პირობის შესრულებისას, სამართლიანი იქნება (1) ტოლობა.

χ^2 შემთხვევითი სიდიდის ცხრილებში, ჩვეულებრივ, მოიცემა

$$P\{\chi^2 \geq \chi_{n,\beta}^2\} = \beta$$

განტოლების ამონახსნი $\chi_{n,\beta}^2$ და მას თავისუფლების n ხარისხის მქონე ხი კვადრატ განაწილების ზედა β-კრიტიკული წერტილი ეწოდება. აქედან გამომდინარე, გვაქვს: $\chi_2^2 = \chi_{n,(1-\gamma)/2}^2 = \chi_{n,\alpha/2}^2$ (შესაბამისად, $\chi_2^2 = \chi_{n-1,(1-\gamma)/2}^2 = \chi_{n-1,\alpha/2}^2$).

χ_1^2-ის საპოვნელად ვისარგებლოთ საწინააღმდეგო ხდომილების ალბათობით. მაშინ მივიღებთ, რომ:

$$P(\chi^2 \geq \chi_1^2) = 1 - (1-\gamma)/2 = (1+\gamma)/2.$$

მიღებული ტოლობიდან ვასკვნით, რომ $\chi_1^2=\chi^2_{n,(1+\gamma)/2}=\chi^2_{n,1-\alpha/2}$ (შესაბამისად, $\chi_1^2=\chi^2_{n-1,(1+\gamma)/2}=\chi^2_{n-1,1-\alpha/2}$).

მას შემდეგ რაც ნაპოვნია χ_1^2-ისა და χ_2^2-ის მნიშვნელობები, გადავწეროთ (1) ტოლება შემდეგი სახით: $P\left(\chi_1^2<nS^2/D\xi<\chi_2^2\right)=\gamma$ (შესაბამისად

$$P\left(\chi_1^2<(n-1)S'^2/D\xi<\chi_2^2\right)=\gamma).$$

უკანასკნელი ტოლობა გადავწეროთ ისეთი ფორმით, რომ განსაზღრული იყოს უცნობი $D\xi$ პარამეტრის ნდობის ინტერვალის საზღერები. პოპულაციის ცნობილი საშუალოს შემთხვევაში გვექნება:

$$P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\alpha/2}}<D\xi<\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\alpha/2}}\right)=1-\alpha,$$

ხოლო უცნობი საშუალოს შემთხვევაში $1-\alpha$ საიმედოობის ნდობის ინტერვალი დიპერსიისათვის იქნება:

$$P\left(\frac{(n-1)S'^2}{\chi^2_{n-1,\alpha/2}}<D\xi<\frac{(n-1)S'^2}{\chi^2_{n-1,1-\alpha/2}}\right)=1-\alpha.$$

აქედან ადვილად მივიღებთ ფორმულას, რომლის მიხედვითაც განისაზღვრება სტანდარტული გადახრის ნდობის ინტერვალი (ამისათვის, თითოეულ უტოლობის სამივე მხარიდან ამოვიღოთ კვადარტული ფესვი).

ამოცანა. ჩავთვალოთ, რომ ხმაური, ერთი და იგივე ტიპის ვერტმფრენის კაბინაში, გარკვეულ რეჟიმში მომუშავე ძრავის დროს, შემთხვევითი სიდიდეა, რომელიც განაწილებულია ნორმალური კანონით. შემთხვევით შერჩეულ იქნა 20 ვერტმფრენი და მოხდა მათში ხმის დონის გაზომვა (დეციბალებში). გაზომვების შესწორებული შერჩევითი დისპერსია აღმოჩნდა 22.5-ის ტოლი. ვიპოვოთ ნდობის ინტერვალი, რომელიც ფარავს მოცემული ტიპის ვერტმფრენების კაბინაში ხმაურის სიდიდის უცნობ სტანდარტულ გადახრას 98%-იანი საიმედოობით.

ამოხსნა. თავისუფლების 19-ის ტოლი ხარისხითა და $(1-0.98)/2=0.01$ ალბათობის საშუალებით χ^2 –ის განაწილების ცხრილიდან ვპოულობთ სიდიდეს: $\chi_2^2=36.2$. ანალოგიურად, $1+0.98)/2=0.99$ ალბათობის საშუალებით ვპოულობთ: $\chi_1^2=7.63$. შესაბამისად, უცნობი საშუალოს შემთხვევაში სტანდარტული გადახრის ნდობის ინტერვალის

$$P\left(\sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1,\alpha/2}}}<\sqrt{D\xi}<\sqrt{\frac{(n-1)S'^2}{\chi^2_{n-1,1-\alpha/2}}}\right)=1-\alpha$$

ფორმულის გამოყენებით ვღებულობთ, რომ საძიებელი ნდობის ინტერვალია: (3.44, 7.49).

ნდობის ინტერვალი ბერნულის სქემაში.

ბერნულის სქემაში (დამუკიდებელ ცდათა სქემაში) უცნობი p ალბათობის წერტილოვანი შეფასებაა ფარდობითი სიხშირე:

$$w_n = S_n / n,$$

სადაც $S_n = \sum_{i=1}^{n} X_i$ ($X_i = 1$, თუ i-ურ ცდაში მოხდა წარმატება, და $X_i = 0$, თუ i-ურ ცდაში მოხდა მარცხი) – წარმატებათა რაოდენობაა n დამოუკიდებელ ცდაში, ამასთან

$$Ew_n = p \text{ და } Dw_n = p(1-p)/n.$$

უცნობი p ალბათობისათვის ნდობის ინტერვალის ასაგებად იყენებენ ფარდობითი სიხშირის სტანდარტიზაციის შედეგად მიღებულ სტატისტიკას:

$$Z_n = \frac{w_n - p}{\sqrt{p(1-p)/n}},$$

რომელიც, ცენტრალური ზღვარითი თეორემის თანახმად, დაახლოებით ნორმალურადაა განაწილებული ნულოვანი საშუალოთი და ერთეულოვანი დისპერსიით, თუ შერჩევის მოცულობა n საკმაოდ დიდია. მაგრამ, სამწუხაროდ, ამ სტატისტიკის გამოსახულების მნიშვნელშიც შედის შესაფასებელი p პარამეტრი, რაც საშუალებას არ იძლევა სტანდარტული გზით მივიღოთ ნდობის ინტერვალი.

არსებობს ასეთი გამოსავალი. შეიძლება გამოვიყენოთ გამარტივებული მიდგომა, რომლის თანახმადაც მნიშვნელში მდგომი უცნობი p ალბათობა უნდა შევცვალოთ მისი w_n შეფასებით და შესაბამისად, Z_n სტატისტიკის ნაცვლად გამოვიყენოთ შემდეგი სტატისტიკა:

$$\hat{Z}_n = \frac{w_n - p}{\sqrt{w_n \cdot (1-w_n)}} \cdot \sqrt{n}.$$

გასაგებია, რომ ამ სტატისტიკას ასიმპტოტურად ექნება იგივე ყოფაქცევა რაც Z_n სტატისტიკას. ამის შემდეგ ნდობის ინტერვალი იგება სტანდარტული გზით, რის შედეგადაც ვღებულობთ, რომ შერჩევის დიდი მოცულობის შემთხვევაში $(1-\alpha)$ ნდობის ალბათობის მქონე ასიმპტოტურ ნდობის ინტერვალს უცნობი p ალბათობისათვის აქვს შემდეგი სახე:

$$\left(w_n - z_{\alpha/2} \cdot \sqrt{\frac{w_n(1-w_n)}{n}},\ w_n + z_{\alpha/2} \cdot \sqrt{\frac{w_n(1-w_n)}{n}}\right), \qquad (1)$$

სადაც $z_{\alpha/2}$ – სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$ კრიტიკული წერტილია (სტანდარტული ნორმალური განაწილების ზედა α კრიტიკული წერტილი ეწოდება ისეთ z_α რიცხვს, რომლისთვისაც

$$P\{N(x;0;1) > z_\alpha\} = \alpha \text{ ანუ } \Phi(z_\alpha) = 1-\alpha).$$

მეორე მიდგომა ეყრდნობა აგრეთვე ნორმალურ აპროქსიმაციას: მოვძებნოთ ისეთი p რიცხვი, რომ სრულდებოდეს უტოლობა

$$P\{-z_{\alpha/2} \le \frac{w_n - p}{\sqrt{p(1-p)/n}} \le -z_{\alpha/2}\} \approx 1-\alpha.$$

რაც ტოლფასია იმისა, რომ ამოვხსნათ p ცვლადის მიმართ განტოლება

$$\frac{w_n - p}{\sqrt{p(1-p)/n}} = z_{\alpha/2}.$$

საბოლოოდ, ამ გზით მიიღებული დაზუსტებული $(1-\alpha)$ ნდობის ალბათობის მქონე ასიმპტოტური ნდობის ინტერვალი უცნობი p ალბათობისათვის იქნება:

$$\left(\frac{w_n + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{w_n(1-w_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}, \frac{w_n + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2}\sqrt{\frac{w_n(1-w_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}\right).$$

ცხადია, რომ თუ n იმდენად დიდია, რომ $\frac{z_{\alpha/2}^2}{n}$-ისა და $\frac{z_{\alpha/2}^2}{n^2}$-ის უგულებელყოფა (ნულთან გატოლება) შეიძლება, მაშინ უკანასკნელი ინტერვალი დაემთხვევა (1) ინტერვალს.

ნდობის ინტერვალი პუასონის მოდელის პარამეტრისათვის.

დავუშვათ, $X_1,...,X_n$ არის n მოცულობის შერჩევა $\Pi(\lambda)$ პოპულაციიდან, ე. ი.

$$P\{X_i = k\} = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \quad k = 0,1,2,..., \quad EX_i = DX_i = \lambda.$$

ცნობილია, რომ თუ n საკმარისად დიდია

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda / n}}$$

სტატისტიკა მიახლოებით ნორმალურადაა განაწილებილი პარამეტრებით 0 და 1. ამიტომ $(1-\alpha)$ ნდობის ალბათობის მქონე ასიმპტოტური ნდობის ინტერვალი პუასონის პოპულაციის უცნობი λ პარამეტრისათვის იქნება:

$$(\overline{X}_n - z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}_n}{n}}, \ \overline{X}_n + z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}_n}{n}}).$$

თავი VI
პარამეტრულ ჰიპოთეზათა შემოწმება

ჰიპოთეზების სტატისტიკური შემოწმება წარმოადგენს მათემატიკური სტატისტიკის უმნიშვნელოვანესს ნაწილს. მათემატიკური სტატისტიკის მეთოდები საშუალებას იძლევა შევამოწმოთ დაშვებები გარკვეული შემთხვევითი სიდიდის (გენერალური ერთობლიობის) განაწილების კანონის შესახებ, ამ კანონის პარამეტრების (მაგალითად, $E\xi$, $D\xi$) მნიშვნელობების შესახებ, ერთი და იგივე გენერალური ერთობლიობის ობიექტების სიმრავლეზე განმარტებულ შემთხვევით სიდიდეებს შორის კორელაციური კავშირის არსებობის შესახებ.

დავუშვათ, რომ გარკვეული მონაცემების მიხედვით, გვაქვს საფუძველი წამოვაყენოთ წინადადება განაწილების კანონის შესახებ ან შემთხვევითი სიდიდის (ან გენერალური ერთობლიობის, რომელთა ობიექტების სიმრავლეზე განმარტებულია მოცემული შემთხვევითი სიდიდე) განაწილების კანონის პარამეტრის შესახებ. ამოცანა მდგომარეობს იმაში, რომ დავადასტუროთ ან უარვყოთ ეს წინადადება შერჩევითი (ექსპერიმენტალური) მონაცემების გამოყენების საფუძველზე.

განაწილების პარამეტრების მნიშვნელობების შესახებ ან ორი განაწილების პარამეტრების სიდიდეების შედარების ჰიპოთეზებს, **პარამეტრული ჰიპოთეზები** ეწოდება. ჰიპოთეზებს განაწილების სახის შესახებ კი **არაპარამეტრული ჰიპოთეზები** ეწოდება.

სტატისტიკური ჰიპოთეზის შემოწმება ნიშნავს, რომ შევამოწმოთ შერჩევიდან მიღებული მონაცემები არის თუ არა შესაბამისობაში მოცემულ ჰიპოთეზასთან (მონაცემები ეთანხმება თუ არა მოცემულ ჰიპოთეზას). შემოწმება ხორციელდება **სტატისტიკური კრიტერიუმის** საშუალებით. *სტატისტიკური კრიტერიუმი – ეს არის შემთხვევითი სიდიდე, რომლის განაწილების კანონი (პარამეტრების მნიშვნელობებთან ერთად) ცნობილია იმ შემთხვევაში, თუ მიღებული ჰიპოთეზა სამართლიანია (ზოგჯერ სტატისტიკურ კრიტერიუმს უბრალოდ* ***სტატისტიკას*** *უწოდებენ).* ამ კრიტერიუმს უწოდებენ აგრეთვე **თანხმობის კრიტერიუმს** (მხედველობაში აქვთ რა მიღებული ჰიპოთეზის თანხმობა შერჩევიდან მიღებულ შედეგებთან).

ჰიპოთეზას, რომელიც წამოყენებულია შერჩევით მონაცემებთან მისი თანხმობის შესამოწმებლად, **ნულოვანი ჰიპოთეზა** ეწოდება და აღინიშნება H_0-ით. H_0 ჰიპოთეზასთან ერთად იხილავენ (წამოაყენებენ) **ალტერნატიულ** ანუ **საწინააღმდეგო** ჰიპოთეზასაც, რომელსაც H_1-ით აღნიშნავენ. მაგალითად:

1) H_0: $E\xi = 0$, H_1: $E\xi \neq 0$ 2) H_0: $E\xi = 0$, H_1: $E\xi > 0$ 3) H_0: $E\xi = 0$, H_1: $E\xi = 2$

დავუშვათ, რომ შემთხვევითი სიდიდე K – არის გარკვეული H_0 ჰიპოთეზის შემოწმების სტატისტიკური კრიტერიუმი. H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში K შემთხვევითი სიდიდის განაწილების კანონი

ხასიათდება გარკვეული, ჩვენთვის ცნობილი განაწილების სიმკვრივით $p_K(x)$. ამოვირჩიოთ გარკვეული მცირე ალბათობა α, რომელიც ტოლია 0.05-ის, 0.01-ის ან კიდევ უფრო მცირეა. განვმარტოთ **კრიტერიუმის კრიტიკული მნიშვნელობა** $K_{კრ}$ როგორც შემდეგი სამი განტოლებიდან ერთ-ერთის ამონახსნი, იმის მიხედვით თუ რა სახისაა ნულოვანი და ალტერნატიული ჰიპოთეზები:

$$P(K>K_{კრ}) = \alpha, \qquad (1)$$

$$P(K<K_{კრ}) = \alpha, \qquad (2)$$

$$P((K< K_{კრ1})\cap(K>K_{კრ2})) = \alpha. \qquad (3)$$

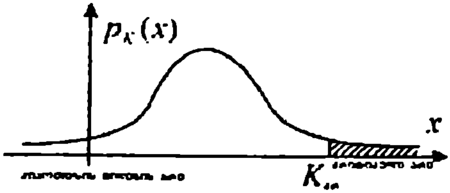
შესაძლებელია სხვა სახის განტოლებებიც, მაგრამ ყველაზე ხშირად გვხვდება სწორედ ასეთები.

(1) განტოლების ამოხსნა (ისევე როგორც (2) და (3) განტოლებების) მდგომარეობს შემდეგში: მოცემული α ალბათობით, ვიცით რა $p_K(x)$ ფუნქცია, რომელიც როგორც წესი მოცემულია ცხრილით, საჭიროა განისაზღვროს $K_{კრ}$.

რას ნიშნავს (1) პირობა?

თუ სამართლიანია H_0 ჰიპოთეზა, მაშინ ალბათობა იმისა, რომ K კრიტერიუმი გადააჭარბებს გარკვეულ $K_{კრ}$ მნიშვნელობას ძალიან მცირეა – 0.05, 0.01 ან კიდევ უფრო მცირე, იმის მიხედვით თუ ჩვენ რას ამოვირჩევთ. თუ $K_{ვ}$ – შერჩევითი მონაცემებით გამოთვლილი K კრიტერიუმის სიმძლავრე მეტია ვიდრე $K_{კრ}$, ეს იმას ნიშნავს, რომ შერჩევითი მონაცემები არ იძლევიან საფუძველს ნულოვანი H_0 ჰიპოთეზის მისაღებად (მაგალითად, თუ $\alpha = 0.01$, მაშინ შეიძლება ითქვას, რომ მოხდა ისეთი ხდომილება, რომელიც H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში საშუალოდ გვხვდება არა უმეტეს ვიდრე ერთჯერ 100 შერჩევიდან). ამ შემთხვევაში ამბობენ, რომ H_0 **ჰიპოთეზა არ ეთანხმება შერჩევით მონაცემებს და ის უნდა იქნეს უკუგდებული.** თუ $K_{ვ}$ არ აღემატება $K_{კრ}$-ს, მაშინ ამბობენ, **რომ შერჩევითი მონაცემები არ ეწინააღმდეგებიან H_0 ჰიპოთეზას**, და არა გვაქვს საფუძველი ამ ჰიპოთეზის უკუგდების.

(1) განტოლების შემთხვევში არეს – $K> K_{კრ}$ ეწოდება **კრიტიკული არე. თუ $K_{ვ}$-ს მნიშვნელობა მოხვდება კრიტიკულ არეში, მაშინ H_0 ჰიპოთეზა უკუგდებულ იქნება.** ასეთ კრიტიკულ არეს მარჯვენა კრიტიკული არე ეწოდება. ქვემოთ მოყვანილია (1) განტოლების საილუსტრაციო ნახაზი:

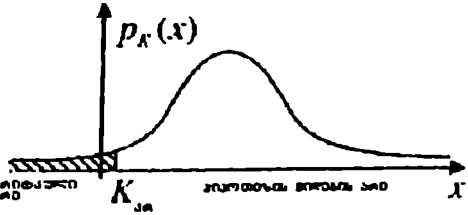


აქ $p_K(x)$ – K შემთხვევითი სიდიდის ცნობილი განაწილების სიმკვრივეა H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში. შევნიშნავთ, რომ დაშტრიხული ფიგურის ფართობი აქ α-ს ტოლია.

დავუშვათ, რომ შერჩეულია გარკვეული მცირე მნიშვნელობა α ალბათობის, ამ მნიშვნელობის მიხედვით განსაზღრულია $K_{კრ}$ და შერჩევითი მონაცემების მიხედვით განსაზღრულია $K_{დ}$-ს მნიშვნელობა, რომელიც მოხვდა კრიტიკულ არეში. ამ შემთხვევაში H_0 ჰიპოთეზა უკუგდებულ იქნება, მაგრამ ის შეიძლება აღმოჩნდეს სამართლიანი. უბრალოდ, შემთხვევით მოდხა ხდომილება, რომელსაც გააჩნია ძალიან მცირე ალბათობა α. ამ აზრით α არის ალბათობა იმისა, რომ უკუგდებულ იქნება სამართლიანი H_0 ჰიპოთეზა.

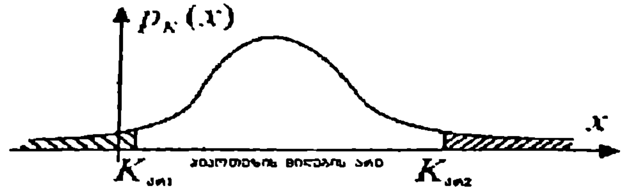
სამართლიანი ჰიპოთეზის უკუგდებას **პირველი გვარის შეცდომა** ეწოდება. α ალბათობას **მნიშვნელოვნების დონე** ეწოდება. ამრიგად, *მნიშვნელოვნების დონე – ეს არის პირველი გვარის შეცდომის დაშვების ალბათობა.*

(2) განტოლება განსაზღვრავს **მარცხენა კრიტიკულ არეს**. მის გამოსახულებას აქვს შემდეგი სახე:



კრიტიკული არის (დაშტრიხული ფიგურის ფართობი) აქაც α-ს ტოლია.

და ბოლოს, (3) განტოლება განსაზღვრავს **ორმხრივ კრიტიკულ არეს**. ასეთი არე გამოსახულია ქვემოთ მოყვანილ ნახაზზე:



აქ კრიტიკული არე შედგება ორი ნაწილისაგან. მისი საზღვრები განისაზღვრება ისე, რომ სრულდებოდეს პირობა:

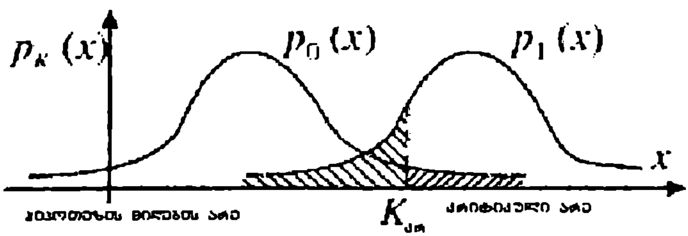
$$P(K \le K_{კრ1}) = P(K \ge K_{კრ2} = \alpha / 2 .$$

ამ შემთხვევაში თითოეული დაშტრიხული ფიგურის ფართობი ტოლია $\alpha / 2$-ის.

კრიტიკული არის სახე დამოკიდებულია იმაზე, თუ როგორია ალტერნატიული ჰიპოთეზაა.

რაც უფრო პატარაა მნიშვნელოვნების დონე, მით უფრო მცირეა ალბათობა იმისა, რომ უკუვაგდოთ შესამოწმებელი H_0 ჰიპოთეზა, როცა ის სამართლიანია, ანუ დავუშვათ პირველი გვარის შეცდომა. მაგრამ, მნიშვნელოვნების დონის შემცირებასთან ერთად ფართოვდება H_0 ჰიპოთეზის მიღების არე და შესაბამისად, იზრდება ალბათობა იმისა, რომ მივიღოთ შესამოწმებელი ჰიპოთეზა, როცა ის არაა სამართლიანი, ანუ მაშინ როცა უპირატესობა უნდა მინიჭოს ალტერნატიულ ჰიპოთეზას.

დავუშვათ რომ H_0 ჰიპოთეზის სამართლიანობისას K სტატისტიკურ კრიტერიუმს გააჩნია სიმკვრივე $p_0(x)$, ხოლო ალტერნატიული H_1 ჰიპოთეზის სამართლიანობისას კი – $p_1(x)$ განაწილების სიმკვრივე. ამ ფუნქციების გრაფიკები გამოსახულია ნახაზზე:



მნიშვნელოვნების გარკვეული დონისათვის ვპოულობთ კრიტიკულ მნიშვნელობას $K_{კრ}$ და მარჯვენა კრიტიკული არეს. თუ შერჩევითი მონაცემების საშუალებით განსაზღვრული $K_{ვ}$-ს მნიშვნელობა აღმოჩნდება უფრო ნაკლები, ვიდრე $K_{კრ}$, მაშინ H_0 ჰიპოთეზა მიიღება. დავუშვათ, რომ სინამდვილეში სამართლიანია H_1 ჰიპოთეზა. მაშინ ალბათობა იმისა, რომ კრიტერიუმი მოხვდება H_0 ჰიპოთეზის მიღების არეში, არის გარკვეული რიცხვი β, რომელიც ტოლია იმ ფიგურის ფართობის, რომელიც შემოსაზღრულია $p_1(x)$ ფუნქციის გრაფიკითა და ჰორიზონტალური საკოორდინატო ღერძის ნახევრადუსასრულო ნაწილით, რომელიც ძევს $K_{კრ}$ წერტილის მარცხნივ. ცხადია, რომ β – არის ალბათობა იმისა, რომ მიღებული იქნება არასამართლიანი H_0 ჰიპოთეზა.

არასამართლიანი ჰიპოთეზის მიღებას **მეორე გვარის შეცდომა** ეწოდება. განსახილველ შემთხვევაში რიცხვი β არის მეორე გვარის შეცდომის ალბათობა. **რიცხვს** $1-\beta$, რომელიც ტოლია ალბათობის იმისა, რომ არ იქნება დაშვებული მეორე გვარის შეცდომა, **კრიტერიუმის სიმძლავრე** ეწოდება. ზემოთ მოყვანილ ნახაზზე, კრიტერიუმის სიმძლავრე ტოლია იმ ფიგურის ფართობის, რომელიც რომელიც შემოსაზღრულია $p_1(x)$ ფუნქციის გრაფიკითა და ჰორიზონტალური საკოორდინატო ღერძის ნახევრადუსასრულო ნაწილით, რომელიც ძევს $K_{კრ}$ წერტილის მარჯვნივ.

სტატისტიკური კრიტერიუმისა და კრიტიკული არის სახის შერჩევა ხდება ისე, რომ კრიტერიუმის სიმძლავრე იყოს მაქსიმალური.

ჰიპოთეზის შემოწმება ლოდინის შესახებ.

სტატისტიკური ჰიპოთეზის შემოწმება ნორმალური განაწილების მათემატიკური ლოდინის შესახებ ცნობილი დისპერსიის შემთხვევაში:

დავუშვათ, რომ მოცემულია ნორმალური კანონით განაწილებული შემთხვევითი სიდიდე ξ, რომელიც განმარტებულია გარკვეული გენერალური ერთობლიობის ობიექტების სიმრავლეზე. ცნობილია, რომ $D\xi=\sigma^2$. მათემატიკური ლოდინი $E\xi$ უცნობია. დავუშვათ, რომ ჩვენ გაგვაჩნია საფუძველი იმისა, რომ დავუშვათ: $E\xi=a$, სადაც a – გარკვეული რიცხვია (ასეთი საფუძველი შეიძლება იყოს ინფორმაცია გენერალური ერთობლიობის ობიექტების შესახებ, მსგავსი ერთობლიობების კვლევის გამოცდილება და სხვა). ვიგულისხმოთ, რომ გვაქვს აგრეთვე მეორე ინფორმაცია, რომელიც გვეიცვენებს, რომ $E\xi=a_1$, სადაც $a_1>a$.

I. ვაყენებთ ნულოვან ჰიპოთეზას – $H_0: E\xi=a$ ალტერნატიული ჰიპოთეზის წინააღმდეგ – $H_1: E\xi=a_1$.

ვაკეთებთ n მოცულობის შერჩევას $x_1,x_2,\ldots,x_n$. შემოწმებას საფუძვლად უდევს ის ფაქტი, რომ შემთხვევითი სიდიდე $\bar{x}$ (შერჩევითი საშუალო) განაწილებულია ნორმალური განაწილების კანონით σ^2/n-ის ტოლ

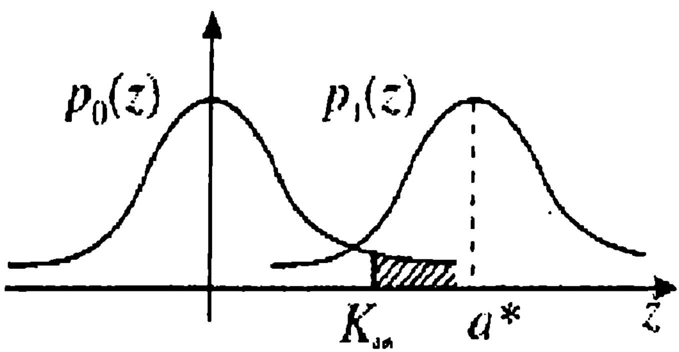
დისპერსიითა და a-ს (შესაბამისად, a_1-ის) ტოლი მათემატიკური ლოდინით H_0 (შესაბამისად, H_1) ჰიპოთეზის სამართლიანობის შემთხვევაში.

ცხადია, რომ თუ სიდიდე $\bar{x}$ აღმოჩნდება საკმარისად მცირე, მაშინ ეს გვაძლევს საფუძველს H_0 ჰიპოთეზას მივანიჭოთ უპირატესობა H_1 ჰიპოთეზასთან შედარებით. მეორეს მხრივ, $\bar{x}$-ის საკმარისად დიდი მნიშვნელობის შემთხვევაში უფრო ალბათურია H_1 ჰიპოთეზის სამართლიანობა. ამოცანა შეიძლება ასე დაისვას: საჭიროა მოიძებნოს გარკვეული კრიტიკული რიცხვი, რომელიც შერჩევითი საშუალოს ყველა შესაძლო მნიშვნელობებს (ამ ამოცანის შემთხვევაში, ეს მთლიანად ნამდვილ რიცხვთა სიმრავლეა) გაყოფს ნახევრადუსასრულო შუალედად. $\bar{x}$ შერჩევითი საშუალოს მარცხენა ინტერვალში მოხვედრისას უნდა მივიღოთ H_0 ჰიპოთეზა, ხოლო $\bar{x}$-ის მარჯვენა ინტერვალში მოხვედრისას უპირატესობა უნდა მიენიჭოს H_1 ჰიპოთეზას. თუმცა სინამდვილეში იქცევიან რამდენადმე სხვანაირად.

სტატისტიკური კრიტერიუმის როლში ირჩევენ შემთხვევით სიდიდეს

$$Z=\left(\overline{X}-a\right)\sqrt{n}/\sigma,$$

რომელიც განაწილებულია ნორმალური განაწილების კანონით, პარამეტრებით: $EZ=0$ და $DZ=1$ H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში. თუ კი სამართლიანია H_1 ჰიპოთეზა, მაშინ $EZ=a^*=(a_1-a)\sqrt{n}/\sigma$ და $DZ=$ $=1$. ქვემოთ მოყვანილია $p_0(z)$ და $p_1(z)$ ფუნქციების გრაფიკები, რომლებიც წარმოადგენენ Z შემთხვევითი სიდიდის განაწილების სიმკვრივის ფუნქციებს შესაბამისად H_0 და H_1 ჰიპოთეზების სამართლიანობისას.



თუ შერჩევითი მონაცემებიდან მიღებული $\bar{x}$-ის მნიშვნელობა შედარებით დიდია, მაშინ z სიდიდეც დიდი იქნება, რაც წარმოადგენს მტკიცებულებას H_1 ჰიპოთეზის სასარგებლოდ. $\bar{x}$-ის შედარებით მცირე მნიშვნ-

ლობებს მივყავართ z-ის მცირე მნიშვნელობებამდე, რაც მეტყველებს H_0 ჰიპოთეზის სასარგებლოდ. აქედან გამომდინარეობს, რომ უნდა შეირჩეს მარჯვენა კრიტიკული არე. არჩეული მნიშვნელოვნების α დონისათვის (მაგალითად, $\alpha = 0.05$), ვისარგებლებთ რა იმ გარემოებით, რომ შემთხვევითი სიდიდე z განაწილებულია ნორმალური განაწილების კანონით, განვსაზღვრავთ $K_{კრ}$-ს შემდეგი თანაფარდობიდან:

$$\alpha = P(K_{კრ} < z < \infty) = \phi(\infty) - \phi(K_{კრ}) = 1 - \phi(K_{კრ}).$$

აქედან $\phi(K_{კრ}) = 1 - \alpha$, და $K_{კრ}$-ს საპოვნელად საჭიროა ვისარგებლოთ სტანდარტული ნორმალური განაწილების ცხრილით. ცხადია, რომ

$$K_{კრ} = x_{1-\alpha} = z_{\alpha}.$$

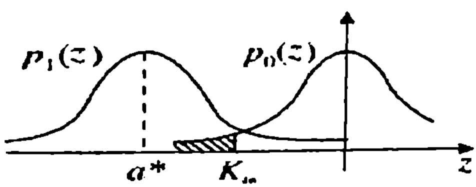
თუ z-ის მნიშვნელობა, გამოთვლილი $\bar{x}$ შერჩევითი საშუალოს მიხედვით, მოხვდება ჰიპოთეზის მიღების არეში ($z < K_{კრ}$), მაშინ H_0 ჰიპოთეზა მიიღება (კეთდება დასკვნა, რომ შერჩევითი მონაცემები არ ეწინააღმდეგება H_0 ჰიპოთეზას). თუ z სიდიდე ხვდება კრიტიკულ არეში, მაშინ H_0 ჰიპოთეზას უკუაგდებენ.

გამოვთვალოთ ამ ამოცანაში კრიტერიუმის სიმძლავრე. გვაქვს:

$$1 - \beta = \phi(\infty) - \phi[K_{კრ} - (a_1 - a)\sqrt{n}/\sigma].$$

აქედან ჩანს, რომ კრიტერიუმის სიმძლავრე მით უფრო დიდია, რაც უფრო დიდია სხვაობა $a_1 - a$.

II. თუ წინა ამოცანაში დავსვამთ სხვა პირობას, კერძოდ, $a_1 < a$, ანუ განვიხილავთ ნულოვან ჰიპოთეზას – $H_0 : E\xi = a$ ალტერნატიული ჰიპოთეზის წინააღმდეგ – $H_1 : E\xi = a_1$, $a_1 < a$, მაშინ ზემოთ მოყვანილი მსჯელობის ანალოგიით გასაგებია, რომ უნდა განვიხილოთ მარცხენა კრიტიკული არე. ნახაზი იქნება შემდეგი სახის:



აქ, ისევე როგორც წინა შემთხვევაში, $a^* = (a_1 - a)\sqrt{n}/\sigma$, ხოლო კრიტიკული რიცხვი $K_{კრ}$ განისაზღვრება შემდეგი თანაფარდობიდან:

$$\alpha = P(-\infty < z < K_{კრ}) = \phi(K_{კრ}) - \phi(-\infty) = \phi(K_{კრ}).$$

თუ ვისარგებლებთ თანაფარდობით $1 - \phi(K_{კრ}) = \phi(-K_{კრ})$, მივიღებთ:

$$\phi(-K_{კრ}) = 1 - \alpha,$$

ე. ი. $-K_{\text{კრ}} = x_{1-\alpha} = z_\alpha$, ანუ $K_{\text{კრ}} = -x_{1-\alpha} = -z_\alpha$ (შევნიშნავთ, რომ ამოცანის შინაარსიდან გამომდინარე, აქ $K_{\text{კრ}}$ უარყოფითი რიცხვია).

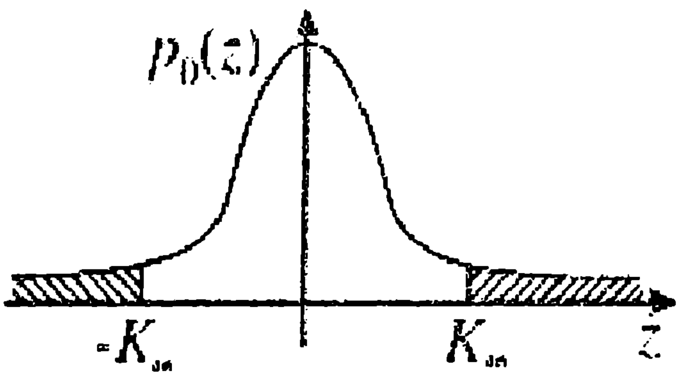
შერჩევითი მონაცემებით გამოთვლილი z-ის ის მნიშვნელობები, რომლებიც მეტია $K_{\text{კრ}}$-ზე ეთანხმება H_0 ჰიპოთეზას. თუ z სიდიდე ხვდება კრიტიკულ არეში ($z < K_{\text{კრ}}$), მაშინ უპირატესობა ენიჭება H_1 ჰიპოთეზას და უკუაგდებენ H_0 ჰიპოთეზას.

III. განვიხილოთ ახლა ასეთი ამოცანა:

$$H_0 : E\xi = a;$$

$$H_1 : E\xi \neq a.$$

ამ შემთხვევაში z სიდიდის დიდი გადახრები ნულისაგან როგორც დადებით, ისე უარყოფით მხარეს, მეტყველებს H_0 ჰიპოთეზის საწინააღმდეგოდ, ანუ აქ საჭიროა განხილულ იქნეს ორმხრივი კრიტიკული არე, ისე როგორც ეს ნაჩვენებია ნახაზზე:



ამ შემთხვევაში კრიტიკული მნიშვნელობა $K_{\text{კრ}}$ განისაზღვრება თანაფარდობიდან:

$$P(-K_{\text{კრ}} < z < K_{\text{კრ}}) = 1-\alpha = \phi(K_{\text{კრ}}) - \phi(-K_{\text{კრ}}) = 2\phi(K_{\text{კრ}}) - 1,$$

საიდანაც ვღებულობთ, რომ:

$$\phi(K_{\text{კრ}}) = 1-\alpha/2,$$

და შესაბამისად,

$$K_{\text{კრ}} = x_{1-\alpha/2} = z_{\alpha/2}.$$

***p*-მნიშვნელობის მეთოდი.**

ჰიპოთეზის შემოწმების მეორე ეკვივალენტური მეთოდი მდგომარეობს შემდეგში: $x_1,\ldots,x_n$ შერჩევის საფუძველზე გამოვთვალოთ კრიტერიუმ-

ის $Z=(\overline{X}-a)\sqrt{n}/\sigma$ სტატისტიკის $z=(\bar{x}-a)\sqrt{n}/\sigma$ მნიშვნელობა და შემდეგი ალბათობა:

$$p=P\{Z\ge z\,|\,H_0\}=P\{\frac{\overline{X}-a}{\sigma/\sqrt{n}}\ge\frac{\bar{x}-a}{\sigma/\sqrt{n}}\,|\,H_0\}=1-\Phi(\frac{\bar{x}-a}{\sigma/\sqrt{n}}).$$

ცხადია, რომ თუ $p\le\alpha$, მაშინ $z\ge z_\alpha$, რაც ჰიპოთეზის გარჩევის მარჯვენა ცალმხრივი კრიტერიუმის შესაბამისად იწვევს H_0 ჰიპოთეზის უარყოფას. აქედან შეიძლება დავასკვნათ, რომ: $x_1,\dots,x_n$ შერჩევის საფუძველზე ნებისმიერი α, $\alpha\ge p$ მნიშვნელოვნების დონით ხდება H_0 ჰიპოთეზის უარყოფა. შესაბამისად, p-მნიშვნელობა არის ის მინიმალური მნიშვნელოვნების დონე, რომლითაც $x_1,\dots,x_n$ მონაცემებით H_0 ჰიპოთეზას უარვყოფთ. ამრიგად, ჰიპოთეზათა გარჩევის p-მნიშვნელობის მეთოდი ასე ყალიბდება:

მოცემული α მნიშვნელოვნების დონისათვის, $x_1,\dots,x_n$ მონაცემებით H_0 ჰიპოთეზას უარვყოფთ, თუ p-მნიშვნელობა ნაკლებია ან ტოლი α-ზე ($p\le\alpha$).

სტატისტიკური ჰიპოთეზის შემოწმება ნორმალური განაწილების მათემატიკური ლოდინის შესახებ უცნობი დისპერსიის შემთხვევაში:

ამ შემთხვევაში სტატისტიკური კრიტერიუმის (სტატისტიკის) როლში იღებენ შემდეგ შემთხვევით სიდიდეს:

$$T=\frac{(\overline{X}-a_0)\sqrt{n}}{S'},$$

სადაც S' – შესწორებული საშუალო კვადრატული გადახრაა. ცნობილია, რომ ამ შემთხვევით სიდიდეს გააჩნია სტიუდენტის განაწილება თავისუფლების ხარისხით $k=n-1$. განვიხილოთ იგივე ალტერნატიული ჰიპოთეზები და შესაბამისი კრიტიკული არეები, რაც გვქონდა ცნობილი დისპერსიის შემთხვევაში. წინასწარ გამოვთვალოთ კრიტერიუმის დაკვირვებული (შერჩევითი) მნიშვნელობა

$$T_{\text{ვ}}=\frac{(\bar{x}_{\text{ვ}}-a_0)\sqrt{n}}{s'}.$$

დავუშვათ, რომ გვაქვს ნულოვანი ჰიპოთეზა – $H_0: E\xi=a_0$ ალტერნატიული ჰიპოთეზის წინააღმდეგ – $H_1: E\xi\ne a_0$. მოცემული α-სა და $k=n-1$-სათვის სტიუდენტის განაწილების ცხრილიდან ვპოულობთ სტიუდენტის ორმხრივ კრიტიკულ წერტილს $t_{\text{კრ}}=t_{n-1,\alpha/2}$. თუ აღმოჩნდა, რომ $|T_{\text{ვ}}|<t_{\text{კრ}}$, მაშინ ნულოვანი ჰიპოთეზა მიიღება. თუ $|T_{\text{ვ}}|>t_{\text{კრ}}$, მაშინ ნულოვანი ჰიპოთეზა უკუგდებულ იქნება.

თუ იგივე ნულოვანი ჰიპოთეზის წინააღმდეგ განვიხილავთ ალტერნატიულ $H_1: E\xi>a_0$ ჰიპოთეზას, მაშინ შესაბამისი ცხრილიდან ვპოულობთ მარჯვენა კრიტიკული არის კრიტიკულ წერტილს $t_{\text{კრ}}(\alpha,k)=t_{n-1,\alpha}$, და მივი-

ღებთ ნულოვან ჰიპოთეზას, თუ $T_g < t_{n}(\alpha,k)$ (წინააღმდეგ შემთხვევაში მიიღება ალტერნატიული ჰიპოთეზა).

ბოლოს, თუ ალტერნატიული ჰიპოთეზაა $H_1: E\xi < a_0$, გვექნება მარცხენა კრიტიკული არე და ნულოვანი ჰიპოთეზა მიიღება იმ შემთხვევაში, როცა $T_g > -t_{n-1,\alpha}$. თუ კი $T_g < -t_{n-1,\alpha}$, მაშინ ნულოვან ჰიპოთეზას უარყოფენ.

ჰიპოთეზის შემოწმება ნორმალური პოპულაციის დისპერსიისათვის.

ნორმალური გენერალური ერთობლიობიდან მიღებული შერჩევისათვის შევამოწმოთ შემდეგი სახის ჰიპოთეზები:

ა). $H_0: \sigma^2 = \sigma_0^2$, ბ). $H_0: \sigma^2 = \sigma_0^2$, გ). $H_0: \sigma^2 = \sigma_0^2$,

$H_1: \sigma^2 > \sigma_0^2$, $H_1: \sigma^2 < \sigma_0^2$, $H_1: \sigma^2 \neq \sigma_0^2$,

სადაც ყველგან σ_0^2 წარმოადგენს რაიმე ცნობილ რიცხვს. შესაბამისად, გვაქვს მარჯვენა ცალმხრივი ალტერნატივა, მარცხენა ცალმხრივი ალტერნატივა და ორმხრივი ალტერნატივა. ამასთანავე, იმ შემთხვევაში, როცა პოპულაციის საშუალო μ ცნობილია, ძირითადი ჰიპოთეზა მარტივია, ხოლო ალტერნატივა რთულია. როცა μ უცნობია, როგორც ნულოვანი, ისე ალტერნატიული ჰიპოთეზები რთულია.

ვინაიდან, უცნობი დისპერსიის შეფასებას წარმოადგენს შერჩევითი დისპერსია, ამიტომ ბუნებრივია ცნობილი μ-ს შემთხვევაში კრიტერიუმის სტატისტიკად ავიღოთ ეს უკანასკნელი:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \qquad (1)$$

ხოლო უცნობი μ-ს შემთხვევაში კი შესწორებული შერჩევითი დისპერსია:

$$S_n'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (2)$$

როგორც ცნობილია, ნულოვანი ჰიპოთეზის სამართლიანობის შემთხვევაში

$$\chi_n^2 \coloneqq \frac{nS_n^2}{\sigma^2} \sim \chi^2(n)$$

და

$$\chi_{n-1}^2 \coloneqq \frac{(n-1)S_n'^2}{\sigma^2} \sim \chi^2(n-1).$$

ვინაიდან (1) და (2) დისპერსიის ჩაუნაცვლებელი შეფასებებია, ამიტომ შერჩევითი (შესაბამისად, შესწორებული შერჩევითი) დისპერსიის მნიშვნელოვანი გადახრა σ_0^2-საგან მეტყველებს ალტერნატიული ჰიპოთეზის სასარგებლოდ. ამიტომ, ფიქსირებული α მნიშვნელოოვნების დონისათვის კრიტიკული არეები შემდეგნაირად აიგება:

ა). ცალმხრივი მარჯვენა ალტერნატივა. კრიტიკულია არეა χ^2 განაწილების α ზომის ზედა არე, ანუ

$U_1 = [\chi_{n,\alpha}^2, +\infty)$, როცა μ ცნობილია,

$U_1 = [\chi_{n-1,\alpha}^2, +\infty)$, როცა μ უცნობია,

სადაც $\chi^2_{n,\alpha}$ არის $\chi^2(n)$ განაწილების ზედა α-კრიტიკული წერტილი.

ბ). ცალმხრივი მარცხენა ალტერნატივა. კრიტიკულია არეა χ^2 განაწილების α ზომის ქვედა არე, ანუ

$$U_1 = (0, \chi^2_{n,1-\alpha}], \text{ როცა } \mu \text{ ცნობილია,}$$

$$U_1 = (0, \chi^2_{n-1,1-\alpha}], \text{ როცა } \mu \text{ უცნობია.}$$

გ). ორმხრივი ალტერნატივა. კრიტიკული არე შედგება ორი ნაწილისაგან: χ^2 განაწილების $\alpha/2$ ზომის ზედა და χ^2 განაწილების $\alpha/2$ ზომის ქვედა არეებისაგან:

$$U_1 = (0, \chi^2_{n,1-\alpha/2}] \cup [\chi^2_{n,\alpha/2}, +\infty), \text{ როცა } \mu \text{ ცნობილია,}$$

$$U_1 = (0, \chi^2_{n-1,1-\alpha/2}] \cup [\chi^2_{n-1,\alpha/2}, +\infty), \text{ როცა } \mu \text{ უცნობია.}$$

შერჩევის მოცემული $x_1, \ldots, x_n$ მნიშვნელობებით ვითვლით შერჩევით დისპერსიასა s^2 (შესაბამისად, შესწორებულ შერჩევით დისპერსიას s'^2) და იმ შემთხვევაში, როცა μ ცნობილია, თუ $\frac{ns^2}{\sigma_0^2} \in U_1$ უარვყოფთ H_0 ჰიპოთეზას (შესაბამისად, როცა μ უცნობია, თუ $\frac{(n-1)s'^2}{\sigma_0^2} \in U_1$ უარვყოფთ H_0 ჰიპოთეზას), წინააღმდეგ შემთხვევაში H_0 ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს.

