საქართველოს ტექნიკური უნივერსიტეტი
სტუ

სტატისტიკა

მონაცემთა შესწავლის ხელოვნება და მეცნიერება

ლექციების კურსი

A. Agresti, Ch. Franklin “Statistics.The Art and Science of Learning from Data”

„აგრონომიის“ და „მეცხოველეობის“

სპეციალობის სტუდენტებისთვის

(სტუ ბიბლიოთეკა)

Statistics

The Art and Science of Learning from Data

Third Edition

Alan Agresti
University of Florida

Christine Franklin
University of Georgia

PEARSON

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

დ. ნატროშვილი თ. ალიაშვილი

სტატისტიკა

ლექციების კურსი მომზადებულია სახელმძღვანელოს

A. Agresti, Ch. Franklin “Statistics.The Art and Science of Learning from Data”

მიხედვით

სტატისტიკა

მონაცემთა შესწავლის ხელოვნება და მეცნიერება

სარჩევი

ლექცია 4

ლექცია 5

ლექცია 6

ლექცია 7

ლექცია 8

ლექცია 9

ლექცია 10

ლექცია 11

ლექცია 12

ლექცია 13

ლექცია 14

ლექცია 15

1

რა არის სტატისტიკა?

როგორ გამოვიყენოთ მონაცემები იმისთვის, რომ პასუხი გავცეთ სტატისტიკურ კითხვებს.

ბიზნესის სამყაროში მენეჯერები სტატისტიკას იყენებენ იმისთვის, რომ ჩაატარონ მარკეტინგული კვლევები ახალ პროდუქციებზე, გააანალიზონ შედეგები, პროგნოზირება გაუკეთონ გაყიდვებს და შეაფასონ თანამშრომელთა მუშაობა. საფინანსო საქმიანობაში სტატისტიკას იყენებენ იმისთვის, რომ შეისწავლონ საფონდო ბირჟების საქმიანობები და საინვესტიციო შესაძლებლობანი. მედიცინის მუშაკები სტატისტიკას იყენებენ იმისთვის, რომ შეაფასონ რამდენად უკეთესია რაიმე კონკრეტული დაავადების მკურნალობის ახალი მეთოდები მანამდე არსებულთან შედარებით. ფაქტია, რომ დღეს პროფესიონალურ მოღვაწეთა უდიდესი ნაწილი ეყრდნობა სტატისტიკურ მეთოდებს. ცხადია აგრეთვე ისიც, რომ კონკურენტულ სამუშაო ბაზარზე სტატისტიკის კარგი ცოდნა უზრუნველყოფს მნიშვნელოვან უპირატესობას.

სტატისტიკის ცოდნა მნიშვნელოვანია მაშინაც კი, თუ არასდროს გამოიყენებთ მას თქვენს საქმიანობაში. მისი ცოდნა მაინც დაგეხმარებათ უკეთესი არჩევანის გაკეთებაში. რატომ? იმიტომ რომ თქვენ ყოველდღიურად იბომბებით სტატისტიკური ინფორმაციებით, ახალ-ახალი ანგარიშებით, განცხადებებითა თუ რეკლამებით, პოლიტიკური კამპანიებითა და კვლევა-გამოკითხვებით. როგორ მივხვდეთ, რომელ მათგანს უნდა მიექცეს ყურადღება და რომელს არა? სწორედ ამაში დაგვეხმარება სტატისტიკური დასაბუთების ცოდნა. ზოგჯერ კი იმის გაგებაშიც დაგვეხმარება, რომ ოფიციალურ განცხადებებში ხშირად ვხვდებით, ისეთ შემთხვევებს, როდესაც ეყრდნობიან ხოლმე სტატისტიკური თვალსაზრისით სრულიად გაურკვეველ დასაბუთებებს. სწორედ ამის გამო, სტატისტიკის ცოდნა საშუალებას იძლევა, შევაფასოთ სამედიცინო კვლევები უფრო ეფექტურად ისე, რომ თან ვიცოდეთ, როდის ვიყოთ სკეპტიკურად განწყობილები და როდის არა. მაგალითად, ასპირინის ყოველდღიური მიღება მართლა ამცირებს კიბოს გაჩენის შანსს თუ არა? ნუთუ მართლა არის ნახშირწყლების შეზღუდული დიეტა წონაში მნიშვნელოვანი დაკლების ნამდვილი მიზეზი? სავარაუდოდ, უფრო მეტი ხალხი გაჩერდება 'Starbucks'-თან, მას შემდეგ რაც ნახეს კომერციული რეკლამა ტელევიზიით?

იმ კვლევებში რომელთაც პასუხი უნდა გასცენ ასეთ კითხვებს, ცენტრალური ადგილი უჭირავს ინფორმაციის შეგროვებას. ინფორმაციას კი ვაგროვებთ ექსპერიმენტებსა კვლევებზე დაყრდნობით, რასაც ერთობლივად ვუწოდებთ *მონაცემებს.* თქვენ უკვე გაქვთ წარმოდგენა იმაზე, თუ რას ნიშნავს სიტყვა *სტატისტიკა*. თქვენ ის გაგიგონიათ სპორტული შეხვედრების დროს (მაგ. რამდენი ქულა მოაგროვა გუნდის თითოეულმა კალათბურთელმა), სტატისტიკური მონაცემები ეკონომიკაზე (საშუალო შემოსავლები, უმუშევრობის დონე) და ა.შ. ასეთ შემთხვევაში სტატისტიკა გაიგება როგორც მხოლოდ რიცხვები, გამოთვლილი რაღაც მონაცემებიდან. მაგრამ სტატისტიკა მხოლოდ ეს როდია, იგი შეიძლება განვიხილოთ როგორც მონაცემთა გააზრების, რაოდენობრივი გარკვეულობებისა თუ გაურკვევლობის ვრცელი ველი და არა როგორც მხოლოდ რიცხვებისა და საშინელი ფორმულების ლაბირინთები.