ჰიპოთეზათა შემოწმება ბერნულის სქემაში.

დავუშვათ, რომ ჩატარებულია n დამოუკიდებელი ექსპერიმენტი (n – საკმაოდ დიდი რიცხვია), რომელთაგან თითოეულში გარკვეული A ხდომილება ხდება ერთი და იგივე, მაგრამ უცნობი p ალბათობით; ნაპოვნია ექსპერიმენტების ამ სერიაში A ხდომილების მოხდენის ფარდობითი სიხშირე $\frac{m}{n}$. მნიშვნელოვნების მოცემული α დონისათვის შევამოწმოთ H_0 ჰიპოთეზა, რომელიც მდგომარეობს იმაში, რომ p ალბათობა ტოლია გარკვეული p_0 რიცხვის.

სტატისტიკური კრიტერიუმის როლში ავიღოთ შემთხვევითი სიდიდე

$$U = \frac{\left(\frac{M}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}},$$

რომელსაც შერჩევის საკმაოდ დიდი მოცულობისათვის (კერძოდ, როცა $n \geq 30$) გააჩნია მიახლოებით ნორმალური განაწილება პარამეტრებით $E(U) = 0$, $\sigma(U) = = 1$. აქ $q_0 = 1 - p_0$. დასკვნა კრიტერიუმის ნორმალურად განაწილებულობის შესახებ გამოდის ლაპლასის თეორემიდან (საკმაოდ დიდი n–ებისათვის ფარდობითი სიხშირე დაახლოებით შეიძლება ჩაითვალოს ნორმალურად განაწილებულად მათემატიკური ლოდინით p და საშუალო

კვადრატული გადახრით $\sqrt{\frac{pq}{n}}$). კრიტიკული არე იგება ალტერნატიული ჰიპოთეზის სახის მიხედვით.

1). თუ $H_0: p = p_0$, ხოლო $H_1: p \neq p_0$, მაშინ კრიტიკული არე უნდა ავაგოთ ისე, რომ კრიტერიუმის ამ არეში მოხვედრის ალბათობა ტოლი იყოს მოცემული α მნიშვნელოვნების დონის. ამასთანავე კრიტერიუმის უდიდესი სიმძლავრე მიიღწევა მაშინ, როცა კრიტიკული არე შედგება ორი ინტერვალისაგან, რომელთაგან თითოეულში მოხვედრის ალბათობაა $\frac{\alpha}{2}$. ვინაიდან U სიმეტრიულია ორდინატთა ღერძის მიმართ, ამიტომ მისი $(-\infty; 0)$ და $(0; +\infty)$ ინტერვალებში მოხვედრის ალბათობებია 0.5. შესაბამისად, კრიტიკული არე აგრეთვე უნდა იყოს სიმეტრიული ორდინატთა ღერძის მიმართ. ამიტომ, $u_{კრ}$ განისაზღვრება ნორმალური განაწილების ცხრილიდან, ისე რომ შესრულდეს პირობა $\Phi(u_{კრ}) = 1 - \frac{\alpha}{2}$, ხოლო კრიტიკულ არეს აქვს სახე: $(-\infty; -u_{კრ}) \cup (u_{კრ}; +\infty)$ (სადაც $u_{კრ} = z_{\alpha/2}$).

შემდგომ უნდა გამოვთვალოთ კრიტერიუმის დაკვირვებული (შერჩევითი) მნიშვნელობა

$$U_{ვ} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}$$

თუ აღმოჩნდა, რომ $|U_{ვ}| \leq u_{კრ}$, მაშინ მიიღება ნულოვანი ჰიპოთეზა, ხოლო თუ $|U_{ვ}| > u_{კრ}$, მაშინ ნულოვან ჰიპოთეზას უკუვაგდებთ.

2). თუ ალტერნატიული ჰიპოთეზა $H_1: p > p_0$ სახისაა, მაშინ კრიტიკული არე განისაზღვრება უტოლობით $U > u_{კრ}$, ე. ი. გვაქვს მარჯვენა კრიტიკული არე, ამასთან $P(U > u_{კრ}) = \alpha$. შესაბამისად,

$$P(0 < U < u_{კრ}) = 1 - \alpha .$$

ამიტომ ნორმალური განაწილების ცხრილიდან ვიპოვით $u_{კრ}$-ს ისე, რომ

$$\Phi(u_{კრ}) = 1 - \alpha ,$$

ე. ი. $u_{კრ} = z_{\alpha}$.

შემდეგ ვითვლით კრიტერიუმის შერჩევით მნიშვნელობას

$$U_{ვ} = \frac{\left(\frac{m}{n} - p_0\right)\sqrt{n}}{\sqrt{p_0 q_0}}.$$

და, თუ აღმოჩნდა, რომ $U_{ვ} \leq u_{კრ}$, მაშინ მიიღება ნულოვანი ჰიპოთეზა. თუ კი $U_{ვ} > u_{კრ}$, მაშინ მიიღება ალტერნატიული ჰიპოთეზა.

3). ალტერნატიული ჰიპოთეზისათვის $H_1: p < p_0$, კრიტიკული არა მარცხენა ცალმხრივია და მოიცემა უტოლობით $U < -u_{კრ}$, სადაც $u_{კრ}$ გამოითვლება ისე, როგორც წინა შემთხვევაში.

თუ $U_{ვ} \geq -z_{\alpha}$, მაშინ მიიღება ნულოვანი ჰიპოთეზა.

თუ $U_{ვ} < -z_{\alpha}$, მაშინ მიიღება ალტერნატიული ჰიპოთეზა.

მაგალითი. დავუშვათ ჩატარებულია 50 დამოუკიდებელი ექსპერიმენტი, A ხდომილების მოხდენის ფარდობითი სიხშირე აღმოჩნდა 0,12. მნიშვნელოვნების $\alpha = 0.01$ დონისათვის შევამოწმოთ ნულოვანი $H_0: p = 0.1$ ჰიპოთეზა ალტერნატიული $H_1: p > 0.1$ ჰიპოთეზის შემთხვევაში.

ამოხსნა. ვიპოვოთ კრიტერიუმის შერჩევითი მნიშვნელობა

$$U_{ვ} = \frac{(0.12 - 0.1)\sqrt{50}}{\sqrt{0.1 \cdot 0.9}} = 0.471.$$

კრიტიკული არე იქნება მარჯვენა ცალმხრივი, ხოლო კრიტიკული წერტილი უნდა ვიპოვოთ პირობიდან

$$\Phi(u_{კრ}) = 1 - 0.01 = 0.99.$$

ნორმალური განაწილების ფუნქციის ცხრილიდან ვპოულობთ, რომ $u_{კრ} = 2.33$. ვინაიდან, $U_{ვ} < u_{კრ}$, ამიტომ მიიღება ჰიპოთეზა იმის შესახებ, რომ $p = 0.1$.

შენიშვნა. მცირე მოცულობის შერჩევისათვის (კერძოდ, თუ $n < 30$) ჰიპოთეზათა შემოწმების პროცედურა ეყრდნობა უშუალოდ კრიტერიუმის სტატისტიკის ზუსტ განაწილებას. ამ შემთხვევაში კრიტერიუმის სტატისტიკად გამოიყენება S_n (წარმატებათა რაოდენობა n დამოუკიდებელ ცდაში) სტატისტიკა, რომელსაც ნულოვანი ჰიპოთეზის სამართლიანობისას აქვს ბინომიალური განაწილება n და p_0 პარამეტრებით.

მაგალითად, ცალმხრივი მარჯვენა ალტერნატივის დროს α მნიშვნელოვნების დონის კრიტიკულ არეს აქვს სახე: $S_n > c_{\alpha}$, სადაც c_{α} მთელი დადებითი რიცხვი უნდა შეირჩეს პირობიდან, რომ:

$$P\{S_n > c_{\alpha} \mid H_0\} = \sum_{k=c_{\alpha}}^{n} C_n^k p_0^k (1 - p_0)^{n-k} = \alpha.$$

თუ ასეთი c_{α} არ არსებობს, მაშინ $c = c_{\alpha}$-ს არჩევენ პირობიდან

$$P\{S_n > c \mid H_0\} \leq \alpha < P\{S_n > c - 1 \mid H_0\},$$

ან რაც იგივეა:

$$\sum_{k=c}^{n} C_n^k p_0^k (1 - p_0)^{n-k} \leq \alpha < \sum_{k=c-1}^{n} C_n^k p_0^k (1 - p_0)^{n-k}.$$

ჰიპოთეზის შემოწმება პუასონის პოპულაციის μ პარამეტრის შესახებ (მცირე მოცულობის შერჩევებისათვის).

პუასონის პოპულაციის μ პარამეტრის შესახებ ჰიპოთეზის შემოწმებისას მცირე მოცულობის შერჩევის შემთხვევაში, იყენებენ p-მნიშვნელობის მეთოდს.

დაეუშვათ, რომ ვამოწმებთ ნულოვან ჰიპოთეზას $H_0 : \mu = \mu_0$ ალტერნატივის $H_1 : \mu \neq \mu_0$ წინააღმდეგ.

აქაც, თუ $x \le \mu_0$, სადაც x არის პუასონის შემთხვევითი სიდიდის დაკვირვებული მნიშვნელობა, მაშინ

$$p/2 = P\{X \le x \mid \mu = \mu_0\} = \sum_{k=0}^{x} \frac{\mu_0^k}{k!} \cdot e^{-\mu_0} , \tag{1}$$

ხოლო თუ $x > \mu_0$, მაშინ

$$p/2 = P\{X \ge x \mid \mu = \mu_0\} = 1 - \sum_{k=0}^{x-1} \frac{\mu_0^k}{k!} \cdot e^{-\mu_0} . \tag{2}$$

თუ p-ს გამოთვლილი მნიშვნელობა აღმოჩნდება 0.05-ზე პატარა, მაშინ ამბობენ, რომ შედეგი (ლაპარაკია x-ზე) ***სტატისტიკურად მნიშვნელოვანია***, წინააღმდეგ შემთხვევაში კი – ***ის სტატისტიკურად უმნიშვნელოა*** და ნულოვან ჰიპოთეზას არ უარყოფენ.

დიდი მოცულობის შემთხვევაში აქაც იყენებენ ნორმალურ აპროქსიმაციას, კერძოდ, იმ ფაქტს, რომ

$$(X - \mu_0)^2 / \mu_0 \cong N^2(0,1) \cong \chi^2(1) .$$

თავი VII
ჰიპოთეზის შემოწმება ორამოკრეფიან ამოცანებში

შესავალი.

სტატისტიკურ გადაწყვეტილებათა თეორიის ერთ-ერთ უმნიშვნელოვანეს ამოცანას წარმოადგენს ორი პოპულაციის შედარება, რათა დადგინდეს, იძლევა თუ არა მოქმედების (ქცევის) ალტერნატიული გზა უკეთეს შედეგს. აქამდე ჩვენ ვამოწმებდით ჰიპოთეზებს ერთი პოპულაციის პარამეტრების კონკრეტული მნიშვნელობის შესახებ, რომელიც შეიძლებოდა ყოფილიყო გარკვეული მოსაზრებებით დასახელებული რიცხვი, ან ცნობილი ყოფილიყო წინა გამოკვლევებიდან ან მოტანილი ყოფილიყო უფრო დიდი (ან მცირე) მოცულობის პოპულაციაზე დაკვირვებებიდან.

ახლა ჩვენ შევუდგებით ე.წ. *ორამოკრეფიანი ამოცანების* განხილვას, რაც გულისხმობს ორი პოპულაციის პარამეტრების შედარებას, ისე, რომ არცერთის კონკრეტული მნიშვნელობა ცნობილი არ არის. მაშასადამე, ამოცანა მდგომარეობს შემდეგში: მოცემულია ორი შერჩევა. ერთი შერჩევა აღებულია A, ხოლო მეორე B პოპულაციებიდან, რომელთა მახასიათებლები (პარამეტრები) ჩვენთვის უცნობია და გვაინტერესებს, არის თუ არა ამ პოპულაციების საშუალოები (ან დისპერსიები, ან პროპორციები) ტოლი. ეს იქნება ძირითადი (ნულოვანი) ჰიპოთეზა, ხოლო ალტერნატივები აქაც შეიძლება იყოს ცალმხრივი (მარჯვენა და მარცხენა) ან ორმხრივი. ჩვენ ხშირად ჩავთვლით, რომ შესადარებელი პოპულაციები ერთმანეთისაგან დამოუკიდებელია, რაც იმას ნიშნავს, რომ ის შემთხვევითი სიდიდეები, რომლებიც ორივე შერჩევისათვის რიცხვითი მონაცემების უკან დგას, არიან დამოუკიდებლები.

ორი პოპულაციის მახასიათებლების შედარების ამოცანა ჩნდება ე. წ. კონტროლირებად ექსპერიმენტებში, როცა შეისწავლება რაიმე ახალი ტექნოლოგიის, მედიკამენტის, რეკლამის თუ სხვა ფაქტორების ზემოქმედების ეფექტურობის საკითხი. ამ ექსპერიმენტს უკავშირდება ორი შერჩევა, რომლებიდანაც ერთი მიღებულია კონსერვატიული (დამკვიდრებული) ქმედების, ხოლო მეორე – შეცვლილი (ახალი) ქმედების შედეგად. მაგალითად, ბანკმა შეცვალა მოლარეები სპეციალური აპარატებით და აინტერესებს თუ რა გავლენას მოახდენს ეს კლიენტების მომსახურებაზე. აქ ბუნებრივად ჩნდება ორი შერჩევა: ერთი შედგება იმ კლიენტების მომსახურების დროებისაგან, რომლებსაც ემსახურებოდნენ მოლარეები, ხოლო მეორე – იმ კლიენტების მომსახურების დროებისაგან, რომლებიც სარგებლობენ სპეციალური აპარატებით. შესაბამისად, კონტროლირებად ექსპერიმენტებში ერთი შერჩევა არის საკონტროლო (მიღებულია ადრინდელ პირობებში), ხოლო მეორე – ექსპერიმენტული. სწორედ ასეთ ამოცანებს უწოდებენ ორამოკრეფიან ამოცანებს.

ამით არ ამოიწურება ორამოკრეფიანი ამოცანების ჩამონათვალი. ხშირად შესადარებელია განსხვავებული სოციალური სტატუსის, სქესის, რეგიონის ან სხვა ფაქტორების მიხედვით განსხვავებული ორი პოპულაცია. განასხვავებენ ორ შემთხვევას: პირველი – გვაქვს ორი დამოუკიდებელი შერჩევა ორი პოპულაციიდან და მეორე – გვაქვს წყვილების შერჩევა.

სადაც თვითონ წყვილები ერთმანეთისაგან დამოუკიდებელია, მაგრამ მათი კომპონენტები შეიძლება არ იყვნენ დამოუკიდებლები (ასეთ მონაცემებს დაწყვილებული მომაცემები ეწოდება).

1. დაწყვილებული მონაცემები.

განვიხილოთ ასეთი მაგალითი: დავუშვათ, ჩვენ გვაინტერესებს მოქმედებს თუ არა გარკვეული ტიპის კონტრაცეპტივის მიღება ქალის სისხლის წნევაზე. ამ მიზნით მოპოვებული მონაცემები წარმოდგენილია შემდეგი ცხრილის სახით:

i	წნევა კონტრაცეპტივის გამოყენებამდე (x_{i1})	წნევა კონტრაცეპტივის გამოყენების შემდეგ (x_{i2})	სხვაობა $d_i = x_{i2} - x_{i1}$
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

ვთქვათ, i-ური ქალის სისხლის წნევა კონტრაცეპტივის მიღებამდე განაწილებულია ნორმალურად საშუალოთი μ_i და დისპერსიით σ^2, ხოლო კონტრაცეპტივის მიღების შემდეგ განაწილებულია ისევ ნორმალურად საშუალოთი $\mu_i + \Delta$ და იმავე σ^2 დისპერსიით.

ნულოვანი ჰიპოთეზა ყალიბდება შემდეგნაირად: $H_0 : \Delta = 0$, რაც იმას ნიშნავს, რომ კონტრაცეპტივის მიღებას წნევა არ შეუცვლია, ხოლო ალტერნატივაა $H_1 : \Delta \neq 0$, რაც იმას ნიშნავს, რომ კონტრაცეპტივის მიღება ცვლის წნევის სიდიდეს ანუ მოქმედებს სისხლის წნევაზე. ცხადია, რომ კრიტერიუმის სტატისტიკა ორივე შერჩევას უნდა ეყრდნობოდეს, რადგან ცალ-ცალკე პოპულაციების პარამეტრები ჩვენთვის უცნობია. სამაგიეროდ, ჰიპოთეზის სამართლიანობის შემთხვევაში, ჩვენთვის ცნობილია, რომ შემთხვევითი სიდიდეები $Y_i := X_{i2} - X_{i1}$ $(i = 1, 2, \ldots, n)$ განაწილებულია ნორმალურად საშუალთი 0 და დისპერსიით, σ_Y^2, რომელიც მართალია არ ვიცით, მაგრამ რომლის შეფასებაც შეგვიძლია გაერთიანებული შერჩევიდან. ამრიგად, კრიტერიუმის სტატისტიკად შეგვიძლია ავიღოთ:

$$T_n = \frac{\sqrt{n} \cdot \overline{Y}}{S_{n,Y}} \qquad (1)$$

შემთხვევითი სიდიდე, სადაც $\overline{Y}=\frac{1}{n}\cdot\sum_{i=1}^{n}Y_i$, ხოლო $S_{n,Y}$ აღნიშნავს სხვაობების შერჩევით სტანდარტულ გადახრას, ანუ

$$S_{n,Y}=\sqrt{\frac{1}{n-1}\cdot\left(\sum_{i=1}^{n}Y_i^2-(1/n)\cdot\left(\sum_{i=1}^{n}Y_i\right)^2\right)} \tag{2}$$

ისევე როგორც ერთამოკრეფიან ამოცანაში (ნორმალური პოპულაციის შესახებ ჰიპოთეზის შემოწმებისას უცნობი დისპერსიის შემთხვევაში), T_n შემთხვევით სიდიდეს ექნება $t(n-1)$ განაწილება (სტიუდენტის განაწილება თავისუფლების ხარისხით $n-1$) და ამიტომ სტატისტიკური კრიტერიუმი შემდეგნაირად ჩამოყალიბდება:

თუ დასახელებული α-სათვის T_n შემთხვევით სიდიდის დაკვირვებული მნიშვნელობა t_n (ანუ ის მნიშვნელობა, რომელიც მიიღება T_n შემთხვევითი სიდიდის გამოსახულებაში Y_i-ების შეცვლისას d_i-ებით) აკმაყოფილებს პირობას $-t_{n-1,\alpha/2}\le t_n\le t_{n-1,\alpha/2}$, მაშინ H_0 ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში კი – მას უარვყოფთ.

დავუბრუნდეთ ჩვენს მაგალითს და გამოვთვალოთ T_n შემთხვევითი სიდიდის დაკვირვებული $t_n=t_{10}$ მნიშვნელობა. ცხადია, რომ

$$\overline{d}_{10}=(13+3-1+9+7+7+6+4+2-2)/10=4.8,$$

$$s_{n,Y}=\sqrt{\frac{13^2+3^2+(-1)^2+9^2+7^2+7^2+6^2+2^2+(-2)^2-10\cdot(4.8)^2}{9}}=$$

$$=\sqrt{20.844}=4.566$$

და ამიტომ $t_{10}=3.17\cdot4.8/4.566=3.32$. $t(9)$-განაწილების ზედა კრიტიკული წერტილების ცხრილებიდან კი, ვპოულობთ, რომ $t_{9,0.025}=2.262$. ვინაიდან $t_{10}=3.32$ მნიშვნელობა არ ეკუთვნის [-2.262, 2.262] რიცხვით ინტერვალს, H_0 ჰიპოთეზას უარვყოფთ. მაშასადამე, შეგვიძლია გავაკეთოთ დასკვნა, რომ კონტრაცეპტივი მოქმედებს ქალის წნევაზე.

დავსვათ ასეთი კითხვა: როგორ ინტერვალშია მოთავსებული სხვაობის ჭეშმარიტი საშუალო μ_Y? ნდობის ინტერვალის ასაგებად ისევ ვიყენებთ იმავე ფაქტს, რომ $T_n=\frac{\sqrt{n}\cdot\overline{Y}}{S_{n,Y}}\cong t(n-1)$ და ამ შემთხვევაში ნდობის ინტერვალს აქვს შემდეგი სახე:

$$\left(\overline{d}_n-t_{n-1,\alpha/2}\cdot s_{n,Y}/\sqrt{n};\overline{d}_n-t_{n-1,\alpha/2}\cdot s_{n,Y}/\sqrt{n}\right). \tag{3}$$

ამიტომ ჩვენს მაგალითში მივიღებთ, რომ:

(4.8-2.262·4.566/3.17, 4.8+2.262·4.566/3.17) = (1.53, 8.07).

მაშასადამე, წნევათა სხვაობის ჭეშმარიტი საშუალო μ_Y მნიშვნელობა 0.95-ის ტოლი ალბათობით მოთავსებულია ინტერვალში 1.53 მმ ვწყ. სვ.-დან 8.07 მმ ვწყ.სვ.-მდე.

2. ორამოკრეფიანი t –კრიტერიუმი ტოლი, უცნობი დისპერსიების შემთხვევაში.

ვთქვათ, მოცემულია ორი დამოუკიდებელი ნორმალურად განაწილებული პოპულაცია საშუალოებით μ_1 და μ_2 და ცნობილია, რომ მათ აქვთ ტოლი დისპერსიები $\sigma_1^2=\sigma_2^2=\sigma^2$, რომელთა საერთო σ^2 მნიშვნელობა უცნობია. ჩვენი ამოცანაა შევამოწმოთ $H_0:\mu_1=\mu_2$ ჰიპოთეზა, $H_1:\mu_1\neq\mu_2$ ალტერნატივის წინააღმდეგ. ცხადია, რომ $\overline{X}_{1n_1}-\overline{X}_{2n_2}\cong N(\mu_1-\mu_2,\sigma^2\cdot(1/n_1+1/n_2))$, სადაც n_1 და n_2 შესაბამისად, პირველი და მეორე შერჩევის მოცულობებია. ამიტომ H_0 ჰიპოთეზის სამართლიანობისას

$$\frac{\overline{X}_{1n_1}-\overline{X}_{2n_2}}{\sigma\cdot\sqrt{1/n_1+1/n_2}}\cong N(0,1). \tag{1}$$

მაგრამ, უკანასკნელი სიდიდე კრიტერიუმის სტატისტიკად არ გამოდგება, ვინაიდან ის შეიცავს უცნობ σ პარამეტრს. ამიტომ, როგორც წესი, იქცევიან შემდეგნაირად: (1) გამოსახულებაში σ-ს მაგივრად სვამენ მის შეფასებას და კრიტერიუმის სტატისტიკად იღებენ ასეთნაირად მიღებულ სიდიდეს. ამ შემთხვევაში ავტომატურად იბადება კითხვა: რომელი შერჩევიდან შევაფასოთ უცნობი σ პარამეტრი? ამ კითხვაზე პასუხის გასაცემად, გავიხსენოთ, რომ შერჩევის მოცულობის ზრდა იწვევს საშუალოს სტანდარტული შეცდომის შემცირებას. ამიტომ იქცევიან ასე: ვინაიდან, ძირითადი ჰიპოთეზის სამართლიანობის შემთხვევაში, ორივე პოპულაცია ერთნაირადაა განაწილებული, აერთიანებენ ამ შერჩევებს (რითაც იზრდება შერჩევის საერთო მოცულობა) და ისე აფასებენ σ-ს მნიშვნელობას, ანუ σ^2-ის შეფასების როლში აიღება შემდეგი სიდიდე:

$$S^2_{n_1,n_2}=\frac{(n_1-1)S^2_{1n_1}+(n_2-1)S^2_{1n_2}}{n_1+n_2-2}. \tag{2}$$

მტკიცდება, რომ ასე აგებული შეფასება წარმოადგენს σ^2–ის ჩაუნაცვლებელ შეფასებას (გავიხსენოთ, რომ ერთი შერჩევის შემთხვევაში შესწორებული შერჩევითი დისპერსია იყო დისპერსიის ჩაუნაცვლებელი შეფასება).

საბოლოოდ, კრიტერიუმის სტატისტიკას ექნება შემდეგი სახე

$$T_{n_1,n_2}=\frac{\overline{X}_{1n_1}-\overline{X}_{2n_2}}{S_{n_1,n_2}\cdot\sqrt{1/n_1+1/n_2}}, \tag{3}$$

რომლისთვისაც მტკიცდება, რომ

$$T_{n_1,n_2}=\frac{\overline{X}_{1n_1}-\overline{X}_{2n_2}}{S_{n_1,n_2}\cdot\sqrt{1/n_1+1/n_2}}\cong t(n_1+n_2-2). \tag{4}$$

ამიტომ სტატისტიკური კრიტერიუმი შემდეგნაირად ყალიბდება:

თუ დასახელებული α-სათვის T_{n_1,n_2} შემთხვევით სიდიდის დაკვირვებული მნიშვნელობა t_{n_1,n_2} აკმაყოფილებს პირობას

$$-t_{n_1+n_2-2,\alpha/2} \le t_{n_1,n_2} \le t_{n_1+n_2-2,\alpha/2}, \tag{5}$$

მაშინ H_0 ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში კი – მას უარვყოფთ.

კრიტერიუმის გამოყენების საილუსტრაციოდ მოვიყვანოთ შემდეგი მაგალითი: დავუშვათ, დაწესებულებაში მომუშავე 35-დან 39 წლამდე ასაკის 8 ქალის, რომლებიც ხმარობენ კონტრაცეპტივს, სისხლის წნევათა საშუალო გამოვიდა 132.86 მმ ვწყ.სვ., სტანდარტული გადახრით 15.34 მმ ვწყ.სვ., ხოლო იმავე დაწესებულებაში მომუშავე 21 იმავე ასაკის ქალისათვის, რომლებიც არ ხმარობენ კონტრაცეპტივს, იგივე მაჩვენებლებია შესაბამისად, 127.44 მმ ვწყ.სვ და 18.23 მმ ვწყ.სვ.. $\alpha = 0.05$ მნიშვნელოვნების დონით რა შეიძლება ითქვას პოპულაციების წნევების საშუალოებს შორის სხვაობაზე?

ამოხსნა. (3) ფორმულაზე დაყრდნობით გამოვთვალოთ T_{n_1,n_2} შემთხვევითი სიდიდის დაკვირვებული t_{n_1,n_2} მნიშვნელობა. ამისათვის ჯერ (2) თანაფარდობიდან გამოვთვალოთ $S^2_{n_1,n_2}$-ის დაკვირვებული $s^2_{n_1,n_2}$ მნიშვნელობა:

$$s^2_{n_1,n_2} = ((8-1)\cdot(15.34)^2+(21-1)\cdot(18.23)^2)/(8+21-2) = 307.18,$$

საიდანაც $s_{n_1,n_2} = \sqrt{307.18} \approx 17.527$.

ამიტომ

$$t_{n_1,n_2} = t_{8,21} = \frac{132.86-127.44}{17.527\cdot\sqrt{1/8+1/21}} \approx 0.74 .$$

მეორეს მხრივ, სტიუდენტის $t(8+21-2)=t(27)$-განაწილების ზედა კრიტიკული წერტილების ცხრილებიდან ვპოულობთ, რომ $t_{27,0.025} = 2.052$. რადგან $t_{8,21} = 0.74$ მნიშვნელობა ეკუთვნის [-2.052, 2.052] რიცხვით ინტერვალს, ამიტომ $\alpha = 0.05$ მნიშვნელოვნების დონით H_0 ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს. მაშასადამე, შეგვიძლია დავასკვნათ, რომ სისხლის წნევის საშუალოები მოცემული პოპულაციებისათვის მნიშვნელოვნად არ განსხვავდება.

რადგან კრიტერიუმის სტატისტიკის განაწილება ცნობილია, ჩვენ შეგვიძლია გამოვთვალოთ $1-\alpha$ ალბათობის მქონე ნდობის ინტერვალი $\mu_1-\mu_2$ სხვაობისათვის, რომელსაც შემდეგი სახე აქვს:

$$\left(\overline{x}_{1n_1} - \overline{x}_{1n_2} - t_{n_1+n_2-2,\alpha/2}\cdot s_{n_1,n_2}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\,;\,\overline{x}_{1n_1} - \overline{x}_{1n_2} + t_{n_1+n_2-2,\alpha/2}\cdot s_{n_1,n_2}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right).$$

ზემოთ განხილულ მაგალითში ნდობის ინტერვალის ქვედა და ზედა საზღვრები შესაბამისად იქნება:

$$132.86-127.44-2.052\cdot17.527\cdot\sqrt{\frac{1}{8}+\frac{1}{21}} = -9.52$$

და

$$132.86-127.44+2.052\cdot17.527\cdot\sqrt{\frac{1}{8}+\frac{1}{21}} = 20.36 .$$

ამრიგად, ნდობის ინტერვალი ყოფილა – [-9.52, 20.36].
როგორც ვხედავთ, ეს ინტერვალი საკმარისად ფართეა და საჭიროა უფრო დიდი მოცულობის შერჩევები, რომ უფრო ზუსტად დაეახასიათოთ $\mu_1-\mu_2$ სხვაობის ჭეშმარიტი მნიშვნელობა. შემდეგ ნაწილში ჩვენ გადავალთ ორი ნორმალური პოპულაციის დისპერსიათა ტოლობის შესახებ ჰიპოთეზის შემოწმებაზე.

3. ორამოკრეფიანი t–კრიტერიუმი არატოლი დისპერსიების შემთხვევაში.

ვთქვათ, მოცემულია ორი დამოუკიდებელი ნორმალურად განაწილებული პოპულაცია საშუალოებით μ_1 და μ_2 და ცნობილია, რომ მათ აქვთ არატოლი დისპერსიები $\sigma_1^2 \neq \sigma_2^2$. ჩვენი ამოცანაა შევამოწმოთ $H_0:\mu_1=\mu_2$ ჰიპოთეზა, $H_1:\mu_1\neq\mu_2$ ალტერნატივის წინააღმდეგ. ცნობილია, რომ

$$\overline{X}_{1n_1}-\overline{X}_{2n_2}\cong N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}),$$

სადაც n_1 და n_2 შესაბამისად, პირველი და მეორე შერჩევის მოცულობებია. აქედან გამომდინარე, თუ ცნობილია σ_1^2 და σ_2^2, მაშინ H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში

$$\frac{\overline{X}_{1n_1}-\overline{X}_{2n_2}}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\cong N(0,1). \qquad (1)$$

ამიტომ, თუ ცნობილია σ_1^2 და σ_2^2, მაშინ (1) შემთხვევითი სიდიდე გამოდგება კრიტერიუმის სტატისტიკად. მაგრამ ეს სიდიდე კრიტერიუმის სტატისტიკად არ გამოგვადგება, როცა ის შეიცავს უცნობ σ_1^2 და σ_1^2 პარამეტრებს. ამ შემთხვევაში მათ მაგიერად (1) გამოსახულებაში ჩავსვათ მათი შეფასებები (შესაბამისი შერჩევითი სტანდარტული გადახრები) და კრიტერიუმის სტატისტიკად ავიღოთ შემდეგი შემთხვევითი სიდიდე:

$$T_{n_1,n_2}\equiv\frac{\overline{X}_{1n_1}-\overline{X}_{2n_2}}{\sqrt{S_{1n_1}^2/n_1+S_{2n_2}^2/n_2}}. \qquad (2)$$

როგორც ვიცით, კრიტერიუმის სტატისტიკა მხოლოდ გამოთვლადი სიდიდე კი არ უნდა იყოს, არამედ კრიტერიუმის ასაგებად, არსებითია მისი ზუსტი ან ასიმპტოტური განაწილების ცოდნაც. როგორაა განაწილებული T_{n_1,n_2} შემთხვევითი სიდიდე? მათემატიკურ სტატისტიკაში ეს პრობლემა ცნობილია ***ბერენს-ფიშერის პრობლემის*** სახელით. მისი გადაჭრის ერთ-ერთი, შედარებით მარტივი გზა ცნობილია ***სატერტვაიტის*** **(Satterthwaite)** ***მეთოდის*** სახელით, რომელიც შემდეგნაირად ყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის T_{n_1,n_2} სტატისტიკის დაკვირვებული მნიშვნელობა t_{n_1,n_2} აკმაყოფილებს შემდეგ უტოლობას

$$-t_{[c],\alpha/2}\le t_{n_1,n_2}\le t_{[c],\alpha/2}, \qquad (3)$$

სადაც [c] აღნიშნავს c რიცხვის მთელ ნაწილს (ანუ c-ს უახლოეს მთელ რიცხვს მარცხნიდან), ხოლო c გამოითვლება შემდეგი წესით

$$c = \frac{(s_{1n_1}^2/n_1 + s_{2n_2}^2/n_2)^2}{(s_{1n_1}^2/n_1)^2/(n_1-1) + (s_{2n_2}^2/n_2)^2/(n_2-1)}, \qquad (4)$$

მაშინ α მნიშვნელოვნების დონით $H_0: \mu_1 = \mu_2$ ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში, მას უარვყოფთ.

კრიტერიუმის გამოყენების საილუსტრაციოდ განვიხილოთ შემდეგი მაგალითი: დავუშვათ, რომ შეისწავლება ქოლესტერინის შემცველობა გულის დაავადებით გარდაცვალებული ადამიანების მემკვიდრეებში. დავუშვათ, რომ 100 ასეთ ბავშვზე დაკვირვებამ მოგვცა ქოლესტერინის საშუალო დონე 207.3 მგ/დლ, ხოლო სტანდარტული გადახრაა 35.6 მგ/დლ. გარდა ამისა, დააკვირდნენ ქოლესტერინის დონეებს იმ ბავშვებში, რომელთა მშობლებიც ცოცხლებია და არა აქვთ გულის დაავადება. ასეთნაირად შერჩეული 74 ბავშვის მონაცემების მიხედვით ქოლესტერინის საშუალო დონე აღმოჩნდა 193.4 მგ/დლ, ხოლო სტანდარტული გადახრა –17.3 მგ/დლ. ჩვენი ამოცანაა $\alpha = 0.05$ მნიშვნელოვნების დონისათვის შევამოწმოთ ჰიპოთეზა პოპულაციების საშუალოების ტოლობის შესახებ.

ამოხსნა. დავუშვათ, რომ პოპულაციები განაწილებულია ნორმალურად, მაგრამ მათი პარამეტრები უცნობია. ეს ამოცანა გავყოთ ორ ეტაპად: პირველ ეტაპზე შევამოწმოთ ჰიპოთეზა დისპერსიების ტოლობის შესახებ. თუ ჰიპოთეზას დისპერსიების ტოლობის შესახებ ვერ უარვყოფთ, შემდეგ შევამოწმოთ ჰიპოთეზა პოპულაციების საშუალოების ტოლობის შესახებ ტოლი დისპერსიების შემთხვევაში. იმ შემთხვევაში თუ არ დადასტურდა დისპერსიათა ტოლობის ჰიპოთეზა, საშუალოთა ტოლობის ჰიპოთეზა შევამოწმოთ ამ პუნქტში განხილული სატერტვაიტის მეთოდით.

განვიხილოთ შერჩევით დისპერსიათა შეფარდება და შევადაროთ ის ფიშერის განაწილების ზედა კრიტიკული წერტილის მნიშვნელობას $F_{99,73,0.025}$. $F_{100,74}$ სტატისტიკის დაკვირვებული მნიშვნელობა ტოლია

$$f_{100,74} = 35.6^2/17.3^2 = 4.23.$$

ვინაიდან ცხრილებში არ არის ფიშერის განაწილების ზედა კრიტიკული წერტილის მნიშვნელობა – $F_{99,73,0.025}$, ამიტომ გამოვთვალოთ p-მნიშვნელობა, რომელიც ალტერნატივის ორმხრივობის გამო ტოლია

$$p = 2 \cdot P\{F_{100,74} \ge 4.23\} \le 0.0002,$$

რაც იმის მაჩვენებელია, რომ შედეგი სტატისტიკურად მნიშვნელოვანია და მაშასადამე, ჰიპოთეზა დისპერსიათა ტოლობის შესახებ უნდა უარვყოთ.

გადავიდეთ მეორე ეტაპზე. რადგან დავადგინეთ, რომ პოპულაციათა დისპერსიები განსხვავებულია, ვიყენებთ სატერტვაიტის მეთოდს საშუალოთა ტოლობის შესახებ ჰიპოთეზის შესამოწმებლად. ჯერ გამოვთვალოთ (2) თანაფარდობითით განსაზღვრული სტატისტიკის დაკვირვებული მნიშვნელობა:

$$t_{100,74} = \frac{207.3 - 193.4}{\sqrt{35.6^2/100 + 17.3^2/74}} = \frac{13.9}{4.089} = 3.4.$$

ახლა გამოვთვალოთ თავისუფლების ხარისხის მიახლოებითი მნიშვნელობა c. (4) ტოლობის მიხედვით მოვიღებთ, რომ:

$$c=\frac{(35.6^2/100+17.3^2/74)^2}{(35.6^2/100)/99+(17.3^2/74)^2/73}=\frac{16.718^2}{1.8465}=151.4,$$

და მაშასადამე, $[c]=151$.

ვინაიდან $t_{100,74}=3.4>t_{120,0.025}=1.98>t_{151,0.025}$, ამიტომ $\alpha=0.05$ მნიშვნელოვნების დონით პოპულაციათა საშუალოების ტოლობის ჰიპოთეზას უარვყოფთ.

აქვე აღვნიშნავთ, რომ $1-\alpha$ ალბათობის მქონე ნდობის ინტერვალს ექნება შემდეგი სახე:

$$\left(\bar{x}_{1n_1}-\bar{x}_{1n_2}-t_{[c],\alpha/2}\cdot\sqrt{\frac{s_{1n_1}^2}{n_1}+\frac{s_{2n_2}^2}{n_2}};\ \bar{x}_{1n_1}-\bar{x}_{1n_2}+t_{[c],\alpha/2}\cdot\sqrt{\frac{s_{1n_1}^2}{n_1}+\frac{s_{2n_2}^2}{n_2}}\right).$$

4. შერჩევათა მოცულობების განსაზღვრა. ორი პოპულაციის საშუალოების შედარების კრიტერიუმის სიმძლავრე.

შერჩევათა მოცულობების განსაზღვრა საჭიროა ექსპერიმენტის წინასწარ დაგეგმვისათვის. ქვემოთ ჩვენ მოვიყვანთ ორამკრეფიანი ამოცანებისათვის შერჩევათა მოცულობების გამოსათვლელ ფორმულებს, რომლებიც საჭიროა პოპულაციათა საშუალოების შედარების ამოცანაში (მოცემული α მნიშვნელოვნების დონით) დასახელებული $1-\beta$ სიმძლავრის მისაღწევად ორმხრივი ალტერნატივების დროს.

დავუშვათ, რომ მოცემულია ორი ნორმალური პოპულაცია საშუალოებით μ_1 და μ_2 და ცნობილი დისპერსიებით σ_1^2 და σ_2^2. შესამოწმებელია $H_0:\mu_1=\mu_2$ ჰიპოთეზა, $H_1:\mu_1\neq\mu_2$ ალტერნატივის წინააღმდეგ. ჩვენი ამოცანაა, განვსაზღვროთ პოპულაციათა ის n_1 და $n_2=k\cdot n_1$ მოცულობები, რომელთათვისაც α მნიშვნელოვნების დონით მიიღწევა მოცემული $1-\beta$ სიმძლავრე. მოცულობების გამოსათვლელ ფორმულებს აქვთ შემდეგი სახე:

$$n_1\geq(\sigma_1^2+\sigma_2^2/k)\cdot(z_{1-\alpha/2}+z_{1-\beta})^2/(\mu_2-\mu_1)^2 \qquad (1)$$

და

$$n_2\geq(k\cdot\sigma_1^2+\sigma_2^2)\cdot(z_{1-\alpha/2}+z_{1-\beta})^2/(\mu_2-\mu_1)^2. \qquad (2)$$

შევნიშნოთ, რომ $k=1$-ის შემთხვევაში $n_2=n_1$ და (1) და (2) შეფასებებიც იძლევა ერთსა და იმავე შედეგს. საილუსტრაციოდ განვიხილოთ შემდეგი მაგალითი: დავუშვათ, რომ 35-დან 39 წლამდე ასაკის ქალების, რომლებიც ხმარობენ კონტრაცეპტივს, სისხლის წნევის საშუალოა 132.86 მმ ვწყ.სვ. და სტანდარტული გადახრა – 15.34 მმ ვწყ.სვ., ხოლო ქალებისათვის, რომლებიც არ ხმარობენ კონტრაცეპტივს, იგივე მაჩვენებლები შესაბამისად, 127.44 და 18.23 მმ ვწყ.სვ.-ის ტოლია. $\alpha=0.05$ მნიშვნელოვნების დონისა და k=2-სათვის განვსაზღვროთ შერჩევათა ის მოცულობები, რომლებიც საშუალოთა გარჩევის ამოცანაში მოგვცემენ 80%-იან სიმძლავრეს?

ამოხსნა. (1) ფორმულიდან მივიღებთ, რომ

$$n_1 = \frac{(15.34^2 + 18.23^2/2)\cdot(1.96+0.84)^2}{(132.86-127.44)^2} = 107.1,$$

საიდანაც ვასკვნით, რომ $n_1 = 108$, და მაშასადამე, $n_2 = 2\cdot n_1 = 216$.

კრიტერიუმის სიმძლავრის გამოსათვლელად $H_0 : \mu_1 = \mu_2$ ჰიპოთეზის წინააღმდეგ განვიხილოთ სპეციფიკური ალტერნატივა: კერძოდ, ალტერნატივა $H_1 : \mu_2 = \mu_1 + \Delta$. მაშინ ორმხრივი ალტერნატივისათვის სიმძლავრე გამოითვლება შემდეგი თანაფარდობიდან:

$$1-\beta = \Phi\left(-z_{\alpha/2} + \frac{\Delta\cdot\sqrt{n_1}}{\sqrt{\sigma_1^2 + \sigma_2^2/k}}\right), \tag{3}$$

სადაც $k = n_2/n_1$.

განვიხილოთ შემდეგი მაგალითი: შეირჩა 35-დან 39 წლამდე ასაკის 100-100 ქალი, რომლებიც ხმარობენ და არ ხმარობენ კონტრაცეპტივს და ორივე შერჩევისათვის სისხლის წნევათა სხვაობის საშუალო გამოვიდა 5 მმ ვწყ.სვ.. გარდა ამისა, მათი სტანდარტული გადახრები შესაბამისად 15.34 მმ ვწყ.სვ. და 18.23 მმ ვწყ.სვ.-ის ტოლია. α = 0.05 მნიშვნელოვნების დონისათვის განვსაზღვროთ კრიტერიუმის სიმძლავრე.

ამოხსნა. ვინაიდან $n_1 = n_2 = 100$, $k = n_2/n_1 = 1$, $\Delta = 5$, $\sigma_1 = 15.34$, $\sigma_2 = 18.23$ და α = 0.05, ამიტომ (3) თანაფარდობიდან მივიღებთ, რომ

$$1-\beta = \Phi(-1.96+2.099) = \Phi(0.139) = 0.555.$$

5. ჰიპოთეზის შემოწმება დისპერსიების ტოლობის შესახებ.

დისპერსიების შესახებ ჰიპოთეზები ძალიან მნიშვნელოვან როლს თამაშობენ ეკონომიკურ-მათემატიკური მოდელირებისას, ვინაიდან ექსპერიმენტული შერჩევითი მონაცემების გაბნევის სიდიდე შესაბამისი პარამეტრების გათვლილი თეორიული მნიშვნელობებიდან, რომელიც ხასიათდება დისპერსიით, შესაძლებლობას გვაძლევს გადავწყვიტოთ იმ მოდელის გამოსადეგობა (ადექვატურობა), რომლის საფუძველზეც იგება თეორია.

დავუშვათ, რომ ნორმალური კანონით განაწილებული ξ შემთხვევითი სიდიდე განსაზღვრულია გარკვეულ სიმრავლეზე, რომელიც ქმნის გენერალურ ერთობლიობას, ხოლო ნორმალური კანონით განაწილებული η შემთხვევითი სიდიდე განმარტებულია სხვა სიმრავლეზე, რომელიც აგრეთვე შეადგენს გენერალურ ერთობლიობას. ორივე ერთობლიობიდან კეთდება შერჩევა: პირველიდან – n_1 მოცულობის მქონე, ხოლო მეორედან – n_2 მოცულობის მქონე (შევნიშნავთ, რომ შერჩევის მოცულობა ყოველთვის არ შეიძლება თავიდანვე განსაზღვრული იყოს, მაგალითად, იმ შემთხვევა-ში, თუ შერჩევა არის ბადეში მოხვედრილი თევზები). თითოეული შერჩევისათვის გამოითვლება შესწორებული შერჩევითი დისპერსია: s_1^2 – შერჩევისათვის პირველი ერთობლიობიდან და s_2^2 – შერჩევისათვის მეორე ერთობლიობიდან.

ამოცანა მდგომარეობს შემდეგში: შერჩევითი მონაცემების საშუალებით შევამოწმოთ სტატისტიკური ჰიპოთეზა $H_0 : D\xi = D\eta$. ალტერნატიული ჰიპოთეზის როლში განვიხილავთ იდეას, რომელიც მდგომარეობს იმაში, რომ იმ ერთობლიობის დისპერსია, რომლის შესწორებული შერჩევითი დისპერსია აღმოჩნდა უდიდესი, მეტია ვიდრე მეორე ერთობლიობის დისპერსია. განიხილება შემდეგი სახის კრიტერიუმი:

$$F = S^{**} / S^{*},$$

სადაც S^{**} – უდიდესია s_1^2 და s_2^2 შეფასებებს შორის, ხოლო S^{*} – კი მათ შორის უმცირესი.

ცნობილია, რომ F კრიტერიუმი განაწილებულია ფიშერის განაწილების კანონით თავისუფლების k_1 და k_2 ხარისხებით, სადაც:

$$k_1 = n_1 - 1, \quad k_2 = n_2 - 1, \text{ თუ } S^{**} = s_1^2;$$

$$k_1 = n_2 - 1, \quad k_2 = n_1 - 1, \text{ თუ } S^{**} = s_2^2.$$

ამ ამოცანაში ბუნებრივია განვიხილოთ მარჯვენა კრიტიკული არე, ვინაიდან F კრიტერიუმის საკმარისად დიდი შერჩევითი მნიშვნელობა მეტყველებს ალტერნატიული ჰიპოთეზის სასარგებლოდ.

მოცემული მნიშვნელოვნების დონისათვის α (ჩვეულებრივ, $\alpha = 0.05$ ან $\beta = 0.01$) კრიტიკული მნიშვნელობა $F_{კრ} = F_{k_1,k_2,\alpha}$ განისაზღვრება ფიშერის განაწილების ცხრილიდან. *იმ შემთხვევაში, როცა $F > F_{კრ}$ ხდება H_0 ჰიპოთე-ზის უარყოფა, ხოლო როცა $F < F_{კრ}$ – H_0 ჰიპოთეზა მიიღება.*

დავუშვათ, რომ გარკვეული ობიექტების ორი სიმრავლე, რომელთაც გააჩნიათ რაოდენობრივი ნიშანი, ექვემდებარება შერჩევით კონტროლს. რაოდენობრივი ნიშნის მნიშვნელობები არიან ნორმალური განაწილების კანონით განაწილებული შემთხვევითი სიდიდეები, რომელთაც ჩვენ ავღნიშნავთ ξ_1-ით და ξ_2-ით შესაბამისად, პირველი და მეორე სიმრავლეებისათვის. პირველი სიმრავლიდან გაკეთებულია $n_1 = 21$ მოცულობის შერჩევა და ნაპოვნია შესწორებული შერჩევითი დისპერსია, რომელიც აღმოჩნდა 0.75-ის ტოლი. მეორე სიმრავლიდან გაკეთებულია $n_2 = 11$ მოცულობის შერჩევა. მისი შესწორებული შერჩევითი დისპერსიაა 0.25. ვაყენებთ ჰიპოთეზას: $H_0 : D\xi_1 = D\xi_2$. ალტერნატიული ჰიპოთეზა მდგომარეობს იმასი, რომ $H_1 : D\xi_1 > D\xi_2$. ამ შემთხვევაში ფიშერის კრიტერიუმის შერჩევითი მნიშვნელობა $F_{შ} = 3$. არჩეული $q = 0.05$ მნიშვნელოვნების დონისათვის, თავისუფლების $k_1 = 20$ და $k_2 = 10$ ხარისხით, ფიშერის განაწილების ცხრილიდან ვპოულობთ, რომ $F_{კრ} = 2.77$. ვინაიდან, $F_{შ} > F_{კრ}$, ჰიპოთეზა დისპერსიების ტოლობის შესახებ უნდა უკუგდებულ იქნეს.

6. ჰიპოთეზის შემოწმება შერჩევითი კორელაციის კოეფიციენტის სტატისტიკური მნიშვნელოვნების შესახებ.

გენერალური ერთობლიობის Δ პარამეტრის შერჩევითი δ შეფასების სტატისტიკური მნიშვნელოვნების შემოწმება ეწოდება $H_0 : \Delta = 0$ სტატი-

სტიკური ჰიპოთეზის შემოწმებას ალტერნატიული $H_1: \Delta \neq 0$ ჰიპოთეზის წინააღმდეგ. თუ მოხდება H_0 ჰიპოთეზის უარყოფა, მაშინ δ შეფასება ითვლება სტატისტიკურად მნიშვნელოვნად.

დავუშვათ, რომ მოცემულია ერთი და იგივე გენერალური ერთობლიობის ობიექტთა სიმრავლეზე განსაზღვრული ნორმალური კანონით განაწილებული ორი შემთხვევითი სიდიდე ξ და η. ჩვენი მიზანია შევამოწმოთ სტატისტიკური ჰიპოთეზა ξ და η შემთხვევით სიდიდეებს შორის კორელაციური კავშირის არ არსებობის შესახებ:

$H_0: \rho(\xi, \eta) = 0$; $\qquad H_1: \rho(\xi, \eta) \neq 0$.

ვატარებთ n მოცულობის შერჩევას და გამოითვლება შერჩევითი კორელაციის კოეფიციენტი r. სტატისტიკური კრიტერიუმის როლში განიხილება შემთხვევითი სიდიდე

$$t = r\sqrt{n-2}/\sqrt{1-r^2},$$

რომელიც განაწილებულია სტიუდენტის განაწილების კანონით თავისუფლების ხარისხით $n-2$.

შევნიშნავთ, რომ შერჩევითი კორელაციის r კოეფიციენტის ყველა შესაძლო მნიშვნელობა მოთავსებულია [-1,1] ინტერვალში. გასაგებია, რომ t სიდიდის შედარებით დიდი გადახრები ნულიდან ნებისმიერ მხარეს მიიღება შედარებით დიდი, ანუ მოდულით 1-თან ახლოს მდგომი r-ის მნიშვნელობებისათვის. ვინაიდან, მოდულით 1-თან ახლოს მდგომი r-ის მნიშვნელობები ეწინააღმდეგებიან H_0 ჰიპოთეზას, ამიტომ ბუნებრივია, რომ აქ განვიხილოთ ორმხრივი კრიტიკული არე t კრიტერიუმისათვის.

მნიშვნელოვნების α დონისა და თავისუფლების ხარისხის $n-2$ რიცხვის მიხედვით სტიუდენტის განაწილების ცხრილიდან ვპოულობთ კრიტიკულ მნიშვნელობას $t_{კრ} = t_{n-2,\alpha/2}$. თუ კრიტერიუმის შერჩევითი მნიშვნელობის $t_{ე}$ მოდული აღემატება $t_{კრ}$-ს, მაშინ ხდება H_0 ჰიპოთეზის უარყოფა და კორელაციის შერჩევითი კოეფიციენტი ითვლება სტატისტიკურად მნიშვნელოვნად. წინააღმდეგ შემთხვევაში, ანუ თუ $|t_{ე}| < t_{კრ}$, მიიღება H_0 ჰიპოთეზა და კორელაციის შერჩევითი კოეფიციენტი ითვლება სტატისტიკურად არა მნიშვნელოვნად.

ამოცანა 1. 12-დან 14 წლამდე ასაკის დაბალ სოციალურ-ეკონომიკურ პირობებში მცხოვრები 25 გოგონას კალციუმის მიღების საშუალო და სტანდარტული გადახრა შესაბამისად, ტოლია 6.56 და 0.64 მგ-ისა. უკეთეს პირობებში მცხოვრები 40 იმავე ასაკის გოგონასათვის კი ანალოგიური მაჩვენებლები შესაბამისად შეადგენს 6.8 და 0.76 მგ-ს. ვთქვათ, $\alpha = 0.05$.

ა). შეამოწმეთ, არის თუ არა მნიშვნელოვანი განსხვავება ორი პოპულაციის დისპერსიებს შორის.

ბ). როგორია საშუალოების შედარების შესაფერისი პროცედურა?

გ). ჩაატარეთ ეს პროცედურა ორნაირად: კრიტიკულ მნიშვნელობასა და p-მნიშვნელობაზე დაყრდნობით.

დ). გამოთვალეთ 95%-იანი ნდობის ინტერვალი პოპულაციების საშუალოთა სხვაობისათვის.

ე). დავუშვათ, რომ შემდგომი შესწავლისათვის საჭიროა დადგინდეს ორი ჯგუფის გოგონათა თანაბარი რაოდენობა, იმისათვის, რომ მივაღწიოთ 80%-იან სიმძლავრეს $\alpha = 0.05$ მნიშვნელოვნების დონისათვის. რამდენი მონაცემია ამისათვის საჭირო?

ე). როგორ შეიცვლება პასუხი ე) ამოცანაში, თუ ავიღებთ α=0.01-ს?

ამოცანა 2. საფიქრებელია, რომ ბავშვის ფილტვების ჯანმრთელობის მდგომარეობას მნიშვნელოვანწილად განსაზღვრავს ოჯახში ეწევიან თუ არა სიგარეტს. ამის შესასწავლად შეირჩა 5-დან 9 წლამდე ასაკის ბავშვების ორი ჯგუფი: 23 იმ ოჯახებიდან, სადაც არ ეწევიან სიგარეტს და 20 იმ ოჯახებიდან, სადაც სიგარეტს ეწევიან. აღმოჩნდა, რომ პირველ ჯგუფში FEV (forced expiratory volume) სიდიდის საშუალომ და სტანდარტულმა გადახრამ შესაბამისად შეადგინა 2.1 ლ და 0.7 ლ, ხოლო მეორე ჯგუფისათვის იგივე მაჩვენებლებმა შესაბამისად შეადგინეს 2.3 ლ და 0.4 ლ. ვთქვათ, $\alpha = 0.05$.

ა). როგორ ჩამოაყალიბებდით ნულოვან და ალტერნატიულ ჰიპოთეზებს?

ბ). როგორია ა)-ში ჩამოყალიბებული ჰიპოთეზების შემოწმების საჭირო პროცედურა?

გ). ჩაატარეთ ეს პროცედურა კრიტიკულ მნიშვნელობაზე დაყრდნობით.

დ). გამოთვალეთ 95%-იანი ნდობის ინტერვალი პოპულაციების საშუალოთა სხვაობისათვის.

ე). დავუშვათ, შემდგომი შესწავლისათვის საჭიროა დადგინდეს ორი ჯგუფის ბავშვების თანაბარი რაოდენობა, იმისათვის, რომ მივაღწიოთ 95%-იან სიმძლავრეს $\alpha = 0.05$ მნიშვნელოვნების დონისათვის. რამდენი მონაცემია ამისათვის საჭირო?

თავი VIII
მრავალამოკრეფიანი ამოცანები. დისპერსიული ანალიზი

შესავალი.

აქამდე ჩვენ ვიხილავდით ორი ნორმალურად განაწილებული პოპულაციის საშუალოების შედარების ამოცანებს. ხშირად საჭიროა უფრო მეტი პოპულაციის საშუალოების შედარება. განვიხილოთ შესაბამისი მაგალითი: ჯანმრთელობის დაცვის სამინისტროს აინტერესებს ახდენს, თუ არა გავლენას და რამდენად, სიგარეტის მოწევა ფილტვების ფუნქციურობაზე. ამისათვის ფილტვების ფუნქციურობა, ანუ FEF (forced mid-expiratory flow) სიდიდის მნიშვნელობები გაზომილ იქნა ადამიანების ექვს ჯგუფში: 1) არამწეველები (პირობითად, NS); 2) პასიური მწეველები (PS); 3) მწეველები, რომლებიც ფილტვებზე არ უშვებენ კვამლს (NI); 4) სუსტი მწეველები (დღეში 1-დან 10 ღერამდე 20 ან მეტი წლის განმავლობაში, LS); 5) ზომიერად მწეველები (დღეში 11-დან 39 ღერამდე 20 ან მეტი წლის განმავლობაში, MS) და 6) მაგარი მწეველები (40 და მეტი ღერი დღეში, 20 ან მეტი წლის განმავლობაში, HS). საინტერესოა FEF-ის საშუალოების შედარება ამ ექვს ჯგუფში. შედეგები მოცემულია ცხრილში:

ჯგუფის ნომერი და სახელი		FEF-ის საშუალო და სტანდარტული გადახრა (ლ/წ)		შერჩევის მოცულობები
1	NS	3.78	0.79	200
2	PS	3.30	0.77	200
3	NI	3.32	0.86	50
4	LS	3.23	0.78	200
5	MS	2.73	0.81	200
6	HS	2.59	0.82	200

t-კრიტერიუმის მეთოდოლოგია, საკმარისად კარგად ზოგადდება ასეთი შემთხვევებისათვის და ამ მეთოდოლოგიას სახელად ჰქვია ***ერთფაქტორიანი დისპერსიული ანალიზი*** (შემოკლებით, დისპერსიული ანალიზისათვის ხმარობენ ANOVA-ს, შესაბამისი ინგლისური ტერმინის შემოკლებიდან – Analysis of variance). ამ ტიპის ამოცანებში მოცემულია n_i, $i=1,2,...,k$, მოცულობის k შერჩევა (ჯგუფი). აღვნიშნოთ y_{ij}–თი i–ური ჯგუფის j–ური მონაცემი, რომლის უკან მდგომი Y_{ij} შემთხვევითი სიდიდისათვის ვუშვებთ, რომ სამართლიანია შემდეგი მოდელი:

$$Y_{ij} = \mu + \alpha_i + X_{ij}, \qquad (1)$$

სადაც μ მუდმივი რიცხვია, α_i – i–ური ჯგუფის მახასიათებელი ფაქტორია (ეფექტია), რომელიც შეიძლება იყოს დეტერმინისტული (არაშემთხვევითი) ან შემთხვევითი და X_{ij} არის შემთხვევითი სიდიდე, რომელსაც შეცდომას უწოდებენ და რომელიც განაწილებულია ნორმალურად საშუალო-

თი 0 და დისპერსიით σ^2. იგულისხმება აგრეთვე, რომ $\sum_{i=1}^{k}\alpha_i = 0$. აქედან გამომდინარე Y_{ij} შემთხვევითი სიდიდე განაწილებულია ნორმალურად საშუალოთი $\mu+\alpha_i$ და დისპერსიით σ^2. ასეთ პირობებში ამბობენ, რომ (1) თანაფარდობით განსაზღვრულია *ერთფაქტორიანი ANOVA მოდელი.*

შეენიშნავთ, რომ ასეთ ამოცანებში ადგილი აქვს ორი ტიპის ვარიაციულობას: ვარიაციულობას როგორც ჯგუფებს შორის, ისე ჯგუფების შიგნით. (1) თანაფარდობით განმარტებულ მოდელში μ გაიაზრება, როგორც ყველა ჯგუფის გაერთიანებით მიღებული შერჩევის შესაბამისი პოპულაციის საშუალო. α_i სიდიდის აზრი ის არის, რომ ის წარმოადგენს განსხვავებას i-ური ჯგუფის საშუალოსა და საერთო საშუალოს შორის, და ბოლოს, X_{ij} წარმოადგენს i-ური ჯგუფის $\mu+\alpha_i$ საშუალოდან შემთხვევითი გადახრას (შეცდომას).

ინტუიციურად, ზემოთ მოყვანილ მაგალითში FEF-ის სიდიდე შეიძლება გავიაზროთ როგორც FEF-ის საერთო საშუალო სიდიდეს დამატებული მწეველთა თითოეული ჯგუფის ეფექტი და დამატებული შემთხვევითი ვარიაციულობა თითოეული ჯგუფის შიგნით.

1. ჰიპოთეზათა შემოწმება ერთფაქტორიან ANOVA მოდელში. დეტერმინისტული ეფექტების შემთხვევა.

ნულოვანი ჰიპოთეზა ამ ამოცანაში ყალიბდება შემდეგნაირად: ყველა ჯგუფს აქვს ერთნაირი საშუალო. შესაბამისად, ჩვენი დაშვების გათვალისწინებით – $\sum_{i=1}^{k}\alpha_i = 0$, ნულოვანი ჰიპოთეზა იქნება:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0,$$

ხოლო ალტერნატივა იქნება, რომ რომელიღაც ჯგუფის საშუალო განსხვავებულია ნულისაგან:

$$H_1 : \text{არსებობს ისეთი } i \ (i = 1,2\dots,k), \text{ რომ } \alpha_i \neq 0.$$

აღვნიშნოთ $\bar{y}_i$ სიმბოლოთი i-ური ჯგუფის შერჩევითი საშუალო $\bar{y}_i = \frac{1}{n_i}\cdot\sum_{j=1}^{n_i} y_{ij}$, ხოლო $\bar{\bar{y}}$ სიმბოლოთი კი შერჩევითი საშუალოების შეწონილი საშუალო – $\bar{\bar{y}} = \frac{1}{n}\cdot\sum_{i=1}^{k} n_i\cdot\bar{y}_i$, ანუ ყველა მონაცემის საერთო საშუალო, სადაც $n = n_1 + n_2 + \cdots + n_k$ – ცხადია, არის ყველა მონაცემის გაერთიანებით მიღებული შერჩევის მოცულობა. ინდივიდუალური მონაცემის გადახრა საერთო საშუალოსაგან, შეიძლება განვიხილოთ, როგორც ორი გადახრის ჯამი, რომლიდანაც პირველი გვიჩვენებს ინდივიდუალურ გადახრას ჯგუფური საშუალოსაგან, ხოლო მეორე – ჯგუფის საშუალოს გადახრას საერთო საშუალოსაგან:

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}}). \qquad (1)$$

თუ ახლა (1) თანაფარდობის ორივე მხარეს ავიყვანთ კვადრატში, მიღებული ტოლობის ორივე მხარეს ავჯამავთ ყველა i და j-ის მიმართ და გავითვალისწინებთ, რომ მარჯვენა მხარის შერეული წევრების ჯამი 0-ის ტოლია (ამაში დასარწმუნებლად უნდა ვისარგებლოთ $\bar{y}_i$-სა და $\bar{\bar{y}}$-ის განმარტებებითა და ჯამისა და ორმაგი ჯამების ელემენტარული თვისებებით), მივიღებთ:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{\bar{y}})^2 \qquad (2)$$

როგორც ვხედავთ, (2)-ის მარცხენა მხარეში დგას ერთიანი საშუალოსაგან (ყველა მონაცემის გაერთიანებით მიღებული შერჩევის $\bar{\bar{y}}$ საშუალოსაგან) გადახრების კვადრატების ჯამი (ფაქტიურად ის არის ყველა მონაცემის გაერთიანებით მიღებული შერჩევის არანორმირებული შერჩევითი დისპერსია). ამ სიდიდეს აღნიშნავენ *tss* აბრევიატურით (*Total Sam of Squares*). ანალოგიურად, თუ დავაკვირდებით (2) თანაფარდობის მარჯვენა მხარის პირველ შესაკრებს, შიდა ჯამი წარმოადგენს i-ური ჯგუფის საშუალოსაგან გადახრების კვადრატების ჯამს. ამიტომ (2)-ის მარჯვენა მხარის პირველ შესაკრებს უწოდებენ გადახრას ჯგუფის შიგნით და მას აღნიშნავენ *wss* აბრევიატურით (*Within Sam of Squares*). და ბოლოს, (2)-ის მარჯვენა მხარის მეორე შესაკრები, რომელიც წარმოადგენს ჯგუფური საშუალოების საერთო საშუალოსაგან გადახრების კვადრატების ჯამს, გაიაზრება, როგორც ჯგუფთაშორის გადახრის საზომი და აღენიშნება *bss* აბრევიატურით (*Between Sam of Squares*). მაშასადამე, გვაქვს:

$$tss = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2, \quad wss = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, \quad bss = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{\bar{y}})^2 \qquad (3)$$

და შესაბამისად, (2)-დან გამომდინარე, ადგილი აქვს შემდეგ თანაფარდობას:

$$tss = wss + bss.$$

შეთანხმების თანახმად, იმ შემთხვევით სიდიდეებს, რომელთა კონკრეტულ რეალიზაციებსაც წარმოადგენს *tss*, *wss*, *bss* რიცხვები, შესაბამისი დიდი ასოებით აღენიშნავენ, ანუ *TSS*, *WSS*, *BSS* წარმოადგენენ შემთხვევით სიდიდეებს, რომლებიც მიიღება თუ (3) გამოსახულებებში ყველა y_{ij} რიცხვებს შევცვლით Y_{ij} შემთხვევითი სიდიდეებით.

გარდა ამისა, გამოთვლების გამარტივების მიზნით, შემოვიღოთ შემდეგი აღნიშვნები:

$$y_{i\bullet} \equiv \sum_{j=1}^{n_i} y_{ij} = n_i \cdot \bar{y}_i$$ – მონაცემების ჯამი i-ურ ჯგუფში; (4)

$$y_{\bullet\bullet} \equiv \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = n \cdot \bar{\bar{y}}$$ – მონაცემების ჯამი ყველა ჯგუფში. (5)

ამ აღნიშვნებში *tss* და *bss* სიდიდეები შემდეგნაირად გადაიწერება:

$$tss=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}^2-\frac{y_{\bullet\bullet}^2}{n}. \qquad (6)$$

$$bss=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i-\bar{\bar{y}})^2=\sum_{i=1}^{k}\frac{y_{i\bullet}^2}{n_i}-\frac{y_{\bullet\bullet}^2}{n}. \qquad (7)$$

იმ შემთხვევაში, როცა მონაცემები დაჯგუფებულია საშუალოებისა და დისპერსიების მიხედვით, მაშინ უფრო მოხერხებულია გამოვიყენოთ შემდეგი ფორმულები:

$$bss=\sum_{i=1}^{k}n_i\cdot\bar{y}_i^2-\frac{1}{n}\cdot\left(\sum_{i=1}^{k}n_i\cdot\bar{y}_i\right)^2. \qquad (8)$$

$$wss=\sum_{i=1}^{k}(n_i-1)\cdot s_i^2. \qquad (9)$$

ფილტვების ფუნქციურობის შესახებ ზემოთ მოყვანილ მაგალითში გამოვთვალოთ ზემოთ შემოღებული სიდიდეების რიცხვითი მნიშვნელობები. ცხადია, რომ:

$$bss=200\cdot3.78^2+200\cdot3.30^2+50\cdot3.32^2+200\cdot3.23^2+200\cdot2.73^2+200\cdot2.59^2-$$

$$-\frac{1}{1050}\cdot(200\cdot3.78+200\cdot3.30+50\cdot3.32+200\cdot3.23+200\cdot2.73+200\cdot2.59)^2=184.38,$$

$$wss=199\cdot0.79^2+199\cdot0.77^2+49\cdot0.86^2+199\cdot0.78^2+199\cdot0.81^2+199\cdot0.82^2=663.87.$$

შემოვიღოთ კიდევ შემდეგი სიდიდეები: $bms=bss/(k-1)$ – ჯგუფთაშორისი კვადრატების საშუალო (*Between Mean Square*) და $wms=wss/(n-k)$ - ჯგუფის შიგნით კვადრატების საშუალო (*Within Mean Square*).

ინტუიციურად გასაგებია, რომ სტატისტიკური კრიტერიუმის ასაგებად შემდეგნაირად უნდა ვიმსჯელოთ: თუ bms/wms შეფარდება "დიდია", მაშინ ჯგუფები "სხვადასხვა საშუალო ყოფაქცევის" ინდივიდებით ყოფილა დაკომპლექტებული და მაშასადამე, ნულოვანი ჰიპოთეზა უნდა უარვყოთ და პირიქით, თუ bms/wms შეფარდება "მცირეა", მაშინ ჯგუფები "ერთნაირი საშუალო ყოფაქცევით" ხასიათდება და მაშასადამე, ნულოვანი ჰიპოთეზის უარყოფის საფუძველი არ უნდა გექონდეს.

სტატისტიკური კრიტერიუმი დამყარებული იქნება სწორედ bms/wms შეფარდების შესაბამის შემთხვევით სიდიდეზე -- $F=BMS/WMS$, რომელსაც მტკიცდება, რომ გააჩნია ფიშერის განაწილება თავისუფლების ხარისხებით $k-1$ და $n-1$ -- $F(k-1,n-1)$.

საბოლოოდ, სტატისტიკური კრიტერიუმი შემდეგნაირად ყალიბდება:

თუ დასახელებული α მნიშვნელოვნების დონისათვის

$$F=BMS/WMS$$

სტატისტიკის დაკვირვებული მნიშვნელობისათვის $f=bms/wms$ შესრულდა თანაფარდობა $f=bms/wms>F_{k-1,n-1,\alpha}$ (სადაც $F_{k-1,n-1,\alpha}$ ფიშერის $F(k-1,n-1)$ განაწილების ზედა α კრიტიკული წერტილია), მაშინ H_0 ჰიპოთეზას უარვყოფთ. წინააღმდეგ შემთხვევაში, H_0 ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს. შესაბამისი p-მნიშვნელობა ტოლია:

$$p=P\{F_{k-1,n-1,\alpha}>f\}.$$

მივუყენოთ ეს კრიტერიუმი ჩვენს მაგალითს. როგორც უკვე ვნახეთ, $bss = 184.38$ და $wss = 663.87$, ამიტომ $bms = bss/(k-1) = 184.38/(6-1) = 36.875$ და $wms = wss/(n-k) = 663.87/(1050-6) = 0.636$. შესაბამისად, კრიტერიუმის სტატისტიკის რიცხვითი მნიშვნელობა იქნება:

$$f = bms / wms = 36.875/0.636 = 57.979 .$$

მეორეს მხრივ, ფიშერის $F(5,120)$ განაწილების ცხრილებიდან ვპოულობთ, რომ $F_{5,120,0.001} = 4.42$. შესაბამისად, გვაქვს, რომ:

$$F_{5,1044,0.001} < F_{5,120,0.001} = 4.42 < 57.979 = f$$

ამიტომ

$$p = P\{F_{5,1044} > f\} = P\{F_{5,120} > f\} = P\{F_{5,120} > 57.979\} < 0.001 ,$$

რაც იმას ნიშნავს, რომ შედეგი სტატისტიკურად მნიშვნელოვანია და ამგვარად, H_0 ჰიპოთეზა $\alpha = 0.001$ მნიშვნელოვნების დონითაც კი უნდა უარვყოთ.

2. ჯგუფთა შედარება ერთფაქტორიან ANOVA მოდელში. წყვილთა შედარების t-კრიტერიუმი.

როგორც ზემოთ იყო აღნიშნული $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, ჰიპოთეზის უარყოფის შემთხვევაში ჩვენ ვაკეთებთ დასკვნას, რომ ჯგუფების საშუალო ყოფაქცევა არ არის ერთნაირი, მაგრამ ჩვენ არ შეგვიძლია იმის თქმა, თუ კერძოდ რომელი ჯგუფებია განსხვავებული. H_0 ჰიპოთეზის უარყოფა ნიშნავს იმას, რომ არსებობს სულ ცოტა ორი ჯგუფი მაინც, რომელთა საშუალოებიც მნიშვნელოვნად განსხვავებულია. დავუშვათ, ესენია i-ური და j-ური ჯგუფები. განსახილველი $Y_{ij} = \mu + \alpha_i + X_{ij}$ მოდელის მიხედვით, ყოველი i-სათვის, $i = 1,2,\dots,k$:

$$\bar{Y}_i = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij}$$

შემთხვევითი სიდიდე ნორმალურადაა განაწილებული საშუალოთი, $\mu + \alpha_i$ და დისპერსიით, σ^2 / n_i. ამიტომ გასაგებია, რომ

$$\bar{Y}_i - \bar{Y}_j \cong N(\alpha_i - \alpha_j, \sigma^2 \cdot (1/n_i + 1/n_j)), \qquad (1)$$

ხოლო H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში კი

$$\bar{Y}_i - \bar{Y}_j \cong N(0, \sigma^2 \cdot (1/n_i + 1/n_j)). \qquad (2)$$

(2) თანაფარდობიდან გამომდინარე, ცნობილი σ^2-ის შემთხვევაში, კრიტერიუმის სტატისტიკად ბუნებრივი იქნებოდა აგვეღო

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sigma \cdot \sqrt{(1/n_i + 1/n_j)}}$$

შემთხვევითი სიდიდე, რომელსაც ცხადია ექნება სტანდარტული ნორმალური განაწილება. მაგრამ, როგორც წესი, σ — უცნობია. ამიტომ σ-ს ადგილას უნდა ჩავსვათ σ-ს შეფასება. ისევე როგორც ორამოკრეფიან ამოცანაში ტოლი, უცნობი დისპერსიების შემთხვევაში საშუალოთა ტოლობის

t-კრიტერიუმის აგებისას, აქაც დისპერსია შევაფასოთ გაერთიანებული შერჩევის მიხედვით (შეცდომის შემცირების მიზნით) შემდეგი სიდიდით:

$$S^2_{n_i,n_j} = \frac{(n_i-1)S^2_{1n_i} + (n_j-1)S^2_{1n_j}}{n_i+n_j-2}.$$

როგორც ვხედავთ, ორი ამოკრეფის შემთხვევაში ჩვენ საქმე გვაქვს თავისუფლების ხარისხებით შეწონილ შერჩევით დისპერსიებთან. ანალოგიურად მოვიქცეთ k ცალი ამოკრეფის შემთხვევაშიც: კერძოდ, შევკრიბოთ ჯგუფების თავისუფლების ხარისხებით შეწონილი დისპერსიები და გავყოთ თავისუფლების ხარისხების ჯამზე, ანუ დისპერსიის შესაფასებლად განვიხილოთ შემდეგი სიდიდე:

$$S^2 := S^2_{n_1,n_2,\dots,n_k} = \frac{(n_1-1)S^2_{1n_1} + (n_2-1)S^2_{2n_2} + \dots + (n_k-1)S^2_{kn_k}}{n_1-1+n_2-1+\dots+n_k-1} = \frac{\sum_{i=1}^{k}(n_i-1)\cdot S^2_{in_i}}{\sum_{i=1}^{k} n_i - k}, \qquad (3)$$

საიდანაც WSS-ისა და WMS-ის განმარტების თანახმად ვღებულობთ, რომ:

$$S^2 = WSS/(n-k) = WMS\,.. \qquad (4)$$

შესაბამისად, კრიტერიუმის სტატისტიკას უცნობი დისპერსიის შემთხვევაში ექნება შემდეგი სახე:

$$T_{i,j} = \frac{\overline{Y}_i - \overline{Y}_j}{\sqrt{WMS\cdot(1/n_i + 1/n_j)}}. \qquad (5)$$

მტკიცდება, რომ შემთხვევით სიდიდეს $T_{i,j}$ გააჩნია სტიუდენტის განაწილება თავისუფლების ხარისხით $n-k$ ($t(n-k)$). ამიტომ $H_0:\alpha_i=\alpha_j$ ჰიპოთეზის შესამოწმებელი კრიტერიუმი $H_1:\alpha_i \neq \alpha_j$ ალტერნატივის წინააღმდეგ შემდეგნაირად ჩამოყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის $T_{i,j}$ სტატისტიკის დაკვირვებული $t_{i,j}$ მნიშვნელობა აკმაყოფილებს პირობას

$$t_{n-k,1-\alpha/2} \le t_{i,j} \le t_{n-k,\alpha/2}\,, \qquad (6)$$

მაშინ α მნიშვნელოვნების დონით $H_0:\alpha_i=\alpha_j$ ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში მას უარვყოფთ. შესაბამისი p-მნიშვნელობა ასე გამოითვლება:

$$p = 2\cdot P\{T_{n-k} < t_{i,j}\},\ \text{თუ}\ t_{i,j} < 0\ \text{და}\ p = 2\cdot P\{T_{n-k} > t_{i,j}\},\ \text{თუ}\ t_{i,j} \ge 0. \qquad (7)$$

დავუბრუნდეთ ფილტვების ფუნქციურობის შესახებ ადრე განხილულ მაგალითს და ვნახოთ არის თუ არა სტატისტიკურად მნიშვნელოვანი განსხვავება არამწეველებსა და პასიურ მწეველებს შორის. ამისათვის გამოვთვალოთ $T_{1,2}$ სტატისტიკის მნიშვნელობა და შესაბამისი p-მნიშვნელობა. ცხადია, რომ:

$$t_{1,2} = \frac{3.78-3.3}{\sqrt{0.636\cdot(1/200+1/200)}} = 6.02$$

და

$$p = 2 \cdot P\{t_{1044} > t_{1,2}\} < 0.001 .$$

შესაბამისად, განსხვავება არამწეველებსა და პასიურ მწეველებს შორის სტატისტიკურად მნიშვნელოვანია.

გასაგებია, რომ ანალოგიურად შედარდება დანარჩენი წყვილებიც. სულ ასეთი წყვილების რაოდენობა იქნება $k \cdot (k-1)/2 = 6 \cdot 5/2 = 15$. ყველა წყვილის შედარების შედეგი მოყვანილია შემდეგ ცხრილში:

შესადარებელი წყვილები	სტატისტიკის რიცხვითი მნიშვნელობა	p-მნიშვნელობა, მნიშვნელოვნება
NS, PS	$t_{1,2} = 6.02$	< 0.001
NS, NI	$t_{1,3} = 3.65$	< 0.001
NS, LS	$t_{1,4} = 6.90$	< 0.001
NS, MS	$t_{1,5} = 13.17$	< 0.001
NS, HS	$t_{1,6} = 14.92$	< 0.001
PS, NI	$t_{2,3} = -0.16$	უმნიშვნელოა
PS, LS	$t_{2,} = 0.88$	უმნიშვნელოა
PS, MS	$t_{2,5} = 7.15$	< 0.001
PS, HS	$t_{2,6} = 8.90$	< 0.001
NI, LS	$t_{3,4} = 0.71$	უმნიშვნელოა
NI, MS	$t_{3,5} = 4.68$	< 0.001
NI, HS	$t_{3,6} = 5.79$	< 0.001
LS, MS	$t_{4,5} = 6.27$	< 0.001
LS, HS	$t_{4,6} = 8.03$	< 0.001
MS, HS	$t_{5,6} = 1.76$	უმნიშვნელოა

შენიშვნა. უცნობი დისპერსია σ^2 ჩვენ შევაფასეთ ყველა ჯგუფის მონაცემების გამოყენებით, და არა მხოლოდ შესადარებელი ჯგუფების მონაცემების საშუალებით. ეს აიხსნება იმით, რომ σ^2 ყველა ჯგუფისათვის ჩვენი მოდელის მიხედვით, ერთი და იგივეა და შერჩევის მოცულობის ზრდა იწვევს საშუალოს სტანდარტული შეცდომის შემცირებას. ამიტომ ყველა ჯგუფის გაერთიანებით მიღებული შეფასება უკეთესია, ვიდრე მხოლოდ ორი ჯგუფით მიღებული. ზოგჯერ, როცა ყოველ ჯგუფს თავისი (სხვადასხვა) დისპერსია აქვს, მაშინ წყვილების შესადარებლად უნდა გამოვიყენოთ წყვილთა t-კრიტერიუმი.

ვინაიდან $T_{i,j}$ შემთხვევით სიდიდეს გააჩნია სტიუდენტის $t(n-k)$ განაწილება, ამიტომ ჩვენ შეგვიძლია ავაგოთ ნდობის ინტერვალი ჯგუფების ჭეშმარიტ საშუალოთა $\mu_i - \mu_j$ სხვაობისათვის. მას ექნება შემდეგი სახე:

$$\left(\bar{Y}_i-\bar{Y}_j-t_{n-k,\alpha/2}\cdot S\cdot\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\,;\bar{Y}_i-\bar{Y}_j+t_{n-k,\alpha/2}\cdot S\cdot\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right), \qquad (8)$$

სადაც $S^2 = WSS/(n-k) = WMS$.

3. მრავლობითი შედარება. ბონფერონის მრავლობითი შედარების მეთოდი.

როგორც ვნახეთ, დისპერსიული ანალიზის ჩატარების შემდეგ H_0 ჰიპოთეზის უარყოფის შემთხვევაში საჭირო ხდება წყვილთა შედარება. იმისათვის, რომ გაგვიგოთ კერძოდ რომელ ჯგუფთა გენერალური საშუალოებია განსხვავებული, k ჯგუფის შემთხვევაში, ჩვენ გვჭირდება $k\cdot(k-1)/2$ წყვილის განხილვა (მაგალითად, 6 ჯგუფის შემთხვევაში დაგვჭირდება 15 წყვილის შედარება), რაც ჯგუფების დიდი რაოდენობის შემთხვევაში საკმაოდ შრომატევადი და მოუხერხებელია. ამიტომ კარგი იქნებოდა ისეთი კრიტერიუმის აგება, რომელიც ჯგუფების კონკრეტული წყვილისათვის კი არ "დაიჭერდა განსხვავებას" გენერალურ საშუალოებს შორის, არამედ, ერთდროულად ყველა წყვილისათვის. ასეთი კრიტერიუმები არსებობს და მათ ***მრავლობითი შედარების კრიტერიუმებს*** ეძახიან. ქვემოთ ჩვენ მოვიყვანთ ერთ ასეთ კრიტერიუმს, რომელიც ბონფერონის კრიტერიუმის სახელითაა.

ეს მეთოდი ფორმალურად ძალიან ჰგავს ზემოთ განხილულ წყვილთა შედარების t-კრიტერიუმს, მაგრამ მათ შორის არის აგრეთვე მნიშვნელოვანი განსხვავებაც. ჯერ ერთი, ჩვენ კვლავ გვინდა $H_0:\alpha_i=\alpha_j$ ჰიპოთეზის შემოწმება $H_1:\alpha_i\neq\alpha_j$ ალტერნატივის წინააღმდეგ, მხოლოდ ახლა უკვე არა კონკრეტული i-ური და j-ური ჯგუფისათვის, არამედ ყველა $i\neq j$-სათვის. კრიტერიუმის სტატისტიკა იგივეა:

$$T_{i,j} := \frac{\bar{Y}_i-\bar{Y}_j}{\sqrt{WMS\cdot(1/n_i+1/n_j)}} \cong t(n-k)\,, \qquad (1)$$

მაგრამ, დასახელებული α მნიშვნელოვნების დონისათვის $T_{i,j}$ სტატისტიკის დაკვირვებული მნიშვნელობა $t_{i,j}$ უნდა შედარდეს $t(n-k)$-განაწილების არა ზედა $\alpha/2$ და $1-\alpha/2$-კრიტიკულ წერტილებს, არამედ შესაბამისად, ზედა $\alpha^*/2$ და $1-\alpha^*/2$-კრიტიკულ წერტილებს, სადაც

$$\alpha^* = \frac{\alpha}{k\cdot(k-1)/2} = \frac{2\alpha}{k\cdot(k-1)}.$$

შესაბამისად, ბონფერონის კრიტერიუმი ასე ყალიბდება:

თუ დასახელებული α მნიშვნელოვნების დონისათვის $T_{i,j}$ სტატისტიკის დაკვირვებული $t_{i,j}$ მნიშვნელობა აკმაყოფილებს პირობას

$$t_{n-k,1-\alpha^*/2} \le t_{i,j} \le t_{n-k,\alpha^*/2}\,, \qquad (2)$$

მაშინ α მნიშვნელოვნების დონით $H_0:\alpha_i=\alpha_j$ ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში მას უარვყოფთ.

(ცხადია, რომ $\alpha^* < \alpha$ (ვინაიდან $k\cdot(k-1)/2>1$, თუ $k\geq 3$). შესაბამისად, $\alpha^*/2<\alpha/2$, ხოლო $1-\alpha^*/2>1-\alpha/2$. ამიტომ $t_{n-k,\alpha^*/2}>t_{n-k,\alpha/2}$ და $t_{n-k,1-\alpha^*/2}<t_{n-k,1-\alpha/2}$. აქედან გამომდინარე, გასაგებია, რომ (2) თანაფარდობაში ინტერვალი უფრო ფართეა, ვიდრე წყვილთა შედარების შესაბამის კრიტერიუმში. გარდა ამისა, აქ ინტერვალის საზღვრები დამოკიდებულია k-ზე და k-ს ზრდასთან ერთად უფრო და უფრო ფართოვდება. ცხადია, რომ ეს ნაკლი მოსალოდნელი იყო, რადგან ახალი საზღვრები "პასუხს აგებს" ყველა i და j–ზე, მაშინ, როცა, წყვილების შედარებისას საზღვრები "პასუხისმგებელი" იყო მხოლოდ კონკრეტულ i და j ჯგუფზე.

ბონფერონის მეთოდის გამოყენების საილუსტრაციოდ ისევ განვიხილოთ ფილტრების ფუნქციურობის შესახებ მოყვანილი მაგალითი და გავაკეთოთ შესაბამისი დასკვნები.

ვიგულისხმოთ, რომ $\alpha=0.05$. მაშინ რადგან ჯგუფების რაოდენობა ექვსია, ანუ $k=6$, ამიტომ

$$\alpha^*=\frac{\alpha}{k\cdot(k-1)/2}=\frac{2\alpha}{k\cdot(k-1)}=\frac{2\cdot 0.05}{6\cdot 5}=0.0033.$$

გარდა ამისა, $n-k=1050-6=1044$. ამიტომ უნდა ვიპოვოთ 1044-ის ტოლი თავისუფლების ხარისხის მქონე სტიუდენტის განაწილების ზედა 0.0033 კრიტიკული წერტილი $t_{n-k,\alpha^*/2}=t_{1044,0.0033}$. ვინაიდან 1044 საკმაოდ დიდი რიცხვია, ამიტომ უნდა გამოვიყენოთ ნორმალური აპროქსიმაცია, რომლის თანახმადაც საძიებელი კრიტიკული წერტილი დაახლოებით ტოლია სტანდარტული ნორმალური განაწილების ზედა 0.0033-კრიტიკული წერტილის და შესაბამისად, $t_{n-k,\alpha^*/2}=t_{1044,0.0033}\approx z_{0.0033}=2.935$. დავუბრუნდეთ წინა ნაწილში მოყვანილ წყვილთა შედარების ცხრილს. ყველა სტატისტიკურად მნიშვნელოვან შემთხვევათა შორის აბსოლუტური მნიშვნელობით ყველაზე პატარა შერჩევითი მნიშვნელობაა 3.65 და ისიც კი მეტია 2.935-ზე, რაც იმას ნიშნავს, რომ ისინი ბონფერონის მეთოდის მიხედვითაც რჩებიან სტატისტიკურად მნიშვნელოვნად. უფრო მეტიც, ის წყვილები, რომელთა შორის განსხვავებაც სტატისტიკურად უმნიშვნელოა, ბონფერონის მეთოდის მიხედვითაც არიან სტატისტიკურად უმნიშვნელო. შევნიშნოთ, რომ იმ ცხრილის აგებისას რომ გამოგვეყენებინა $\alpha=0.05$ მნიშვნელოვნების დონე, მაშინ კრიტიკული არე იქნებოდა $|t|>1.96$, მაშინ, როცა როგორც ვნახეთ, ბონფერონის მეთოდის შემთხვევაში, ეს არეა $|t|>2.935$.

4. ჰიპოთეზათა შემოწმება ერთფაქტორიან ANOVA მოდელში შემთხვევითი ეფექტების დროს.

როგორც აღენიშნული იყო, ერთფაქტორიან ANOVA მოდელში – $Y_{ij}=\mu+\alpha_i+X_{ij}$, i–ური ჯგუფის მახასიათებელი α_i ფაქტორი (ეფექტი), შეიძლება იყოს დეტერმინისტული ან შემთხვევითი. აქამდე ჩვენ ვიხილავდით დეტერმინისტული α_i ეფექტების შემთხვევას. ახლა ჩვენ შევისწავლით ერთფაქტორიან ANOVA მოდელს, რომელშიც α_i ეფექტები შემთხვევი-

თი სიდიდეებია, რომლებიც დამოუკიდებელია X_{ij} შემთხვევითი სიდიდეებისაგან და რომლებიც განაწილებულია ნორმალურად საშუალოთი 0 და დისპერსიით σ_A^2. ასეთ მოდელს სახელად ჰქვია ***შემთხვევით ეფექტებიანი, ერთფაქტორიანი ANOVA მოდელი.***

შევნიშნავთ, რომ ამ მოდელში μ საერთო საშუალოს ემატება ნორმალური შემთხვევითი სიდიდე α_i და მაშასადამე, სხვადასხვა დაკვირვებაში გაჩნდება α_i-ის სხვადასხვა შერჩევითი მნიშვნელობა. ამიტომ დაკვირვებად ობიექტებს (ინდივიდებს) შორის განსხვავება დამოკიდებულია ამ შემთხვევითი ეფექტის გაფანტულობის ხარისხზე, ანუ α_i შემთხვევითი სიდიდის σ_A^2 დისპერსიის მნიშვნელობაზე: თუ ის დიდია, მაშინ ინდივიდებს შორის განსხვავება დიდია და თუ ის მცირეა, მაშინ ინდივიდები ფაქტიურად "არ განსხვავდებიან". ეს ეფექტი სწრაფად შესამჩნევი რომ გახდეს, გამოვთვალოთ ახალ მოდელში Y_{ij} შემთხვევითი სიდიდის დისპერსია:

$$DY_{ij} = D\mu + D\alpha_i + DX_{ij} = \sigma_A^2 + \sigma^2 .$$

როგორც ვხედავთ, დეტერმინისტულ ეფექტებიანი მოდელისაგან განსხვავებით, დისპერსიას გაუჩნდა დამატებითი შესაკრები σ_A^2. ამიტომ სტატისტიკური ამოცანა ამ შესაკრების მიმართ ყალიბდება: ნულოვანი ჰიპოთეზა მდგომარეობს იმაში, რომ $H_0:\sigma_A^2=0$, ანუ შემთხვევითი ეფექტები უმნიშვნელოა, ხოლო ალტერნატივა კი – $H_1:\sigma_A^2>0$, ანუ შემთხვევითი ეფექტები მნიშვნელოვანია. ამ მოდელში განიხილავენ ორ შემთხვევას: როცა ყველა ჯგუფის შერჩევითი მოცულობები ტოლია (ე.წ. ***ბალანსირებული შემთხვევა***) და როცა ისინი განსხვავებულია (ე.წ. ***არაბალანსირებული შემთხვევა***).

მტკიცდება, რომ ANOVA-ს ამ მოდელში WMS-ის მათემატიკური ლოდინი σ^2-ის ტოლია:

$$E(WMS)=\sigma^2 . \qquad (1)$$

გარდა ამისა, ბალანსირებულ შემთხვევაში

$$E(BMS)=\sigma^2+m\cdot\sigma_A^2, \qquad (2)$$

სადაც $m:=n_1=n_2=\cdots=n_k$, ხოლო არაბალანსირებულ შემთხვევაში

$$E(BMS)=\sigma^2+n_0\cdot\sigma_A^2, \qquad (3)$$

სადაც

$$n_0=\frac{1}{n(k-1)}\cdot\left(n^2-\sum_{i=1}^{k} n_i^2\right), \qquad (4)$$

სადაც $n:=n_1+n_2+\cdots+n_k$.