სტატისტიკა არის ხელოვნება და მეცნიერება მონაცემთა ანალიზის, კვლევისა და შესწავლის შესახებ. მისი მთავარი მიზანია შევისწავლით და გარდავქმნათ მონაცემები ისე, რომ მეტად გასაგები გახდეს გარემომცველი სამყარო. მოკლედ რომ ვთქვათ, სტატისტიკა არის ხელოვნება და მეცნიერება მონაცემებზე დაყრდნობით პრობლემის შესწავლის შესახებ.

სტატისტიკური მეთოდები გვეხმარება საკითხების ობიექტურ შეწავლაში. სტატისტიკური პრობლემის კვლევა მოიცავს ოთხ კომპონენტს:

1) პრობლემის ფორმულირება,
2) მონაცემთა შეგროვება,
3) მონაცემთა ანალიზი, და
4) შედეგების ინტერპრეტაცია.

ქვემოთ წარმოდგენილ ამოცანებში ისმება კითხვები, რომლებზედაც პასუხის გაცემა შესაძლებელი იქნება მხოლოდ სტატისტიკური კვლევების დახმარებით.

სცენარი 1: არჩევნების პროგნოზირება Exit Poll-ის საშუალებით. როგორც იცით, ტელევიზიით ხშირად აცხადებენ ხოლმე არჩევნებში გამარჯვებულს იმაზე კარგა ხნით ადრე, ვიდრე მოხდება ხმების საბოლოო დათვლა. ისინი ამას აკეთებენ ამომრჩეველთა გამოკითხვებზე დაყრდნობით, მას შემდეგ რაც მათ უკვე გააკეთეს არჩევანი, ანუ როცა ისინი ტოვებენ სააჩევნო უბანს. Exit Poll-ის საშუალებით ძალიან ხშირად შესაძლებელია სწორად გამოვიცნოთ არჩევნებში გამარჯვებული მხოლოდ

რამდენიმე ათასი გამოკითხულის პასუხების მიხედვით, როცა ამომრჩეველთა რაოდენობა მილიონობითაა.

2010 წელს კალიფორნიის გუბერნატორობისთვის კენჭს იყრიდნენ დემოკრატების კანდიდატი ჯერი ბრაუნი და რესპუბლიკელების კანდიდატი მეგ უიტმანი. 3889 გამოკითხულიდან 53.1%-მა ხმა მისცა ჯერი ბრაუნს. იყო თუ არა ეს საკმარისი იმისთვის, რომ ბრაუნი გამოცხადებულიყო გამარჯვებულად, მაშინ როცა განსხვავება ასეთი მცირეა და არჩევნებში მონაწილეობდა კალიფორნიის 9.5 მილიონი ამომრჩეველი?

სცენარი 2: სამედიცინო კვლევების დასკვნები. სამედიცინო კვლევების უდიდეს ნაწილს საფუძვლად უდევს სტატისტიკური მეთოდები. განვიხილოთ შემთხვევები, რომელთანაც სტატისტიკა არის უშუალო კავშირში.

გულის დაავადება ერთერთი ყველაზე ხშირად გავრცელებული სიკვდილიანობის მიზეზია ინდუსტრიულად განვითარებულ ქვეყნებში. შეერთებულ შტატებსა და კანადაში ყოველწლიურად სიკვდილის მიზეზი, დაახლოებით 30%, არის გულის დაავადებები, ძირითადად კი ინფარქტები. ასპირინის რეგულარული მოხმარება ამცირებს გულის ინფარქტით გამოწვეულ სიკვდილიანობას? ჰარვარდის სამედიცინო სკოლამ ჩაატარა საეტაპო კვლევა. კვლევაში მონაწილე ადამიანები რეგულარულად ღებულობდნენ ასპირინს ან პლაცებოს (ტაბლეტები აქტიური ინგრედიენტების გარეშე). მათგან ვინც იღებდა ასპირინს, 0.9%-ს აღენიშნა ინფარქტი კვლევის პროცესში. ხოლო მათგან ვინც იღებდა პლაცებოს, ინფარქტი აღენიშნა 1.7% -ს, ანუ დაახლოებით ორჯერ მეტს.

შეგიძლიათ დაასკვნათ, რომ ასპირინის რეგულარული მიღება სასარგებლოა? ან, როგორ ავხსნათ განსხვავება? ან, როგორ გადაწყდა ვის უნდა მიეღო ასპირინი და ვის პლაცებო? შეიძლება ისინი, ვინც იღებდნენ ასპირინს, მხოლოდ იმიტომ ჰქონდათ უკეთესი შედეგი, რომ ზოგადად უფრო ჯანმრთლები იყვნენ ვიდრე ისინი, ვინც იღებდნენ პლაცებოს? ან, ისინი ვინც იღებდნენ ასპირინს, იქნებ უკეთეს დიეტაზე იყვნენ, ან საშუალოდ უფრო მეტს ვარჯიშობდნენ?