ცხადია, რომ თუ ყველა $n_i=m$-ს, მაშინ (4) ტოლობიდან მივიღებთ, რომ $n_0=m$, სხვა სიტყვებით რომ ვთქვათ, (3) წარმოადგენს (2)-ის განზოგადებას. ზოგად შემთხვევაში, მტკიცდება, რომ $n_0<\overline{n}:=n/k$. თუმცა, ჩვეულებრივ, ამ სიდიდეებს შორის განსხვავება მცირეა. კრიტერიუმის სტატისტ-

იკად ისევე ვიყენებთ $F = BMS/WMS$ სტატისტიკას, რომელსაც აქვს ფიშერის განაწილება თავისუფლების ხარისხებით $k-1$ და $n-k$ – $F(k-1,n-k)$.
რაც შეეხება σ^2-სა და σ_A^2-ს, მათთვის იგება შეფასებები. ამ შეფასებების ასაგებად, ვიყენებთ (1) და (2) ტოლობებს ბალანსირებულ და (1) და (3) ტოლობებს არაბალანსირებულ შემთხვევაში. მართლაც, თუ შევხედავთ ფორმულათა ამ წყვილებს, როგორც ორუცნობიან განტოლებათა სისტემებს σ^2-ისა და σ_A^2-ის მიმართ, მაშინ ამოხსნის შემდეგ მივიღებთ, რომ:

$\sigma^2 = E(WMS)$ და $\sigma_A^2 = E(BMS-WMS)/m$ ბალანსირებულ შემთხვევაში და

$\sigma^2 = E(WMS)$ და $\sigma_A^2 = E(BMS-WMS)/n_0$ არაბალანსირებულ შემთხვევაში.

ამიტომ ორივე შემთხვევაში, σ^2-ის შეფასებად ავიღებთ $\overline{\sigma}^2 = WMS$-ს და σ_A^2-ის შეფასებად კი $\overline{\sigma}_A^2 = \max((BMS-WMS)/m, 0)$-ს ბალანსირებულ შემთხვევაში, ხოლო $\overline{\sigma}_A^2 = \max((BMS-WMS)/n_0, 0)$-ს არაბალანსირებულ შემთხვევაში.

საბოლოოდ, სტატისტიკური კრიტერიუმი ისევე ყალიბდება, როგორც დეტერმინისტული ეფექტების დროს:

თუ დასახელებული α მნიშვნელოვნების დონისათვის სტატისტიკის დაკვირვებული $f = bms/wms$ მნიშვნელობისათვის სრულდება პირობა

$$f = bms/wms > F_{k-1,n-k,\alpha}$$

მაშინ H_0 ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს. შესაბამისი p-მნიშვნელობა ტოლია $p = P\{F_{k-1,n-k} > f\}$.

კრიტერიუმის გამოყენების საილუსტრაციოდ გავარჩიოთ შემდეგი მაგალითი: ცხრილში მოცემულია პლაზმა ესტრადიოლის (plasma estradiol) მენოპაუზის შემდგომი ლოგარითმების მნიშვნელობები ხუთი ქალის (გადიის) სისხლში. ექსპერიმენტი ტარდებოდა შემდეგნაირად: ყოველი სისხლის სინჯი დაიყო ორ ტოლ ნაწილად და ათივე სინჯი გაიგზავნა ანალიზისათვის ლაბორატორიაში, ისე, რომ იქ არ სცოდნოდათ, თუ ვისი სისხლის ანალიზს აკეთებდნენ. მიიღეს შემდეგი შედეგები:

ინდივიდი	სინჯი 1	სინჯი 2	სხვაობის მოდული	საშუალოები
1	25.5	30.4	4.9	27.95
2	11.1	15.0	3.9	13.05
3	8.0	8.1	0.1	8.05
4	20.7	16.9	3.8	18.80
5	5.8	8.4	2.6	7.10

შეიძლება თუ არა ინდივიდუალური და ინდივიდებს შორის ვარიაციების სიდიდეების შეფასება ამ მონაცემებით?

ამოხსნა. გამოთვლების შედეგად მივიღებთ: $bss = 2.65775$, $wss = 0.15$, $bms = 2.65775/(9-5) = 0.66444$, $wms = 0.15/5 = 0.3$ და $f = 0.66444/0.03 = 22.15$.
ფიშერის $F(4,5)$ განაწილების ცხრილებიდან ვპოულობთ, რომ

$$P\{F_{4,5} > 22.15\} < 0.0022,$$

საიდანაც ვასკვნით, რომ დაკვირვებადი სიდიდის ინდივიდუალური საშუალო მნიშვნელობებს შორის განსხვავება სტატისტიკურად მნიშვნელოვანია. თუ გავიხსენებთ σ_A^2-ის $\overline{\sigma}_A^2 = \max((BMS - WMS)/m, 0)$, მივიღებთ, რომ $\overline{\sigma}_A^2 = \max((0.6644 - 0.03)/2, 0) = 0.317$ და მაშასადამე, ინდივიდებს შორის ვარიაციების სიდიდე დაახლოებით 10–ჯერ მეტია ინდივიდუალური (ინდივიდში) ვარიაციის $wms = 0.03$ სიდიდეზე.

ამოცანა 1. ცხრილში მოცემულია პროტეინის მიღების სიდიდეები ქალებში, რომლებიც იცავენ სამი ტიპის დიეტას, მენოპაუზის შემდგომ პერიოდში.

დიეტის ტიპი	საშუალო	სტანდ. გადახრა	მოცულობა
STD	75	9	10
LAC	57	13	10
VEG	47	17	6

ა). შეადარეთ ჯგუფების საშუალოები კრიტიკული მნიშვნელობებით;
ბ). გააკეთეთ იგივე p–მნიშვნელობაზე დაყრდნობით;
გ). შეადარეთ საშუალოები ჯუფთა ყოველი წყვილისათვის t–რიტერიუმით
დ). მრავლობითი შედარების მეთოდით განსაზღვრეთ რომელი პოპულაციების საშუალოებია განსხვავებული.

ამოცანა 2. ჩატარდა გამოკვლევა იმის დასდგენად, თუ რომელი მეთოდით ჯობია სისხლის წნევის განსაზღვრა, სტანდარტულით, თუ ახლით. გაზომვები აღირიცხა 4 სხვადასხვა ადგილზე და მიიღეს შემდეგი შედეგები:

	გაზომვის ახალი მეთოდი			სტანდარტული მეთოდი		
ადგილი	საშუალო	გადახრა	მოცულ	საშუალო	გადახრა	მოცულ
A	142.5	21.0	98	142.0	18.1	98
B	134.1	22.5	84	133.6	23.2	84
C	147.9	20.3	98	133.9	18.3	98
D	135.4	16.7	62	128.5	19.0	62

ა). რომელი მიდგომა უფრო მიესადაგება ამ ამოცანას, დეტერმინიტული, თუ შემთხვევითი ეფექტების?
ბ). შეამოწმეთ არის თუ არა მნიშვნელოვანი განსხვავება ოთხივეგან.

თავი IX
ჰიპოთეზათა შემოწმება. კატეგორული მონაცემები

შესავალი.

ზემოთ ჩვენ განვიხილეთ ჰიპოთეზათა შემოწმების ამოცანები ერთი, ორი და მეტი შერჩევისათვის, სადაც ძირითადად დაშვებული იყო, რომ მონაცემების შესაბამისი პოპულაციები ნორმალურია (კერძოდ, ისინი უწყვეტი ტიპისაა). ახლა ჩვენ გადავდივართ დისკრეტულ მოდელებში ჰიპოთეზათა შემოწმების საკითხებზე ერთზე მეტ ამოკრეფიან ამოცანებში. ამასთანავე განვიხილავთ არა მხოლოდ რიცხვით, არამედ ე.წ. კატეგორულ მონაცემებსაც, რომლებიც თავისი ბუნებით სწორედაც რომ დისკრეტულები არიან და არა უწყვეტი. დისკრეტული ტიპის შემთხვევით სიდიდეებს შორის უმნიშვნელოვანესია ბინომური შემთხვევითი სიდიდე. ამიტომ, ქვემოთ ჩვენ შევისწავლით ორი და მეტი ბინომური პოპულაციის შედარების საკითხს.

მაგალითი. იმის დასადგენად, დამოკიდებულია თუ არ მკერდის კიბო ქალებში მათი პირველი მშობიარობის ასაკზე ჩატარდა საერთაშორისო გამოკვლევა, რომელშიც მონაწილეობდა აშშ, საბერძნეთი, იუგოსლავია, ბრაზილია ტაივანი და იაპონია. ქალებს, მიუხედავად იმისა, ჰქონდათ მათ ეს დაავადება, თუ – არა, ეკითხებოდნენ მათი პირველი მშობიარობის ასაკს. აღმოჩნდა, რომ 683 ქალს 3220-დან, რომელსაც ჰქონდა მკერდის კიბო, პირველი მშობიარობა ჰქონდათ 30-ზე მეტი წლის ასაკში, ხოლო იმ ქალებს შორის, რომლებსაც ეს დაავადება არ ჰქონიათ, 1498 ქალის პირველი მშობიარობის ასაკი იყო 30 წელზე მეტი 10245 შემთხვევიდან. არის თუ არა მნიშვნელოვანი განსხვავება ამ მონაცემებს შორის?

ამ მაგალითიდან ჩანს, რომ დაკვირვებადი სიდიდეები წარმოადგენენ გარკვეული ხდომილობების მოხდენათა რაოდენობებს და გასაგებია, რომ მათ უკან დისკრეტული შემთხვევითი სიდიდეები დგას. ჩვენი მიზანია გავარკვიოთ როგორია ეს შემთხვევითი სიდიდეები, როგორი განაწილება შეიძლება მათ ჰქონდეს და როგორ ჩამოვაყალიბოთ და შემოწმოთ შესაბამისი ჰიპოთეზები?

1. ორამოკრაფიანი ამოცანები ბინომური პროპორციებისათვის.

დავუბრუნდეთ მაგალითს მკერდის კიბოს მშობიარობის ასაკთან კავშირის შესახებ. ალბათობა იმისა, რომ შემთხვევით ამორჩეული ქალის ასაკი პირველი მშობიარობისას მეტია ან ტოლი 30 წელზე დაავადებულთა და ჯანმრთელთა შორის შესაბამისად, აღვნიშნოთ p_1-ითა და p_2-ით. ჩვენ გვაინტერესებს არის თუ არა p_1 და p_2 ერთმანეთის ტოლი? ცხადია, რომ ის რაოდენობები, რომლებიც აღწერილია მაგალითში (683 და 1498), შეიძლება სულ სხვა ყოფილიყო, თუ ჩვენ გამოვკითხავდით სხვა 3220 და შესაბამისად, 10245 ქალს. ამიტომ ჩვენ ვუშვებთ, რომ არსებობს შემთხვევითი სიდიდეები S_1 და S_2, რომელთა კონკრეტული გათამაშების შედეგადაც მივიღეთ სწორედ ეს რიცხვები, ანუ $x_1=683$ და $x_2=1498$. ჩვენ გვაინტერესებს როგორაა განაწილებული S_1 და S_2 შემთხვევითი სიდიდეები?

ბუნებრივია დავუშვათ, რომ ყოველი ქალის პირველი მშობიარობის ასაკის მოხვედრა ამა თუ იმ ინტერვალში დანარჩენებისაგან დამოუკიდებელია. მაშინ ცხადია, რომ S_1 და S_2 სიდიდეები წარმოადგენს ბინომურად განაწილებულ შემთხვევით სიდიდეებს ცდათა სხვადასხვა რიცხვით (ამ შემთხვევაში 3220-ითა და 10245-ით, საზოგადოდ, კი, n_1-ითა და n_2-ით). გარდა ამისა, ბუნებრივია, ვიგულისხოთ, რომ S_1 და S_2 დამოუკიდებელი შემთხვევითი სიდიდეებია. მაშინ ნულოვანი და ალტერნატიული ჰიპოთეზა ასე ყალიბდება: $H_0: p_1 = p_2$ და $H_1: p_1 \neq p_2$.

ცნობილია, რომ ბინომური განაწილების p პარამეტრის საუკეთესო წერტილოვან შეფასებას წარმოადგენს ფარდობითი სიხშირე, $\overline{p}_n = S/n$. ამ შემთხვევაში გვექნება: $\overline{p}_{1n_1} = S_1/n_1$ და $\overline{p}_{2n_2} = S_2/n_2$. ვინაიდან ამ სიდიდეების სიახლოვე p_1-ისა და p_2-ის სიახლოვის მაჩვენებელია, ამიტომ კრიტერიუმის სტატისტიკა უნდა აიგოს ამ სიდიდეების საშუალებით. განვიხილოთ შემთხვევით სიდიდეთა სხვაობა $\overline{p}_{1n_1} - \overline{p}_{2n_2} = S_1/n_1 - S_2/n_2$. ზოგადად, ამ შემთხვევითი სიდიდის განაწილების გამოთვლა საკმაოდ შრომატევადი და რთულია, მაგრამ როცა n_1 და n_2 საკმაოდ დიდია, მაშინ ისინი უშვებენ ნორმალურ აპროქსიმაციას. კერძოდ, როცა $n_1 p_1(1-p_1) \geq 5$ და $n_2 p_2(1-p_2) \geq 5$ მაშინ შეგვიძლია ჩავთვალოთ, რომ

$$\overline{p}_{1n_1} \cong N(p_1, p_1(1-p_1)/n_1) \text{ და } \overline{p}_{2n_2} \cong N(p_2, p_2(1-p_2)/n_2).$$

ცნობილია, რომ დამოუკიდებელ ნორმალურ შემთხვევით სიდიდეთა სხვაობა ისევ ნორმალურია და ამიტომ $H_0: p_1 = p_2$ ჰიპოთეზის სამართლიანობის შემთხვევაში გვაქვს:

$$\overline{p}_{1n_1} - \overline{p}_{2n_2} \cong N(0, p(1-p)\cdot(1/n_1 + 1/n_2)), \qquad (1)$$

სადაც $p := p_1 = p_2$. აქედან გამომდინარე სტანდარტიზაცია გვაძლევს, რომ:

$$(\overline{p}_{1n_1} - \overline{p}_{2n_2})/\sqrt{p(1-p)\cdot(1/n_1 + 1/n_2)} \cong N(0.1). \qquad (2)$$

როგორც ვხედავთ აქ მონაწილეობს უცნობი p პარამეტრი. ამიტომ ვიქცევით სტანდარტული გზით: გადავდივართ მის შეფასებაზე. ჰიპოთეზის მიხედვით, ორივე შერჩევა ამოკრეფილია ერთი და იმავე პოპულაციიდან და ამიტომ შესაძლებელია მათი გაერთიანებით მიღებული შერჩევის საფუძველზე შევაფასოთ p პარამეტრი (როგორც ცნობილია შერჩევის მოცულობის ზრდა იწვევს სტანდარტული შეცდომის შემცირებას) შემდეგი სიდიდით:

$$\overline{p}_{n_1,n_2} = \frac{S_1 + S_2}{n_1 + n_2}. \qquad (3)$$

საბოლოოდ, ჩვენ მივდივართ შემდეგ სტატისტიკამდე:

$$T_{n_1,n_2} := (\overline{p}_{1n_1} - \overline{p}_{2n_2})/\sqrt{\overline{p}_{n_1,n_2}\cdot(1-\overline{p}_{n_1,n_2})\cdot(1/n_1 + 1/n_2)} \cong N(0,1), \qquad (4)$$

ხოლო კრიტერიუმი შემდეგნაირად ყალიბდება:

თუ დასახელებული α მნიშვნელოვნების დონისათვის T_{n_1,n_2} სტატისტიკის დაკვირვებული t_{n_1,n_2} მნიშვნელობა აკმაყოფილებს პირობას

$$-z_{\alpha/2} \le t_{n_1,n_2} \le z_{\alpha/2}, \tag{5}$$

მაშინ α მნიშვნელოვნების დონით $H_0: p_1 = p_2$ ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს, წინააღმდეგ შემთხვევაში, მას უარვყოფთ. შესაბამისი p-მნიშვნელობა გამოითვლება ფორმულით

$$p = 2\cdot(1-\Phi(t_{n_1,n_2})),\ \text{თუ}\ t_{n_1,n_2} \ge 0\ \text{და}\ p = 2\cdot\Phi(t_{n_1,n_2}),\ \text{თუ}\ t_{n_1,n_2} < 0. \tag{6}$$

დავუბრუნდეთ მაგალითს და შევამოწმოთ სრულდება თუ არა ნორმალური მიახლოებისათვის საჭირო პირობები $\overline{p}_{n_1,n_2}$-სათვის. ჯერ გამოვთვალოთ $\overline{p}_{n_1,n_2}$ შემთხვევითი სიდიდის დაკვირვებული რიცხვითი მნიშვნელობა:

$$\overline{p}_{n_1,n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{683+1498}{3220+10245} = \frac{2181}{13465} = 0.162.$$

ამიტომ გვაქვს:

$$n_2 \cdot \overline{p}_{n_1,n_2}\cdot(1-\overline{p}_{n_1,n_2}) \ge n_1 \cdot \overline{p}_{n_1,n_2}\cdot(1-\overline{p}_{n_1,n_2}) = 3220\cdot 0.162 \cdot 0.838 = 437.1 \ge 5.$$

შესაბამისად, ნორმალური მიახლოება შესაძლებელია.

გამოვთვალოთ ახლა T_{n_1,n_2} სტატისტიკის დაკვირვებული მნიშვნელობა:

$$t_{n_1,n_2} = \frac{683/3220 - 1498/10245}{\sqrt{0.162\cdot 0.838\cdot(1/3220+1/10245)}} = 8.9.$$

ვინაიდან $t_{n_1,n_2} = 8.9 > 0$, ამიტომ კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია $p = 2\cdot(1-\Phi(t_{n_1,n_2})) = 2\cdot(1-\Phi(8.9)) = 2\cdot(1-0.99999...)) < 0.001$ და შესაბამისად, $H_0: p_1 = p_2$ ჰიპოთეზას უარვყოფთ $\alpha = 0.001$ მნიშვნელოვნების დონის შემთხვევაშიც კი. მაშასადამე, პირველი მშობიარობა 30-ზე მეტი წლის ასაკში ჰქონდათ მკერდის კიბოთი დაავადებული ქალების უფრო მაღალ პროცენტს, (683 / 3220) 100% = 21.2%, ვიდრე კიბოს დაავადების არ მქონე ქალებს, (1498/10245)·100% = =14.6%.

თუ ძირითადი $H_0: p_1 = p_2$ ჰიპოთეზის ალტერნატივად განვიხილავთ $H_1: p_1 > p_2$ (შესაბამისად, $H_1: p_1 < p_2$) ჰიპოთეზას, მაშინ კრიტიკული არე იქნება $U_1 = \{t_{n_1,n_2} > z_{\alpha/2}\}$ (შესაბამისად, $U_1 = \{t_{n_1,n_2} < -z_{\alpha/2}\}$), სადაც, ისევე როგორც ზემოთ, $z_{\alpha/2}$ სტანდარტული ნორმალური განაწილების ზედა $\alpha/2$-კრიტიკული წერტილია.

ახლა ავაგოთ α მნიშვნელოვნების დონის მქონე ნდობის ინტერვალი დამოუკიდებელ პოპულაციათა უცნობი ალბათობების სხვაობისათვის $p_1 - p_2$. განვიხილოთ სტატისტიკა

$$\overline{T}_{n_1,n_2} := \frac{(\overline{p}_{1n_1} - \overline{p}_{2n_2}) - (p_1 - p_2)}{\sqrt{\dfrac{\overline{p}_{1n_1}\cdot(1-\overline{p}_{1n_1})}{n_1} + \dfrac{\overline{p}_{2n_2}\cdot(1-\overline{p}_{2n_2})}{n_2}}}.$$

მტკიცდება, რომ $\overline{T}_{n_1,n_2}$ შემთხვევით სიდიდეს მიახლოებით სტანდარტული ნორმალური განაწილება აქვს, როცა პოპულაციათა მოცულობები დიდია. ამიტომ $\overline{T}_{n_1,n_2}$ სტატისტიკის გამოყენებით სტანდარტული გზით იგება ასიმპტოტური ნდობის ინტერვალი პოპულაციათა პროპორციების სხვაობისათვის $p_1 - p_2$.

α მნიშვნელოვნების ნდობის ინტერვალი ($p_1 - p_2$) სხვაობისათვის:

$$((\overline{p}_{1n_1} - \overline{p}_{2n_2}) - z_{\alpha/2}\sqrt{\frac{\overline{p}_{1n_1}\cdot\overline{q}_{1n_1}}{n_1} + \frac{\overline{p}_{2n_2}\cdot\overline{q}_{2n_2}}{n_2}}, (\overline{p}_{1n_1} - \overline{p}_{2n_2}) + z_{\alpha/2}\sqrt{\frac{\overline{p}_{1n_1}\cdot\overline{q}_{1n_1}}{n_1} + \frac{\overline{p}_{2n_2}\cdot\overline{q}_{2n_2}}{n_2}}),$$

სადაც $\overline{p}_{1n_1} = S_1/n_1$, $\overline{p}_{2n_2} = S_2/n_2$, $\overline{q}_{1n_1} = 1 - \overline{p}_{1n_1}$ და $\overline{q}_{2n_2} = 1 - \overline{p}_{2n_2}$

2. ფიშერის ზუსტი კრიტერიუმი.

ზემოთ ჩვენ განვიხილეთ ჰიპოთეზათა შემოწმებისა და ნდობის ინტერვალის აგების მეთოდები ნორმალური აპროქსიმაციის საშუალებით. ახლა გავარკვიოთ როგორ მოვიქცეთ იმ შემთხვევაში, როცა $n_1\overline{p}_{1n_1}(1-\overline{p}_{1n_1}) \geq 5$ ან $n_2\overline{p}_{2n_2}(1-\overline{p}_{2n_2}) \geq 5$ პირობა არ სრულდება? განვიხილოთ მაგალითი: იმის გასარკვევად, მოქმედებს თუ არა ადამიანის გულსისხლძარღვთა სისტემის (CVD) დაავადებით გარდაცვალებაზე ე.წ. "მლაშე" დიეტა, მოპოვებულ იქნა ცხრილში წარმოდგენილი მონაცემები:

გარდაცვალების მიზეზი	დიეტის ტიპი: მლაშე	დიეტის ტიპი: არა მლაშე	
არა CVD	$a = 2$	$b = 23$	$a + b = 25 = n_1$
CVD	$c = 5$	$d = 30$	$c + d = 35 = n_2$
	$a + c = 7 = m_1$	$b + d = 53 = m_2$	$n_1 + n_2 = 60 = n$

არის თუ არა დიეტის ტიპსა და CVD –თი გარდაცვალებას შორის კავშირი, თუ ისინი დამოუკიდებელია?

ამ შემთხვევაში გვაქვს:

$$n_1\overline{p}_{1n_1}(1-\overline{p}_{1n_1}) = 25\cdot(2/25)\cdot(1-2/25) = 46/25 = 1.84$$

და

$$n_2\overline{p}_{2n_2}(1-\overline{p}_{2n_2}) = 35\cdot(5/35)(1-5/35) = 30/7 \approx 4.33.$$

როგორც ვხედავთ, ორივე ეს რიცხვი 5-ზე ნაკლებია, ამიტომ ნორმალურ აპროქსიმაციაზე აგებული კრიტერიუმი აქ არ გამოდგება. ასეთ შემთხვევაში იყენებენ ფიშერის ზუსტ კრიტერიუმს, რომელიც გულისხმობს

2×2 ცხრილში დაკვირვებული მნიშვნელობების მიღების ალბათობის ზუსტი მნიშვნელობის გამოთვლას. ამ ალბათობების ერთობლიობა წარმოადგენს ე. წ. *ჰიპერგეომეტრიული განაწილებას* (თუ ორი ტიპის ობიექტების საერთო რაოდენობაა N, მათ შორის I ტიპისაა M, ხოლო II ტიპის – $N-M$, მაშინ ალბათობა იმისა, რომ შემთხვევით შერჩეულ, დაბრუნების გარეშე, n ობიექტს შორის იქნება ზუსტად m I ტიპის ობიექტი ტოლია:

$$P\{X=m\}=\frac{C_M^m\cdot C_{N-M}^{n-m}}{C_N^n},\ m=0,1,...,\min(M,n).$$

ამ ალბათობათა ერთობლიობას უწოდებენ ***ჰიპერგეომეტრიული განაწილებას.*** ამ განაწილების მახასიათებლებია:

$$EX=\frac{n\cdot M}{N}\ \text{და}\ DX=\frac{n\cdot M\cdot(N-M)\cdot(N-n)}{N^2\cdot(N-1)}).$$

ამ მაგალითში:

$$P\{X=a\}=\frac{C_{m_1}^a\cdot C_{n_2}^{m_1-a}}{C_{n_1+n_2}^{m_1}},\ EX=\frac{n_1\cdot m_1}{n}\ \text{და}\ DX=\frac{n_1\cdot n_2\cdot m_1\cdot m_2}{n^2\cdot(n-1)}.$$

გამოთვლებისათვის უფრო მოხერხებულია გამოსახულება:

$$P\{X=a\}=\frac{(a+b)!\cdot(a+c)!\cdot(b+c)!\cdot(b+d)!}{a!\cdot b!\cdot c!\cdot d!\cdot n!}. \qquad (1)$$

მოვიყვანოთ ფიშერის კრიტერიუმი. ნულოვანი ჰიპოთეზა ასე ყალიბდება: $H_0: p_1=p_2$, სადაც p_1 და p_2 შესაბამისად, პირველი და მეორე სტრიქონის პროპორციებია. ალტერნატივა შეიძლება იყოს ორმხრივი ან ცალმხრივი (მარცხენა და მარჯვენა): ამიტომ ალტერნატიული ჰიპოთეზისათვის გვაქვს შემდეგი სამი შემთხვევა:

1). $H_0: p_1\neq p_2$, 2). $H_1: p_1<p_2$, 3). $H_1: p_1>p_2$

ფიშერის ზუსტი კრიტერიუმი ძალიან ჰგავს ერთამოკრეფიანი ბინომური პროპორციისა და პუასონის ინტენსივობისთვის მარტივი ჰიპოთეზის შემოწმების ზუსტ კრიტერიუმებს, დამყარებულს p–მნიშვნელობის გამოთვლაზე. დავუშვათ, ცხრილის (1,1) უჯრედში დგას a რიცხვი. მაშინ ალტერნატივების მიხედვით p–მნიშვნელობა შემდეგნაირად ითვლება:

1). $p=2\cdot\min(P\{X\le a\},P\{X\ge a\},1/2)=2\cdot\min(P\{X\le a\},1-P\{X\le a-1\},1/2)$;

2). $p=P\{X\le a\}$;

3). $p=P\{X\ge a\}=1-P\{X\le a-1\}$.

სამივე შემთხვევაში სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ p–მნიშვნელობა საკმარისად მცირეა (მაგალითად, მცირეა 0.05-ზე, 0.01-ზე და ა.შ.), მაშინ შედეგი სტატისტიკურად მნიშვნელოვანია და მასასადამე, $H_0: p_1=p_2$ ჰიპოთეზა უნდა უარვყოთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს.

ამ კრიტერიუმის გამოყენება ცხადია, მაშინაცაა შესაძლებელი, როცა ნორმალური აპროქსიმაცია დასაშვებია, მაგრამ ვინაიდან მისი ერთადერთი სირთულე ალბათობათა გამოანგარიშებაშია, ამიტომ ამჯობინებენ ნორმალურ აპროქსიმაციას, როცა ის შესაძლებელია, მით უმეტეს, რომ

ასეთ შემთხვევაში ორივე გზით მიღებული შედეგები საკმარისად ახლოსაა ერთმანეთთან.

ცხადია, რომ $P\{X \le a\} = P\{X=0\} + P\{X=1\} + \cdots + P\{X=a\}$. გამოთვლების გამარტივების მიზნით, შევნიშნოთ, რომ სამართლიანია შემდეგი რეკურენტული თანაფარდობა:

$$P\{X=k+1\} = P\{X=k\} \cdot \frac{n_1-k}{k+1} \cdot \frac{m_1-k}{n_2-m_1+k+1}. \qquad (2)$$

ამიტომ ჩვენს მაგალითში გვაქვს:

$$P\{X=1\} = P\{X=0\} \cdot \frac{25-0}{0+1} \cdot \frac{7-0}{35-7+0+1} = 6.034 \cdot P\{X=0\},$$

$$P\{X \le 1\} = P\{X=0\} + P\{X=1\} = P\{X=0\} + 6.034 \cdot P\{X=0\} = 7.034 \cdot P\{X=0\},$$

$$P\{X=2\} = P\{X=1\} \cdot \frac{25-1}{1+1} \cdot \frac{7-1}{35-7+1+1} = \cdots = 14.482 \cdot P\{X=0\}.$$

საიდანაც გვაქვს:

$$P\{X \le 2\} = 21.517 \cdot P\{X=0\}.$$

მეორეს მხრივ, (1) ფორმულის თანახმად:

$$P\{X=0\} = \frac{C_{25}^0 \cdot C_{35}^{7-0}}{C_{60}^7} = \frac{35! \cdot 53!}{28! \cdot 60!} = \frac{527}{30267} \approx 0.0174$$

ამიტომ საბოლოოდ გვაქვს:

$$P\{X \le 2\} = 21.517 \cdot P\{X=0\} = 21.517 \cdot 0.0174 \approx 0.3744,$$

$$P\{X \ge 2\} = 1 - P\{X \le 1\} = 1 - 7.034 \cdot P\{X=0\} \approx 0.878,$$

$$p = 2 \cdot \min(0.3744, 0.878, 1/2) = 2 \cdot 0.3744 = 0.7488.$$

აქედან ვასკვნით, რომ p_1 და p_2 პროპორციები მნიშვნელოვნად არ განსხვავდებიან და შესაბამისად, ჩვენ არ შეგვიძლია იმის თქმა, რომ მლაშე დიეტასა და სიკვდილის მიზეზს შორის მნიშვნელოვანი კავშირია.

3. მაკნემარის კრიტერიუმი პროპორციებისათვის დაწყვილებულ მონაცემებში.

ჩვენ ახლა გვაინტერესებს პროპორციების შედარების საკითხი, მხოლოდ არა დამოუკიდებელი პოპულაციებისათვის, არამედ დაწყვილებული მონაცემებისათვის. სხვა სიტყვებით რომ ვთქვათ, ბერნულის შემთხვევითი სიდიდეები, $X_1,\ldots,X_n$ და $Y_1,\ldots,Y_n$ (რომელთა შეკრებითაც მიიღება დაკვირვებული S_1 და S_2 ბინომური შემთხვევითი სიდიდეები, $S_1 = X_1+\cdots+X_n$, $S_2 = Y_1+\cdots+Y_n$), აღარ არიან წყვილ-წყვილად დამოუკიდებელი, რის გამოც, აღარაა დამოუკიდებელი თვითონ S_1 და S_2 ბინომური შემთხვევითი სიდიდეები. განვიხილოთ მაგალითი: უნდოდათ შეედარებინათ მკერდის კიბოს ორი ქიმიოთერაპიული რეჟიმი მასტექტომიის (mastectomy) შემდეგ. ამისათვის შეირჩა ავადმყოფთა ორი ჯგუფი და ისინი დააწყვილეს ისეთნაირად, რომ წყვილების ასაკებს შორის არ ყოფილიყო დიდი სხვაობა (არაუმეტეს 5 წელი) და მათ ჰქონოდათ თითქმის ერთნაირი წინასწარ განსაზღვრული მონაცემები. ამის შემდეგ, ყოველი წყვილის შემთხვევით არჩეულ წევრს უტარდებოდა A ტიპის (პირველი კვირის განმავლობაში მასტექტომიიდან

და ასე 6 თვის განმავლობაში), მეორეს კი – B ტიპის ქიმიოთერაპია (მხოლოდ პირველი კვირის განმავლობაში მასტექტომიიდან). ავადმყოფების მკურნალობაზე დაკვირვება მიმდინარეობდა 5 წლის განმავლობაში და მიიღეს შემდეგი მონაცემები:

ქიმიოთერაპიის ტიპი	გარდაიცვალა 5 წლის მანძილზე		
	არა	კი	
A	526	95	621
B	515	106	621
სულ	1041	201	1242

ცხადია, რომ ასეთნაირად შედგენილი შეუღლების ცხრილი A და B ტიპის ქიმიოთერაპიის ეფექტურობაზე ვერაფერს გვეტყვის, რადგან გარდაცვლილთა (ან პირიქით, გადარჩენილთა) რაოდენობები საკმარისად ახლოსაა ერთმანეთთან, უფრო სწორად, ერთმანეთთან ახლოსაა პროპორციათა შეფასებები: $\bar{p}_{1n} = 95/621 = 0.153$ და $\bar{p}_{2n} = 106/621 = 0.171$ გარდა ამისა, შევნიშნოთ, რომ აქ მონაცემთა საერთო რაოდენობაა $n = 1242$, რაც არ შეესაბამება ექსპერიმენტს, რომელშიც სინამდვილეში აკვირდებოდნენ მონაცემთა წყვილებს, რომელთა რაოდენობაა $n = 621$. ამიტომ სწორი დასკვნების გასაკეთებლად საჭიროა შესაბამისი (სწორი) ცხრილის აგება. მას შემდეგი სახე აქვს:

გარდაიცვალა 5 წლის მანძილზე A ტიპის თერაპია	გარდაიცვალა 5 წლის მანძილზე B ტიპის თერაპია		
	არა	კი	
არა	510	16	526
კი	5	90	95
სულ	515	106	621

ამ ცხრილის უჯრედებში ჩაწერილი რიცხვები გვიჩვენებს უკვე იმ წყვილების რაოდენობას, რომლის წევრებიც ხუთი წლის შემდეგ ორივე ცოცხალია ((1,1) უჯრედი), A ტიპის თერაპიით ცოცხალია, B-თი – არა ((1,2) უჯრედი), B ტიპის თერაპიით ცოცხალია, A-თი – არა ((2,1) უჯრედი) და წყვილის ორივე წევრი გარდაიცვალა ((2,2) უჯრედი).

რადგან შერჩევები არაა დამოუკიდებელი, პროპორციათა შესადარებლად χ^2 –კრიტერიუმი არ გამოდგება. ასეთ შემთხვევებში იყენებენ ***მაკნემარის კრიტერიუმს***. ამ კრიტერიუმის აღსაწერად ხმარობენ ასეთ ტერმინებს: წყვილს, რომელშიც აღირიცხა ერთიდაიგივე შედეგი შეთანხმებული, ანუ *კონკორდანტული* (concordant), ხოლო წყვილს, სადაც აღირიცხა სხვადასხვა შედეგი – შეუთანხმებელი, ანუ *დისკორდანტული* (discordant) წყვილი ეწოდება. ცხადია, რომ კონკორდანტული წყვილები არაფერს გვეუბნება მკურნალობის ტიპის ეფექტურობის შესახებ და ამიტომაც ისინი არ

მონაწილეობს კრიტერიუმის სტატისტკის გამოსახულებაში, და პირიქით, სწორედ დისკორდანტულ წყვილებში აღმოჩენილი განსხვავება მეტყველებს მკურნალობის ეფექტურობაზე და ამიტომაც მაკნემარის სტატისტიკაც მათზეა აგებული. შევნიშნოთ, რომ დისკორდანტული წყვილები ორნაირია, პირობითად, (A,B^C) და (A^C,B). გასაგებია, რომ თუ ამ ორი ტიპის წყვილების რაოდენობები ახლოსაა ერთმანეთთან, მაშინ ძნელი უნდა იყოს იმის გარჩევა თუ მკურნალობის რომელი ტიპი რომელს სჯობია. ფარდობითი სიხშირეების ტერმინებში ეს იმას ნიშნავს, რომ თითოეული დისკორდანტული ტიპის წყვილების ფარდობითი სიხშირე ახლოსაა 0.5-თან. ამიტომაც ჰიპოთეზა და ალტერნატივა სწორედ დისკორდანტული ტიპის წყვილების ალბათობის ტერმინებში გამოითქმება.

აღვნიშნოთ (A,B^C) ტიპის დისკორდანტული შემთხვევის მოხდენის ალბათობა p-თი. მაშინ ნულოვანი ჰიპოთეზა ასე ყალიბდება: $H_0: p=0.5$, ხოლო ალტერნატივაა $H_1: p\neq 0.5$. აღვნიშნოთ n_D-თი დისკორდანტული წყვილების საერთო რაოდენობა, ხოლო $X_{D,A}$-თი დისკორდანტული (A,B^C) ტიპის წყვილების რაოდენობა, რომელიც განაწილებულია ბინომურად პარამეტრებით n_D და p. ამიტომ ცხადია, რომ საზოგადოდ, $EX_{D,A}=n_D\cdot p$ და $DX_{D,A}=n_D\cdot p\cdot(1-p)$, ხოლო $H_0: p=0.5$ ჰიპოთეზის სამართლიანობის შემთხვევაში კი — $EX_{D,A}=n_D/2$ და $DX_{D,A}=n_D/4$ ამიტომ როცა $n_D/4\geq 5$ ანუ $n_D\geq 20$, შეგვიძლია გამოვიყენოთ ბინომური განაწილების ნორმალური აპროქსიმაცია, წინააღმდეგ შემთხვევაში კი — ზუსტი კრიტერიუმი დამყარებული p-მნიშვნელობაზე, როგორც ამას ვაკეთებდით ბინომური განაწილების პარამეტრის შესახებ მარტივი ჰიპოთეზის შემოწმების ერთამოკრეფიან ამოცანაში.

ნორმალური აპროქსიმაციის შემთხვევაში (ანუ როცა $n_D\geq 20$), მაკნემარის (შესწორებულ) კრიტერიუმის სტატისტიკას წარმოადგენს სიდიდე

$$T_D := \frac{(|X_{D,A}-n_D/2|-1/2)^2}{n_D/4}, \qquad (1)$$

რომელსაც ნულოვანი ჰიპოთეზის სამართლიანობის შემთხვევაში გააჩნია ხი-კვადრატ განაწილება თავისუფლების ხარისხით 1 ($\chi^2(1)$). ამიტომ კრიტერიუმი ასე ჩამოყალიბდება:

თუ T_D სტატისტიკის დაკვირვებული $t_D := (|n_A-n_D/2|-1/2)^2/(n_D/4)$ მნიშვნელობისათვის, სადაც n_A აღნიშნავს ცხრილში დისკორდანტული (A,B^C) ტიპის წყვილების რაოდენობას და მოცემული α მნიშვნელოვნების დონისათვის სრულდება პირობა

$$t_D > \chi^2_{1,\alpha}, \qquad (2)$$

მაშინ α მნიშვნელოვნების დონით $H_0: p=0.5$ ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს.

თუ ნორმალური აპროქსიმაციის დაუშვებელია, ანუ $n_D < 20$, მაშინ ვითვლით p-მნიშვნელობას შემდეგნაირად:

$$p = 2 \cdot \sum_{k=0}^{n_A} C_{n_D}^k \cdot 0.5^k \text{, როცა } n_A \le n_D / 2,$$

$$p = 2 \cdot \sum_{k=n_A}^{n_D} C_{n_D}^k \cdot 0.5^k \text{, როცა } n_A > n_D / 2 \qquad (3)$$

და თუ ის α-ზე ნაკლებია, ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს.

განხილულ მაგალითში ცხადია, რომ $n_D = 21 > 20$, ანუ ნორმალური აპროქსიმაცია დასაშვებია. ცხადია, აგრეთვე, რომ $n_A = 5$ და კრიტერიუმის დაკვირვებული მნიშვნელობაა

$$t_D = (|5 - 21/2| - 1/2)^2 / (21/4) = 4.76.$$

ავიღოთ $\alpha = 0.05$. მაშინ ხი-კვადრატ განაწილების ზედა კრიტიკული წერტილების ცხრილებიდან ვპოულობთ, რომ $\chi^2_{1,0.05} = 3.8415$. ვინაიდან $4.76 > 3.8415$, ამიტომ $\alpha = 0.05$ მნიშვნელოვნების დონით $H_0 : p = 0.5$ ჰიპოთეზას უარვყოფთ, ანუ $p \neq 0.5$. გასაგებია ისიც, რომ $p < 0.5$, ანუ A ტიპის ქიმიოთერაპიით მკურნალობა ჯობია B-თი მკურნალობას.

თავი X
თანხმობის კრიტერიუმები

აქამდე ჩვენ ვიხილავდით ჰიპოთეზებს, რომლებშიც გენერალური ერთობლიობის განაწილების კანონი ითვლებოდა, რომ იყო ცნობილი. ახლა ჩვენ შევუდგებით ჰიოპოთეზების შემოწმებას უცნობი განაწილების კანონის სავარაუდო სახის შესახებ, ე. ი. შევამოწმებთ ნულოვან ჰიპოთეზას იმის შესახებ, რომ გენერალური ერთობლიობა განაწილებულია გარკვეული ცნობილი კანონის მიხედვით. ასეთი ჰიპოთეზების შემოწმების სტატისტიკურ კრიტერიუმებს, ჩვეულებრივ, **თანხმობის კრიტერიუმებს** უწოდებენ.

პირსონის კრიტერიუმი (ხი კვადრატ კრიტერიუმი).

პირსონის კრიტერიუმის საშუალებით შესაძლებელია სხვადასხვა განაწილების კანონის შესახებ ჰიპოთეზების შემოწმება.

I. ჰიპოთეზის შემოწმება განაწილების ნორმალურობის შესახებ.

ვიგულისხმოთ, რომ მიღებულია საკმარისად დიდი n მოცულობის შერჩევა განსხვავებული ვარიანტების დიდი რიცხვით. მისი დამუშავების მოხერხებულობის მიზნით ვარიანტების უმცირესი მნიშვნელობიდან უდიდეს მნიშვნელობამდე ინტერვალი დავყოთ s ტოლ ნაწილად და ჩავთვალოთ, რომ ვარიანტების მნიშვნელობები, რომლებიც მოხვდნენ ცალკეულ ინტერვალში დაახლოებით ტოლია ამ ინტერვალის შუაწერტილის მომცემი რიცხვის. დავთვალოთ თოთოეულ ინტერვალში მოხვედრილი ვარიანტების რაოდენობა და შევადგინოთ ე. წ. დაჯგუფებული შერჩევა

ვარიანტები	x_1	x_2		x_n
სიხშირე	n_1	n_2		n_k

სადაც x_i – ინტერვალის შუაწერტილის მნიშვნელობაა, ხოლო n_i – ვარიანტების რიცხვია, რომლებიც მოხვდნენ i –ურ ინტერვალში (ემპირიული სიხშირეები).

მიღებული მონაცემებით გამოვთვალოთ შერჩევითი საშუალო $\overline{x}_\text{ə}$ და შერჩევითი საშუალო კვადრატული გადახრა $\sigma_\text{ə}$. შევამოწმოთ წინადადება, რომ გენერალური ერთობლიობა განაწილებულია ნორმალური კანონით პარამეტრებით $E\xi = \overline{x}_\text{ə}$ და $D\xi = \sigma_\text{ə}$. მაშინ ჩვენ შეგვიძლია დავითვალოთ რიცხვების რაოდენობა n მოცულობის შერჩევიდან, რამდენიც უნდა აღმოჩნდეს თითოეულ ინტერვალში ამ დაშვების დროს (ე. ი. თეორიული სიხშირეები). ამ მიზნით, ნორმალური განაწილების ფუნქციის ცხრილიდან ვპოულობთ i –ურ ინტერვალში მოხვედრის ალბათობას:

$$p_i = \Phi\left(\frac{b_i - \overline{x}_\text{ə}}{\sigma_\text{ə}}\right) - \Phi\left(\frac{a_i - \overline{x}_\text{ə}}{\sigma_\text{ə}}\right)$$

სადაც a_i და b_i -- i -ური ინტერვალის საზღვრებია. მიღებული ალბათობების შერჩევის მოცულობაზე გამრავლებით ვპოულობთ თეორიულ სიხშირეებს: $n_i' = n \cdot p_i$. ჩვენი მიზანია – შევადაროთ ემპირიული და თეორიული სიხშ-

ირეები, რომლებიც, რა თქმა უნდა, განსხვავდებიან ერთმანეთისაგან, და გავარკვიოთ, არიან თუ არა ეს განსხვავებები არაარსებითი, რომლებიც არ უარყოფენ ჰიპოთეზას გამოსაკვლევი შემთხვევითი სიდიდის ნორმალური განაწილების შესახებ, ან ეს განსხვავებები იმდენად დიდია, რომ ეწინააღმდეგებიან ამ ჰიპოთეზას. ამ მიზნით გამოიყენება კრიტერიუმი შემდეგი შემთხვევითი სიდიდის სახით

$$\chi^2 = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'} . \qquad (1)$$

ამ კრიტერიუმის აღების აზრი შემდეგში მდგომარეობს: იკრიბება ის წილები, რასაც შეადგენს ემპირიული სიხშირეების თეორიული სიხშირეებისაგან გადახრის კვადრატები, შესაბამისი თეორიული სიხშირეებისაგან. შეიძლება დამტკიცდეს, რომ გენერალური ერთობლიობის რეალური განაწილების კანონისაგან დამოუკიდებლად (1) შემთხვევითი სიდიდის განაწილების კანონი უახლოვდება (მიისწრაფის) χ^2 განაწილებისაკენ თავისუფლების ხარისხით $k = s - 1 - r$, (როცა $n \to \infty$), სადაც r – შერჩევის მონაცემებით შესაფასებული სავარაუდო განაწილების პარამეტრების რაოდენობაა.

ნორმალური განაწილება ხასიათდება ორი პარამეტრით, ამიტომ $k = s - 3$. არჩეული კრიტერიუმისათვის იგება მარჯვენა ცალმხრივი კრიტიკული არე, რომელიც განისაზღვრება უტოლობით

$$p(\chi^2 > \chi^2_{კრ}(\alpha,k)) = \alpha , \qquad (2)$$

სადაც α –მნიშვნელოვნების დონეა. შესაბამისად, კრიტიკული არე მოიცემა უტოლობით $\chi^2 > \chi^2_{კრ}(\alpha,k)$, ხოლო ჰიპოთეზის მიღების არეა – $\chi^2 < \chi^2_{კრ}(\alpha,k)$.