დიდი ხნის განმავლობაში კამათი იყო იმაზეც, რომ დიდი რაოდენობით ვიტამინ C-ს მიღება სასარგებლოა. ზოგიერთმა კვლევამ აჩვენა, რომ ეს მართლაც ასეა. მაგრამ ზოგიეთი მეცნიერი აკრიტიკებდა ამ კვლევას და ამტკიცებდა, რომ შემდგომი სტატისტიკური ანალიზი არის უაზრობა. როგორ გავიგოთ, როდის დავუჯეროთ სტატისტიკური კვლევის შედეგებს მედიცინაში, თუკი ამას აცხადებენ მედიაში?

სცენარი 3: ხალხის რწმენის გამოკვლევა. რამდენად მსგავსია თქვენი აზრები და ცხოვრების წესი სხვებთან შედარებით? ამის გარკვევა შედარებით ადვილია. ყოველ მეორე წელს ჩიკაგოს უნივერსიტეტთან არსებული ეროვნული აზროვნების კვლევის ცენტრი ატარებს ზოგად სოციალურ კვლევას (GSS). კვლევის დროს ხდება რამდენიმე ათასი ზრდასრული ამერიკელის გამოკითხვა. გამოკითხვა წრმოადგენს მონაცემებს ამერიკული საზოგადოების აზროვნებისა და ცხოვრების წესის შესახებ. კითხვები არის საკმაოდ ბევრი და ძალიან ფართო სპექტრის. მაგ., „გჯერათ თუ არა სიკვდილის შემდგომი ცხოვრების?“, „გაქვთ სურვილი, გადაიხადოთ უფრო მეტი გადასახადები, იმისთვის რომ დავიცვათ გარემო?“, „რამდენ დროს ატარებთ ტელევიზორთან ყოველდღიურად?“ მსგავსი კვლევები ტარდება სხვა ქვეყნებშიც. ასეთია, მაგალითად, ევრობარომეტრის კვლევები ევროკავშირის მასშტაბით. ამ და სხვა მსგავსი კვლევებიდან ჩვენ გამოვიყენებთ მონაცემებს სტატისტიკური მეთოდების სწორად გამოყენების საილუსტრაციოდ.

მიზეზები იმისა, თუ რატომ უნდა გამოვიყენოთ სტატისტიკური მეთოდები. ახლახანს მოყვანილი სცენარები ილუსტრაციაა სტატისტიკის სამი მთავარი კომპონენტის, რომელთა საშუალებითაც პასუხი უნდა გაეცეს სტატისტიკურ კითხვებს. ეს კომპონენტებია:

- **დიზაინი**: დაგეგმვა, თუ როგორ მივიღოთ მონაცემები, იმისთვის რომ ვუპასუხოთ დასმულ კითხვებს.
- **აღწერა**: მიღებული მონაცემების შეჯამება და ანალიზი.
- **დასკვნა**: გადწყვეტილების მიღება და მონაცემებზე დაფუძნებული პროგნოზების გაკეთება დასმულ კითხვებზე საპასუხოდ.

დიზაინი არის დაგეგმვა იმისა, თუ როგორ უნდა მივიღოთ მონაცემები, რომლებიც შუქს მოფენს ინტერესის ობიექტს. როგორ ჩაატარებდით ექსპერიმენტს იმისთვის, რომ დაგედგინათ მართლა სასარგებლოა C ვიტამინის რეგულარულად დიდი დოზა? მარკეტინგში, როგორ შეარჩევთ გამოსაკვლევ ხალხს, რომ შედეგად მიიღოთ მონაცემები, რომლებიც განაპირობებენ მომავალი გაყიდვების კარგ პროგნოზს?

აღწერა ნიშნავს მონაცემებში კანონზომიერების შესწავლას და შეჯამებას. ზოგადად, მონაცემთა ფაილები ძალიან დიდია. მაგალითად, უკვე დიდი ხანია ზოგად სოციალურ გამოკვლევას აქვს მონაცემები დაახლოებით ას მახასიათებელზე და ეს ყველაფერი მრავალ ათას კაცზე. ასეთი საწყისი მონაცემები ძნელი შესაფასებელია. ასეთ შემთხვევაში ჩვენ ამ მახასიათებლებს მარტივად შევუსაბამებთ რიცხვებს. უფრო

მეტად ინფორმაციულია, ვიმოქმედოთ მცირე რაოდენობის რიცხვებზე ან გამოვიყენოთ გრაფიკები მონაცემთა შესაჯამებლად. მაგალითად, როგორიცაა ტელევიზორის ყურების საშუალო ხანგრძლივობა, ან გრაფიკი, რომელიც გამოსახავს დამოკიდებულებას ტელევიზორის ყურების საშუალო დროსა და კვირაში ვარჯიშის საშუალო დროს შორის.

დასკვნა ნიშნავს მონაცემებზე დაყრდნობით მივიღოთ გადაწყვეტილება ან გავაკეთოთ პროგნოზები. როგორც წესი გადაწყვეტილება ან პროგნოზირება ეხება ადამიანების დიდ ჯგუფს და არა მხოლოდ მათ, ვინც მონაწილეობდა კვლევაში. მაგალითად, Exit poll-ის კვლევაში, რომელიც აღწერილია სცენარი 1-ში, ამორჩევა შეადგენდა 3889 ამომრჩეველს, რომელთაგან 53.1%-მა თქვა, რომ ხმა მისცა ჯერი ბრაუნს. ამ მონაცემებზე დაყრდნობით, შეიძლება დავასკვნათ, რომ 9.5 მილიონი ამომრჩევლის უდიდესმა ნაწილმა მას მისცა ხმა. დავადგენთ რა პროცენტს *აღწერილ* 3889 ამომრჩეველში, *ვასკვნით* შედეგს 9.5 მილონი ამომრჩევლის შემთხვევაში.