ამრიგად, იმისათვის, რომ შევამოწმოთ ნულოვანი ჰიპოთეზა H_0: გენერალური ერთობლიობა განაწილებულია ნორმალურად – უნდა გამოვთვალოთ შერჩევის მიხედვით კრიტერიუმის დაკვირვებული მნიშვნელობა:

$$\chi^2_{დ} = \sum_{i=1}^{s} \frac{(n_i - n_i')^2}{n_i'} , \qquad (3)$$

ხოლო χ^2 განაწილების კრიტიკული წერტილების ცხრილიდან ვიპოვოთ კრიტიკული წერტილი $\chi^2_{კრ}(\alpha,k)$ ცნობილი α და $k = s - 3$ მნიშვნელობებისათვის. თუ აღმოჩნდა, რომ $\chi^2_{დ} < \chi^2_{კრ}(\alpha,k)$ – ვღებულობთ ნულოვან ჰიპოთეზას, თუ $\chi^2_{დ} > \chi^2_{კრ}(\alpha,k)$, მაშინ – უკუვაგდებთ.

II. ჰიპოთეზის შემოწმება თანაბარი განაწილების შესახებ.

პირსონის კრიტერიუმის გამოყენებისას გენერალური ერთობლიობის თანაბარი განაწილების შესახებ ჰიპოთეზის შემოწმებისას სავარაუდო განაწილების სიმკვრივით

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a,b) \\ 0, & x \notin (a,b) \end{cases}$$

აუცილებელია არსებული შერჩევის მიხედვით გამოვთვალოთ შერჩევითი საშუალო $\bar{x}_{დ}$ და შევაფასოთ a და b პარამეტრები ფორმულებით:

$$a^* = \bar{x}_B - \sqrt{3}\sigma_B, \quad b^* = \bar{x}_B + \sqrt{3}\sigma_B, \tag{4}$$

სადაც a^* და b^* – a-სა და b-ს შეფასებებია. მართლაც, ვინაიდან თანაბარი განაწილებისათვის:

$$E\xi = \frac{a+b}{2}, \quad \sigma(x) = \sqrt{D(X)} = \sqrt{\frac{(a-b)^2}{12}} = \frac{a-b}{2\sqrt{3}},$$

აქედან შეგვიძლია მივიღოთ განტოლებათა სისტემა a^*-სა და b^*-სათვის:

$$\begin{cases} \dfrac{b^*+a^*}{2} = \bar{x}_B \\ \dfrac{b^*-a^*}{2\sqrt{3}} = \sigma_B \end{cases}$$

რომლის ამოხსნასაც წარმოადგენს სწორედ (4) გამოსახულებები.

შემდეგ, ვუშვებთ, რომ $f(x) = \dfrac{1}{b^*-a^*}$ და ვპოულობთ თეორიულ სიხშირეებს ფორმულებიდან:

$$n_1' = np_1 = nf(x)(x_1 - a^*) = n \cdot \frac{1}{b^*-a^*}(x_1 - a^*);$$

$$n_2' = n_3' = \ldots = n_{s-1}' = n \cdot \frac{1}{b^*-a^*}(x_i - x_{i-1}), \quad i = 1,2,\ldots,s-1;$$

$$n_s' = n \cdot \frac{1}{b^*-a^*}(b^* - x_{s-1}).$$

აქ s – იმ ინტერვალების რიცხვია, რამდენ ინტერვალადაც გაიყო შერჩევა.

პირსონის კრიტერიუმის დაკვირვებული მნიშვნელობა გამოითვლება (3) ფორმულიდან, ხოლო კრიტიკული წერტილი $\chi^2_{კრ}(\alpha,k)$ – χ^2 განაწილების კრიტიკული წერტილების ცხრილიდან თავისუფლების ხარისხის $k = s-3$ რიცხვის გათვალისწინებით. ამის შემდეგ ვიქცევით ისე, როგორც წინა შემთხვევაში. კერძოდ, თუ აღმოჩნდა, რომ $\chi^2_{д} < \chi^2_{კრ}(\alpha,k)$ -- ვღებულობთ ნულოვან ჰიპოთეზას, თუ $\chi^2_{д} > \chi^2_{კრ}(\alpha,k)$, მაშინ – უკუვაგდებთ.

III. ჰიპოთეზის შემოწმება მაჩვენებლიანი (ექსპონენციალური) განაწილების შესახებ.

ამ შემთხვევაში მოცემულ შერჩევას ვყოფთ თანაბარი სიგრძის ინტერვალებად და ვიხილავთ ვარიანტების მიმდევრობას $x_i^* = \dfrac{x_i + x_{i+1}}{2}$, რომლებიც თანაბრად დაშორებული არიან ერთმანეთისაგან (ითვლება, რომ ყველა ვარიანტი, რომელიც მოხვდა i-ურ ინტერვალში ღებულობს მნიშვნელობას, რომელიც ემთხვევა ამ ინტერვალის შუაწერტილს), და შესაბამისი n_i სიხშირეების მიმდევრობას (i-ურ ინტერვალში მიხვედრილი ვარიანტების რიცხვი). ამ მონაცემებით გამოვთვალოთ შერჩევითი საშუალო $\bar{x}_B$ და მივიღოთ λ პარამეტრის შეფასებად $\lambda^* = \dfrac{1}{\bar{x}_B}$. მაშინ თეორიული სიხშირეები გამოითვლება ფორმულით

$$n_i' = n_i p_i = n_i p(x_i < X < x_{i-1}) = n_i(e^{-\lambda x_i} - e^{-\lambda x_{i+1}}).$$

შემდეგ პირსონის კრიტერიუმის დაკვირვებული მნიშვნელობა გამოითვლება (3) ფორმულიდან, ხოლო კრიტიკული წერტილი $\chi^2_{კრ}(\alpha,k)$ – χ^2 განაწილების კრიტიკული წერტილების ცხრილიდან თავისუფლების ხარისხის $k = s-2$ რიცხვის გათვალისწინებით.

თუ აღმოჩნდა, რომ $\chi^2_{დ} < \chi^2_{კრ}(\alpha,k)$ -- ვღებულობთ ნულოვან ჰიპოთეზას, თუ $\chi^2_{დ} > \chi^2_{კრ}(\alpha,k)$, მაშინ – უკუვაგდებთ.

IV. ჰიპოთეზის შემოწმება ბინომიალური განაწილების შესახებ.

ჩვენი მიზანია შევამოწმოთ ჰიპოთეზა იმის შესახებ, რომ გენერალური ერთობლიობა განაწილებულია ბინომიალური კანონით $Bi(N,p)$. ავღნიშნოთ ν_i-თი იმ x-ების რაოდენობა $x_1,...,x_n$ შერჩევიდან, რომელთათვისაც $x=i$, $i=0,1,...,N$. ჰიპოთეზის სამართლიანობის შემთხვევაში:

$$p_i = P(X_j = i) = C_N^i p^i (1-p)^{N-i}, i=0,1,...,N; j=1,...,n.$$

ბინომიალური განაწილების ცხრილებიდან მოცემული p-სათვის მიღებული p_i ალბათობებით გამოვიანგარიშოთ პირსონის სტატისტიკის რიცხვითი მნიშვნელობა:

$$\chi^2_n(k) = \frac{1}{n}\sum_{j=1}^{k}\frac{\nu_j^2}{p_j} - n.$$

თუ აღმოჩნდა, რომ $\chi^2_n(k) < \chi^2_{კრ}(\alpha,k-1)$, მაშინ ჰიპოთეზას ვღებულობთ, თუ $\chi^2_n(k) > \chi^2_{კრ}(\alpha,k-1)$, მაშინ – უკუვაგდებთ.

იმ შემთხვევაში, როცა წარმატების p ალბათობა უცნობია, p_i ალბათობების როლში უნდა ავიღოთ მათი შეფასებები:

$$\overline{p}_i = P(X_j = i) = C_N^i \overline{p}^i (1-\overline{p})^{N-i}, \text{ სადაც } \overline{p} = \frac{1}{Nn}\sum_{j=1}^{n} x_j,$$

და, თუ აღმოჩნდა, რომ $\chi^2_n(k) < \chi^2_{კრ}(\alpha,k-2)$ – მივიღოთ ჰიპოთეზა, ხოლო წინააღმდეგ შემთხვევაში კი – უკავაგდოთ.

კოლმოგოროვ-სმირნოვის კრიტერიუმი.

მცირე შერჩევის დროს მიზანშეწონილია ისეთი კრიტერიუმის გამოყენება, რომელიც (განსხვავებით χ^2 კრიტერიუმისაგან) დაეყრდნობა ინდივიდუალურ და არა დაჯგუფებულ მონაცემებს. ერთ-ერთი ასეთი უმნიშვნელოვანესი კრიტერიუმია კოლმოგოროვის კრიტერიუმი. იგი გამოიყენება H_0 ჰიპოთეზის შესამოწმებლად იმის შესახებ, რომ დამოუკიდებელ და ერთნაირად განაწილებულ $X_1, X_2, \ldots, X_n$ შემთხვევით სიდიდეებს გააჩნიათ მოცემული უწყვეტი $F(x)$ განაწილების ფუნქცია. განვიხილოთ ჰიპოთეზა

$H_0: F(x) = F_0(x)$ ორმხრივი ალტერნატივის წინააღმდეგ

$H_1: \max\limits_{|x|<\infty} |F(x) - F_0(x)| > 0.$

განვიხილოთ აგრეთვე ცალემხრივი ალტერნატიული ჰიპოთეზები

$$H_1^+ : \max_{|x|<\infty}(F(x) - F_0(x)) > 0 \text{ და } H_1^- : \max_{|x|<\infty}(F(x) - F_0(x)) < 0 .$$

H_0 ჰიპოთეზის შესამოწმებლად H_1, H_1^+ და H_1^- ალტერნატივების წინააღმდეგ გამოიყენება კოლმოგოროვისა და სმირნოვის კრიტერიუმები, რომელთა შესაბამისი სტატისტიკებია:

$$D_n = \max_{|x|<\infty} | F_n(x) - F(x) |, \quad D_n^+ = \max_{|x|<\infty}(F_n(x) - F(x)) \text{ და } D_n^- = -\min_{|x|<\infty}(F_n(x) - F(x)) .$$

ვიპოვოთ ემპირიული განაწილების ფუნქცია $F_n(x)$ და ორმხრივი კრიტიკული არის საზღვრები მოვძებნოთ პირობიდან:

$$D_n = \sup_{|x|<\infty} | F_n(x) - F(x) | > \lambda_n . \qquad (1)$$

ა. კოლმოგოროვმა დაამტკიცა, რომ H_0 ჰიპოთეზის სამართლიანობის შემთხვევაში D_n სტატისტიკის განაწილება არ არის დამოკიდებული $F(x)$ ფუნქციაზე, და როცა $n \to \infty$, ადგილი აქვს კრებადობას:

$$p(\sqrt{n} D_n < \lambda) \to K(\lambda), \ \lambda > 0, \qquad (2)$$

სადაც

$$K(\lambda) = \sum_{m=-\infty}^{\infty} (-1)^m e^{-2m^2\lambda^2} \text{ --}$$

არის კოლმოგოროვის კრიტერიუმი, რომლის მნიშვნელობების პოვნა შესაძლებელია შესაბამისი ცხრილებიდან. კრიტერიუმის კრიტიკული მნიშვნელობა $\lambda_n(\alpha)$ გამოითვლება მოცემული მნიშვნელოვნების α დონის მიხედვით, როგორც $p(D_n \geq \lambda) = \alpha$ განტოლების ამონახსნი.

მტკიცდება, რომ კრიტერიუმის კრიტიკული მნიშვნელობა $\lambda_n(\alpha)$ გამოითვლება შემდეგი მიახლოებითი ფორმულით:

$$\lambda_n(\alpha) \approx \sqrt{\frac{z}{2n}} - \frac{1}{6n},$$

სადაც z – არის $1 - K\left(\sqrt{\dfrac{\lambda}{2}}\right) = \alpha$ განტოლების ამონახსნი.

პრაქტიკულ ამოცანებში D_n სტატისტიკის გამოსათვლელად გამოიყენება თანაფარდობა:

$$D_n = \max(D_n^+, D_n^-),$$

სადაც

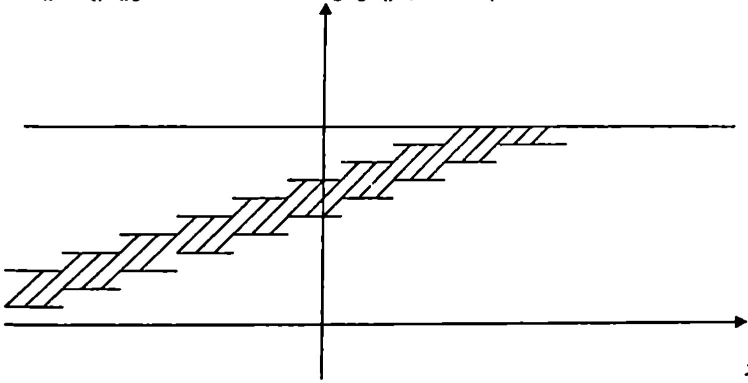
$$D_n^+ = \max_{1 \leq m \leq n}\left(\frac{m}{n} - F(X_{(m)})\right), \quad D_n^- = \max_{1 \leq m \leq n}\left(F(X_{(m)}) - \frac{m-1}{n}\right),$$

ხოლო $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ – ვარიაციული მწკრივია, აგებული $X_1, X_2, \ldots, X_n$ შერჩევის მიხედვით.

თუ H_0 ჰიპოთეზა სამართლიანია, მაშინ D_n^+ და D_n^- სტატისტიკები ერთნაირად არიან განაწილებული. ცნობილია, რომ თუ $\alpha < 0.2$, მაშინ დიდი სიზუსტით $\lambda_n^+(\alpha) \approx \lambda_n(2\alpha)$, სადაც $\lambda_n^+(\alpha)$ არის D_n^+ კრიტერიუმის კრიტიკული მნიშვნელობა.

ჰიპოთეზების შემოწმების წესი შემდეგში მდგომარეობს: ა). H_0 ჰიპოთეზის შემოწმებისას H_1 ალტერნატივის წინააღმდეგ ვიწუნებთ H_0 ჰიპოთეზას, როცა $D_n > \lambda_n(\alpha)$: ბ). თუ $D_n^* > \lambda_n^*(\alpha)$, მაშინ H_0 ჰიპოთეზას უარვყოფთ H_1^* ალტერნატივის სასარგებლოდ.

კოლმოგოროვის კრიტერიუმს შეიძლება მიეცეს შემდეგი გეომეტრიული ინტერპრეტაცია: თუ საკოორდინატო სიბრტყეზე გამოვსახავთ $F_n(x)$ და $F_n(x) \pm \lambda_n(\alpha)$ ფუნქციების გრაფიკებს, მაშინ H_0 ჰიპოთეზა სამართლიანია, თუ $F(x)$ ფუნქციის გრაფიკი არ გამოდის $F_n(x) - \lambda_n(\alpha)$ და $F_n(x) + \lambda_n(\alpha)$ ფუნქციების გრაფიკებს შორის მოთავსებული არიდან:



შენიშვნა. აღსანიშნავია, რომ კოლმოგოროვ-სმირნოვის ტიპის სტატისტიკების კვლევაში დიდი წვლილი მიუძღვით ქართველ მეცნიერებს. 1949-1951 წლებში პროფ. გ. მანიამ დაადგინა აღნიშნული სტატისტიკების ზღვარითი განაწილება და გამოთვალა კრიტიკული მნიშვნელობები. 1964-1965 წლებში პროფ. ე. ნადარაიამ აჩვენა უცნობი განაწილების სიმკვრივის გულოვანი შეფასების კრებადობა თეორიული სიმკვრივისაკენ და დაადგინა შეფასების სიზუსტე. ე. ნადარაიას მიერ შემოთავაზებული იყო აგრეთვე უცნობი რეგრესიის ფუნქციისათვის გულოვანი შეფასებები, რომელიც ლიტერატურაში ნადარაია-ვატსონის შეფასების სახელითაა ცნობილი.

განაწილების ნორმალურობის შემოწმების მიახლოებითი მეთოდი.

განვიხილოთ განაწილების ნორმალურობის შემოწმების მიახლოებითი მეთოდი, რომელიც დაკავშირებულია ასიმეტრიისა და ექსცესის კოეფიციენტების შეფასებებთან. განვმარტოთ ემპირიული განაწილებისათვის ასიმეტრიისა და ექსცესის კოეფიციენტები თეორიული განაწილების შესაბამისი ცნებების ანალოგიურად.

ემპირიული (შერჩევითი) განაწილების **ასიმეტრია** განიმარტება თანაფარდობით:

$$a_\text{ე} = \frac{m_3}{\sigma_\text{ე}^3},$$

სადაც m_3 — მესამე რიგის ცენტრალური ემპირიული მომენტია.

ემპირიული (შერჩევითი) განაწილების ექსცესი განიმარტება ტოლობით:

$$e_{ვ} = \frac{m_4}{\sigma_{ვ}^4} - 3,$$

სადაც m_4 – მეოთხე რიგის ცენტრალური ემპირიული მომენტია.

ცნობილია, რომ ნორმალურად განაწილებული შემთხვევითი სიდიდისათვის ასიმეტრია და ექსცესი ნულია. ამიტომ, თუ შესაბამისი ემპირიული სიდიდეები საკმარისად მცირეა, შეიძლება დავუშვათ, რომ გენერალური ერთობლიობა განაწილებულია ნორმალური განაწილების კანონით.

თავი XI
დამოუკიდებლობისა და ერგვაროვნების ჰიპოთეზათა შემოწმება

დამოუკიდებლობის ჰიპოთეზის შემოწმება.

განვიხილოთ ერთი პოპულაციის ორი სხვადასხვა ნიშნის (ან ფაქტორის) ერთმანეთთან დამოკიდებულების საკითხი. დავუშვათ, რომ პოპულაციიდან აღებულია n მოცულობის შერჩევა და ამ შერჩევის ელემენტები კლასიფიცირებულია ორი A და B ნიშნის მიხედვით. დავუშვათ, რომ დაკვირვებათა ყველა შესაძლო შედეგი დაყოფილია A ნიშნით $A_1,\dots,A_k$, ხოლო B ნიშნით $B_1,\dots,B_r$ კატეგორიებად. პოპულაციის ყოველი ელემენტი ეკუთვნის ზუსტად ერთ კატეგორიას A ნიშნის შესაბამისი რომელიმე კლასიდან და ასევე ზუსტად ერთ რომელიმე კატეგორიას B ნიშნის მიხედვით. ამიტომ დაკვირვებული მონაცემები იყოფა $r\times k$ რაოდენობის A_iB_j არათაესებად ჯგუფად. თუ n_{ij}-ით ავღნიშნავთ იმ მონაცემთა რაოდენობას, რომლებიც ერთდროულად ეკუთვნიან A ნიშნის i-ურ და B ნიშნის j-ურ კატეგორიას და ჩავწერთ ამ სიდიდეს ცხრილის i-ური სვეტისა და j-ური სტრიქონის გადაკვეთაზე, მივიღებთ ორგანზომილებიან ნიშანთა შეუღლების ქვემოთ მოყვანილ ცხრილს, რომელსაც იყენებენ A და B ნიშნების დამოუკიდებლობის ჰიპოთეზის შესამოწმებლად.

$A \backslash B$	B_1	B_2		B_j		B_r	Σ
A_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1r}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2r}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ir}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kr}	$n_{k\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet j}$		$n_{\bullet r}$	n

ამ ცხრილში ($n_{i\bullet}$, $1\le i\le k$) და ($n_{\bullet j}$, $1\le j\le r$) სიდიდეები აღნიშნავს A და B ნიშნების შესაბამის მარგინალურ სიხშირეებს. $n_{i\bullet}$ წარმოადგენს შერჩევის იმ ელემენტთა სიხშირეს, რომლებიც მოხვდნენ i-ურ კლასში A ნიშნის მიხედვით, ხოლო $n_{\bullet j}$ არის შერჩევის იმ ელემენტთა რაოდენობა, რომლებიც მოხვდნენ j-ურ კლასში B ნიშნით. ამასთანავე

$$n_{i\bullet}=\sum_{j=1}^{r} n_{ij}\,,\quad n_{\bullet j}=\sum_{i=1}^{k} n_{ij}\,,\quad \sum_{i=1}^{k} n_{i\bullet}=\sum_{j=1}^{r} n_{\bullet j}=\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}=n.$$

ავღნიშნოთ P_{ij} სიმბოლოთი ალბათობა იმისა, რომ პოპულაციიდან შემთხვევით ამორჩეული ელემენტი აღმოჩნდება ერთდროულად A ნიშნის i-ურ და B ნიშნის j-ურ კატეგორიაში. მაშინ $P_{i\bullet}=\sum_{j=1}^{r} P_{ij}$ იქნება ალბათობა იმისა, რომ პოპულაციის ელემენტი აღმოჩნდება i-ურ კატეგორიაში A ნიშნის მიხედვით, ხოლო $P_{\bullet j}=\sum_{i=1}^{k} P_{ij}$ – ალბათობა იმისა, რომ პოპულაციის ელემენტი მოხვდება j-ურ კატეგორიაში B ნიშნით.

ჩვენ შეგვიძლია ჩავთვალოთ, რომ ზემოთ მოყვანილი ცხრილი წარმოადგენს n დამოუკიდებელი დაკვირვების შედეგს ალბათურ მოდელზე, რომლის ერთობლივი განაწილების კანონია:

$X \backslash Y$	B_1	B_2		B_j		B_r
A_1	P_{11}	P_{12}	...	P_{1j}	...	P_{1r}
A_2	P_{21}	P_{22}	...	P_{2j}	...	P_{2r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	P_{i1}	P_{i2}	...	P_{ij}	...	P_{ir}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_k	P_{k1}	P_{k2}	...	P_{kj}	...	P_{kr}

სადაც $A_1,...,A_k$ – არის A ნიშნის შესაძლო შედეგი, ხოლო $B_1,...,B_r$ – არის B ნიშნის შესაძლო შედეგი.

როდესაც P_{ij} ალბათობები მოცემულია, მაშინ მარტივად გამოითვლება n დამოუკიდებელ ცდაში შესაძლო შედეგთა სავარაუდო სიხშირეები:

$$n_{ij}^{*}=n\cdot P_{ij},\quad i=1,2,...,k;\quad j=1,2,...,r\,.$$

განვიხილოთ დამოუკიდებლობის შემდეგი ჰიპოთეზის შემოწმების ამოცანა.

$H_0: P_{ij}=P_{i\bullet}\cdot P_{\bullet j},\quad i=1,2,...,k;\quad j=1,2,...,r;$

$H_1: H_0$ არ არის მართებული.

როცა A და B ნიშნები დამოუკიდებელია, მაშინ $\forall i,j$-სათვის უნდა შესრულდეს ტოლობა:

$$P_{ij}=P_{i\bullet}\cdot P_{\bullet j},\quad i=1,2,...,k;\quad j=1,2,...,r\,,$$

სადაც

$$\sum_{i=1}^{k} P_{i\bullet}=1 \text{ და } \sum_{j=1}^{r} P_{\bullet j}=1\,.$$

ჩვენ განვიხილავთ დამოუკიდებლობის ჰიპოთეზის შემოწმების ამოცანას, როცა $P_{i\bullet}$ და $P_{\bullet j}$ მარგინალური განაწილებები უცნობია. ამ შემთხვევაში H_0 ჰიპოთეზა არ აზუსტებს უცნობ პარამეტრთა მნიშვნელობას და

საჭიროა მათი შეფასება შერჩევის საშუალებით. შეფასების როლში ავიღოთ ფარდობითი სიხშირე:

$$\overline{P}_{i\cdot}=\frac{n_{i\cdot}}{n},\quad \overline{P}_{\cdot j}=\frac{n_{\cdot j}}{n},\ i=1,2,\dots,k;\quad j=1,2,\dots,r,$$

მაშინ ჰიპოთეტური სიხშირეები გამოითვლება ფორმულებით

$$n_{ij}^{*}=\frac{n_{i\cdot}\cdot n_{\cdot j}}{n},\ i=1,2,\dots,k;\quad j=1,2,\dots,r. \qquad (1)$$

ნიშანთა დამოუკიდებლობის ჰიპოთეზის შესამოწმებლად გამოიყენება

$$\hat{\chi}^2=\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{\left(n_{ij}-n_{ij}^{*}\right)^2}{n_{ij}^{*}}$$

სტატისტიკა, რომელიც მიახლოებით χ^2 კანონით არის განაწილებული თავისუფლების ხარისხით $(k-1)(r-1)$. მოცემული მნიშვნელოვნების α დონისათვის χ^2 განაწილების ცხრილიდან ვპოულობთ $\chi^2_{\alpha,(k-1)(r-1)}$ კრიტიკულ წერტილს. თუ აღმოჩნდა, რომ $\hat{\chi}^2\ge\chi^2_{\alpha,(k-1)(r-1)}$, მაშინ ნულოვან ჰიპოთეზას უარვყოფთ. წინააღმდეგ შემთხვევაში ვასკვნით, რომ A და B ნიშნები დამოუკიდებელია.

მაგალითი. სოციოლოგს სურს 385 ოჯახზე დაკვირვებით მიღებული შერჩევის საფუძველზე შეამოწმოს ჰიპოთეზა იმის შესახებ, რომ ოჯახში ბავშვების რაოდენობა არ არის დამოკიდებული ოჯახის შემოსავალზე:

ბავშვების რაოდენობა	0–6 A ჯგუფი	6–12 B ჯგუფი	12–18 C ჯგუფი	18-ზე მეტი D ჯგუფი
0	10	9	18	24
1	8	12	25	31
2	24	28	23	28
3	26	24	20	6
4 ან მეტი	32	22	18	7

ამოხსნა. ავიღოთ 0.01-ის ტოლი ნდობის ალბათობა. (1) ფორმულების თანახმად გვექნება ჰიპოთეტური სიხშირეების შემდეგი ცხრილი:

ბავშვების რაოდენობა	0–6 A ჯგუფი	6–12 B ჯგუფი	12–18 C ჯგუფი	18-ზე მეტი D ჯგუფი
0	15.44	14.67	16.06	14.83
1	19.24	18.28	20.01	18.47
2	26.08	24.77	27.12	25.03
3	19.24	18.28	20.01	18.47
4 ან მეტი	20.00	19.00	20.80	19.20

გამოთვლების მოხერხებულობის მიზნით χ^2 სტატისტიკის დაკვირვებული მნიშვნელობების გამოსათვლელად ვისარგებლოთ შემდეგი ცხრილით:

	n^0_{ij}	$n^{\bullet}_{ij}$	$n^0_{ij}-n^{\bullet}_{ij}$	$\left(n^0_{ij}-n^{\bullet}_{ij}\right)^2$	$\left(n^0_{ij}-n^{\bullet}_{ij}\right)^2/n^{\bullet}_{ij}$
A0	10	15.44	-5.44	29.63	1.92
A1	8	19.24	-11.24	126.35	6.57
A2	24	26.08	-2.08	4.31	0.17
A3	26	19.24	6.76	45.69	2.37
A4	32	20.00	12.00	144.00	7.20
B0	9	14.67	-5.67	32.16	2.19
B1	12	18.28	-6.28	39.42	2.16
B2	28	24.77	3.23	10.42	0.42
B3	24	18.28	5.72	32.74	1.79
B4	22	19.00	3.00	9.00	0.47
C0	18	16.06	1.94	3.76	0.23
C1	25	20.01	4.99	24.90	1.24
C2	23	27.12	-4.12	16.97	0.63
C3	20	20.01	-0.01	0.00	0.00
C4	18	20.80	-2.80	7.84	0.38
D0	24	14.83	9.17	84.17	5.68
D1	31	18.47	12.53	156.98	8.50
D2	28	25.03	2.97	8.80	0.35
D3	6	18.47	-12.47	155.52	8.42
D4	7	19.20	-12.20	148.84	7.75
					$\sum = 58.44$

როგორც ვხედავთ სტატისტიკის დაკვირვებული მნიშვნელობაა $\hat{\chi}^2 = 58.44$. მეორეს მხრივ, რადგან თავისუფლების ხარისხია $(r-1)(k-1)=12$, ამიტომ (0.01 მნიშვნელოვნების დონისათვის) კრიტიკული მნიშვნელობაა $\chi^2_{12,0.01}=26.217$. ვინაიდან, $\hat{\chi}^2 > \chi^2_{12,0.01}$, სოციოლოგი დაასკვნის, რომ ოჯახში ბავშვების რაოდენობა და ოჯახის შემოსავალი დამოკიდებულია ერთმანეთზე.

ერთგვაროვნების ჰიპოთეზის შემოწმება.

დავუშვათ, რომ მოცემულია k რაოდენობის სხვადასხვა პოპულაცია და ყოველი პოპულაციიდან, ერთმანეთისაგან დამოუკიდებლად, აღებულია $n_1,\dots,n_k$ მოცულობის შერჩევები. ვიგულისხმოთ, რომ ყველა პოპულაცია კლასიფიცირებულია ერთი და იგივე A ნიშნის $A_1,\dots A_r$ კატეგორიის მიხედვით. i-ური შერჩევის იმ ელემენტთა სიხშირე, რომლებსაც აღმოაჩნდათ j-ური კატეგორია აღვნიშნოთ n_{ij} სიმბოლოთი. მაშინ მონაცემები განლაგდება ნიშანთა შეუღლების შემდეგ ცხრილში:

	კატეგორიები						$\sum$
	A_1	A_2	...	A_j	...	A_r	
შერჩევა I პოპულაციიდან	n_{11}	n_{12}	...	n_{1j}	...	n_{1r}	n_1
შერჩევა II პოპულაციიდან	n_{21}	n_{22}	...	n_{2j}	...	n_{2r}	n_2
	...	...	...	...	...	...	...
შერჩევა j-ური პოპულაციიდან	n_{i1}	n_{i2}		n_{ij}		n_{ir}	n_i
	...	...	...	...	...	...	...
შერჩევა k-ური პოპულაციიდან	n_{k1}	n_{k2}		n_{kj}		n_{kr}	n_k
$\sum$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet r}$	n

ასეთ შემთხვევაში ხშირად ჩნდება პოპულაციათა ერთგვაროვნების (შერჩევები, რომ აღებულია ერთი და იგივე გენერალური ერძთობლობიდან) ჰიპოთეზის შემოწმების აუცილებლობა. ასეთი ჰიპოთეზა ექვივალენტურია ჰიპოთეზისა, რომ პოპულაციიდან შემთხვევით არჩეული ელემენტის ყოველ A_j კლასში მოხვედრის P_j ალბათობა ერთი და იგივეა ყველა პოპულაციისათვის.

ცხრილში n_i არის i-ური პოპულაციიდან აღებული შერჩევის მოცულობა

$$n_i = \sum_{j=1}^{r} n_{ij}, \quad i = 1,2,...,k.$$

$n_{\bullet j}$ სიმბოლოთი აღნიშნულია ყველა შერჩევის იმ ელემენტთა რაოდენობა, რომლებსაც აღმოაჩნდათ A ნიშნის j-ური კატეგორია ცხადია, რომ n_i სიდიდეებისგან განსხვავებით, $n_{\bullet j}$ სიდიდეები (შერჩევის აღებამდე) შემთხვევით სიდიდეებს წარმოადგენენ და

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}, \quad j = 1,2,...,r.$$

ალბათობა იმისა, რომ i-ური პოპულაციიდან შემთხვევით არჩეულ ელემენტს აღმოაჩნდება j-ური კატეგორია (ან i-ური პოპულაციიდან j-ური კატეგორიის ელემენტთა პროპორცია) აღვნიშნოთ P_{ij} სიმბოლოთი.

$$\sum_{j=1}^{r} P_{ij} = 1.$$

განვიხილოთ შემდეგი ჰიპოთეზები.

$H_0 : P_j \equiv P_{1j} = P_{2j} = \cdots = P_{kj}, \quad j = 1,2,...,r;$

$H_1 : H_0$ არ არის მართებული.

H_0 ჰიპოთეზის დროს მოსალოდნელი რაოდენობა i-ური შერჩევის ელემენტებისა, რომლებიც j-ური კატეგორიის აღმოჩნდნენ ტოლია:

$$n_{ij}^{*} = n_i \cdot P_j.$$

P_j პარამეტრის შეფასების როლში ავიღოთ

$$\overline{P}_j = \frac{n_{\cdot j}}{n}, \quad j = 1,2,\dots,r.$$

მაშინ, H_0 ჰიპოთეზის დროს i-ური შერჩევის j-ური კატეგორიის ელემენტების მოსალოდნელი რაოდენობა იქნება:

$$n_{ij}^{\bullet} = \frac{n_i \cdot n_{\cdot j}}{n}.$$

დაკვირვებულ n_{ij} სიდიდეებსა და H_0 ჰიპოთეზის დროს მათ მოსალოდნელ $n_{ij}^{\bullet}$ მნიშვნელობებს შორის გადახრის საზომად აიღება

$$\hat{\chi}^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - n_{ij}^{\bullet}\right)^2}{n_{ij}^{\bullet}}$$

სტატისტიკა. ნულოვანი ჰიპოთეზის უარყოფის არეა $\chi^2 \geq \chi^2_{\alpha,(k-1)(r-1)}$. წინააღმდეგ შემთხვევაში ვასკვნით, რომ პოპულაცია ერთგვაროვანია.

მაგალითი. ქვემოთ მოყვანილია სამი სხვადასხვა საწარმოს მიერ წარმოებულ ერთი და იგივე ტიპის პროდუქციაშივარგის და უვარგის ნაწარმთა რაოდენობები:

	ვარგისი	უვარგისი	სულ
I საწარმო	240	10	250
II საწარმო	191	9	200
III საწარმო	139	11	150
სულ	570	30	600

არის თუ არა განსხვავება ამ საწარმოთა მიერ გამოშვებული პროდუქციის ხარისხში?

ამოხსნა. განვიხილოთ ნულოვანი ჰიპოთეზა: შერჩევები ერთგვაროვანია. ამ ჰიპოთეზის დროს სავარაუდო სიხშირეებია

237.5	12.5
190	10
142.5	7.5

ცხრილის საშუალებით გამოვთვალოთ ხი კვადრატ სტატისტიკის დაკვირვებული მნიშვნელობები:

	n_{ij}	n^*_{ij}	$n_{ij} - n^*_{ij}$	$(n_{ij} - n^*_{ij})^2$	$(n_{ij} - n^*_{ij})^2 / n^*_{ij}$
I საწარმო/ვარგისი	240	237.5	2.5	6.25	0.026
II საწარმო/ვარგისი	191	190	1	1	0.005
III საწარმო/ვარგისი	138	142.5	-4.5	20.25	0.142
I საწარმო/უვარგისი	10	12.5	-2.5	6.25	0.500
II საწარმო/უვარგისი	9	10	-1	1	0.100
III საწარმო/უვარგისი	11	7.5	3.5	12.25	1.633
					$\hat{\chi}^2=2.4$

ვიპოვოთ, მნიშვნელოვნების 0.1 დონისათვის თავისუფლების ხარისხით (3-1)(2-1)=2, კრიტერიუმის კრიტიკული მნიშვნელობა: $\chi^2_{2;0.1}=4.60517$.

რადგანაც $\hat{\chi}^2=2.4<4.60517$, ამიტომ აღნიშნული მონაცემები არ იძლევა ერთგვაროვნების ჰიპოთეზის უარყოფის საფუძველს.

თავი XII
დასკვნითი სტატისტიკის არაპარამეტრული მეთოდები

1. ნიშნების კრიტერიუმი.

სტატისტიკის პარამეტრულ ამოცანებში ცნობილია მოდელის განაწილების სახე და სტატისტიკური დასკვნები კეთდება ამ განაწილების უცნობი პარამეტრების შესახებ. თუ მოდელის განაწილების სახე უცნობია, ან ცენტრალური ზღვარითი თეორემის პირობები დარღვეულია, მაშინ ამოცანების ამოსახსნელად იყენებენ *სტატისტიკის არაპარამეტრულ მეთოდებს*, რომელთა მიხედვითაც მოდელის შესახებ კეთდება საკმაოდ სუსტი დაშვებები და ამდენად, ჰიპოთეზები მოწმდება განაწილების ფუნქციათა ფართო კლასისათვის. მაგალითად, აქაც შეიძლება გვაინტერესებდეს ორი პოპულაციის საშუალოების შედარების ამოცანა. მაგრამ, რადგან აღარ კეთდება პოპულაციათა ნორმალურობის დაშვება, ცხადია, რომ პოპულაციათა შესაბამისი განაწილებები შეიძლება იყოს ტოლი საშუალოების მქონე ნებისმიერი განაწილებები. გარდა ამისა, ამ ამოცანისათვის წყვილთა t-კრიტერიუმი ოპტიმალური იყო იმ დაშვებაში, რომ პოპულაციები ნორმალურადაა განაწილებული. თუ განაწილებათა ნორმალურობა არა გვაქვს და გადახრა ამ დაშვებისაგან დიდია, მაშინ ოპტიმალური აღმოჩნდება ე.წ. *რანგობრივი კრიტერიუმები*. ჩვენ შევისწავლით რამდენიმე ასეთ კრიტერიუმს.

ნორმალური აპროქსიმაცია. განვიხილოთ მაგალითი: საჭიროა შევადაროთ მზით დამწვრობის საწინააღმდეგო ორი A და B ტიპის საცხის ეფექტურობა. ორივე ტიპის საცხი წაუსვეს 45 ადამიანს შემთხვევით ამორჩეულ მკლავზე (ერთი, ერთ მკლავზე, ხოლო მეორე - მეორეზე). აღმოჩნდა, რომ 22 შემთხვევაში B ტიპის საცხწასმული მკლავი უფრო წითელი იყო, ვიდრე A ტიპისა, 18 შემთხვევაში პირიქით და მხოლოდ 5 შემთხვევაში, A და B ტიპის საცხებს ჰქონდათ ერთნაირი ეფექტი. შეიძლება თუ არა ამ მონაცემების მიხედვით დავასკვნათ, რომ დამწვრობის საწინააღმდეგოდ A ტიპის საცხი ჯობია B ტიპისას?

აღვნიშნოთ, i-ური ინდივიდისათვის დამწვრობის ხარისხი A ტიპის საცხის ხმარებისას x_i-ით, ხოლო იგივე სიდიდე B ტიპის საცხის ხმარებისას იყოს y_i. ერთი შეხედვით, საქმე გვაქვს დაწყვილებულ მონაცემებთან, მაგრამ როგორც ვხედავთ, ჩვენთვის ცნობილი არაა არც x_i და არც y_i სიდიდეთა მნიშვნელობები თავისთავად და მაშასადამე, არც მათი სხვაობები, $d_i = x_i - y_i$, რომელიც გამოიყენება წყვილთა t-კრიტერიუმის ასაგებად დაწყვილებულ მონაცემებში. ერთადერთი, რაც ჩვენთვისაა ცნობილი, ის არის, თუ რამდენ შემთხვევაში იყო ეს სიდიდე უარყოფითი, $d_i < 0$, რამდენჯერ – დადებითი, $d_i > 0$ და რამდენჯერ 0-ის ტოლი, ანუ $d_i = 0$. შემოვიღოთ შესაბამისი შემთხვევითი სიდიდეები X_i, Y_i და $D_i = X_i - Y_i$ და შევეცადოთ მათ ტერმინებში გამოვთქვათ ის, რისი შემოწმებაც გვინდა.

ბუნებრივია იქნება დავუშვათ, რომ $D_1, D_2, \ldots, D_n$ დამოუკიდებელი და ერთნაირად განაწილებული შემთხვევითი სიდიდეებია, თუმცა ჩვენ არაფერ

რო ვიცით მათი F_D განაწილების სახის შესახებ. სინამდვილეში განაწილების სახე ჩვენ არ გვჭირდება, ვინაიდან ჩვენი ამოცანა მდგომარეობს მხოლოდ იმის დადგენაში, აქვს თუ არა ეფექტი რომელიმე საცხს, ანუ მნიშვნელოვანია, თუ არა განსხვავება დამწერიბათა საწინააღმდეგო ეფექტების რაოდენობებს (22–სა და 18-ს) შორის. A და B ტიპის საცხების ხმარებისას. ცხადია, რომ რაც უფრო ახლოს აღმოჩნდება ეს რაოდენობები (ან შესაბამისი ფარდობითი სიხშირეები) ერთმანეთთან, მით უფრო ვერ გავარჩევლით განსხვავებას საცხებს შორის. მაგრამ რადგან ფარდობითი სიხშირეები ახლოსაა შესაბამის ხდომილობათა ალბათობებთან, ამიტომ ეს იქნებოდა $\{D<0\}$ და $\{D>0\}$ ხდომილობათა ალბათობების ტოლობის მაჩვენებელი. ამიტომ, გასაგებია, რომ ჩვენი ამოცანა მდგომარეობს $H_0: m_D = 0$ ნულოვანი ჰიპოთეზის შემოწმებაში, სადაც m_D – F_D განაწილების მედიანაა (ანუ ისეთი წერტილი, რომლისთვისაც $F_D(m_D) = 1/2$).

ცხადია, რომ ჩვენ საქმე გვაქვს ორმხრივ ალტერნატივასთან: $H_1: m_D \neq 0$.

კრიტერიუმის ასაგებად შემოვიღოთ შემთხვევითი სიდიდე

$$T = \sum_{i=1}^{n} I\{D_i > 0\}, \tag{1}$$

სადაც n ღნიშნავს შერჩევის მოცულობას (0-საგან განსხვავებულ D_i-ების რაოდენობას) და $I\{D_i > 0\}$ არის შემთხვევითი სიდიდე, რომელსაც $\{D_i > 0\}$ *ხდომილობის ინდიკატორს* უწოდებენ და რომელიც ასე განიმარტება:

$$I\{D_i > 0\} = \begin{cases} 1, & \text{თუ } D_i > 0 \\ 0, & \text{თუ } D_i < 0 \end{cases}$$

მაშინ გასაგებია, რომ T გვიჩვენებს შერჩევის იმ D_i-ების რაოდენობას, რომლებიც დადებითია. ამიტომ T შემთხვევით სიდიდეს $H_0: m_D = 0$ ჰიპოთეზის სამართლიანობის დროს აქვს ბინომური განაწილება ცდათა რიცხვით n და განაწილების პარამეტრით $p = P\{D_i > 0\} = 1/2$. იმ შემთხვევაში თუ $np(1-p) = n/4 \geq 5$, ანუ $n \geq 20$, ბინომური განაწილებისათვის დასაშვებია ნორმალური აპროქსიმაცია, ანუ

$$\frac{T - ET}{\sqrt{DT}} = \frac{(T - n/2)}{\sqrt{n/4}} \cong N(0,1) \tag{2}$$

და ამიტომ სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ დასახელებული α მნიშვნელოვნების დონისათვის T სტატისტიკის დაკვირვებული t მნიშვნელობა აკმაყოფილებს პირობას

$$t > n/2 - 1/2 - z_{\alpha/2} \cdot \sqrt{n/4} \text{ ან } t < n/2 + 1/2 + z_{\alpha/2} \cdot \sqrt{n/4}, \tag{3}$$

მაშინ α მნიშვნელოვნების დონით $H_0: m_D = 0$ ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს. ამასთანავე, კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია

$$p = 2 \cdot [1 - \Phi(\frac{t - n/2 - 1/2}{\sqrt{n/4}})], \text{ როცა } t \geq n/2 \text{ და}$$

$$p = 2\cdot\Phi\left(\frac{t-n/2+1/2}{\sqrt{n/4}}\right), \text{ როცა } t<n/2. \qquad (4)$$

განხილულ მაგალითში: $n=45-5=40$, $t=18<n/2=20$. დავუშვათ, $\alpha=0.05$. მაშინ $z_{\alpha/2}=z_{0.025}=1.96$ და

$$t_1 := n/2-1/2-z_{\alpha/2}\cdot\sqrt{n/4} = 20-1/2-1.96\cdot\sqrt{40/4}=13.3,$$

$$t_2 := n/2+1/2+z_{\alpha/2}\cdot\sqrt{n/4} = 20+1/2+1.96\cdot\sqrt{40/4}=26.7$$

ვინაიდან $13.3\le t=18\le 26.7$, ამიტომ კრიტერიუმის თანახმად, $\alpha=0.05$ მნიშვნელოვნების დროით, $H_0: m_D=0$ ჰიპოთეზის უარყოფის საფუძველი არა გვაქვს. ე. ი. A და B ტიპის საცხებს აქვს ერთნაირი მზით დამწვრობის საწინააღმდეგო ეფექტი.

ცხადია, იგივე პასუხს მოგვცემს p-მნიშვნელობის გამოთვლაც. მართლაც, ვინაიდან $t=18<n/2=20$, ამიტომ (4) თანაფარდობის მეორე ტოლობის თანახმად p-მნიშვნელობა იქნება:

$$p = 2\cdot\Phi\left(\frac{t-n/2+1/2}{\sqrt{n/4}}\right) = 2\cdot\Phi\left(\frac{20-18+1/2}{\sqrt{40/4}}\right) = 0.635 > 0.05.$$

ზუსტი მეთოდი. როგორც ყოველთვის, ზუსტი მეთოდი გულისხმობს p-მნიშვნელობის გამოთვლას ზუსტად და ის უპირატესად გამოიყენება იმ შემთხვევაში, როცა დაუშვებელია ბინომური განაწილების ნორმალურით აპროქსიმაცია, ანუ როცა $n<20$. ამ შემთხვევაში, p-მნიშვნელობა გამოითვლება ფორმულით:

$$p = 2\cdot P\{bin(n,1/2)\ge t\} = \sum_{k\ge t} C_n^k\cdot(1/2)^n, \text{ როცა } t>n/2$$

და

$$p = 2\cdot P\{bin(n,1/2)\le t\} = \sum_{k\le t} C_n^k\cdot(1/2)^n, \text{ როცა } t<n/2$$

2. უილკოქსონის ნიშნიანი რანგების კრიტერიუმი.

განვიხილოთ მაგალითი: დავუშვათ, რომ A და B ტიპის საცხების შესახებ ზემოთ მოყვანილ მაგალითში შედარება ხდება დამწვრობის ხარისხების მიხედვით, რომელიც იზომება 10 ბალიანი სკალით და მონაცემები წარმოდგენილია ცხრილის სახით.