სტატისტიკური აღწერა და დასკვნა წამოადგენენ მონაცემთა ანალიზის ურთიერთდამატებით ხერხებს. სტატისტიკური აღწერა შეიცავს სასარგებლო მითითებებს მონაცემებში არსებული კანონზომიერების საპოვნელად, მაშინ როცა დასკვნა გვეხმარება გავაკეთოთ პროგნოზები და გადავწყვიტოთ, აქვს თუ არა აზრი შემჩნეული კანონზომიერებების გამოყენებას. თქვენ შეგიძლიათ გამოიყენოთ ორივე მათგანი საზოგადოებისთვის მნიშვნელოვანი საკითხების გამოსაკვლევად.

და ბოლოს, თემა, რომელიც ჯერ არ გვიხსენებია, მაგრამ არის სტატისტიკური დასკვნების საფუძველი, არის **ალბათობა**. იგი არის დასკვნებისგამოტანის ძირითადი დასაყრდენი, ვინაიდან რიცხობრივად გამოსახავს სხვადასხვა შედეგების დადგომის შესაძლებლობებს. ჩვენ ვსწავლობთ ალბათობას იმიტომ, რომ ის გვეხმარება პასუხი გავცეთ, მაგალითად, ასეთ კითხვებს: „თუ ბრაუნი განწირული იყო არჩევნების წაგებისთვის (ანუ მას მხარი დაუჭირა ნახევარზე ნაკლებმა ამომრჩეველმა), მაშინ რა იყო შანსი იმისა, რომ Exit poll-ის 3889 ამომრჩეველიდან 53.1% -მა ამომრჩეველმა მას დაუჭურა მხარი?“ თუ ამის შანსი იქნებოდა ძალიან მცირე, მაშინ ჩვენ უფრო კომფორტულად ვიგრძნობდით თავს გვეთქვა, რომ მის გადარჩევას 9.5 მილიონი ამომრჩევლის უდიდესმა ნაწილმა დაუჭირა მხარი.

აქტიურობა 1: ინტრენეტიდან მონაცემების ჩამოტვირთვა. ადვილია მივიღოთ General Social Survey (GSS)-ის მონაცემთა აღწერითი შეჯამება. ამის დემონსტრირებას გავაკეთებთ ერთი კითხვის მაგალითზე, რომელიც დასმული იყო

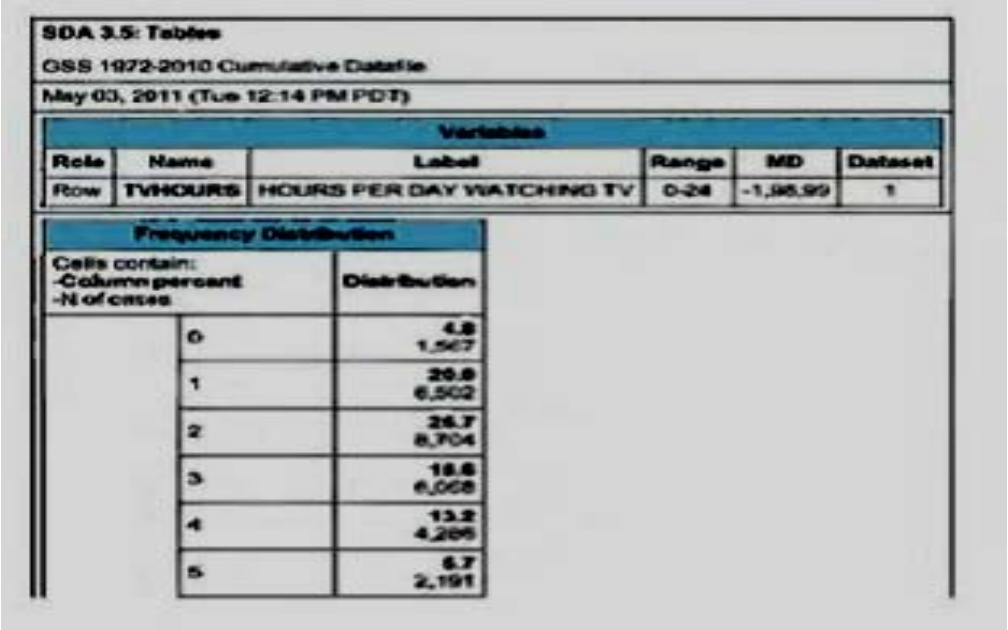
ბოლო გამოკითხვებზე, „ჩვეულებრივ დღეს სავარაუდოდ რამდენ საათს უყურებთ ტელევიზორს?“

- Go to the Web site sda.berkeley.edu/GSS.
- Click on GSS- with *No Weight*as the default weight selection.
- The GSS name for the number of hours of TV watching is

TVHOURS. Type TVHOURS as the row variable name.

- In the Weight menu, make sure that *No Weight* is selected.

Click on *Run the Table*.



SDA 3.5: Tables

GSS 1972-2010 Cumulative Datafile

May 03, 2011 (Tue 12:14 PM PDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	TVHOURS	HOURS PER DAY WATCHING TV	0-24	-1,98,99	1

Frequency Distribution

Cells contain: -Column percent -N of cases	Distribution
0	4.8 1,567
1	20.0 6,502
2	26.7 8,704
3	18.6 6,058
4	13.2 4,286
5	6.7 2,191

GSS 1972-2010 შემაჯამებელი მონაცემთა ფაილი

3 მაისი, 2011 (12:14 სამშაბათი, PDT)

ცვლადები

როლი	სახელი	მონიშვნა	სპექტრი		მონაცემები
ხაზი	ტვ საათები	დღეში ტვ-ს ყურების სთ	0-24	- 1,98,99	1

სიხშირეთა განაწილება

		განაწილება
	0	**4,8** 1,567
	1	**20,0** 6,502
	2	**26,7** 8,704
	3	**18,6** 6,058
	4	**13,2** 4,286
	5	**6,7** 2,191

ცხრილი გვიჩვენებს ადამიანთა რაოდენობას და მუქი შრიფტით კი მათ პროცენტს - ვინც თითოეულ კითხვას გასცა შესაძლო პასუხი. ყველა იმ წლების განმავლობაში, როცა ეს კითხვები იყო დასმული, საერთო პასუხი რომ დღეში 2 საათს უყურებენ ტელევიზორს, უპასუხა 27%-მა.

რამდენ %-ს შეადგენენ ის გამოკითხულები, ვინც განაცხადა, რომ სულ არ უყურებენ, ანუ 0 საათი დღეში? რამდენმა ადმიანმა უპასუხა, რომ ტელევიზორს უყურებს 24 საათის განმავლობაში?

GSS-ის მიერ დასმული სხვა კითხვა არის შემდეგი: „თუ ავიღებთ ერთად ყველაფერს, შეგიძლიათ თქვათ, რომ თქვენ ხართ ძალიან ბედნიერი, საკმაოდ ბედნიერი თუ, არც თუ ისე ბედნიერი?“ GSS-მ კვლევების ამ დეტალს უწოდა HAPPY. ადამიანების რამდენმა პრცენტმა განაცხადა, რომ არის ძალიან ბედნიერი? GSS-ის გამოყენება შეიძლება იმის გამოსაკვლევად, თუ რომელი ჯგუფის ადამიანები არიან უფრო ბედნიერნი. ვისაც ჰქონდა ბედნიერი ქორწინება? ვისაც უკეთესი ჯანმრთელობა აქვს? ვისაც ჰყავს ბევრი მეგობარი?

პრაქტიკული ნაწილი.

1.1 ასპირინი და გულის შეტევები. ჰარვარდის სამედიცინო სკოლის კვლევაში (სცენარი 2) ჩართული იყო დაახლოებით 22 000 მამაკაცი ექიმი. ვის უნდა მიეღო ასპირინი და ვის პლაცებო დადგინდა მონეტის აგდებით. შედეგად დაახლოებით 11 000 ექიმს დაენიშნა ასპირინი და დაახლოებით ამდენივეს პლაცებო. დამკვირვებლებმა ექსპერიმენტის შედეგები შეაფასეს პროცენტულად. მათ ვინც მიიღეს ასპირინი, გულის შეტევა ჰქონდათ 0.9%-ს, ხოლო ვინც მიიღო პლაცებო 1.7%-ს. დაკვირვების შედეგებზე დაყრდნობით, კვლვის ავტორებმა დაასკვნეს, რომ ასპირინის მიღება ამცირებს გულის შეტევის რისკს. აღწერეთ ეს კვლევა შემდეგ ასპექტში: (ა) დიზაინი, (ბ) აღწერა და (გ) დასკვნა.

1.2 სიღარიბე და რასა. Current Population Survey (CPS) არის აშშ მოსახლეობის აღწერის შრომის სტატისტიკის ბიუროს კვლევა. იგი უზრუნველყოფს სრულყოფილი მონაცემების არსებობას სამუშაო ძალებზე, უმუშევრობაზე, სიმდიდრეზე, სიღარიბეზე და ა.შ. მონაცემები შეიძლება ნახოთ ინტერნეტში შემდეგ მისამართზე: www.census.gov/hhes/www/cpstc/cps_table_creator.html. 2009 წლის ანგარიშში წარმოდგენილია დაახლოებით 50 000 ოჯახი, ისეთი სადაც იმყოფება ერთი წევრი მაინც 18 წელზე ნაკლები ასაკის. ანგარიშში მითითებულია, რომ თეთრკანიანი ოჯახების 14.7%-ს, შავკანიანი ოჯახების 30.4%-ს და აზიური ოჯახების 11.1%-ს წლიური შემოსავალი აქვთ სიღარიბის ზღვარს ქვემოთ.

მიებულ შედეგებზე დაყრდნობით კვლევის ავტორებმა დაასკვნეს, რომ *ყველა* შავკანიანი ოჯახის პროცენტული მაჩვენებელი, რომელთა წლიური შემოსავალი სიღარიბის ზღვარზე ნაკლებია, მერყეობს 28.6%-დან 32.2%-მდე. აღწერეთ ეს კვლევა შემდეგ ასპექტში: (ა) დიზაინი, (ბ) აღწერა და (გ) დასკვნა.