ამ ცხრილის მონაცემები ასე უნდა გავიგოთ: მაგალითად, $d_i=-6$ და $f_i=2$ ნიშნავს, რომ ორ ადამიანს A ტიპის საცხის ხმარებისას ცალ ხელზე და B ტიპისას – მეორეზე, აღმოაჩნდა დამწვრობის ხარისხებს შორის სხვაობა 6-ის ტოლი, თანაც A ტიპის საცხის სასარგებლოდ. ჩვენი ამოცანა კვლავ მდგომარეობს ამ მონაცემებზე დაყრდნობით გავარკვიოთ საცხების ეფექტურობის ტოლობის საკითხი. ე. ი. ვიხილავთ $H_0: m_D=0$ ჰიპოთეზის შემოწმების ამოცანას $H_1: m_D\ne 0$ ალტერნატივის წინააღმდეგ.

სხვაობა d_i	სიხშირე f_i	სხვაობა d_i	სიხშირე f_i
-10	0	10	0
-9	0	9	0
-8	1	8	0
-7	3	7	0
-6	2	6	0
-5	2	5	0
-4	1	4	0
-3	5	3	2
-2	4	2	6
-1	4	1	10
Σ	22	Σ	18
0	5		

ერთი შეხედვით შეიძლება მოგვეჩვენოს, რომ აქ შესაძლებელია წყვილთა t-კრიტერიუმის გამოყენება. აქ მონაცემები გაზომილია ე.წ. *ორდინალურ სკალაში*, ანუ ისინი დალაგებულია, მაგრამ არ გააჩნიათ კონკრეტული რიცხვითი მნიშვნელობა. სხვა სიტყვებით რომ ვთქვათ, $d_i = -6$ სულაც არ ნიშნავს იმას, რომ 6 ბალიანი სხვაობა 6-ჯერ მეტია 1 ბალიან $d_i = -1$ სხვაობაზე. ეს დალაგება მხოლოდ იმას გულისხმობს, რომ A ტიპის საცხის სასარგებლოდ -6 ჯობია -5-ს, -5 ჯობია -4-ს და ა.შ. ასეთ შემთხვევაში გამოიყენება **უილკოკსონის ნიშნიანი რანგების კრიტერიუმი**, რომელიც შემდეგნაირად იგება: პირველ რიგში, გამოვრიცხოთ ის მონაცემები, სადაც $d_i = 0$. შემდეგ შევხედოთ ერთნაირი მოდულების მქონე d_i-ების ჯამურ სიხშირეებს. მაგალითად, იმ მონაცემების რაოდენობა, სადაც $|d_i| = 1$, არის $4 + 10 = 14$. იმ მონაცემების რაოდენობა, სადაც $|d_i| = 2$, არის $4 + 6 = 10$ და ა.შ. იმ მონაცემების ჯგუფს, სადაც $|d_i| = 1$ და მონაცემების რაოდენობა არის 14, შეესაბამება რანგების დიაპაზონი 1-დან 14-მდე, ამიტომ ამ ჯგუფს ენიჭება საშუალო რანგი $(1+14)/2 = 7.5$. ანალოგიურად, იმ ჯგუფისათვის, სადაც $|d_i| = 2$ | და მონაცემების რაოდენობა არის 10, შეესაბამება რანგების დიაპაზონი $14+1=15$-დან $14+10=24$-მდე, ამიტომ ამ ჯგუფს ენიჭება საშუალო რანგი $(15+24)/2 = 19.5$ და ა.შ. ასეთი წესით მიღებული რიცხვები ჩავწეროთ ცხრილის სახით:

$\lvert d_i \rvert$	სიხშირე	რანგის დიაპაზონი	საშუალო რანგი
10	0 + 0 = 0	–	–
9	0 + 0 = 0	–	–
8	1 + 0 = 1	40	40.0
7	3 + 0 = 3	37 – 39	38.0
6	2 + 0 = 2	35 – 36	35.5
5	2 + 0 = 2	33 – 34	33.5
4	1 + 0 = 1	32	32.0
3	5 + 2 = 7	25 – 31	28.0
2	4 + 6 = 10	15 – 24	19.5
1	4 + 10 = 14	1 – 14	7.5

კრიტერიუმის სტატისტიკად იღებენ დადებითნიშნიანი d_i სხვაობების შესაბამისი რანგების ჯამს, მას აღნიშნავენ R^+ სიმბოლოთი. მისი რიცხვითი მახასიათებლებია:

$$ER^+ = n(n+1)/4 \quad \text{და} \quad DR^+ = n(n+1)(2n+1)/24.$$

მტკიცდება, რომ თუ არანულოვანი d_i სხვაობების რაოდენობა ≥ 16, მაშინ R^+ სტატისტიკის განაწილება აპროქსიმირდება ნორმალური განაწილებით, ანუ

$$T = \frac{|R^+ - ER^+| - 1/2}{\sqrt{DR^+}} = \frac{|R^+ - n(n+1)/4| - 1/2}{\sqrt{n(n+1)(2n+1)/24}} \cong N(0.1). \qquad (1)$$

შევნიშნოთ, რომ ეს თანაფარდობა სამართლიანია მაშინ, როდესაც არა გვაქვს ერთნაირი მოდულების მქონე სხვაობათა ჯგუფები. იმ შემთხვევაში, როცა ეს პირობა დარღვეულია (1) თანაფარდობაში დისპერსიის გამოსახულება უნდა შესწორდეს შემდეგი სახით:

$$DR^+ = n(n+1)(2n+1)/24 - \sum_{i=1}^{k}(t_i^3 - t_i)/2, \qquad (2)$$

სადაც t_i სიმბოლოთი აღნიშნულია i-ურ ჯგუფში სხვაობათა რაოდენობა, რომლებსაც ერთნაირი მოდულები აქვთ, ხოლო k აღნიშნავს ასეთი ჯგუფების საერთო რაოდენობას.

სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის T სტატისტიკის დაკვირვებული t მნიშვნელობა აკმაყოფილებს პირობას

$$|t| > z_{\alpha/2},$$

მაშინ α მნიშვნელოვნების დონით $H_0: m_D = 0$ ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს. კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია

$$p = 2 \cdot (1 - \Phi(t)).$$

ჩვენს მაგალითში $n = 45 - 5 = 40 \geq 16$, და შესაბამად, დასაშვებია ნორმალური აპროქსიმაცია. უკანასკნელი ცხრილიდან გამოვთვალოთ R^+-ის

დაკვირვებული მნიშვნელობა: $r^{*} = 10\cdot 7.5 + 6\cdot 19.5 + 2\cdot 28.0 = 248$. (2) ტოლობის მიხედვით გამოვთვალოთ შესწორებული დისპერსიის მნიშვნელობა (ვინაიდან, ამ შემთხვევაში გვაქვს ერთნაირი მოდულების მქონე სხვაობათა ჯგუფები):

$$DR^{*} = n(n+1)(2n+1)/24 - \sum_{i=1}^{k}(t_i^3 - t_i)/2 = 40\cdot 41\cdot 81/84 -$$

$$-\frac{14^3 - 14 + 10^3 - 10 + 7^3 - 7 + 2^3 - 2 + 2^3 - 2 + 3^3 - 3}{2} = 3489.$$

ამიტომ (1) ფორმულის მიხედვით გამოთვლილი T სტატისტიკის დაკვირვებული t მნიშვნელობა იქნება:

$$t = \frac{|r^{*} - n(n+1)/4| - 1/2}{\sqrt{DR^{*}}} = \frac{|248 - 40\cdot 41/4|}{\sqrt{3489}} = 2.73,$$

ხოლო კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია:

$$p = 2\cdot(1-\Phi(t)) = 2\cdot(1-\Phi(2.73)) = 2\cdot(1-0.997)) = 0.006,$$

რაც მეტყველებს შედეგის სტატისტიკურ მნიშვნელოვნებაზე. შესაბამისად, $H_0 : m_D = 0$ ჰიპოთეზას უარვყოფთ და დავასკვნით, რომ A ტიპის საცხი ჯობია B ტიპისას. როგორც ვხედავთ, წინა პარაგრაფისაგან განსხვავებით, მივიღეთ საწინააღმდეგო დასკვნა. რაც მეტყველებს იმაზე, რომ რაც უფრო მეტი ინფორმაციაა ხელმისაწვდომი, დასკვნა მით უფრო ზუსტი იქნება: წინა პარაგრაფში მოცემული ინფორმაციით ლაპარაკი იყო მხოლოდ იმაზე, თუ დამწვრობის ხარისხებს შორის სხვაობა დადებითია თუ უარყოფითი და ამიტომაც იქ უკეთესი დასკვნის მიღება შეუძლებელია, მაშინ როცა ამ შემთხვევაში დამწვრობის ხარისხები ხასიათდება კიდევ მათი სიახლოვით, და შესაბამისად, ინფორმაციის დამატებამ თვისობრივად შეცვალა სტატისტიკური დასკვნა.

3. უილკოქსონის რანგთა ჯამის კრიტერიუმი.

უილკოქსონის ნიშნიანი რანგების კრიტერიუმი, გარკვეული აზრით, არის t-კრიტერიუმის არაპარამეტრული ანალოგი დაწყვილებული მონაცემებისათვის. ახლა შევისწავლოთ იმავე t-კრიტერიუმის არაპარამეტრული ანალოგი დამოუკიდებელი შერჩევებისათვის. ამ კრიტერიუმს ეწოდება ***უილკოქსონის რანგთა ჯამის კრიტერიუმი***.

განვიხილოთ შემდეგი მაგალითი: ითვლება, რომ სხვადასხვა გენოტიპს აქვს თვალის ბადურის ანთების (RP – Retinitis Pigmentosa) განვითარების სხვადასხვა სისწრაფე: დომინანტური ფორმა, რომლის დროსაც დაავადება ნელა ვითარდება, რეცესიული ფორმა, რომლის დროსაც დაავადება უფრო ნელა ვითარდება და SL (Sex-linked) ფორმა, რომლის დროსაც დაავადება ვითარდება სწრაფად. ამ ჰიპოთეზის შესამოწმებლად დააკვირდნენ 10-დან 19 წლამდე ასაკის მოზრდებში VA (vizual acuity) სიდიდეებს, რომლებსაც ჰქონდათ RP-ის სხვადასხვა გენოტიპი და მიიღეს ცხრილში წარმოდგენილი მონაცემები:

VA	დომინანტური ფორმა	SL ფორმა
20 – 20	5	1
20 – 25	9	5
20 – 30	6	4
20 – 40	3	4
20 – 50	2	8
20 – 60	0	5
20 – 70	0	2
20 – 80	0	1
	$\Sigma = 25$	$\Sigma = 30$

როგორ შეიძლება ეს მონაცემები გამოვიყენოთ იმის შესამოწმებლად, რომ VA სიდიდეების განაწილებათა მედიანები ორი ფორმისათვის სხვადასხვაა?

აღვნიშნოთ m_D და m_{SL} სიმბოლოებით შესაბამისად პირველი და მეორე ამოკრეფის შესაბამისი პოპულაციების მედიანები. მაშინ ნულოვანი და ალტერნატიული ჰიპოთეზა ასე ყალიბდება: $H_0: m_D = m_{SL}$ და $H_1: m_D \neq m_{SL}$. წინა პარაგრაფში განხილული ამოცანის მსგავსად, აქაც არ შეიძლება t-კრიტერიუმის გამოყენება, იმის გამო, რომ VA სიდიდეების რიცხვითი მნიშვნელობები ჩვენ არ ვიცით. ასეთ შემთხვევებში გარდაუვალია არაპარამეტრული მეთოდის გამოყენება. უილკოკსონის რანგთა ჯამის კრიტერიუმი, წარმოადგენს სწორედ ასეთ მეთოდს. ის დამყარებულია მონაცემთა რანგებზე და შემდეგნაირად იგება:

გავაერთიანოთ ორივე ამოკრეფა და მივაწეროთ რანგები VA სიდიდეების ზრდადობის მიხედვით (საუკეთესო VA სიდიდიდან (20-20), ყველაზე უარესამდე (20-80)). მაგალითად, გაერთიანებული შერჩევისათვის პირველი სტრიქონიდან მივიღებთ, რომ მონაცემების რაოდენობაა $5+1=6$, ამიტომ მას მიეწერება რანგობრივი 1 – 6 ინტერვალი და საშუალო რანგი იქნება $(1+6)/2=3.5$, მეორე სტრიქონისათვის შესაბამისად გვექნება $9+5=14$, რანგების ინტერვალია 7 – 20 (წინა ინტერვალის უდიდესს + 1, ანუ 6 + 1 = 7 და წინა ინტერვალის უდიდესს + 14, ანუ 6 + 14 = 20) და ა.შ. საბოლოოდ, მიღებული შედეგები წარმოვადგინოთ შემდეგი ცხრილის სახით:

გაერთიანებული შერჩევა	რანგების დიაპაზონი	საშუალო რანგი
5 + 1 = 6	1 – 6	3.5
9 + 5 = 14	7 – 20	13.5
6 + 4 = 10	21 – 30	25.5
3 + 4 = 7	31 – 37	34.0
2 + 8 = 10	38 – 47	42.5
0 + 5 = 5	48 – 52	50.0
0 + 2 = 2	53 – 54	53.5
0 + 1 = 1	55	55.0
$\Sigma = 55$		

კრიტერიუმის სტატისტიკად იღებენ სიდიდეს R_1, რომელიც წარმოადგენს I ამოკრეფის შესაბამისი რანგების ჯამს. მისი რიცხვითი მახასიათებლებია:

$$ER_1 = n_1\cdot(n_1+n_2+1)/2 \quad \text{და} \quad DR_1 = n_1\cdot n_2\cdot(n_1+n_2+1)/12. \qquad (1)$$

მტკიცდება, რომ თუ უმცირესი შერჩევის მოცულობა მეტია ან ტოლი 10-ზე ($\min(n_1,n_2)\ge 10$) და შერჩევათა შესაბამისი პოპულაციები უწყვეტადაა განაწილებული, მაშინ R_1-ის განაწილება აპროქსიმირდება ნორმალური განაწილებით, ანუ

$$T=\frac{|R_1-ER_1|-1/2}{\sqrt{DR_1}}=\frac{|R_1-n_1\cdot(n_1+n_2+1)/2|-1/2}{\sqrt{n_1\cdot n_2\cdot(n_1+n_2+1)/12}}\cong N(0,1). \qquad (2)$$

შევნიშნავთ, რომ ეს თანაფარდობა სამართლიანია მაშინ, როდესაც გაერთიანებულ ამოკრეფაში ელემენტი ხვდება ერთხელ. წინააღმდეგ შემთხვევაში (2) თანაფარდობაში დისპერსიის გამოსახულება უნდა შესწორდეს. კერძოდ, ამ შემთხვევაში

$$DR_1=\frac{n_1\cdot n_2}{12}\cdot\left[n_1+n_2+1-\frac{\sum_{i=1}^{k}(t_i^3-t_i)}{(n_1+n_2)\cdot(n_1+n_2-1)}\right], \qquad (3)$$

სადაც t_i სიმბოლოთი აღნიშნულია გაერთიანებული შერჩევის i-ურ სტრიქონში მდგომი რიცხვი, ხოლო k-თი – სტრიქონების საერთო რაოდენობა. თუ ყველა $t_i=1$, მაშინ შესწორებული დისპერსია დაემთხვევა შეუსწორებელ დისპერსიას.

სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის T სტატისტიკის დაკვირვებული t მნიშვნელობა აკმაყოფილებს პირობას

$$|t|> z_{\alpha/2},$$

მაშინ α მნიშვნელოვნების დონით $H_0: m_I = m_{II}$ ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს. კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია

$$p=2\cdot(1-\Phi(t)).$$

მოყვანილ მაგალითში: $n_2=30>n_1=25\ge 10$ და შესაბამისად, დასაშვებია ნორმალური აპროქსიმაცია. უკანასკნელი ცხრილიდან გამოთვლილი R_1-ის დაკვირვებული მნიშვნელობა იქნება:

$$r_1=5\cdot3.5+9\cdot13.5+6\cdot25.5+3\cdot34.0+2\cdot42.5=479.$$

გარდა ამისა, $ER_1=n_1\cdot(n_1+n_2+1)/2=25\cdot(25+30+1)/12=700$. გამოვთვალოთ ახლა შესწორებული დისპერსია (3) თანაფარდობის მიხედვით:

$$DR_1=\frac{25\cdot30}{12}\cdot\left[25+30+1-\frac{6^3-6+14^3-14+10^3-10+7^3-7+10^3-10+5^3-5+2^3-2}{(25+30)\cdot(25+30-1)}\right]=3386.74.$$

ამიტომ (2) ფორმულის მიხედვით გამოთვლილი T სტატისტიკის დაკვირვებული t მნიშვნელობა ტოლია:

$$t = \frac{|r_1 - ER_1| - 1/2}{\sqrt{DR_1}} = \frac{|479 - 700| - 1/2}{\sqrt{3386.74}} = 3.79.$$

ამასთანავე, კრიტერიუმის შესაბამისი p-მნიშვნელობა იქნება:

$$p = 2\cdot(1-\Phi(t)) = 2\cdot(1-\Phi(3.79)) = 2\cdot(1-0.99...)) < 0.001,$$

რაც ცხადია, მეტყველებს შედეგის სტატისტიკურ მნიშვნელოვნებაზე. შესაბამისად, $H_0: m_D = m_{SL}$ ჰიპოთეზას უარვყოფთ და დავასკვნით, რომ დაავადების დომინანტური და SL ჯგუფები მნიშვნელოვნად განსხვავდებიან: დომინანტურ ჯგუფს აქვს უკეთესი VA.

4. კრასკელ-უოლისის კრიტერიუმი.

ზოგჯერ საჭიროა ორზე მეტი პოპულაციის საშუალოს შედარება, მაგრამ დისპერსიული ანალიზის (ANOVA) მოდელებისაგან განსხვავებით პოპულაციათა განაწილებები შეიძლება არ იყოს ნორმალური. ასეთ შემთხვევებში, დისპერსიული ანალიზის მეთოდები აღარ გამოდგება და პოპულაციათა საშუალოების შესადარებლად საჭიროა პოპულაციათა განაწილებებისაგან თავისუფალი (არაპარამეტრული) პროცედურის აგება. სწორედ ასეთ პროცედურას წარმოადგენს უილკოქსონის კრიტერიუმის განზოგადება, რომელიც ლიტერატურაში ***კრასკელ-უოლისის კრიტერიუმის*** სახელითაა ცნობილი. ძირითად დაშვებად ისევ რჩება პოპულაციათა უწყვეტად განაწილებულობის დაშვება (თუმცა როგორც უილკოქსონის რანგთა ჯამის, ისე კრასკელ-უოლისის კრიტერიუმის გამოყენება შესაძლებელია დაჯგუფებული მონაცემების შემთხვევაშიც).

განვიხილოთ შემდეგი მაგალითი: ცნობილია, რომ თვალის მეტაბოლიზმზე გავლენას ახდენს არაქიდონის მჟავა (Arachidonic acid). კერძოდ, ცნობილია, რომ სხვა ეფექტებთან ერთად, ის იწვევს თვალის ქუთუთოს დახურვას, itching და მის გათავისუფლებას. ოთხი სხვადასხვა წამლის ეფექტიანობის შესადარებლად შეისწავლებოდა მათი მოქმედება ალბინოსურ კურდღლებში არაქიდონის მჟავის შეყვანის შემდეგ. ყოველ ჯგუფში აერთიანებდნენ ექვს კურდღელს და თითოეულ ცხოველს ერთ თვალში აწვეთებდნენ შესაბამის წამალს, ხოლო მეორეში – უვნებელ საშუალებას. ათი წუთის შემდეგ ორივე თვალში აწვეთებდნენ არაქიდონის მჟავას. ამის შემდეგ ახდენდნენ ორივე თვალის კონტროლს 15 წუთის შემდეგ ქუთუთოს დახურვიდან. ყოველ ცდაში აღირიცხებოდა ორივე თვალის ქუთუთო ოთხ ბალიანი, 0-დან 3-მდე სკალით, სადაც 0 ნიშნავდა – თვალი მთლიანად ღიაა, 3 ნიშნავდა, რომ თვალი მთლიანად დახურულია, ხოლო 1 და 2 კი ნიშნავდა გარკვეულ შუალედურ მდგომარეობებს. ეფექტურობის საზომად (x) იყენებდნენ სხვაობას: საცდელ (წამლიან) თვალში თვალის დაფარვის ხარისხს გამოკლებული იგივე სიდიდის მნიშვნელობა მეორე თვალში. x-ის დიდი მნიშვნელობა მიუთითებს წამლის ეფექტიანობაზე. მონაცემები მოყვანილია შემდეგ ცხრილში:

ცხოველის რიგითი ნომერი	წამლის ტიპი							
	I ტიპი		II ტიპი		III ტიპი		IV ტიპი	
	ქულა	რანგი	ქულა	რანგი	ქულა	რანგი	ქულა	რანგი
1	+2	13.5	+1	9.0	+3	20.0	+1	9.0
2	+3	20.0	+3	20.0	+1	9.0	0	4.0
3	+3	20.0	+1	9.0	+2	13.5	0	4.0
4	+3	20.0	+2	13.5	+1	9.0	0	4.0
5	+3	20.0	+2	13.5	+3	20.0	0	4.0
6	0	4.0	+3	20.0	+3	20.0	-1	1.0

კრიტერიუმის ასაგებად აერთიანებენ ყველა ჯგუფს და თითოეულ ინდივიდს მიაწერენ შესაბამის რანგს, ან ტოლი მონაცემების შემთხვევაში, რანგების დიაპაზონს და შემდეგ ითვლიან რანგების საშუალო $\bar{R}_i$ მნიშვნელობას იმავე წესით, როგორც ეს გაკეთებული იყო უილკოქსონის კრიტერიუმის აგებისას (ცხრილში ეს უკვე გაკეთებულია). კრიტერიუმის ასაგებად ადარებენ ჯგუფების ჯამურ R_i რანგებს. კრიტერიუმის სტატისტიკას აქვს სახე:

$$T=\left(\frac{12}{n(n+1)}\cdot\sum_{i=1}^{l}\frac{R_i^2}{n_i}-3\cdot(n+1)\right)\Big/\left(1-\frac{1}{n^3-n}\cdot\sum_{j=1}^{k}(t_j^3-t_j)\right), \qquad (1)$$

სადაც l აღნიშნავს შერჩევათა რაოდენობას, k – ტოლი რანგების მქონე დაკვირვებათა საერთო რაოდენობას, t_j აღნიშნავს მონაცემების რაოდენობას j-ურ კლასტერულ (ტოლი რანგობრივი დიაპაზონის მქონე) ჯგუფში და ბოლოს, $n=\sum_{i=1}^{l}n_i$ - არის გაერთიანებული შერჩევის მოცულობა.

შევნიშნოთ, რომ თუ ყველა $t_j=1$, მაშინ (1) სტატისტიკა ღებულობს შემდეგ სახეს:

$$T=\frac{12}{n(n+1)}\cdot\sum_{i=1}^{l}\frac{R_i^2}{n_i}-3\cdot(n+1)\ . \qquad (2)$$

თუ შერჩევათა მოცულობები ტოლია, მაშინ $n_i=n/l$ და (1) სტატისტიკა ასე გადაიწერება:

$$T=\left(\frac{12l}{n^2\cdot(n+1)}\cdot\sum_{i=1}^{l}R_i^2-3\cdot(n+1)\right)\Big/\left(1-\frac{1}{n^3-n}\cdot\sum_{j=1}^{k}(t_j^3-t_j)\right). \qquad (3)$$

შესამოწმებელია ჰიპოთეზა – H_0 პოპულაციები ერთგვაროვანია (ანუ საშუალოები ტოლია).

მტკიცდება, რომ H_0 ჰიპოთეზის სამართლიანობისას T სტატისტიკას აქვს ხი-კვადრატ განაწილება თავისუფლების ხარისხით $l-1$, $T\cong\chi^2(l-1)$.

სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის T სტატისტიკის დაკვირვებული t მნიშვნელობა აკმაყოფილებს პირობას

$$t > \chi^2_{l-1,\alpha},$$

მაშინ α მნიშვნელოვნების დონით H_0 ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს. კრიტერიუმის შესაბამისი p-მნიშვნელობა ტოლია $p = P\{\chi^2(l-1) > t\}$.

კრიტერიუმი გამოიყენება მაშინ, როცა ყველა $n_i \geq 5$.

განსახილველ მაგალითში კრიტერიუმის სტატისტიკის მნიშვნელობის გამოსათვლელად მოხერხებულია მონაცემები წარმოვადგინოთ შემდეგი ცხრილის სახით:

ქულა	სიხშირე	რანგების დიაპაზონი	საშუალო რანგი
-1	1	1	1.0
0	5	2 – 6	4.0
+1	5	7 – 11	9.0
+2	4	12 – 15	13.5
+3	9	16 – 24	20.0

გამოვთვალოთ ჯგუფების ჯამური R_i რანგები. გვაქვს:

$R_1 = 13.5 + 20.0 + 20.0 + 20.0 + 20.0 + 4.0 = 97.5$,

$R_2 = 9.0 + 20.0 + 9.0 + 13.5 + 13.5 + 20.0 = 85.0$,

$R_3 = 20.0 + 9.0 + 13.5 + 9.0 + 20.0 + 20.0 = 91.5$,

$R_4 = 9.0 + 4.0 + 4.0 + 4.0 + 4.0 + 1.0 = 26.0$.

შესაბამისად, სტატისტიკის მრიცხველი იქნება:

$$\frac{12l}{n^2\cdot(n+1)}\cdot\sum_{i=1}^{l} R_i^2 - 3\cdot(n+1) = \frac{12\cdot 4}{24^2\cdot 25}\cdot(97.5^2 + 85^2 + 91.5^2 + 26.0^2) - 3\cdot 25 = 10.932,$$

ხოლო მნიშვნელი ტოლია:

$$1 - \frac{1}{n^3 - n}\cdot\sum_{j=1}^{k}(t_j^3 - t_j) = 1 - \frac{1}{24^3 - 24}\cdot(5^3 - 5 + 5^3 - 5 + 4^3 - 4 + 9^3 - 9) = 0.926$$

ამიტომ T სტატისტიკის დაკვირვებული მნიშვნელობა იქნება: $t = 10.932/0.926 = 11.804$. მეორეს მხრივ, ხი-კვადრატ განაწილების ზედა კრიტიკული წერტილების ცხრილიდან $\alpha = 0.01$ მნიშვნელოვნების დონისათვის ვპოულობთ $\chi^2_{l-1\,\alpha} = \chi^2_{3,0.01} = 11.3449$. ვინაიდან $11.804 = t > \chi^2_{3,0.01} = 11.3449$, H_0 ჰიპოთეზას უარვყოფთ და დავასკვნით, რომ წამლებს გააჩნიათ ანთების საწინააღმდეგო სხვადასხვა ეფექტი.

იმის დასადგენად, არის თუ არა i-ური და j-ური ჯგუფების საშუალოები ტოლი, იყენებენ შემდეგ სტატისტიკას

$$Z_{i,j} = (\overline{R}_i - \overline{R}_j)\Big/\sqrt{\frac{n(n+1)}{12}\cdot\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}, \qquad (4)$$

სადაც $\overline{R}_i := \dfrac{R_i}{n_i}$.

სტატისტიკური კრიტერიუმი ასე ყალიბდება:

თუ მოცემული α მნიშვნელოვნების დონისათვის $Z_{i,j}$ სტატისტიკის დაკვირვებული $z_{i,j}$ მნიშვნელობა აკმაყოფილებს პირობას

$$|z_{i,j}| > z_{\alpha^*} \quad (\text{სადაც } \alpha^* := \frac{\alpha}{l(l-1)}),$$

მაშინ α მნიშვნელოვნების დონით უარვყოფთ ჰიპოთეზას i-ური და j-ური ჯგუფების საშუალოების ტოლობის შესახებ, წინააღმდეგ შემთხვევაში, ამის საფუძველი არა გვაქვს.

ჩვენს მაგალითში ყველა $n_i = 6$. თუ ავიღებთ $\alpha = 0.05$-ს, მაშინ

$$\alpha^* = \frac{\alpha}{l(l-1)} = \frac{0.05}{4 \cdot 3} = 0.0042.$$

$\overline{R}_1 = \dfrac{R_1}{n_1} = \dfrac{97.5}{6} = 16.25$; $\overline{R}_2 = \dfrac{R_2}{n_2} = \dfrac{85}{6} = 14.7$; $\overline{R}_3 = \dfrac{R_3}{n_3} = \dfrac{91.5}{6} = 15.25$ და $\overline{R}_4 = \dfrac{R_4}{n_4} = \dfrac{26}{6} = 4.33$. ამიტომ $Z_{i,j}$ სტატისტიკის დაკვირვებული მნიშვნელობები იქნება:

$$z_{1,2} = \frac{16.25 - 14.17}{\sqrt{24 \cdot 25 \cdot (1/6 + 1/6)}} = 0.51; \quad z_{1,3} = 0.24; \quad z_{1,4} = 2.92; \; z_{2,3} = -0.27,$$

$$z_{2,4} = 2.41, \quad z_{3,4} = 2.67$$

ნორმალური განაწილების ცხრილიდან ვპოულობთ

$$z_{\alpha^*} = z_{0.0042} = 2.635$$

ვინაიდან ამ კრიტიკულ მნიშვნელობაზე აბსოლუტური სიდიდით მხოლოდ $z_{1,4}$ და $z_{3,4}$ მნიშვნელობებია მეტი, ვასკვნით, რომ I და III ტიპის წამლებს IV ტიპის წამალთან შედარებით გააჩნიათ მნიშვნელოვანი ეფექტები, ხოლო დანარჩენი შედეგები სტატისტიკურად უმნიშვნელოა.

თავი XIII
რეგრესია და კორელაცია

1. მარტივი წრფივი რეგრესია.

ადამიანის პრაქტიკული მოღვაწეობის მრავალ დარგში აუცილებელი ხდება ორ მახასიათებელს (სიდიდეს, ცვლადს) შორის დამოკიდებულების შესწავლა. ორ ცვლადს შორის კავშირის უმარტივესი ფორმაა ისეთი კავშირი, როდესაც ერთი ცვლადის (არგუმენტის) ნებისმიერ კონკრეტულ მნიშვნელობას შეესაბამება მეორე ცვლადის (ფუნქციის) ერთი და მხოლოდ ერთი მნიშვნელობა.

მაგალითი 1. დავუშვათ გამყიდველის ყოველდღიური ხელფასი Y წარმოადგენს ფიქსირებული ხელფასის B_0 და ყოველთვიური x ნავაჭრის B_1 ნაწილის ჯამს. მაშინ, ბუნებრივია x-ს ვუწოდოთ დამოუკიდებელი, ხოლო Y-ს დამოკიდებული ცვლადი. ყოველთვიური გაყიდვების ნებისმიერი კონკრეტული მნიშვნელობისათვის, ჩვენ ცალსახად განვსაზღვრავთ ხელფასს შემდეგი წრფივი მოდელის საფუძველზე:

$$Y_i = B_0 + B_1 x_i \,. \qquad (1)$$

ახლა განვიხილოთ შებრუნებული ამოცანა. დავუშვათ მოდელის B_0 და B_1 პარამეტრები უცნობია, მაგრამ ჩვენ შეგვიძლია დავაფიქსიროთ დამოუკიდებელი და შესაბამისი დამოკიდებული ცვლადების ყოველთვიური მნიშვნელობები. მაშინ x და Y ცვლადებზე ორი დაკვირვებული მნიშვნელობების (x_i, Y_i), $i = 1, 2$, $x_1 \neq x_2$ საფუძველზე (ორი ორუცნობიანი განტოლებათა სისტემის ამოხსნით) მარტივად განისაზღვრება B_0 და B_1 პარამეტრები.

მაგრამ პრაქტიკულ ამოცანებში ორ ცვლადს შორის კავშირი უმეტეს შემთხვევაში არ არის დეტერმინისტული. მაგალითად, არ უნდა მოველოდეთ, რომ ერთი და იგივე x შემოსავლის მქონე ოჯახების სამომხმარებლო ხარჯები Y ერთნაირი იქნება. მაგრამ შესაძლებელია x შემოსავლის მქონე ოჯახების სამომხმარებლო ხარჯების დაჯგუფება x-ზე დამოკიდებული რაიმე მნიშვნელობის სიახლოვეში.

მაგალით 1-ში დავუშვათ, რომ მაღაზიის დირექტორს დამატებით დაწესებული აქვს ე. წ. "პრემია-ჯარიმების" წესი: იგი ყოველთვიურად აჯარიმებს ან აჯილდოებს გამყიდველს რაიმე "შემთხვევითი" ε თანხით, ანუ გამყიდველის ხელფასი გამოითვლება შემდეგი მოდელის მიხედვით:

$$Y = B_0 + B_1 x + \varepsilon \,. \qquad (2)$$

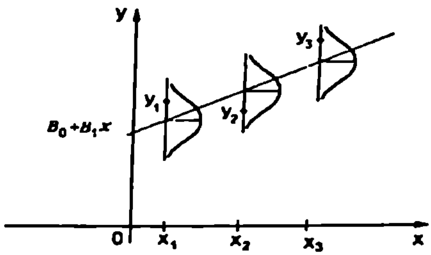
იგულისხმება, რომ დამოუკიდებელი x ცვლადის მნიშვნელობები ზუსტადაა ცნობილი მანამდე, სანამ დავაკვირდებით შესაბამის დამოკიდებული ცვლადის მნიშვნელობებს. x-ის მნიშვნელობები განიხილება როგორც ზუსტი რიცხვები და გადახრები მიეწერება მხოლოდ Y ცვლადს.

ზოგადად ამოცანა ასე აღიწერება: მოცემულია (x_i, Y_i), $i = 1, 2, \ldots, n$ წყვილები, რომელთა შორის შემდეგი დამოკიდებულებაა

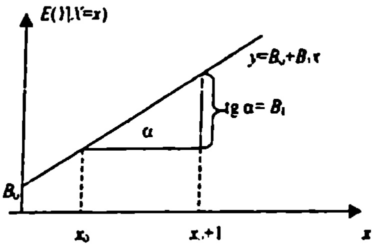
$$Y_i = B_0 + B_1 x_i + \varepsilon_i \,, \quad i = 1, 2, \ldots, n \,, \qquad (3)$$

სადაც B_0 და B_1 უცნობი მუდმივებია, ხოლო ε_i – ნორმალურად განაწილებული შემთხვევითი სიდიდეებია საშუალოთი ნული და უცნობი σ^2 დისპერსიით, $\varepsilon_i \cong N(0,\sigma^2)$. ამ დაშვებაში, გასაგებია, რომ Y_i, $i=1,2,...,n$ აგრეთვე ნორმალურად განაწილებული შემთხვევითი სიდიდეებია, $EY_i = B_0 + B_1x_i$ და $DY_i = \sigma^2$.

$y = B_0 + B_1x$ წრფეს **რეგრესიის წრფე** ეწოდება. B_0 და B_1 კოეფიციენტებს – **რეგრესიის კოეფიციენტები**, ხოლო ε_i შემთხვევით სიდიდეს – **შემთხვევითი გადახრა** ეწოდება. ε_i ის შემთხვევითი სიდიდეა, რომელიც განაპირობებს $y = B_0 + B_1x$ წრფიდან გაბნევას. x ცვლადს უწოდებენ *ამხსნელ* (დამოუკიდებელ) ცვლადს ან **პრედიქტორს**, ხოლო Y-ს კი – *მოპასუხე* (დამოკიდებულ) ცვლადს.



B_1 კოეფიციენტს შეიძლება მიეცეთ შემდეგი ინტერპრეტაცია: ამხსნელი x ცვლადის ერთი ერთეულით ცვლილება იწვევს დამოკიდებული Y ცვლადის ცვლილებას საშუალო B_1 სიდიდით. B_0 კოეფიციენტი წარმოადგენს რეგრესიის წრფის y ღერძთან გადაკვეთის წერტილს.



ჩვენი შემდგომი მიზანია $(x_1,Y_1),(x_2,Y_2),...,(x_n,Y_n)$ დამოუკიდებელი დაკვირვებების საფუძველზე ავაგოთ უცნობი B_0 და B_1 კოეფიციენტებისა და σ^2 დისპერსიის შეფასებები.

2. უმცირეს კვადრატთა მეთოდი.

B_0 და B_1 კოეფიციენტების შეფასებები აღვნიშნოთ შესაბამისად b_0 და b_1 ასოებით და მოვძებნოთ ისინი ე. წ. უმცირეს კვადრატთა მეთოდით, რომელიც დამუშავებული იყო ლაგრანჟისა და გაუსის მიერ XIX საუკუნის დასაწყისში. ამ მეთოდის თანახმად b_0 და b_1 შეფასებების აგება ხდება შემდეგი გამოსახულების მინიმიზაციით:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2 \to \min . \qquad (1)$$

ე. ი. ვეძებთ ისეთ $y = b_0 + b_1 x$ წრფეს, რომლიდანაც Y-ის ვერტიკალური გადახრების კვადრატების ჯამი უმცირესია ყველა სხვა წრფესთან შედარებით.

ვისარგებლოთ მრავალი ცვლადის ფუნქციის ექსტრემუმის მოძებნის ზოგადი წესით: გავაწარმოოთ (1) გამოსახულება b_0 და b_1 ცვლადებით და წარმოებულები გავუტოლოთ ნულს. მივიღებთ, რომ:

$$\sum_{i=1}^{n} 2(Y_i - b_0 - b_1 x_i) = 0 ,$$

$$\sum_{i=1}^{n} 2(Y_i - b_0 - b_1 x_i) x_i = 0 .$$

შევკვეცოთ 2-ზე დაავჯამოთ წევრ-წევრად, შემდეგ ცნობილი და უცნობი წევრები დავალაგოთ ტოლობის სხვადასხვა მხარეს:

$$n b_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i ,$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i .$$

ამოვხსნათ მიღებულ განტოლებათა სისტემა b_0-ისა და b_1-ის მიმართ შეკრება-გამოკლების ხერხით:

$$b_1 \left(n \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \right) = n \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i ,$$

საიდანაც

$$b_1 = \frac{n \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \qquad (2)$$

და

$$b_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - b_1 \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{Y} - b_1 \overline{x} . \qquad (3)$$

შემოვიღოთ აღნიშვნები:

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 , \tag{4}$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} . \tag{5}$$

მაშინ (3) ფორმულა გადაიწერება შემდეგი სახით:

$$b_1 = \frac{SS_{XY}}{SS_X} . \tag{6}$$

$$\hat{Y} = b_0 + b_1 x \tag{7}$$

წრფეს ეწოდება **რეგრესიის წრფის შეფასება**, ან უმცირეს კვადრატთა მეთოდით აგებული წრფე.

შევნიშნავთ, რომ დამოუკიდებელი და დამოკიდებული ცვლადების დაკვირვებული $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ წყვილების საფუძველზე მარტივად გამოითვლება შერჩევითი კორელაციის რიცხვითი მნიშვნელობა. კერძოდ,

$$r = \frac{SS_{XY}}{\sqrt{SS_X \cdot SS_Y}} . \tag{8}$$

მარტივი გარდაქმნების შემდეგ შეფასებული რეგრესიის წრფის განტოლება შეიძლება გადავწეროთ შემდეგი სახით:

$$y = \bar{Y} + b_1(x - \bar{x}) . \tag{9}$$

ამ განტოლებაში შეიძლება შეიძლება ჩაისვას x-ის ნებისმიერი მნიშვნელობა დაკვირვებული x_i, $i = 1, 2, \dots, n$ სიდიდეების ცვალებადობის ინტერვალიდან ($x_{(1)} = \min\{x_1, \dots, x_n\}$-დან $x_{(n)} = \max\{x_1, \dots, x_n\}$-მდე). როგორც აღნიშნული იყო x ცვლადს ამხსნელ ცვლადს უწოდებენ. მისი ჩასმა რეგრესიის განტოლებაში ახსნის Y ცვლადის გარკვეულ ნაწილს, ხოლო გარკვეული ნაწილი Y-ს ცვლილებისა აუხსნელი დარჩება.

რეგრესიის წრფის შეფასებას გააჩნია შემდეგი ორი თვისება:

1. ეს წრფე გადის $(\bar{x}, \bar{Y})$ წერტილზე.

2. $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$.

მიღებული რომელიც $Y = B_0 + B_1 x + \varepsilon$ ფორმულითაა მოცემული გულისხმობს, რომ გაბნევის დიაგრამაზე წერტილების გაბნევა განპირობებულია Y ცვლადის მნიშვნელობათა გაბნევით რეგრესიის მრუდიდან, ხოლო x ცვლადის მნიშვნელობები განიხილება ზუსტ მნიშვნელობებად. თუ რეგრე-

სიის წრფის განტოლებაში x-ის ნაცვლად ჩავსვამთ დაკვირვებულ $x_1,...,x_n$ მნიშვნელობებს, მაშინ შესაბამის $\hat{Y}_1 = b_0 + b_1x_1,..., \hat{Y}_n = b_0 + b_1x_n$ მნიშვნელობებს ეწოდება Y ცვლადის პროგნოზირებული მნიშვნელობები, ხოლო $e_1 = Y_1 - \hat{Y}_1,...,e_n = Y_n - \hat{Y}_n$ სიდიდეებს ნაშთები ეწოდება. სიდიდეს

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1x_i)\right]^2 \qquad (10)$$

ეწოდება *კვადრატების ნარჩენი ჯამი* (ნაშთთა კვადრატების ჯამი) და აღინიშნება აბრევიატურით *SSE* (*sum of square for eror*), $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ახლა (10) თანაფარდობა შეიძლება ასე გადაიწეროს

$$SSE = SS_Y - \frac{SS_{XY}^2}{SS_X}. \qquad (11)$$

გასაგებია, რომ $\sum_{i=1}^{n} e_i^2$ წარმოადგენს $\sum_{i=1}^{n} \varepsilon_i^2$-ის გარკვეულ შეფასებას. მას დიდი მნიშვნელობა აქვს რეგრესიის წრფის დასახასიათებლად.

ადვილი დასანახია, რომ:

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \qquad (12)$$

აღვნიშნოთ უკანასკნელი თანაფარდობის მარცხენა მხარე, Y-ის დაკვირვებული Y_i, $i = 1,2,...,n$ მნიშვნელობების მათი საშუალო მნიშვნელობებიდან გადახრების კვადრატების სრული ჯამური სიდიდე (*total sum of squares*), *SST* აბრევიატურით

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \text{ у} \qquad (13)$$

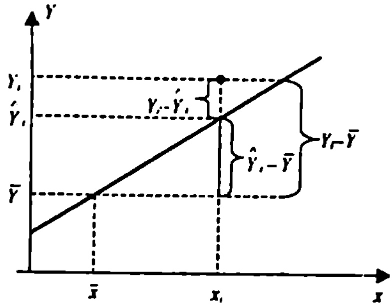
მას *სრულ ვარიაციას* უწოდებენ. მტკიცდება, რომ $\frac{SST}{\sigma^2} \sim \chi^2(n-1)$. გარდა ამისა შემოვიღოთ შემდეგი აღნიშვნები:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \quad SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{ у}$$

SSR არის გაბნევის ის ნაწილი, რომელის ახსნილია რეგრესიის წრფით (*regression sum of squares*). მტკიცდება, რომ $\frac{SSR}{\sigma^2} \sim \chi^2(1)$. *SSE* არის სრული

გაბნევის ის ნაწილი, რომელიც არ აიხსნება რეგრესიის საშუალებით. მტკიცდება, რომ $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$.

საბოლოოდ, ჩვენ გვაქვს: $\underset{n-1}{SST} = \underset{n-2}{SSE} + \underset{1}{SSR}$.



დეტერმინაციის კოეფიციენტი. რეგრესიის ხარისხის დასახასიათებლად შემოვიყვანოთ ე. წ. შერჩევითი დეტერმინაციის კოეფიციენტი:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{SSR}{SST}, \quad \left(\frac{SSR}{SST} = 1 - \frac{SSE}{SST}\right) \tag{14}$$

ე. ი. R^2 წარმოადგენს რეგრესიით ახსნილ გაბნევის წილს სრულ გაბნევაში. მტკიცდება, რომ R^2 – ახალს არაფერს იძლევა, ვინაიდან $R^2=r^2$, სადაც r შერჩევითი კორელაციის კოეფიციენტია.

თუ r^2 მცირეა, ე.ი. რეგრესიით ახსნილია სრული გაბნევის მცირე წილი, მაშინ მოსაძებნია სხვა ალტერნატიული მოდელი (მაგალითად, არაწრფივი ან მრავლობითი რეგრესიის მოდელი), რომელიც უფრო ეფექტურად ახსნიდა Y ცვლადის დაკვირვებული მნიშვნელობების სრულ გაბნევას თავისი საშუალო მნიშვნელობების მიმართ.

კორელაციის კოეფიციენტის საშუალებით შეიძლება შემდეგი ფორმულების დაწერა:

$$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 = R^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = r^2\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2,$$

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = (1-r^2)\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \; y$$

3. სტატისტიკური დასკვნები რეგრესიის კოეფიციენტების შესახებ.

მტკიცდება, რომ b_0 და b_1 B_0 და B_1 პარამეტრების ჩაუნაცვლებელი შეფასებებია ($Eb_0=B_0$, $Eb_1=B_1$) და

$$\sigma_{b_0}^2 := Db_0 = \sigma^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad \sigma_{b_1}^2 := Db_1 = \sigma^2 \cdot \frac{n}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

b_0 და b_1 შეფასებათა საშუალოები და დისპერსიები გამოითვლება (2) და (3) ფორმულებით, იმ ფაქტის გათვალისწინებით, რომ დაკვირვებული Y_i მნიშვნელობები დამოუკიდებელი, ნორმალურად განაწილებული შემთხვევითი სიდიდეებია $EY_i = B_0 + B_1 x_i$ და $DY_i = \sigma^2$ პარამეტრებით.