1.3 GSS და ცათა საუფეველი. შედით General Social Survey-ს ვებგვერდზე, http://sda.berkeley.edu/GSS. შედით HEAVEN, როგორც ცვლადების რიგი, შემდეგ დააწკაპუნეთ Run the Table. კითხვაზე, გჯერათ თუ არა ზეციური ცხოვრების, გამოკითხულთა რამდენმა პროცენტმა უპასუხა: ა) დიახ, ნამდვილად. ბ) დიახ, ალბათ. გ) ალბათ არა. დ) არა, ნამდვილად არ არის. (მონაცემები აღებულია CSM-დან, ბერკლის კალიფორნიის უნივერსიტეტი).

1.4 GSS და სამოთხე და ჯოჯოხეთი. იხილეთ წინა სავარჯიშო. თქვენ შეგიძლიათ მიიღოთ კვლევის მონაცემები ყოველი კონკრეტული წლისთვის. მაგალითად, 2008 წლის - ასეთნაირად: შეიყვანეთ YEAR(2008) „Selection Filter"-ში, სანამ დააწკაპუნებთ *Run the Table*-ზე.

ა) გააკეთეთ ეს ჯერ HEAVEN წელი 2008 და ნახეთ პროცენტული მაჩვენებელი ოთხივე შესაძლო შედეგზე.

ბ) შეაჯამეთ 2008 წლის მოსაზრებები 2008 ჯოჯოხეთის შესახებ (ცვლადი HELL). შეადარეთ პასუხები ერთმანეთს სამოთხის და ჯოჯოხეთის შემთხვევაში.

1.5 GSS იმ სანების შესახებ, რომელსაც ირჩევთ. GSS-ს ვებსაიტზე დააწკაპუნეთ *Standard Codebook* და შემდეგ *Sequential Variable List*. მოძებნეთ საგნები, რომლებიც თქვენ გაინტერესებთ და მოძებნეთ GSS კოდური სახელი ცვლადი სტიქონის შესაყვანად. შეაჯამეთ მიღებული შედეგები.

2

პოპულაცია.
კალკურატორებისა და კომპიუტერების გამოყენება.

ჩვენ ვაკვირდებით საგნებს, მაგრამ გვაინტერესებს პოპულაცია. ობიექტებს, რომელთაც ვზომავთ შესწავლის პროცეში ვუწოდებთ **საგნებს**. ზოგადად, საგნები არიან, მაგალითად, ადამიანები, რომლებიც მონაწილეობენ ზოგად სოციალურ გამოკითხვებში. საგნები შეიძლება იყოს სკოლები, ქვეყნები ან დღეები. შეიძლება გავზომოთ მათი მახასიათებლები, ვთქვათ, როგორიცაა:

- ყოველი სკოლისთვის: დანახარჯი თითოეულ მოსწავლეზე, საკლასო ოთახების საშუალო ფართობი, სტუდენტების მიერ ტესტირებაზე მიღებული საშუალო ქულა.
- ყოველი ქვეყნისთვის, მაგალითად, შემდეგი პროცენტული მაჩვენებელები: შობადობა, უმუშევრობა, სიღარიბეში მცხოვრებთა რაოდენობა, კომპიუტერის მცოდნეთა რაოდენობა და ა.შ.
- ყოველდღიურად ინტერნეტ-კაფეში: დანახარჯები ყავაზე, ჭამაზე, ინტერნეტით საგებლობაზე.

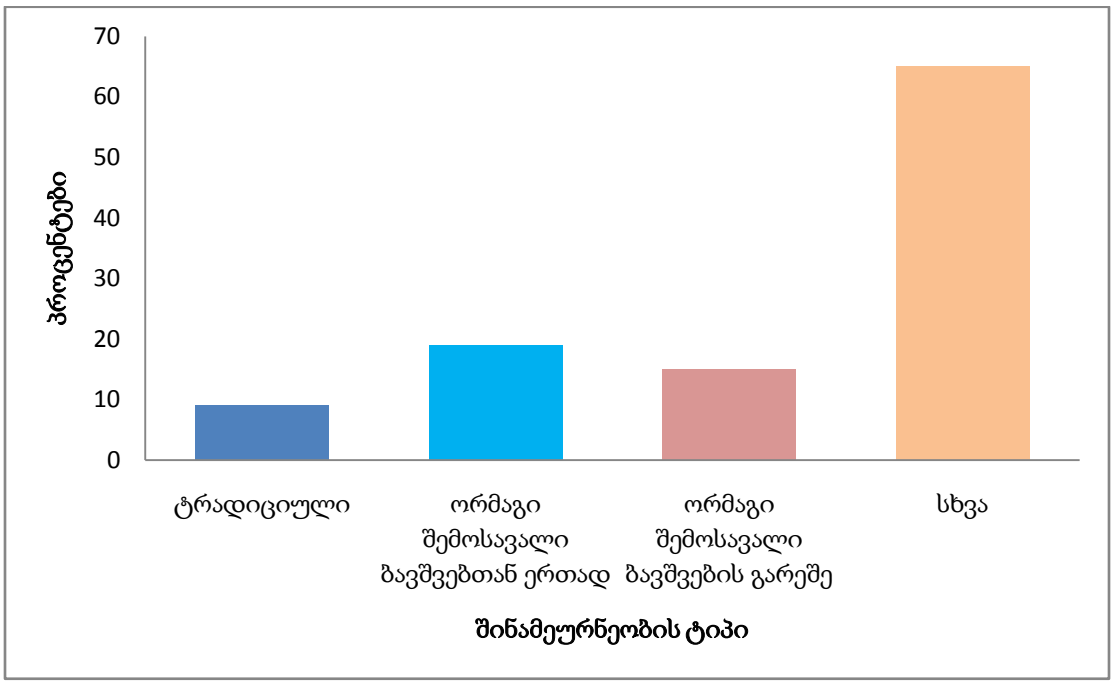
პოპულაცია არის ყველა იმ საგნის სიმრავლე, რომლებიც ჩვენი ინტერესის ობიექტებს წარმოადგენენ. როგორც წესი, პრაქტიკაში გვაქვს ხოლმე მონაცემები მხოლოდ *ზოგიერთ* საგნებზე, რომლებიც მიეკუთვნებიან პოულაციას. ამ საგნების ერთობლიობას ეწოდება **შერჩევა**, ან ამორჩევა, ან ნიმუში (sample). ცხადია, რომ შერჩევა არის პოპულაციის ქვესიმრავლე, რომლის ელემენტებზეც გვაქვს მონაცემები. ეს ელემენტები ხშირად შემთხვევითაა შერჩეული.

2008 წელს ზოგად სოციალურ კვლევაში (GSS), შერჩევაში მოხვდა 2023 ადამიანი, ვინც მონაწილეობა მიიღო გამოკითხვებში. პოპულაცია კი წარმოადგენდა მთელი ამერიკის ზრდასრულ მოსახლების სიმრავლეს, რომელიც იმ დროს მეტი იყო 200 მილიონზე.

აღწერითი სტატისტიკა და დასკვნითი სტატისტიკა. შეგროვილი მონაცემების შეჯამების ერთერთი მეთოდია **აღწერითი სტატისტიკა.** მონაცემები მოცემული გვაქვს შერჩევაზე ან ზოგჯერ მთელ პოპულაციაზე. შეჯამება (რეზიუმე),

როგორც წესი, შედგება გრაფიკებისა და რიცხვებისგან, როგორიცაა, მაგალითად, საშუალო ან პროცენტი.

აღწერითი სტატისტიკური ანალიზი, როგორც წესი, მოიცავს გრაფიკულ და რიცხვით რეზიუმეს. მაგალითად, ნახატი 2.1 არის სვეტოვანი დიაგრამა, რომელიც გვაძლევს პროცენტულ მაჩვენებელს ამრიკის სხვადასხვა ტიპის შინამეურნეობაზე 2005 წელს. მასში მოკლედაა შეჯამებული შეერთებული შტატებს აღწერის ბიუროს მონაცემები ამერიკის 50 000 შინამეურნეობის კვლევის შედეგებზე. აღწერითი სტატისტიკის ძირითადი მიზანია შეამციროს მონაცემები, ისე რომ არ დაირღვეს ან არ დაიკარგოს დიდი ან მნიშვნელოვანი ინფორმაცია.



ნახ. 2.1. *აშშ-ს შინამეურნეობის ტიპები 2005 წელს.* (წყარო: შეერთებული შტატების აღწერის ბიუროს მონაცემები).

პროცენტების და საშუალოების გრაფიკების და რიცხვების აღქმა გაცილებით ადვილია ვიდრე მონაცემთა მთელი სიმრავლის. აგრეთვე გაცილებით იოლია, წარმოდგენა ვიქონიოთ მონაცემებზე გრაფიკზე დაკვირვებით, ვიდრე წავიკითხოთ 50 000 შინამეურნეობის მიერ შევსებული კითხვარები. ეს გრაფიკი ასევე ცხადად გვიჩვენებს, რომ „ტრადიციული“ შინამეურნეობები ანუ ისეთი შინამეურნეობები,

რომლებიც შედგებიან ოჯახისგან (მამა, დედა და შვილები), სადაც მხოლოდ ქმარია სამუშაო ძალა, მაინც და მაინც აღარაა გავრცელებული შეერთებულ შტატებში. სინამდვილეში „სხვა“ ტიპის შინამეურნეობებია უფრო გავრცელებული, კერძოდ ისეთები, რომლებსაც სათავეში უდგანან ქალები, ან ზრდასრული ახალგაზრდები, ან მოხუცები, რომლებიც არ ცხოვრობენ მეუღლეებთან ერთად.

აღწერითი სტატისტიკა სასარგებლოა მაშინაც, როცა მონაცემები გვაქვს მთელი პოპულაციისთვის, როგორიცაა მაგალითად მოსახლეობის აღწერა. ამის საპირისპიროდ, დასკვნითი სტატისტიკა გამოიყენება მაშინ როცა მონაცემები გვაქვს მხოლოდ შერჩევისთვის, მაგრამ გვინდა, რომ მივიღოთ გადაწყვეტილება ან გამოვთქვათ პროგნოზი მთელი პოლულაციისთვის.

დასკვნითი სტატისტიკა არის მთელი პოპულაციისთვის გადწვეტილების მიღების ან პროგნოზის გაკეთების მეთოდი, რომელიც დამყარებულია მხოლოდ შერჩევისთვის მიღებულ მონაცემებზე, რომელიცამ პოპულაციიდანაა აღებული.

კვლევების უდიდეს ნაწილში ჩვენ გვაქვს მონაცები შერჩვისთვის და არა მთელი პოპულაციისთვის. აღწერით სტატისტიკას ვიყენებთ შერჩევის მონაცემთა განზოგადებისა და პროგნოზირებისთვის მთელი პოპულაციისთვის.