უცნობი σ^2-ის ჩაუნაცვლებელი შეფასება მოიცემა თანაფარდობით

$$S^2 = (\sum_{i=1}^{n} e_i^2)/(n-2) = (\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2)/(n-2) = SSE/(n-2).$$

სიდიდეს $S = \sqrt{SSE/(n-2)}$ – სტანდარტული შეცდომა ეწოდება.

b_0 და b_1 სტატისტიკების დისპერსიების შეფასებებია შესაბამისად

$$S_{b_0}^2 = S^2 \cdot \frac{\sum x_i^2}{n \cdot SS_X}, \quad S_{b_1}^2 = S^2 \cdot \frac{1}{SS_X}.$$

მტკიცდება, რომ $T_{b_j} = \dfrac{b_j - B_j}{S_{b_j}} \cong t(n-2)$, $j = 0,1$. აქედან გამომდინარე იგება ნდობის ინტერვალი B_j-სათვის. α მნიშვნელოვნების დონის მქონე ნდობის ინტერვალი B_j-სათვის შემდეგი სახისაა:

$$(b_j - s_{b_j} \cdot t_{n-2,\alpha/2}, b_j + s_{b_j} \cdot t_{n-2,\alpha/2}).$$

განვიხილოთ ძირითადი $H_0 : B_j = B_j^0$ ჰიპოთეზა $H_1 : B_j \neq B_j^0$ ალტერნატივის წინააღმდეგ. ამ შემთხვევაში კრიტერიუმის სტატისტიკა იქნება

$$T_{b_j} = (b_j - B_j^0) / S_{b_j},$$

რომელსაც ძირითადი ჰიპოთეზის სამართლიანობის შემთხვევაში გააჩნია სტიუდენტის განაწილება თავისუფლების ხარისხით $n-2$. ამიტომ გადაწყვეტილების მიღების წესი შემდეგია:

თუ T_{b_j} სტატისტიკის დაკვირვებული მნიშვნელობისათვის t_{b_j} და α მნიშვნელოვნების დონისათვის სრულდება პირობა

$$\left|t_{b_j}\right| > t_{n-k-1,\alpha/2},$$

მაშინ უარვყოფთ $H_0 : B_j = B_j^0$ ჰიპოთეზას (შესაბამისად, ვღებულობთ $H_1 : B_j \neq B_j^0$ ჰიპოთეზას), წინააღმდეგ შემთხვევაში H_0-ის უარყოფის საფუძველი არ გაგვაჩნია.

შევნიშნავთ, რომ თუ $B_1 = 0$, მაშინ რეგრესიის წრფე აბსცისთა ღერძის პარალელურია, ე. ი. x-ის ცვლილებით Y არ იცვლება. თუ ჩვენ გვინდა, α მნიშვნელოვნების დონით, დავრწმუნდეთ, რომ არ არსებობს რაიმე (დადებითი ან უარყოფითი) წრფივი კავშირი x და Y სიდიდეებს შორის, ნულოვანი ჰიპოთეზა $H_0 : B_1 = 0$ უნდა შემოწმდეს ორმხრივი $H_1 : B_1 \neq 0$ ალტერნატივის წინააღმდეგ. თუ გვსურს დავრწმუნდეთ დადებითი (შესაბამისად, უარყოფითი) წრფივი კავშირის არსებობაში, მაშინ ვიხილავთ მარჯვენა ცალმხრივ $H_1 : B_1 > 0$ (შესაბამისად, მარცხენა ცალმხრივ $H_1 : B_1 < 0$) ალტერნატივას. ყველა შემთხვევაში ჰიპოთეზის შემოწმება ხდება

$$T_1 := T_{b_1} = \frac{b_1}{S_{b_1}} \cong t(n-2)$$

სტატისტიკის გამოყენებით. გადაწყვეტილების მიღების წესი ასე ყალიბდება:

ა). ორმხრივი ალტერნატივის $H_1 : B_1 \neq 0$ შემთხვევაში $H_0 : B_1 = 0$ ჰიპოთეზას უარვყოფთ, თუ $|t_1| > t_{n-2,\alpha/2}$;

ბ). მარჯვენა ცალმხრივი ალტერნატივის $H_1 : B_1 > 0$ შემთხვევაში $H_0 : B_1 = 0$ ჰიპოთეზას უარვყოფთ, თუ $t_1 > t_{n-2,\alpha}$;

გ). მარცხენა ცალმხრივი ალტერნატივის $H_1 : B_1 < 0$ შემთხვევაში $H_0 : B_1 = 0$ ჰიპოთეზას უარვყოფთ, თუ $t_1 < -t_{n-2,\alpha}$,

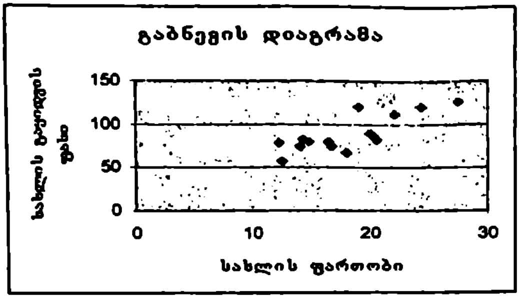
სადაც t_1 არის T_1 სტატისტიკის დაკვირვებული მნიშვნელობა, ხოლო $t_{n-2,\alpha}$ – სტიუდენტის $t(n-2)$ განაწილების ზედა α-კრიტიკული წერტილი.

მაგალითი 1. უძრავი ქონების კომპანიის აგენტს სურს განჭვრიტოს სახლის გასაყიდი ფასი. მისი აზრით სახლის ფასი ყველაზე მჭიდრო კავშირშია მის ფართობთან. მან შემთხვევით შეარჩია ადრე გაყიდული 15 სახლის მონაცემები, რომლებიც მოყვანილია ცხრილის სახით (სადაც x სახლის ფართობია ($\times 10\partial^2$), ხოლო Y – სახლის გაყიდვის ფასი (1000\$)). აგენტის მიზანია სახლის ფასის (დამოუკიდებელი ფასი) პროგნოზირება მისი ფართობის (პრედიქტორის) მიხედვით.

x	Y	x	Y	x	Y
20.0	89.5	14.3	82.5	22.0	112.6
14.8	79.9	27.5	126.3	19.0	120.8
20.5	83.1	16.5	79.3	12.3	78.5
12.5	56.9	24.3	119.9	14.0	74.3
18.0	66.6	20.2	87.6	16.7	74.8

ამოხსნა.

ნაბიჯი 1. შევადგინოთ გაბნევის დიაგრამა



აქედან ჩანს, რომ Y ცვლადის X-ისაგან დამოკიდებულებაში იკვეთება წრფივობის ტენდენცია და შესაძლებელია, რომ მარტივმა წრფივმა რეგრესიამ კარგად აღწეროს იგი. ამიტომ მათემატიკურ მოდელად შეგვიძლია ავირჩიოთ

$$Y_i = B_0 + B_1 x_i + \varepsilon_i .$$

ნაბიჯი 2. გამოვთვალოთ სიდიდეები

$$\sum_{i=1}^{n} x_i ,\ \sum_{i=1}^{n} Y_i ,\ \sum_{i=1}^{n} x_i^2 ,\ \sum_{i=1}^{n} Y_i^2 ,\ \sum_{i=1}^{n} x_i Y_i .$$

ამისათვის შევადგინოთ შემდეგი ცხრილი:

n	x_i	Y_i	x_i^2	Y_i^2	$x_i Y_i$
1	20.0	89.5	400.00	8010.25	1790.00
2	14.8	79.9	219.25	6384.01	1182.52
3	20.5	83.1	420.25	6905.61	1703.55
4	12.5	56.9	156.25	3237.61	711.25
5	18.0	66.6	324.00	4435.56	1198.80
6	14.3	82.5	204.49	6806.25	1179.75
7	27.5	126.3	756.25	15951.69	3473.25
8	16.5	79.3	272.25	6288.49	1308.45
9	24.3	119.9	590.49	14376.01	2913.57
10	20.2	87.6	408.04	7673.76	1769.52
11	22.0	112.6	484.00	12678.76	2477.20
12	19.0	120.8	361.00	14592.64	2295.20
13	12.3	78.5	151.29	6162.25	965.55
14	14.0	74.3	196.00	5520.49	1040.20
15	16.7	74.8	278.89	5595.04	1249.16
$\sum$	172.6	1332.6	5222.24	124618.42	25257.97

ნაბიჯი 3. გამოვთვალოთ b_0-ისა და b_1-ის გამოსათვლელად საჭირო

სიდიდეები:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{272.6}{15}=18.17,\quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1332.6}{15}=88.84 \text{ და}$$

$$SS_x=\sum_{i=1}^{n}x_i^2-n\bar{x}^2=268.19,\ SS_Y=\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2=6230.24.$$

$$SS_{xY}=\sum_{i=1}^{n}x_iY_i-n\overline{xY}=1040.18.$$

ნაბიჯი 4. გამოვთვალოთ b_0 და b_1 შეფასებები:

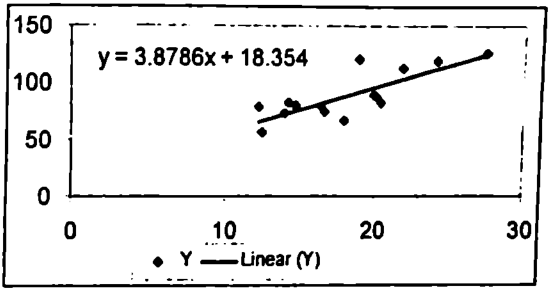
$$b_1=\frac{SS_{xY}}{SS_x}=\frac{1040.18}{268.19}=3.8786.$$

$$b_0=\bar{Y}-b_1\bar{x}=88.84-3.8786\cdot 18.17=18.354.$$

ამიტომ რეგრესიის წრფის შეფასებაა:

$$\hat{Y}=18.354+3.8786x.$$

მიღებული მონაცემები გვაძლევს შემდეგ გრაფიკულ სურათს :



დასკვნა. b_1-ის მიღებული მნიშვნელობა $b_1=3.8786$, ნიშნავს, რომ სახლის ფართობის ყოველი ერთი ერთეულით გაზრდა იწვევს სახლის ფასის საშუალოდ $3.8786\cdot 1000\$=3878.6\$$-ით გაზრდას.

ნაბიჯი 5. გამოვთვალოთ ESS. გვაქვს:

$$SSE=SS_Y-\frac{SS_{xY}^2}{SS_x}=6230.24-\frac{(1040.18)^2}{268.19}=21995.88,$$

საიდანაც მიიღება უცნობი დისპერსიის σ^2-ის შეფასება

$$S^2=\frac{SSE}{n-2}=\frac{21995.88}{13}=168.91.$$

ნაბიჯი 6. გამოვთვალოთ $S_{b_0}^2$ და $S_{b_1}^2$:

$$S_{b_0}^2=S^2\cdot\frac{\sum x_i^2}{n\cdot SS_x}=168.91\cdot\frac{5222.24}{15\cdot 268.19}=219.26,$$

$$S_{b_1}^2 = S^2 \cdot \frac{1}{SS_X} = 168.91 \cdot \frac{1}{268.19} = 0.6298,$$

შესაბამისად, $S_{b_1} = 0.7936$.

ნაბიჯი 7. გამოვთვალოთ დეტერმინაციის კოეფიციენტი:

$$R^2 = 1 - \frac{SSE}{SST} = 0.648.$$

დეტერმინაციის კოეფიციენტის გამოთვლილი მნიშვნელობა გვიჩვენებს, რომ მარტივი წრფივი რეგრესიით აიხსნება სახლის გასაყიდი ფასის ვარიაციის მხოლოდ 64.8%, დანარჩენი 35.2% აუხსნელი რჩება. საჭიროა შეირჩეს სხვა მოდელი (არაწრფივი რეგრესია, ან მრავლობითი რეგრესია).

მაგალითი 2. მაგალით 1-ში:

ა). ავაგოთ $\alpha = 0.05$ მნიშვნელოვნების დონის მქონე ნდობის ინტერვალი B_1-სათვის;

ბ). შევამოწმოთ $H_0: B_1 = 0$ ჰიპოთეზა $H_1: B_1 \neq 0$ ალტერნატივის წინააღმდეგ.

ამოხსნა. ა). ცნობილია, რომ $T_1 := \frac{b_1 - B_1}{S_{b_1}} \cong t(n-2)$, სტიუდენტის ზედა კრიტიკული წერტილების ცხრილში ვპოულობთ $t_{n-2,\alpha/2} = t_{13,0.025} = 2.16$. ამიტომ საძიებელი ნდობის ინტერვალი იქნება:

$$3.88 - 2.16 \cdot 0.7936 < B_1 < 3.88 + 2.16 \cdot 0.7936,$$

$$\text{ანუ } 2.1658 < B_1 < 5.5942.$$

ბ). T_1 სტატისტიკის გამოთვლილი მნიშვნელობა t_1 იქნება:

$$t_1 = \frac{b_1 - B_1}{S_{b_1}} = \frac{3.88 - 0}{0.7936} = 4.8891 > t_{13,0.025} = 2.16.$$

ამიტომ $H_0: B_1 = 0$ ჰიპოთეზას უარვყოფთ (ვინაიდან $|t_1| > t_{n-2,\alpha/2}$).

ამრიგად, $\alpha = 0.05$ მნიშვნელოვნების დონით სარწმუნოა მტკიცება, რომ სახლის ფასსა და მის ფართობს შორის არსებობს წრფივი კავშირი.

4. მრავლობითი რეგრესია.

განვიხილოთ წრფივი მრავლობითი რეგრესიის მოდელი

$$Y = B_0 + B_1 X_1 + \cdots + B_k X_k + \varepsilon, \tag{1}$$

სადაც Y – დამოკიდებული ცვლადია, $X_1, \ldots, X_k$ – დამოუკიდებელი ცვლადებია (მათ ამხსნელ ცვლადებს, პრედიქტორებს, ან რეგრესორებს უწოდებენ), ε – შეშფოთებაა, რომელსაც ხშირად ჭეშმარიტ გადახრას ან შეცდომას უწოდებენ. ε შემთხვევითი სიდიდეა, რომლისთვისაც $E\varepsilon = 0$, ხოლო $D\varepsilon = \sigma^2 > 0$. იგულისხმება, რომ დამოუკიდებელი $X_1, \ldots, X_k$ ცვლადების მნიშვნელობები ზუსტადაა ცნობილი და გადახრები მიეწერება მხოლოდ Y ცვლადს.

ცხადია, რომ Y შემთხვევითი სიდიდის მათემატიკური ლოდინი პირობაში, რომ $X_1=x_1,\dots,X_k=x_k$ მოიცემა ფორმულით:

$$E(Y\mid X_1=x_1,\dots,X_k=x_k)=B_0+B_1x_1+\dots+B_kx_k, \qquad (2)$$

ხოლო დისპერსია იმავე პირობებში იქნება:

$$D(Y\mid X_1=x_1,\dots,X_k=x_k)=\sigma^2.$$

k ცვლადის $y=B_0+B_1x_1+\dots+B_kx_k$ ფუნქციას **ჭეშმარიტი რეგრესიის ფუნქცია** ეწოდება, ხოლო $B_0,B_1,\dots,B_k$ სიდიდეებს კი – **რეგრესიის კოეფიციენტები**.

საბოლოოდ, მრავლობითი რეგრესიის ამოცანა შემდეგნაირად აღიწერება: მოცემულია n მოცულობის შერჩევა $(Y_i,x_{1i},\dots,x_{ki})$, $i=1,2,\dots,n$, $n>k$, სადაც

$$Y_i=B_0+B_1x_{1i}+\dots+B_kx_{ki}+\varepsilon_i,\quad i=1,2,\dots,n, \qquad (3)$$

ხოლო ε_i შემთხვევითი სიდიდეა, რომელიც განაპირობებს Y_i ცვლადის გადახრას ჭეშმარიტი რეგრესიის ფუნქციის $B_0+B_1x_{1i}+\dots+B_kx_{ki}$ მნიშვნელობიდან. ქვემოთ ჩვენ ვიგულისხმებთ, რომ $\varepsilon_1,\dots,\varepsilon_n$ ნორმალურად განაწილებული დამოუკიდებელი შემთხვევითი სიდიდეებია, $\varepsilon_i \cong N(0,\sigma^2)$, $E(\varepsilon_i\cdot\varepsilon_j)=0$, $i\neq j$.

ისევე როგორც მარტივი რეგრესიის შემთხვევაში, პირველ ეტაპზე ხდება რეგრესიის თეორიული $B_0,B_1,\dots,B_k$ კოეფიციენტებისა და უცნობი σ^2 დისპერსიის შეფასება. რეგრესიის კოეფიციენტების შეფასება ისევ წარმოებს უმცირეს კვადრატთა მეთოდის გამოყენებით. მას შემდეგ რაც მიღებულია თეორიული რეგრესიის კოეფიციენტების $b_0,b_1,\dots,b_k$ შეფასებები,

$$\hat{Y}=b_0+b_1x_1+\dots+b_kx_k$$

ფუნქცია განსაზღვრავს რეგრესიის ფუნქციის შეფასებას. სიდიდეს

$$e_i=Y_i-\hat{Y}_i=Y_i-b_0-b_1x_{1i}-\dots-b_kx_{ki}$$

ნაშთი ეწოდება.

ისევე როგორც მარტივი რეგრესიის მოდელში, უცნობი $\sigma^2=D\varepsilon_i$ დისპერსიის შეფასება ეყრდნობა ნაშთების (შესწორებების) კვადრატების ჯამს (*sum of square errors (residuals) – SSE*):

$$SSE:=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2.$$

ვინაიდან ამ გამოსახულებაში შედის $(k+1)$ შეფასებული პარამეტრი, მისი თავისუფლების ხარისხია $n-(k+1)$ და ამიტომ დისპერსიის შეფასებად იღებენ სიდიდეს:

$$S^2=\frac{SSE}{n-(k+1)}=MSE.$$

რადგანაც, დაშვების თანახმად $\varepsilon_i \cong N(0,\sigma^2)$, ამიტომ Y_i ნორმალურებია და როგორც მათი წრფივი კომბინაცია, ასევე ნორმალური იქნება b_j შეფასებებიც. მტკიცდება, რომ

$$T_j := \frac{b_j - B_j}{S_{b_j}} \cong t(n-(k+1)).$$

დეტერმინაციის კოეფიციენტი. დავწეროთ დისპერსიული თანაფარდობა

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2.$$

შემოვიღოთ აღნიშვნები:

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2, \ SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \text{ და } SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2.$$

მაშინ გვაქვს:

$$\underset{n-1}{SST} = \underset{n-k-1}{SSE} + \underset{k}{SSR},$$

სადაც SSR -- არის სრული გაბნევის რეგრესიით ახსნილი ნაწილი, SSE -- სრული გაბნევის ის ნაწილი, რომელიც არ აიხსნება რეგრესიით. თითოეულ წევრს ქვევით მიწერილი აქვს თავისუფლების ხარისხი.

ე. წ. მრავლობითი დეტერმინაციის კოეფიციენტი წარმოადგენს სრული ვარიაციის იმ ნაწილს, რომელიც ახსნილია მრავლობითი რეგრესიის მოდელით:

$$R^2 = 1 - \frac{SSE}{SST} \left(= \frac{SSR}{SST}\right).$$

გარდა ამისა, შემოაქვთ ე. წ. თავისუფლების ხარისხთან შეთანხმებული (*adjusted*) დეტერმინაციის კოეფიციენტი:

$$adjusted\ R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)},$$

რაც აიხსნება იმ გარემოებით, რომ თუ დამოუკიდებელ ცვლადთა რაოდენობა k შედარებადია n-თან, R^2-ის მნიშვნელობა შეიძლება არარეალურად დიდი გამოვიდეს და მიგვიყვანოს მცდარ დასკვნამდე. იმ შემთხვევაში, როცა n მნიშვნელოვნად დიდია k-სთან შედარებით, მაშინ ეს კოეფიციენტები დაახლოებით ტოლია, საზოგადოდ კი $adjusted\ R^2 < R^2$.

მტკიცდება, რომ R^2 სტატისტიკა ე. წ. შერჩევითი მრავლობითი კორელაციის კოეფიციენტის კვადრატის ტოლია $R^2 = r^2$, რომელიც წარმოადგენს შერჩევითი კორელაციის კოეფიციენტს Y და $\hat{Y}$ შემთხვევით სიდიდეებს შორის:

$$r = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(\hat{Y}_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \cdot \sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}}.$$

მოდელის ვარგისიანობის შემოწმება. მოდელის ვარგისიანობის შემოწმება მრავლობითი რეგრესიის შემთხვევაში მდგომარეობს

$$H_0 : B_1 = B_2 = \cdots = B_k$$

ჰიპოთეზის შემოწმებაში

$$H_1 : \text{ერთი მაინც } B_i \neq 0, i = 1, 2, ..., k$$

ალტერნატივის წინააღმდეგ.

გასაგებია, რომ თუ H_0 ჰიპოთეზა დაწუნებული არ იქნება, არც ერთი დამოუკიდებელი ცვლადი (რეგრესორი) არ არის წრფივად დაკავშირებული Y-თან და ამიტომ მოდელი უვარგისია პროგნოზირების მიზნებისათვის. მაშინ როცა, თუ ერთი მაინც $B_i \neq 0$. მოდელი რაიმე თვალსაზრისით მაინც გამოსადეგია.

მტკიცდება, რომ ნულოვანი ჰიპოთეზის სამართიანობისას

$$F = \frac{SSR/k}{SSE/(n-(k+1))} = \frac{MSR}{MSE}$$

სტატისტიკას გააჩნია ფიშერის განაწილება k და $n-(k+1)$ თავისუფლების ხარისხებით, $F \cong F(k, n-(k+1))$.

F-სტატისტიკა შეიძლება ჩაიწეროს მრავლობითი დეტერმინაციის კოეფიციენტის საშუალებით:

$$F = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{n-(k+1)}{k} \cdot \frac{R^2}{1-R^2},$$

ე. ი. F-სტატისტიკა $\frac{n-(k+1)}{k}$ თანამამრავლის სიზუსტით ემთხვევა $\frac{R^2}{1-R^2}$-სტატისტიკას, რომელიც წარმოადგენს ახსნილი ვარიაციის ფარდობას აუხსნელთან და თუ ეს ფარდობა საკმარისად დიდია, ბუნებრივია უარვყოთ H_0 ჰიპოთეზა H_1 ალტერნატივის სასარგებლოდ. შესაბამისად, კრიტიკულ არე იქნება – $F > F_{კრიტ}$ სახის.

გადაწყვეტილების მიღების წესი შემდეგია:

თუ F სტატისტიკის დაკვირვებული მნიშვნელობისათვის f და α მნიშვნელოვნების დონისათვის სრულდება პირობა

$$f = \frac{n-(k+1)}{k} \cdot \frac{R^2}{1-R^2} > F_{k,n-k-1,\alpha}$$

მაშინ უარვყოფთ H_0 ჰიპოთეზას, წინააღმდეგ შემთხვევაში H_0-ის უარყოფის საფუძველი არ გაგვაჩნია.

5. სტატისტიკური დასკვნები მრავლობითი მოდელის პარამეტრების შესახებ.

თუ H_0 ჰიპოთეზა დაწუნებულია, მაშინ ვამოწმებთ ჰიპოთეზას ცალკეული რეგრესიის კოეფიციენტის ნულთან ტოლობის შესახებ. ამრიგად, ყოველი j-სათვის ($j = 1, 2, ..., k$) მოწმდება $H_0 : B_j = 0$ ჰიპოთეზა $H_1 : B_j \neq 0$

ალტერნატივის წინააღმდეგ. ამ შემთხვევაში ტესტის სტატისტიკაა $T_{b_j}=b_j/S_{b_j}$, რომელსაც ნულოვანი ჰიპოთეზის სამართლიანობისას აქვს სტიუდენტის განაწილება თავისუფლების ხარისხით $n-(k+1)$:

$$T_{b_j}=\frac{b_j}{S_{b_j}}\cong t(n-(k+1)).$$

გადაწყვეტილების მიღების წესი შემდეგია:

თუ T_{b_j} სტატისტიკის დაკვირვებული მნიშვნელობისათვის t_{b_j} და α მნიშვნელოვნების დონისათვის სრულდება პირობა

$$|t_{b_j}|>t_{n-k-1,\alpha/2}$$

მაშინ H_0 ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში H_0-ის უარყოფის საფუძველი არ გაგვაჩნია.

α მნიშვნელოვნების დონის ნდობის ინტერვალი B_j კოეფიციენტისათვის შემდეგია:

$$b_j-s_{b_j}\cdot t_{n-k-1,\alpha/2}<B_j<b_j+s_{b_j}\cdot t_{n-k-1,\alpha/2}, \qquad (4)$$

სადაც $t_{n-k-1,\alpha/2}$ – თავისუფლების $n-(k+1)$ ხარისხის მქონე სტიუდენტის განაწილების ზედა $\alpha/2$-კრიტიკული წერტილია.

დავუბრუნდეთ უძრავი ქონების აგენტის ამოცანას, რომელსაც ჩვენ ვიხილავდით მარტივი წრფივი რეგრესიის კონტექსტში. როგორც ვნახეთ, სახლის ფასის ცვალებადობის ახსნა მხოლოდ ფართობის ცვალებადობის ხარჯზე არ იძლევა დამაკმაყოფილებელ შედეგს (დეტერმინაციის კოეფიციენტი იყო $R^2=0.648$, ანუ ახსნილია ცვალებადობის მხოლოდ 64.8%). აგენტმა გადაწყვიტა განეხილა მრავლობითი წრფივი რეგრესიის მოდელი:

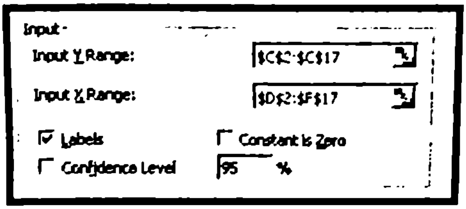
$$Y=B_0+B_1X_1+B_2X_2+B_3X_3+\varepsilon,$$

სადაც Y – სახლის გაყიდვის ფასია (×1000$-ში), X_1 – სახლის ფართობია (×10მ^2-ში), X_2 – სახლის ექსპლოატაციის ხანგრძლივობაა (წლებში), ხოლო X_3 – სახლის კუთვნილი მიწის ნაკვეთის ფართობია (×100მ^2-ში). შეგროვებული მონაცემები ჩაწერილია Excel-ის დავთრის ფურცელში (ქვეპროგრამა: **Tools/Data Analysis/Regression**):

	A	B	C	D	E
1	სახლის ნომერი	Y	X_1	X_2	X_3
2	1	89.5	20	5	4.1
3	2	79.9	14.8	10	6.8
4	3	83.1	20.5	8	6.3
5	4	56.9	12.5	7	5.1
6	5	66.6	18	8	4.2
7	6	82.5	14.3	12	8.6
8	7	126.3	27.5	1	4.9

9	8	79.3	16.5	10	6.2
10	9	119.9	24.3	2	7.5
11	10	87.6	20.2	8	5.1
12	11	112.6	22	7	6.3
13	12	120.8	19	11	12.9
14	13	78.5	12.3	16	9.6
15	14	74.3	14	12	5.7
16	15	74.8	16.7	13	4.8

დეტერმინაციის კოეფიციენტი. მოცემული მონაცემები შევიყვანოთ კომპიუტერში შემდეგი წესით: საწყისი (შემავალი) მონაცემების Y ცვლადის არეში ჩავწეროთ დამოკიდებული ცვლადის რიცხვითი მნიშვნელობების მისამართი $\$C\$2:\$C\$17$, X ცვლადის არეში – $\$D\$2:\$F\$17$:



ქვეპროგრამა **Regression** გვაძლევს შემდეგ ამონაბეჭდს:

SUMMARY OUTPUT

Regression Statistics	
Multiple R / შერჩევითი კორელაციის კოეფიციენტი	0.957
R Square / დეტერმინაციის კოეფიციენტი	0.916
Adjusted R Square / შეთანხმებული დეტერმინაციის კოეფიციენტი	0.893
Standard Error / სტანდარტული შეცდომა	6.894
Observations / დაკვირვებები	15.000

როგორც ვხედავთ, ამ შემთხვევაში დეტერმინაციის კოეფიციენტი (R-square) გამოვიდა $R^2 = 0.916$ ანუ 91.6%, ხოლო *adjusted* $R^2 = 0.893$ ანუ 89.3%. გავიხსენოთ, რომ მარტივი წრფივი რეგრესიის შემთხვევაში $R^2 = 0.648$. ამრიგად, ორი დამატებითი ცვლადის შემოყვანამ საგრძნობლად გაზარდა დეტერმინაციის კოეფიციენტი, სახლის ფასის ვარიაციის 91.6% აიხსნა სამი დამოუკიდებელი ცვლადით და აუხსნელი დარჩა მხოლოდ 8.4%.

მოდელის ვარგისიანობის შემოწმება. კომპიუტრული ამონაბეჭდის მეორე ბლოკი (ANOVA) წარმოადგენს დისპერსიული ანალიზისათვის საჭირო სიდიდეების რიცხვით მნიშენელობებს:

ANOVA
დისპერსიული ანალიზი

	df თავისუფლების ხარისხი	SS კვადრ. ჯამი	MS საშ.კვადრ.გადახრა	F F სტატისტიკა	Significance F მნიშვნელოვნების დონე
Regression	3	SSR=5707.43	$MSR=\dfrac{SSR}{k}$ =1902.479	$f=\dfrac{MSR}{MSE}=$ 40.029	0.000
Residual	11	SSE=522.79	$MSE=\dfrac{MSE}{n-k-1}$ =47.527		
Total	14	SST=6230.23			

აქ ბოლო სვეტში მითითებულია $\alpha = P\{F > f \mid H_0\}$-ის რიცხვითი მნიშვნელობა, ანუ მნიშვნელოვნების ის მინიმალური დონე, რომლითაც არსებული მონაცემების საფუძველზე ხდება H_0 ჰიპოთეზის უარყოფა.

თუ ავიღებთ მაგალითად, $\alpha = 0.05$, მაშინ $F_{k,n-k-1,\alpha} = F_{3,11,0.05} = 8.76$ და ვინაიდან

$$40.029 = f > F_{3,11,0.05} = 8.76,$$

ამიტომ ძირითად ჰიპოთეზას უკუვაგდებთ $\alpha = 0.05$ მნიშვნელოვნების დონით და ვასკვნით, რომ ერთი მაინც რეგრესიის კოეფიციენტი განსხვავებულია ნულისაგან.

სტატისტიკური დასკვნები მრავლობითი მოდელის პარამეტრების შესახებ. ამონაბეჭდის მესამე ბლოკიდან ვიღებთ ინფორმაციას უცნობი რეგრესიის კოეფიციენტების შეფასების შესახებ:

	Coefficients კოეფიციენტი	Standard Error სტანდარტული შეცდომა	t Stat t ტესტი	P-value P მნიშვნელობა	Lower 95%	Upper 95%
Intercept (b_0)	-16.058	19.071	-0.842	0.418	-58.033	25.917
x1 (b_1)	4.146	0.751	5.520	0.000	2.493	5.800
x2 (b_2)	-0.236	0.881	-0.268	0.794	-2.176	1.703
x3 (b_3)	4.831	0.901	5.361	0.000	2.848	6.814

აქ პირველ სვეტში მოცემულია b_0, b_1, b_2, b_3 შეფასებები; მეორეში – მათი სტანდარტული გადახრები, s_{b_j}; მესამეში – $T_{b_j} = b_j / S_{b_j}$ სტატისტიკის დაკვირვებული მნიშვნელობები, $t_{b_j} = b_j / s_{b_j}$; მეოთხე სვეტში – T_{b_j} სტატისტიკებისათვის p-მნიშვნელობები შესაბამისი $H_0 : B_j = 0$ ჰიპოთეზის შესაბ-

ოწმებლად $H_1 : B_j \neq 0$ ალტერნატივის წინააღმდეგ, ხოლო ბოლო ორ სვეტში კი მითითებულია 95%-იანი ნდობის ინტერვალები B_j-სათვის.

$b_0 = -16.058$ მნიშვნელობა მიუთითებს საშუალო გასაყიდ ფასს, როცა $X_1 = X_2 = X_3 = 0$ და მისი ინტერპრეტაცია აზრს მოკლებულია; $b_1 = 4.146$ მნიშვნელობა მიუთითებს იმაზე, რომ ფართობის ყოველი დამატებითი ერთეულისათვის სახლის გაყიდვის ფასი იზრდება საშუალოდ 4146 დოლარით; $b_2 = -0.236$ მნიშვნელობა მიუთითებს იმაზე, რომ სახლის ექსპლოატაციის ყოველი დამატებითი წლისათვის ფასი იკლებს 236 დოლარით; $b_3 = 4.831$ მნიშვნელობა კი მიუთითებს იმაზე, რომ მიწის ნაკვეთის ყოველი დამატებითი 100 კვ. მეტრისათვის სახლის ფასი იზრდება 4831 დოლარით.

ტესტის სტატისტიკის $T_{b_2} = b_2 / S_{b_2}$ დაკვირვებული მნიშვნელობაა $t_2 = -0.268$, ხოლო p-მნიშვნელობაა 0.794, რაც იმის მაჩვენებელია, რომ ნულოვანი ჰიპოთეზის უარყოფა ხდება მნიშვნელოვნების მხოლოდ ისეთი α დონით, რომელიც არანაკლებია 0.794-ზე. ამიტომ მნიშვნელოვნების $\alpha = 0.05$ დონით (0.05<0.794) H_0 ჰიპოთეზას არ უარვყოფთ.

B_3-ის შემთხვევაში p-მნიშვნელობა ნულია. ე. ი. ნებისმიერი $\alpha > 0$ დონით (კერძოდ, როცა $\alpha = 0.05$) $H_0 : B_3 = 0$ ჰიპოთეზას უარვყოფთ $H_1 : B_3 \neq 0$ ალტერნატივის წინააღმდეგ.

ცხრილის თანახმად B_3 კოეფიციენტის 95%-იანი ნდობის ინტერვალია (2.848, 6.814). ვნახოთ, რომ იგივე შედეგს მივიღებდით უშუალო გამოთვლებითაც. მართლაც, (4) თანაფარდობის თანახმად, 95%-იანი ნდობის ინტერვალი B_3 კოეფიციენტისათვის იქნება:

$$b_j - s_{b_j} \cdot t_{n-k-1,\alpha/2} < B_j < b_j + s_{b_j} \cdot t_{n-k-1,\alpha/2}, \quad \alpha/2 = (1-0.95)/2 = 0.025;$$

$$b_3 - s_{b_3} \cdot t_{11,0.025} < B_3 < b_3 + s_{b_3} \cdot t_{11,0.025}, \quad t_{11,0.025} = 2.201;$$

$$4.831 - 0.901 \cdot 2.201 < B_3 < 4.831 + 0.901 \cdot 2.201.$$

ამიტომ საბოლოოდ მივიღებთ, რომ B_3 კოეფიციენტის 95%-იანი ნდობის ინტერვალია (2.848, 6.814).

6. რანგობრივი კორელაცია.

განვიხილოთ ინდივიდთა ან ობიექტთა ერთობლიობა, რომლის ყველა წევრი ხასიათდება რაიმე ორი ნიშნით. ვიგულისხმოთ, რომ ორი ნიშნით დახასიათების შედეგები ჩაწერილია შემდეგი სახით

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}, \tag{1}$$

სადაც x_i და y_i არის i-ური ობიექტის დამახასიათებელი ნიშნები. ამ ნიშნებს შეიძლება არც ჰქონდეთ ზუსტი რიცხობრივი მნიშვნელობები, მაგრამ

შესაძლებელი იყოს მათი ისე დალაგება, რომ თითოეული ნიშნის მიხედვით ყოველ ობიექტს მიეცეს რიგითი ნომერი – რანგი.

მაგალითი 1. ქვემოთ მოყვანილია სხვადასხვა კომპანიების მიერ გამოშვებული ერთი და იგივე სახის პროდუქციის რანჟირების შედეგები ჩატარებული ორი *A* და *B* ექსპერტის მიერ. პირველ სვეტში მოცემულია შესამოწმებელი პროდუქციის პირობითი ნომერი, ხოლო II და III სვეტში *A* და *B* ექსპერტების შეფასებები. მაგალითად, *A* ექსპერტი საუკეთესოდ მიიჩნევს პროდუქციის №3 ნიმუშს, მაშინ როცა *B* ექსპერტი მას მეოთხე ადგილს მიაწერს. თუ რამოდენიმე ნიმუშს ერთნაირად აფასებენ, მაშინ თითოეულს მიეწერება მიწერილი რანგების საშუალო არითმეტიკული. მაგალითად, *A* ექსპერტი ერთნაირად აფასებს პროდუქციის №5 და №7 ნიმუშებს, ისინი იყოფენ მეორე-მესამე ადგილს, ამიტომ ორივეს მიეწერება $(2+3)/2=2.5$-ის ტოლი რანგი. IV სვეტში *A* ექსპერტის რანგები დალაგებულია ზრდის მიხედვით, ხოლო მომდევნო V სვეტში მას მიწერილი აქვს პროდუქციის იგივე ნიმუშისათვის *B* ექსპერტის მიერ მიკუთვნებული რანგი.

პროდუქციის პირობითი ნომერი	*A* ექსპერტის შეფასება	*B* ექსპერტის შეფასება	R_A	R_B	$R_A - R_B$	$(R_A - R_B)^2$
1	4	8	1	4	-3	9
2	9	7	2	1	1	1
3	1	4	3	5	-2	4
4	7	6	4	8	-4	16
5	2.5	5	5	3	2	4
6	8	10	6	2	4	16
7	2.5	1	7	2	-6	36
8	6	2	8	1	7	49
9	5	3	9	10	1	1
10	10	9	10	5	5	25

იმისათვის, რომ შევისწავლოთ *A* და *B* ექსპერტების კრიტერიუმებს შორის ურთიერთდამოკიდებულება, გამოვიყენოთ **სპირმენის ე. წ. რანგობრივი კორელაციის კოეფიციენტი**, რომელიც შემდეგი ფორმულით გამოითვლება:

$$\rho=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}\left(R_A-R_B\right)^2$$

თუ ექსპერტების მიერ გამოყენებულ კრიტერიუმებს შორის სრული თანხვედრაა, მაშინ $\sum_{i=1}^{n}\left(R_A-R_B\right)^2=0$ და $\rho=1$; თუ უარყოფითი კავშირია, მაშინ $\rho=-1$ და თუ კავშირი არ არსებობს, მაშინ $\rho=0$.

განხილულ მაგალითში $\rho=0$.

სპირმენის რანგობრივი კორელაციის კოეფიციენტი გამოიყენება შემდეგი ჰიპოთეზების შესამოწმებლად:

H_0: ნიშნები დამოუკიდებელია, ე. ი $\rho = 0$,

H_1: ნიშნებს შორის დადებითი კავშირია, ე. ი. $\rho > 0$.

ცნობილია, რომ დიდი მოცულობის შემთხვევაში სტატისტიკა

$$t = \sqrt{n-2}\,/\sqrt{1-\rho^2}$$

განაწილებულია სტიუდენტის კანონით, თავისუფლების ხარისხით $n-2$. ამიტომ კრიტერიუმი ასე ყალიბდება:

თუ t სტატისტიკის დაკვირვებული t_g მნიშვნელობისათვის და მოცემული α მნიშვნელოვნების დონისათვის სრულდება პირობა

$$t_g > t_{n-2,\alpha},$$

მაშინ H_0 ჰიპოთეზას უარვყოფთ, წინააღმდეგ შემთხვევაში ამის საფუძველი არა გვაქვს.

გასაგებია, რომ ორმხრივი ალტერნატივის შემთხვევაში — $H_1: \rho \neq 0$, კრიტიკული არე იქნება: $U_1 = \{|t| > t_{n-2,\alpha/2}\}$.

მაგალითი 2. დღისა და საღამოს სწავლების ეკონომისტ სტუდენტებს ეკითხებოდნენ თუ როგორ აფასებენ ისინი სხვადასხვა პროფესიის პრესტიჟს 8 ბალიანი სისტემით. მიიღეს შემდეგი ცხრილი:

პროფესია	დღის	საღამოს
ბუღალტერი	6	3
პროგრამისტი	7	2
ბანკის მენეჯერი	2	6
ადმინისტრატორი	5	4
სტატისტიკოსი	1	7
მარკეტინგის მენეჯერი	4	8
ფინანსისტ-ანალიტიკოსი	3	5
წარმოების მენეჯერი	8	1

იპოვეთ კორელაციის კოეფიციენტი.

ამოხსნა. ამ შემთხვევაში გვაქვს პროფესიებისაგან შედგენილი ერთობლიობა, ხოლო ნიშნების როლში განიხილება დღისა და საღამოს სწავლების სტუდენტთა მიერ ამ პროფესიების შეფასებები. გავაკეთოთ შემდეგი ცხრილი:

პროფესია	დღის	საღამოს	$R_A - R_B$	$(R_A - R_B)^2$
ბუღალტერი	6	3	3	9
პროგრამისტი	7	2	5	25
ბანკის მენეჯერი	2	6	–4	16
ადმინისტრატორი	5	4	1	1
სტატისტიკოსი	1	7	–6	36
მარკეტინგის მენეჯერი	4	8	–4	16
ფინანსისტ-ანალიტიკოსი	3	5	–2	4
წარმოების მენეჯერი	8	1	7	49
$\sum$				156

$$\rho = 1 - \frac{6S_\rho}{n(n^2-1)} = 1 - \frac{6 \cdot 156}{8(8^2-1)} = -0.857.$$

თავი XIV
შემთხვევით სიდიდეთა მოდელირება. მონტე-კარლოს მეთოდი

მონტე-კარლოს მეთოდი გამოიყენება შემდეგი ამოცანის ამოსახსნელად: საჭიროა მოიძებნოს შესასწავლი შემთხვევითი სიდიდის მნიშვნელობა a. მისი განსაზღვრისათვის ირჩევენ X შემთხვევით სიდიდეს, რომლის მათემატიკური ლოდინი ტოლია a–სი, და X შემთხვევითი სიდიდის n ცალი მნიშვნელობის შერჩევიდან, რომელიც მიიღება n ექსპერიმენტში, გამოითვლება შერჩევითი საშუალო:

$$\bar{x} = \frac{\sum x_i}{n},$$

რომელიც მიიღება საძიებელი a რიცხვის შეფასებად:

$$a \approx a^* = \bar{x}.$$

ეს მეთოდი მოითხოვს ექსპერიმენტების დიდი რიცხვის ჩატარებას, ამიტომ მას სხვანაირად **სტატისტიკური ექსპერიმენტების მეთოდი** ეწოდება. მონტე-კარლოს მეთოდის თეორია იკვლევს: როგორ უფრო მიზანშეწონილია აირჩეს X შემთხვევითი სიდიდე, როგორ უნდა ვიპოვოთ მისი შესაძლო მნიშვნელობები, როგორ შევამციროთ გამოყენებული შემთხვევითი სიდიდეების დისპერსია, რათა ცდომილება a–ს a^* –თი შეცვლისას იყოს რაც შეიძლება მცირე.

X შემთხვევითი სიდიდის შესაძლო მნიშვნელობების მოძებნას უწოდებენ **შემთხვევითი სიდიდის გათამაშებას (მოდელირებას)**. ქვემოთ ჩვენ განვიხილავთ შემთხვევითი სიდიდის მოდელირების ზოგიერთ მეთოდს და გავარკვევთ თუ როგორ შევაფასოთ ამ დროს დაშვებული შეცდომა.

თუ ჩვენ გვინდა განვსაზღვროთ დაშვებული შეცდომის ზედა საზღვარი მოცემული საიმედოობის γ ალბათობით, ანუ მოვძებნოთ δ რიცხვი, რომლისთვისაც $p(|\bar{X} - a| \le \delta) = \gamma$, ჩვენ ვღებულობთ გენერალური ერთობლიობის მათემატიკური ლოდინისათვის ნდობის ინტერვალის მოძებნის ცნობილ ამოცანას. ამიტომ ჩვენ ამ ამოცანაზე ცალკე არ შევჩერდებით.

განმარტება 1. (0; 1) ინტერვალზე თანაბრად განაწილებული R შემთხვევითი სიდიდის შესაძლო r მნიშვნელობებს **შემთხვევითი რიცხვები** ეწოდება.

დისკრეტული შემთხვევითი სიდიდის მოდელირება. დავუშვათ, რომ გასათამაშებელია დისკრეტული X შემთხვევითი სიდიდე, ე. ი. X შემთხვევითი სიდიდის ცნობილი განაწილების კანონის მიხედვით მივიღოთ მისი შესაძლო მნიშვნელობების მიმდევრობა:

$$\begin{array}{llll} X & x_1 & x_2 & \dots x_n \\ p & p_1 & p_2 & \dots p_n . \end{array}$$

განვიხილოთ (0; 1) ინტერვალზე თანაბრად განაწილებული R შემთხვევითი სიდიდე და დავყოთ (0, 1) ინტერვალი p_1, $p_1 + p_2$, ..., $p_1 + p_2 + \dots + p_{n-1}$ კოორდინატების მქონე წერტილებით n ქვეინტერვალად: $\Delta_1, \Delta_2, \dots, \Delta_n$, რომელთა სიგრძეები ტოლია შესაბამისი ინდექსის მქონე ალბათობების.

თეორემა 1. თუ ნებისმიერ შემთხვევით რიცხვს $r_j (0 \le r_j < 1)$, რომელიც მოხვდა Δ_i ინტერვალში, შევუსაბამებთ x_i შესაძლო მნიშვნელობას, მაშინ გასათამაშებელ სიდიდეს ექნება მოცემული განაწილების კანონი:

$$X \quad x_1 \quad x_2 \ldots x_n$$

$$p \quad p_1 \quad p_2 \ldots p_n .$$

დამტკიცება. შემთხვევითი სიდიდის შესაძლო მნიშვნელობები ემთხვევა $\{x_1, x_2, \ldots, x_n\}$ სიმრავლეს, რადგანაც ინტერვალების რაოდენობა ტოლია n-ის, და r_j-ს Δ_i ინტერვალში მოხვედრისას შემთხვევით სიდიდეს შეუძლია მიიღოს მხოლოდ ერთი $x_1, x_2, \ldots, x_n$ მნიშვნელობებიდან. ვინაიდან R განაწილებულია თანაბრად, ამიტომ მისი თითოეულ ინტერვალში მოხვედრის ალბათობა ტოლია ამ ინტერვალის სიგრძის, საიდანაც გამოდის, რომ ნებისმიერ x_i მნიშვნელობას შეესაბამება ალბათობა p_i. ამრიგად, გასათამაშებელი შემთხვევითი სიდიდეს გააჩნია მოცემული განაწილების კანონი.

მაგალითი. გავათამაშოთ 10 მნიშვნელობა დისკრეტული X შემთხვევითი სიდიდის, რომლის განაწილების კანონია:

$$X \quad 2 \quad 3 \quad 6 \quad 8$$

$$p \quad 0.1 \quad 0.3 \quad 0.5 \quad 0.1.$$

ამოხსნა. დავყოთ (0. 1) ინტერვალი ქვეინტერვალებად: Δ_1- (0; 0.1), Δ_2 – (0.1; 0.4), Δ_3 - (0.4; 0.9), Δ_4 – (0.9; 1). შემთხვევითი რიცხვების ცხრილიდან ამოვწეროთ 10 რიცხვი: 0.09; 0.73: 0.25; 0.33: 0.76; 0.52; 0.01; 0.35; 0.86; 0.34. პირველი და მეშვიდე რიცხვი ძევს Δ_1 ინტერვალში, შესაბამისად, ამ ორ შემთხვევაში გასათამაშებელი შემთხვევითი სიდიდე მიიღებს მნიშვნელობას $x_1 = 2$; მე-3, მე-4, მე-8 და მე-10 რიცხვები ჩავარდნენ Δ_2 იმტერვალში, რასაც შეესაბამება $x_2 = 3$; მე-2, მე-5, მე-6 და მე-9 რიცხვები აღმოჩნდნენ Δ_3 ინტერვალში, ამასთანავე $X = x_3 = 6$; და ბოლოს, უკანასკნელ ინტერვალში არ ჩავარდა არც ერთი რიცხვი. ამრიგად, X შემთხვევითი სიდიდის გათამაშებული მნიშვნელობებია: 2, 6, 3, 3, 6. 6, 2, 3, 6, 3.

საწინააღმდეგო ხდომილებების მოდელირება. დავუშვათ, რომ უნდა გავითამაშოთ ექსპერიმენტები, რომელთაგან თითოეულში A ხდომილება ჩნდება (ხდება) ცნობილი p ალბათობით. განვიხილოთ დისკრეტული X შემთხვევითი სიდიდე, რომელიც ღებულობს მნიშვნელობას 1 (იმ შემთხვევაში, როცა ხდება A) ალბათობით p და მნიშვნელობას 0 (თუ არ მოხდა A) ალბათობით $q = 1 - p$., შემდეგ ვათამაშებთ ამ შემთხვევით სიდიდეს, ისე როგორც ეს იყო წინა პუნქტში.

მაგალითი. გავათამაშოთ 10 ექსპერიმენტი, რომელთაგან თითოეულში A ხდომილება ხდება ალბათობით 0.3.

ამოხსნა. X შემთხვევითი სიდიდისათვის განაწილების კანონით

$$X \quad 1 \quad 0$$

$$p \quad 0.3 \quad 0.7$$

მივიღებთ ინტერვალებს Δ_1 – (0; 0,3) и Δ_2 – (0,3; 1). გამოვიყენოთ შემთხვევითი რიცხვების იგივე შერჩევა, რაც გვქონდა წინა მაგალითში: 0.09; 0.73; 0.25; 0.33; 0.76; 0.52; 0.01; 0.35; 0.86; 0.34. Δ_1 ინტერვალში მოხვდება პირველი,

მე-3 და მე-7 რიცხვი, ხოლო დანარჩენი კი – Δ_2 ინტერვალში. შესაბამისად, შეგვიძლია ჩავთვალოთ, რომ A ხდომილება მოხდა პირველ, მე-3 და მე-7 ექსპერიმენტში, ხოლო დანარჩენებში კი – არ მოხდა.

ხდომილებათა სრული სისტემის მოდელირება. თუ ხდომილებები A_1, $A_2, \ldots, A_n$, რომელთა ალბათობებია შესაბამისად p_1 , $p_2 , \ldots p_n$, ქმნიან ხდომილებათა სრულ ჯგუფს, მაშინ მათი მოდელირებისათვის (ე. ი. ექსპერიმენტების სერიაში მათი გამოჩენის მიმდევრობის მოდელირება) უნდა გავათამაშოთ დისკრეტული X შემთხვევითი სიდიდე განაწილების კანონით:

$$\begin{array}{cccc} X & 1 & 2 & n \\ p & p_1 & p_2 \ldots & p_n \end{array}$$

ამასთანავე ითვლება, რომ თუ X მიიღებს მნიშვნელობას $x_i = i$, მაშინ ამ ექსპერიმენტში მოხდა A_i ხდომილება.

უწყვეტი შემთხვევითი სიდიდის მოდელირება.

ა). *შებრუნებული ფუნქციების მეთოდი.* დავუშვათ, რომ უნდა გავათამაშოთ უწყვეტი X შემთხვევითი სიდიდე, ე. ი. უნდა მივიღოთ მისი შესაძლო მნიშვნელობების მიმდევრობა x_i ($i = 1, 2, \ldots, n$), როცა ცნობილია მისი განაწილების ფუნქცია $F(x)$.

თეორემა 2. თუ r_i – შემთხვევითი რიცხვია, მაშინ მოცემული მკაცრად ზრდადი $F(x)$ განაწილების ფუნქციის მქონე გასათამაშებელი უწყვეტი X შემთხვევითი სიდიდის შესაძლო x_i მნიშვნელობა, რომელიც შეესაბამება r_i –ს, წარმოადგენს შემდეგი განტოლების ამონახსნს

$$F(x_i) = r_i . \qquad (1)$$

დამტკიცება. ვინაიდან $F(x)$ მკაცრად იზრდება ინტერვალში 0-დან 1-მდე, ამიტომ მოიძებნება (ამასთანავე ერთადერთი) არგუმენტის ისეთი მნიშვნელობა x_i, რომლის დროსაც განაწილების ფუნქცია მიიღებს მნიშვნელობას r_i, ანუ (1) განტოლებას გააჩნია ერთადერთი ამონახსნი: $x_i = F^{-1}(r_i)$, სადაც F^{-1}– არის F ფუნქციის შექცეული ფუნქციაა. ვაჩვენოთ, რომ (1) განტოლების ამონახსნი წარმოადგენს განსახილველი X შემთხვევითი სიდიდის შესაძლო მნიშვნელობას.

წინასწარ ვაჩვენოთ, რომ თუ x_i – შესაძლო მნიშვნელობაა გარკვეული ξ შემთხვევითი სიდიდის, მაშინ ξ შემთხვევითი სიდიდის (c, d) ინტერვალში მოხვედრის ალბათობაა $F(d) - F(c)$. მართლაც, $F(x)$ ფუნქციის მონოტონურობის გამო, $F(x_i) = r_i$ ტოლობის გათვალისწინებით გვაქვს:

$$c < x_i < d \Leftrightarrow F(c) < r_i < F(d).$$

ამიტომ

$$c < x_i < d \Leftrightarrow F(c) < r_i < F(d),$$

შესაბამისად,

$$p(c < \xi < d) = p(F(c) < R < F(d)) = F(d) - F(c).$$

ე. ი. ξ შემთხვევითი სიდიდის (c, d) ინტერვალში მოხვედრის ალბათობა ტოლია ამ ინტერვალზე $F(x)$ განაწილების ფუნქციის ნაზრდის, შესაბამისად, $\xi = X$.

მაგალითი. გავათამაშოთ (5; 8) ინტერვალზე თანაბრად განაწილებული უწყვეტი X შემთხვევითი სიდიდის 3 შესაძლო მნიშვნელობა.

ამოხსნა. გასაგებია, რომ

$$F(x)=\frac{x-5}{3}.$$

ამიტომ უნდა ამოვხსნათ განტოლება $\frac{x_i-5}{3}=r_i$, საიდანაც $x_i=3r_i+5$. ავირჩიოთ 3 შემთხვევითი რიცხვი: 0.23; 0.09; 0.56 და ჩავსვათ ისინი ამ განტოლებაში. მივიღებთ X შემთხვევითი სიდიდის შესაბამის შესაძლო მნიშვნელობებს: $x_1=5.69$; $x_2=5.27$; $x_3=6.68$.

ბ). *სუპერპოზიციის მეთოდი.* თუ გასათამაშებელი შემთხვევითი სიდიდის განაწილების ფუნქცია შეიძლება წარმოდგეს ორი განაწილების ფუნქციის წრფივი კომბინაციის სახით:

$$F(x)=C_1F_1(x)+C_2F_2(x)\quad (C_1,C_2>0),$$

მაშინ $C_1+C_2=1$, ვინაიდან, $F(x)\to 1$, როცა $x\to\infty$.

შემოვიღოთ დამხმარე დისკრეტული შემთხვევითი სიდიდე Z განაწილების ფუნქციით:

Z	1	2
p	C_1	C_2

ავირჩიოთ 2 დამოუკიდებელი შემთხვევითი რიცხვი r_1 და r_2 გავათამაშოთ Z შემთხვევითი სიდიდე r_1 რიცხვის მიხედვით. თუ $Z=1$, მაშინ X –ის შესაძლო მნიშვნელობას ვეძებთ განტოლებიდან $F_1(x)=r_2$, ხოლო თუ $Z=$ $=2$, მაშინ ვხსნით განტოლებას $F_2(x)=r_2$. შეიძლება დამტკიცდეს, რომ ამ შემთხვევაში გასათამაშებელი შემთხვევითი სიდიდის განაწილების ფუნქცია ტოლია მოცემული განაწილების ფუნქციის.

გ). *ნორმალური შემთხვევითი სიდიდის მიახლოებითი გათამაშება.* ვინაიდან (0, 1) ინტერვალში თანაბრად R განაწილებული შემთხვევითი სიდიდისათვის: $M(R)=\frac{1}{2}$, $D(R)=\frac{1}{12}$, ამიტომ (0, 1) ინტერვალზე თანაბრად განაწილებული დამუკიდებელი R_j $(j=1,2,...,n)$ შემთხვევითი სიდიდეების ჯამისათვის $\sum_{j=1}^{n}R_j$:

$$E\left(\sum_{j=1}^{n}R_j\right)=\frac{n}{2},\ D\left(\sum_{j=1}^{n}R_j\right)=\frac{n}{12},\ \sigma=\sqrt{\frac{n}{12}}.$$

ამიტომ ცენტრალური ზღვარითი თეორემის თანახმად ნორმირებულ შემთხვევით სიდიდეს $(\sum_{j=1}^{n}R_j-\frac{n}{2})/\sqrt{n/12}$, როცა $n\to\infty$ ექნება ნორმალურთან ახლოს მყოფი განაწილება, პარამეტრებით $a=0$ და $\sigma=1$. კერძოდ, საკმაოდ კარგი მიახლოება მიიღება, როცა

$$n=12:\ \sum_{j=1}^{12}R_j-6.$$

ამრიგად, იმისათვის, რომ გავაითამაშოთ ნორმირებული ნორმალური შემთხვევითი სიდიდის შესაძლო მნიშვნელობა, უნდა შევკრიბოთ 12 დამოუკიდებელი შემთხვევითი რიცხვი და ჯამს გამოვაკლოთ 6.

ინტეგრალის გამოთვლა მონტე-კარლოს მეთოდით. ვნახოთ, თუ როგორ შეიძლება

$$\int_0^1 f(x)dx \tag{2}$$

ინტეგრალის მიახლოებითი გამოთვლა, სადაც $f:[0,1]\to[0,1]$ უწყვეტი ფუნქციაა. განვიხილოთ [0,1] შუალედში თანაბრად განაწილებულ დამოუკიდებელ შემთხვევით სიდიდეთა მიმდევრობა $X_1, X_2,...$ და ავაგოთ ახალი მიმდევრობა:

$$Z_i = f(X_i),\ i\ge 1. \tag{3}$$

მტკიცდება, რომ $Z_i, i\ge 1$ აგრეთვე დამოუკიდებელ ერთნაირად განაწილებულ შემთხვევით სიდიდეთა მიმდევრობაა და $\forall i$:

$$EZ_i = \int_0^1 f(x)dx.$$

ამიტომ დიდ რიცხვთა კანონის თანახმად ადგილი აქვს კრებადობას:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \to \int_0^1 f(x)dx \ \text{(ალბათობით 1).}$$

შესაბამისად, (2) ინტეგრალის მიახლოებითი გამოთვლისათვის უნდა დამოდელირდეს შემთხვევით სიდიდეთა $(X_i, Y_i), i\ge 1$ მიმდევრობა და გამოითვალოს (3) წესით შედგენილ სიდიდეთა საშუალო არითმეტიკული.

დანართი (სტატისტიკური ცხრილები)

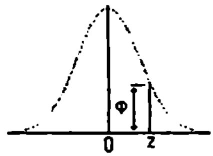
პუასონის განაწილების ცხრილები ($P(k)=\frac{\lambda^k}{k!}e^{-\lambda}$)

	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$
p(0)	0.9048	0.8187	0.7408	0.6703		0.5488	0.4966	0.4493	0.4066
p(1)	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659
p(2)	0.0045	0.0164	0.0333	0.0536	0.0758	0.0988	0.1217	0.1438	0.1647
p(3)	0.0002	0.0011	0.0033	0.0072	0.0126	0.0198	0.0284	0.0383	0.0494
p(4)		0.0001	0.0003	0.0007	0.0016	0.0030	0.0050	0.0077	0.0111
p(5)				0.0001	0.0002	0.0004	0.0007	0.0012	0.0020
p(6)							0.0001	0.0002	0.0003

	$\lambda=1.0$	$\lambda=1.5$	$\lambda=2.0$	$\lambda=2.5$	$\lambda=3.0$	$\lambda=3.5$	$\lambda=4.0$	$\lambda=4.5$	$\lambda=5.0$
p(0)	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
p(1)	0.3679	0.3347	0.2707	0.2052	0.1494	0.1057	0.0733	0.0500	0.0337
p(2)	0.1839	0.2510	0.2707	0.2565	0.2240	0.1850	0.1465	0.1125	0.0842
p(3)	0.0613	0.1255	0.1804	0.2138	0.2240	0.2158	0.1954	0.1687	0.1404
p(4)	0.0153	0.0471	0.0902	0.1336	0.1680	0.1888	0.1954	0.1898	0.1755
p(5)	0.0031	0.0141	0.0361	0.0668	0.1008	0.1322	0.1563	0.1708	0.1755
p(6)	0.0005	0.0035	0.0120	0.0278	0.0504	0.0771	0.1042	0.1281	0.1462
p(7)	0.0001	0.0008	0.0034	0.0099	0.0216	0.0385	0.0595	0.0824	0.1044
p(8)		0.0001	0.0009	0.0031	0.0081	0.0169	0.0298	0.0463	0.0653
p(9)			0.0002	0.0009	0.0027	0.0066	0.0132	0.0232	0.0363
p(10)				0.0002	0.0008	0.0023	0.0053	0.0104	0.0181
p(11)					0.0002	0.0007	0.0019	0.0043	0.0082
p(12)					0.0001	0.0002	0.0006	0.0016	0.0034
p(13)						0.0001	0.0002	0.0006	0.0013
p(14)							0.0001	0.0002	0.0005
p(15)								0.0001	0.0002

სტანდარტული ნორმალური განაწილების სიმკვრივის

$(\varphi(z) = \frac{1}{2\pi} e^{-z^2/2})$ მნიშვნელობები

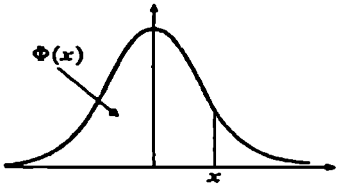


Z	0	1	2	3	4	5	6	7	8	9
0.0	.398942	.398922	.398862	.398763	.398623	.398444	.398225	.397966	.397668	.397330
0.1	.396953	.396536	.396080	.395585	.395052	.394479	.393868	.393219	.392531	.391806
0.2	.391043	.390242	.389404	.388529	.387617	.396668	.385683	.384663	.383606	.382515
0.3	.381388	.380226	.379031	.377801	.376537	.375240	.373911	.372548	.371154	.369728
0.4	.368270	.366782	.365263	.363714	.362135	.360527	.358890	.357225	.355533	.353812
0.5	.352065	.350292	.348493	.346668	.344818	.342944	.341046	.339124	.337180	.335213
0.6	.333225	.331215	.329184	.327133	.325062	.322972	.320864	.318737	.316593	.314432
0.7	.312254	.310060	.307851	.305627	.303389	.301137	.298872	.296595	.294305	.292004
0.8	.289692	.287369	.285036	.282694	.280344	.277985	.275618	.273244	.270864	.268477
0.9	.266085	.263688	.261286	.258881	.256471	.254059	.251644	.249228	.246809	.244390
Z	0	1	2	3	4	5	6	7	8	9
1.0	.241971	.239551	.237132	.234714	.232297	.229882	.227470	.225060	.222653	.220251
1.1	.217852	.215458	.213069	.210686	.208308	.205936	.203571	.201214	.198863	.196520
1.2	.194186	.191860	.189543	.187235	.184937	.182649	.180371	.178104	.175847	.173602
1.3	.171369	.169147	.166937	.164740	.162555	.160383	.158225	.156080	.153948	.151831
1.4	.149727	.147639	.145564	.143505	.141460	.139431	.137417	.135418	.133435	.131468
1.5	.129518	.127583	.125665	.123763	.121878	.120009	.118157	.116323	.114505	.112704
1.6	.110921	.109155	.107406	.105675	.103961	.102265	.100586	.098925	.097282	.095657
1.7	.094049	.092459	.090887	.089333	.087796	.086277	.084776	.083293	.081828	.080380
1.8	.078950	.077538	.076143	.074766	.073407	.072065	.070740	.069433	.068144	.066871
1.9	.065616	.064378	.063157	.061952	.060765	.059595	.058441	.057304	.056183	.055079
Z	0	1	2	3	4	5	6	7	8	9
2.0	.053991	.052919	.051864	.050824	.049800	.048792	.047800	.046823	.045861	.044915
2.1	.043984	.043067	.042166	.041280	.040408	.039550	.038707	.037878	.037063	.036262
2.2	.035475	.034701	.033941	.033194	.032460	.031740	.031032	.030337	.029655	.028985
2.3	.028327	.027682	.027048	.026426	.025817	.025218	.024631	.024056	.023491	.022937
2.4	.022395	.021862	.021341	.020829	.020328	.019837	.019356	.018885	.018423	.017971
2.5	.017528	.017095	.016670	.016254	.015848	.015449	.015060	.014678	.014305	.013940
2.6	.013583	.013234	.012892	.012558	.012232	.011912	.011600	.011295	.010997	.010706
2.7	.010421	.010143	3z98712	3z96058	3z93466	3z90936	3z88465	3z86052	3z83697	3z81398
2.8	3z79155	3z76965	3z74829	3z72744	3z70711	3z68728	3z66793	3z64907	3z63067	3z61274
2.9	3z59525	3z57821	3z56160	3z54541	3z52963	3z51426	3z49929	3z48470	3z47050	3z45666

$\varphi(z)$-ის მნიშვნელობები (გაგრძელება)

z	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
3.0	3z44318	3z43007	3z41729	3z40486	3z39276	3z38098	3z36951	3z35836	3z34751	3z33695
3.1	3z32668	3z31669	3z30698	3z29754	3z28835	3z27943	3z27075	3z26231	3z25412	3z24615
3.2	3z23841	3z23089	3z22358	3z21649	3z20960	3z20290	3z19641	3z19010	3z18397	3z17803
3.3	3z17226	3z16666	3z16122	3z15595	3z15084	3z14587	3z14106	3z13639	3z13187	3z12748
3.4	3z12322	3z11910	3z11510	3z11122	3z10747	3z10383	3z10030	4z96886	4z93577	4z90372
3.5	4z87268	4z84263	4z81352	4z78534	4z75807	4z73166	4z70611	4z68138	4z65745	4z63430
3.6	4z61190	4z59024	4z56928	4z54901	4z52941	4z51046	4z49214	4z47443	4z45731	4z44077
3.7	4z42478	4z40933	4z39440	4z37998	4z36605	4z35260	4z33960	4z32705	4z31494	4z30324
3.8	4z29195	4z28105	4z27053	4z26037	4z25058	4z24113	4z23201	4z22321	4z21473	4z20655
3.9	4z19866	4z19105	4z18371	4z17664	4z16983	4z16326	4z15693	4z15083	4z14495	4z13928
z	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
4.0	4z13383	4z12858	4z12352	4z11864	4z11395	4z10943	4z10509	4z10090	5z96870	5z92993
4.1	5z89262	5z85672	5z82218	5z78895	5z75700	5z72626	5z69670	5z66828	5z64095	5z61468
4.2	5z58943	5z56516	5z54183	5z51942	5z49788	5z47719	5z45731	5z43821	5z41988	5z40226
4.3	5z38535	5z36911	5z35353	5z33856	5z32420	5z31041	5z29719	5z28449	5z27231	5z26063
4.4	5z24942	5z23868	5z22837	5z21848	5z20900	5z19992	5z19121	5z18286	5z17486	5z16719
4.5	5z15984	5z15280	5z14605	5z13959	5z13340	5z12747	5z12180	5z11636	5z11116	5z10618
4.6	5z10141	6z96845	6z92477	6z88297	6z84298	6z80472	6z76812	6z73311	6z69962	6z66760
4.7	6z63698	6z60771	6z57972	6z55296	6z52739	6z50295	6z47960	6z45728	6z43596	6z41559
4.8	6z39613	6z37755	6z35980	6z34285	6z32667	6z31122	6z29647	6z28239	6z26895	6z25613
4.9	6z24390	6z23222	6z22108	6z21046	6z20033	6z19066	6z18144	6z17265	6z16428	6z15629
5.0	6z14867	6z14141	6z13450	6z12791	6z12162	6z11564	6z10994	6z10451	7z99339	7z94414

სტანდარტული ნორმალური განაწილების ფუნქციის ($\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$) მნიშვნელობები

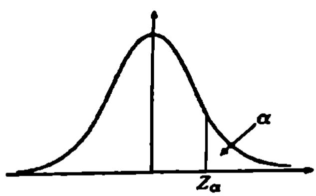


x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.500	**0.33**	0.629	**0.66**	0.745	**0.99**	0.838	**1.32**	0.906	**1.65**	0.950
0.01	0.503	**0.34**	0.633	**0.67**	0.748	**1.00**	0.841	**1.33**	0.908	**1.66**	0.951
0.02	0.507	**0.35**	0.636	**0.68**	0.751	**1.01**	0.843	**1.34**	0.909	**1.67**	0.952
0.03	0.511	**0.36**	0.640	**0.69**	0.754	**1.02**	0.846	**1.35**	0.911	**1.68**	0.953
0.04	0.515	**0.37**	0.644	**0.70**	0.758	**1.03**	0.848	**1.36**	0.913	**1.69**	0.954
0.05	0.519	**0.38**	0.648	**0.71**	0.761	**1.04**	0.850	**1.37**	0.914	**1.70**	0.955
0.06	0.523	**0.39**	0.651	**0.72**	0.764	**1.05**	0.853	**1.38**	0.916	**1.71**	0.956
0.07	0.527	**0.40**	0.655	**0.73**	0.767	**1.06**	0.855	**1.39**	0.917	**1.72**	0.957
0.08	0.531	**0.41**	0.659	**0.74**	0.770	**1.07**	0.857	**1.40**	0.919	**1.73**	0.958
0.09	0.535	**0.42**	0.662	**0.75**	0.773	**1.08**	0.859	**1.41**	0.920	**1.74**	0.959
0.10	0.539	**0.43**	0.666	**0.76**	0.776	**1.09**	0.862	**1.42**	0.922	**1.75**	0.959
0.11	0.543	**0.44**	0.670	**0.77**	0.779	**1.10**	0.864	**1.43**	0.923	**1.76**	0.960
0.12	0.547	**0.45**	0.673	**0.78**	0.782	**1.11**	0.866	**1.44**	0.925	**1.77**	0.961
0.13	0.551	**0.46**	0.677	**0.79**	0.785	**1.12**	0.868	**1.45**	0.926	**1.78**	0.962
0.14	0.555	**0.47**	0.680	**0.80**	0.788	**1.13**	0.870	**1.46**	0.927	**1.79**	0.963
0.15	0.559	**0.48**	0.684	**0.81**	0.791	**1.14**	0.872	**1.47**	0.929	**1.80**	0.964
0.16	0.563	**0.49**	0.687	**0.82**	0.793	**1.15**	0.874	**1.48**	0.930	**1.81**	0.964
0.17	0.567	**0.50**	0.691	**0.83**	0.796	**1.16**	0.876	**1.49**	0.931	**1.82**	0.965
0.18	0.571	**0.51**	0.694	**0.84**	0.799	**1.17**	0.879	**1.50**	0.933	**1.83**	0.966
0.19	0.575	**0.52**	0.698	**0.85**	0.802	**1.18**	0.881	**1.51**	0.934	**1.84**	0.967
0.20	0.579	**0.53**	0.701	**0.86**	0.805	**1.19**	0.882	**1.52**	0.935	**1.85**	0.967
0.21	0.583	**0.54**	0.705	**0.87**	0.807	**1.20**	0.884	**1.53**	0.936	**1.86**	0.968
0.22	0.587	**0.55**	0.708	**0.88**	0.810	**1.21**	0.886	**1.54**	0.938	**1.87**	0.969
0.23	0.590	**0.56**	0.712	**0.89**	0.813	**1.22**	0.888	**1.55**	0.939	**1.88**	0.969
0.24	0.594	**0.57**	0.715	**0.90**	0.815	**1.23**	0.890	**1.56**	0.940	**1.89**	0.970
0.25	0.598	**0.58**	0.719	**0.91**	0.818	**1.24**	0.892	**1.57**	0.941	**1.90**	0.971
0.26	0.602	**0.59**	0.722	**0.92**	0.821	**1.25**	0.894	**1.58**	0.942	**1.91**	0.971
0.27	0.606	**0.60**	0.725	**0.93**	0.823	**1.26**	0.896	**1.59**	0.944	**1.92**	0.972
0.28	0.610	**0.61**	0.729	**0.94**	0.826	**1.27**	0.897	**1.60**	0.945	**1.93**	0.973
0.29	0.614	**0.62**	0.732	**0.95**	0.828	**1.28**	0.899	**1.61**	0.946	**1.94**	0.973
0.30	0.617	**0.63**	0.735	**0.96**	0.831	**1.29**	0.901	**1.62**	0.947	**1.95**	0.974
0.31	0.621	**0.64**	0.738	**0.97**	0.833	**1.30**	0.903	**1.63**	0.948	**1.96**	0.975
0.32	0.625	**0.65**	0.742	**0.98**	0.836	**1.31**	0.904	**1.64**	0.949	**1.97**	0.975

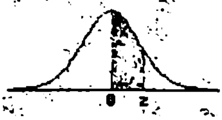
$\Phi(x)$-ის მნიშვნელობები (გაგრძელება)

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1.98	0.976	2.26	0.988	2.54	0.994	2.82	0.997	3.10	0.999
1.99	0.976	2.27	0.988	2.55	0.994	2.83	0.997	3.11	0.999
2.00	0.977	2.28	0.988	2.56	0.994	2.84	0.997	3.12	0.999
2.01	0.977	2.29	0.988	2.57	0.994	2.85	0.997	3.13	0.999
2.02	0.978	2.30	0.989	2.58	0.995	2.86	0.997	3.14	0.999
2.03	0.978	2.31	0.989	2.59	0.995	2.87	0.997	3.15	0.999
2.04	0.979	2.32	0.989	2.60	0.995	2.88	0.998	3.16	0.999
2.05	0.979	2.33	0.990	2.61	0.995	2.89	0.998	3.17	0.999
2.06	0.980	2.34	0.990	2.62	0.995	2.90	0.998	3.18	0.999
2.07	0.980	2.35	0.990	2.63	0.995	2.91	0.998	3.19	0.999
2.08	0.981	2.36	0.990	2.64	0.995	2.92	0.998	3.20	0.999
2.09	0.981	2.37	0.991	2.65	0.995	2.93	0.998	3.21	0.999
2.10	0.982	2.38	0.991	2.66	0.996	2.94	0.998	3.22	0.999
2.11	0.982	2.39	0.991	2.67	0.996	2.95	0.998	3.23	0.999
2.12	0.983	2.40	0.991	2.68	0.996	2.96	0.998	3.24	0.999
2.13	0.983	2.41	0.992	2.69	0.996	2.97	0.998	3.25	0.999
2.14	0.983	2.42	0.992	2.70	0.996	2.98	0.998	3.26	0.999
2.15	0.984	2.43	0.992	2.71	0.996	2.99	0.998	3.27	0.999
2.16	0.984	2.44	0.992	2.72	0.996	3.00	0.998	3.28	0.999
2.17	0.985	2.45	0.992	2.73	0.996	3.01	0.998	3.29	0.999
2.18	0.985	2.46	0.993	2.74	0.996	3.02	0.998	3.30	0.999
2.19	0.985	2.47	0.993	2.75	0.997	3.03	0.998	3.31	0.999
2.20	0.986	2.48	0.993	2.76	0.997	3.04	0.998	3.32	0.999
2.21	0.986	2.49	0.993	2.77	0.997	3.05	0.998	3.33	0.999
2.22	0.986	2.50	0.993	2.78	0.997	3.06	0.998	3.34	0.999
2.23	0.987	2.51	0.993	2.79	0.997	3.07	0.998	3.35	0.999
2.24	0.987	2.52	0.994	2.80	0.997	3.08	0.998	3.36	0.999
2.25	0.987	2.53	0.994	2.81	0.997	3.09	0.999	3.37	0.999

სტანდარტული ნორმალური განაწილების ზედა α-კრიტიკული წერტილები (z_α)



α	0.1	0.05	0.025	0.125	0.01	0.005	0.0025	0.001
z_α	1.28	1.64	1.96	2.24	2.33	2.57	2.81	3.08

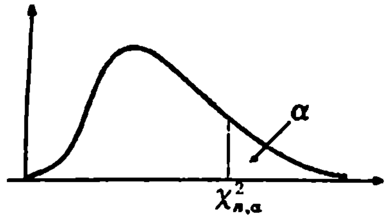


$$\Phi_0(z)=\frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{t^2}{2}}dt$$

ფუნქციის ცხრილები

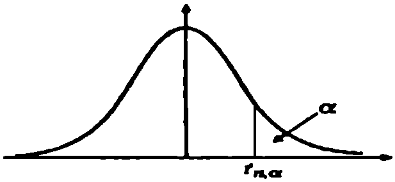
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

χ^2 (ხი კვადრატ) განაწილების ზედა α-კრიტიკული წერტილები ($\chi^2_{\alpha,n}$)



	α							
n	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01
1	0.0002	0.0010	0.0039	0.0158	2.7055	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	0.2107	4.6052	5.9915	7.3778	9.2104
3	0.1148	0.2158	0.3518	0.5844	6.2514	7.8147	9.3484	11.3449
4	0.2971	0.4844	0.7107	1.0636	7.7794	9.4877	11.1433	13.2767
5	0.5543	0.8312	1.1455	1.6103	9.2363	11.0705	12.8325	15.0863
6	0.8721	1.2373	1.6354	2.2041	10.6446	12.5916	14.4494	16.8119
7	1.2390	1.6899	2.1673	2.8331	12.0170	14.0671	16.0128	18.4753
8	1.6465	2.1797	2.7326	3.4895	13.3616	15.5073	17.5345	20.0902
9	2.0879	2.7004	3.3251	4.1682	14.6837	16.9190	19.0228	21.6660
10	2.5582	3.2470	3.9403	4.8652	15.9872	18.3070	20.4832	23.2093
11	3.0535	3.8157	4.5748	5.5778	17.2750	19.6752	21.9200	24.7250
12	3.5706	4.4038	5.2260	6.3038	18.5493	21.0261	23.3367	26.2170
13	4.1069	5.0087	5.8919	7.0415	19.8119	22.3620	24.7356	27.6882
14	4.6604	5.6287	6.5706	7.7895	21.0641	23.6848	26.1189	29.1412
15	5.2294	6.2621	7.2609	8.5468	22.3071	24.9958	27.4884	30.5780
16	5.8122	6.9077	7.9616	9.3122	23.5418	26.2962	28.8453	31.9999
17	6.4077	7.5642	8.6718	10.0852	24.7690	27.5871	30.1910	33.4087
18	7.0149	8.2307	9.3904	10.8649	25.9894	28.8693	31.5264	34.8052
19	7.6327	8.9065	10.1170	11.6509	27.2036	30.1435	32.8523	36.1908
20	8.2604	9.5908	10.8508	12.4426	28.4120	31.4104	34.1696	37.5663
21	8.8972	10.2829	11.5913	13.2396	29.6151	32.6706	35.4789	38.9322
22	9.5425	10.9823	12.3380	14.0415	30.8133	33.9245	36.7807	40.2894
23	10.1957	11.6885	13.0905	14.8480	32.0069	35.1725	38.0756	41.6383
24	10.8563	12.4011	13.8484	15.6587	33.1962	36.4150	39.3641	42.9798
25	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3140
26	12.1982	13.8439	15.3792	17.2919	35.5632	38.8851	41.9231	45.6416
27	12.8785	14.5734	16.1514	18.1139	36.7412	40.1133	43.1945	46.9628
28	13.5647	15.3079	16.9279	18.9392	37.9159	41.3372	44.4608	48.2782
29	14.2564	16.0471	17.7084	19.7677	39.0875	42.5569	45.7223	49.5878
30	14.9535	16.7908	18.4927	20.5992	40.2560	43.7730	46.9792	50.8922

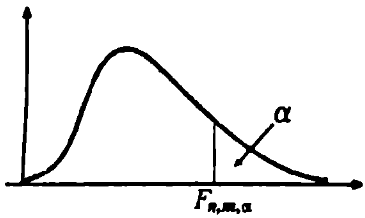
t (სტიუდენტის) განაწილების ზედა α-კრიტიკული წერტილები ($t_{\alpha,n}$)



	α						
n	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
1	3.078	6.314	12.706	31.821	63.656	127.321	318.289
2	1.886	2.920	4.303	6.965	9.925	14.089	22.328
3	1.638	2.353	3.182	4.541	5.841	7.453	10.214
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5	1.476	2.015	2.571	3.365	4.032	4.773	5.894
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385

$F(n,m)$ (ფიშერის) განაწილების ზედა $\alpha = 0.05$ კრიტიკული წერტილები

($F_{n,m,\alpha}$)

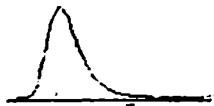


2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	50	10
199	215	224	230	234	236	238	240	241	243	243	244	245	245	246	246	247	247	248	251	2
19.0	19.1	19.2	19.3	19.3	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19
9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66	8.58	8.5
6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80	5.70	5.6
5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56	4.44	4.4
5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87	3.75	3.7
4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44	3.32	3.2
4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15	3.02	2.9
4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94	2.80	2.76
4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77	2.64	2.59
3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65	2.51	2.46
3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54	2.40	2.35
3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46	2.31	2.26
3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39	2.24	2.19
3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33	2.18	2.12
3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28	2.12	2.07
3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23	2.08	2.02
3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19	2.04	1.98
3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16	2.00	1.94
3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12	1.97	1.91
3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.80	1.78	1.60	1.52
3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.75	1.73	1.71	1.69	1.68	1.48	1.39

ფიშერის განაწილების ზედა $\alpha = 0.025$ კრიტიკული წერტილები

df2/df1	1	2	3	4	5	6	7	8	9	10
1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.627 4
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879	3.5257
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.3880	3.3120	3.2497
14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093	3.1469
15	6.1995	4.7650	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227	3.0602

ფიშერის განაწილების ზედა $\alpha = 0.01$ კრიტიკული წერტილები



$F_{(.01, df1, df2)}$

df2/df1	1	2	3	4	5	6	7	8	9	10
1	4052 181	4999 500	5403 352	5624 583	5763 650	5858 986	5928 356	5981 070	6022 473	6055 847
2	98 503	99 000	99 166	99 249	99 299	99.333	99 356	99 374	99 388	99 399
3	34 116	30.817	29 457	28 710	28 237	27 911	27 672	27 489	27 345	27 229
4	21 198	18 000	16 694	15 977	15 522	15 207	14 976	14 799	14 659	14 546
5	16 258	13.274	12 060	11 392	10 967	10 672	10 456	10.289	10 158	10 051
6	13.745	10 925	9 780	9 148	8 746	8 466	8 260	8 102	7 976	7.874
7	12 246	9.547	8 451	7 847	7.460	7 191	6 993	6 840	6 719	6 620
8	11 259	8 649	7 591	7 006	6 632	6 371	6.178	6 029	5.911	5 814
9	10 561	8 022	6 992	6 422	6 057	5 802	5 613	5.467	5 351	5 257
10	10 044	7.559	6 552	5 994	5 636	5 386	5 200	5 057	4 942	4.849
11	9 646	7 206	6.217	5 668	5 316	5 069	4.886	4 744	4 632	4 539
12	9.330	6 927	5 953	5 412	5 064	4 821	4 640	4 499	4 388	4 296
13	9 074	6.701	5 739	5 205	4 862	4 620	4 441	1 302	4 191	4 100
14	8 862	6 515	5 564	5 035	4 695	4 456	4 278	4 140	4.030	3 939
15	8 683	6.359	5 417	4.893	4 556	4.318	4 142	4 004	3 895	3 805

კოლმოგოროვის განაწილების ($K(\lambda) = \sum_{m=-\infty}^{\infty} (-1)^m e^{-2m^2\lambda^2}$) კვანტილები

α	0.20	0.10	0.05	0.02	0.01
$K_{1-\alpha}$	1.073	1.224	1.358	1.517	1.628

D_n სტატისტიკის კრიტიკული $k_{n,\alpha}$ მნიშვნელობები

n	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
1	0.975	0.990	0.995
5	0.563	0.627	0.668
10	0.409	0.456	0.489
15	0.338	0.377	0.404
20	0.294	0.329	0.352
25	0.264	0.295	0.326
30	0.243	0.270	0.290
35	0,224	0.250	0.268
40	0.210	0.234	0.252
45	0.198	0.221	0.232
50	0.188	0.210	0.226
55	0.180	0.201	0.216
60	0.172	0.193	0.207
70	0.160	0.179	0.192
80	0.150	0.167	0.179
90	0.141	0.158	0.169
100	0.134	0.150	0.161

39634 62349 74088 65564 16379 19713 39153 69459 17986 24537
14595 35050 40469 27478 44526 67331 93365 54526 22356 93208
30734 71571 83722 79712 25775 65178 07763 82928 31131 30196
64628 89126 91254 24090 25752 03091 39411 73146 06089 15630
42831 95113 43511 42082 15140 34733 68076 18292 69486 80468

80583 70361 41047 26792 78466 03395 17635 09697 82447 31405
00209 90404 99457 72570 42194 49043 24330 14939 09865 45906
05409 20830 01911 60767 55248 79253 12317 84120 77772 50103
95836 22530 91785 80210 34361 52228 33869 94332 83868 61672
65358 70469 87149 89509 72176 18103 55169 79954 72002 20582

72249 04037 36192 40221 14918 53437 60571 40995 55006 10694
41692 40581 93050 48734 34652 41577 04631 49184 39295 81776
61885 50796 96822 82002 07973 52925 75467 86013 98072 91942
48917 48129 48624 48248 91465 54898 61220 18721 67387 66575
88378 84299 12193 03785 49314 39761 99132 28775 45276 91816

77800 25734 09801 92087 02955 12872 89848 48579 06028 13827
24028 03405 01178 06316 81916 40170 53665 87202 88638 47121
86558 84750 43994 01760 96205 27937 45416 71964 52261 30781
78545 49201 05329 14182 10971 90472 44682 39304 19819 55799
14969 64623 82780 35686 30941 14622 04126 25498 95452 63937

58697 31973 06303 94202 62287 56164 79157 98375 24558 99241
38449 46438 91579 01907 72146 05764 22400 94490 49833 09258
62134 87244 73348 80114 78490 64735 31010 66975 28652 36166
72749 13347 65030 26128 49067 27904 49953 74674 94617 13317
81638 36566 42709 33717 59943 12027 46547 61303 46699 76243

შემთხვევითი რიცხვების ცხრილის გაგრძელება

46574	79670	10342	89543	75030	23428	29541	32501	89422	87474
11873	57196	32209	67663	07990	12288	59245	83638	23642	61715
13862	72778	09949	23096	01791	19472	14634	31690	36602	62943
08312	27886	82321	28666	72998	22514	51054	22940	31842	54245
11071	44430	94664	91294	35163	05494	32882	23904	41340	61185
82509	11842	86963	50307	07510	32545	90717	46856	86079	13769
07426	67341	80314	58910	93948	85738	69444	09370	58194	28207
57696	25592	91221	95386	15857	84645	89659	80535	93233	82798
08074	89810	48521	90740	02687	83117	74920	25954	99629	78978
20128	53721	01518	40699	20849	04710	38989	91322	56057	58573
00190	27157	83208	79446	92987	61357	38752	55424	94518	45205
23798	55425	32454	34611	39605	39981	74691	40836	30812	38563
85306	57995	68222	39055	43890	36956	84861	63624	04961	55439
99719	36036	74274	53901	34643	06157	89500	57514	93977	42403
95970	81452	48873	00784	58347	40269	11880	43395	28249	38743
56651	91460	92462	98566	72062	18556	55052	47614	80044	60015
71499	80220	35750	67337	47556	55272	55249	79100	34014	17037
66660	78443	47545	70736	65419	77489	70831	73237	14970	23129
35483	84563	79956	88618	54619	24853	59783	47537	88822	47227
09262	25041	57862	19203	86103	02800	23198	70639	43757	52064

მილიონი შემთხვევითი რიცხვის სიხშირეები

No.	0	1	2	3	4	5	6	7	8	9	χ^2
1	4923	5013	4916	4951	5109	4993	5055	5080	4986	4974	7.556
2	4870	4956	5080	5097	5066	5034	4902	4974	5012	5009	10.132
3	5065	5014	5034	5057	4902	5061	4942	4946	4960	5019	6.078
4	5009	5053	4966	4891	5031	4895	5037	5062	5170	4886	15.004
5	5033	4982	5180	5074	4892	4992	5011	5005	4959	4872	13.846
6	4976	4993	4932	5039	4965	5034	4943	4932	5116	5070	7.076
7	5011	5152	4990	5047	4974	5107	4869	4925	5023	4902	14.116
8	5003	5092	5163	4936	5020	5069	4914	4943	4914	4946	13.051
9	4860	4899	5138	4959	5089	5047	5030	5039	5002	4937	13.410
10	4998	4957	4964	5124	4909	4995	5053	4946	4995	5059	7.212
11	4948	5048	5041	5077	5051	5004	5024	4886	4917	5004	7.142
12	4958	4993	5064	4987	5041	4984	4991	4987	5113	4882	6.992
13	4968	4961	5029	5038	5022	5023	5010	4988	4936	5025	2.162
14	5110	4923	5025	4975	5095	5051	5035	4962	4942	4882	10.172
15	5094	4962	4945	4891	5014	5002	5038	5023	5179	4852	16.261
16	4957	5035	5051	5021	5036	4927	5022	4988	4910	5053	4.856
17	5088	4989	5042	4948	4999	5028	5037	4893	5004	4972	5.347
18	4970	5034	4996	5008	5049	5016	4954	4989	4970	5014	1.625
19	4998	4981	4984	5107	4874	4980	5057	5020	4978	5021	6.584
20	4963	5013	5101	5084	4956	4972	5018	4971	5021	4901	6.584
$\sum$	99802	100050	100641	100311	100094	100214	99942	99559	100107	99280	13.316

1. ლ. გოკიელი. მათემატიკის საფუძვლები. თსუ, თბილისი, 1958.
2. გ. მანია. ალბათობის თეორია და მათემატიკური სტატისტიკა. სახელმძღვანელო ეკონომიკური ფაკულტეტის სტუდენტებისათვის. თსუ, თბილისი, 1976
3. გ. მანია, ნ. ანთელავა, ა. ქლიბერიძე. ალბათობის თეორიისა და მათემატიკური სტატისტიკის ამოცანათა კრებული. თსუ, თბილისი, 1980.
4. ბ. დოჭვირი. ალბათობის თეორია და მათემატიკური სტატისტიკა. ლექციების კურსი ეკონომიკური ფაკულტეტის სტუდენტებისათვის, ნაწილი I, II. თსუ, თბილისი, 1984.
5. 14. მარი გ., მოსიძე ა., ციგროშვილი ზ., სტატისტიკა. დამხმარე სახელმძღვანელო ESM-თბილისის სტუდენტებისათვის, ESM-თბილისი, 1996.
6. ნ. ლაზრიევა, მ. მანია, გ. მარი, ა. მოსიძე, ა. ტორონჯაძე, თ. ტორონჯაძე, თ. შერვაშიძე. ალბათობის თეორია და მათემატიკური სტატისტიკა ეკონომისტებისათვის. ფონდი «ევრაზია», თბილისი, 2000.
7. თ. ტორონჯაძე, ზ. ციგროშვილი. ბიო-სამედიცინო სტატისტიკა (ლექციების კურსი). ფონდი "ღია საზოგადოება საქართველო", თბილისი, 2001.
8. ო. ფურთუხია. ალბათობის თეორია და მათემატიკური სტატისტიკა. თბილისი, 2007.
9. Allan G. Bluman. Ementary Statistics: a brief version. second edition. Published by McGraw-Hill. New York. 2003.
10. P. Newbold, W. L. Carlson, B. M. Thorne. Statistics for Business and Economics, sixth edition. Prentice Hall. Upper Saddle River. New Jersey, 2007.
11. В. Феллер. ВВедение в теорию вероятностей и ее приложения. Москва. 1967.
12. У. Кокрен. Методы выборочного исследования. Москва "Статистика", 1976.
13. А. Г. Дьячков. Теория вероятностей. Москва. 1980.
14. А. Н. Колмогоров. И. Г Журбенко. А. В. Прохоров. ВВедение в теорию вероятностей. Москва. 1982.
15. В. К. Захаров. Б. А. Севастьянов, В. П. Чистяков. Теория вероятностей. Москва. 1988.
16. В. Е. Гмурман. Руководство к решению задач по теории вероятностей и математической статистике. Москва. 1988.
17. Н. Д. Выск. Конспект лекций по теории вероятностей и математической статистике. Москва, 2003.
18. А. И. Орлов. Математика случая (Вероятность и статистика – основные факты). Москва. 2004.