მოსაზრებები იარაღის კონტროლზე. დავუშვათ, გვინდა ვიცოდეთ, რას ფიქრობენ ადმიანები პისტოლეტის გაყიდვის კონტროლზე. განვიხილოთ რას ფიქრობს ხალხი, ვთქვათ, ფლორიდაში, შტატში სადაც კრიმინალის დონე შედარებით მაღალია. ჩვენთვის საინტერესო პოპულაცია არის შტატის 10 მილიონიანი ზრდასრული მოსახლეობის სიმრავლე. იმის გამო, რომ შეუძლებელია ეს საკითხი განვიხილოთ შტატის მთელ მოსახლეობასთან, შევისწავლოთ მონაცემები, რომლებიც გაკეთებულია ფლორიდის 834 მცხოვრების გამოკითხვის შედეგად. კვლევა ჩატარებულია ფლორიდის საერთაშორისო უნივერსიტეტთან არსებული საზოგადოებრი აზრის შემსწავლელი ინსტიტუტის მიერ. ამ გამოკითხვაში შერჩევის საგნების 54.0%-მა თქვა, რომ ისინი მომხრენი არიან კონტროლისა იარაღის გაყიდვაზე. საგაზეთო სტატია იუწყება, რომ „ცდომილება“ რამდენად ზუსტად ემთხვევა ეს მონაცემი სინამდვილეში არსებულს, არის 3.4%. შეიძლება ვივარაუდოთ ნდობის მაღალი ალბათობით (დაახლოებით 95% -ით), რომ პოცენტული მაჩვენებელი ფლორიდის *მთელი ზრდასრული* მოსახლეობის, რომელთა სურვილია იყოს კონტროლი იარაღის გაყიდვაზე, მოხვდება შუალედში 50.4%-დან 57.4%-მდე, ანუ ინტერვალში (54.0%-3.4%; 54.0%+3.4%).

კითხვა: დაადგინეთ ამ ანალიზში რა არის აღწერითი სტატისკის ანალიზი და რა არის დასკვნითი სტატისტიკის ანალიზი.

სტატისტიკური დასკვნის ერთერთი მნიშვნელოვანი ასპექტია მონაცემთა ისეთნაირად წარმოდგენა, რომ პროგნოზები იყოს ზედმიწევნით ზუსტი. ფაქტია, რომ დასკვნითი სტატისტიკური ანალიზის საშუალებით შესაძლებელია საკმაოდ კარგად გავაკეთოთ პროგნოზი მთელი პოპულაციის პარამეტრების შესახებ ნიმუშების კარგი შერჩევით, რომელთა ზომები გაცილებით მცირეა, ვიდრე მთელი პოპულაციის. გასაოცარია, რომ ჩინეთის მოსახლეობიდან, რომელიც დაახლოებით ოთხჯერ მეტია აშშ-ს მოსახლეობაზე, შემთხვევით აღებული ამორჩევა, რომელიც შედგება 1000 ადამიანისგან ჩინეთის მოსახლეობიდან, და შემთხვევით აღებული ამორჩევა, რომელიც შედგება 1000 ადმიანისგან აშშ-ს მოსახლეობიდან, კვლევის პროგნოზები აღწევენ სიზუტის ერთსა და იმავე დონეს. აი რატომ არის, რომ გამოკითხვათა უდიდესი ნაწილი იყენებს დაახლოებით ათას კაციან შერჩევას, თუნდაც რომ მოსახლეობა მილიონებს ითვლიდეს.

შერჩევის სტატისტიკა და პოპულაციის პარამეტრები. მოსაზრებები იარაღის კონტროლზე არის **შერჩევის სტატისტიკის** მაგალითი. მეტად მნიშვნელოვანია, დავინახოთ განსხვავება შერჩევის სტატისტიკასა და პოპულაციის შესაბამის მნიშვნელობებს შორის. ტემინი **პარამეტრი** გამოიყენება პოპულაციის რიცხვითი შეფასებისთვის. **სტატისტიკა** კი არის პოულაციიდან აღებული შერჩევის რიცხვითი შეფასება. მაგალითად, პროცენტული მაჩვენებელი ფლორიდის ზრდასრული მოსახლეობის, ვინც იარაღის გაყიდვის კონტროლს უჭერს მხარს, არის პარამეტრი. ჩვენ ვცდილობთ, კარგად შევისწავლოთ პარამეტრები, რომ უკეთ გავიგოთ პოპულაცია. თუმცა პარამეტრი თითქმის ყოველთვის უცნობია. ამიტომ ვიყენებთ შერჩევის სტატისტიკას პარამეტრების მნიშვნელობათა შესაფასებლად.

შემთხვევითობა და ცვლადობა. შემთხვევითობაზე ხშირად ფიქრობენ, როგორც ქაოსურზე, არაორგანიზებულზე, მაგრამ შემთხვევითობა არის მძლავრი იარაღი კარგი შერჩევის მიღებისთვის და კარგი ექსპერიმენტის ჩატარებისთვის. შერჩევას გააჩნია ტენდენცია, იყოს კარგი ანარეკლი პოპულაციისა, თუ პოპულაციის ყოველ საგანს აქვს ამორჩევაში მოხვედრის ერთნაირი შანსი. ეს არის საფუძველი **შემთხვევითი შერჩევის**, რომელიც განკუთვნილია იმისთვის, რომ შერჩევა იყოს წარმომადგენლობითი მთელი პოპულაციიდან. შემთხვევითი ამორჩევის მარტივი მაგალითია, როცა მასწავლებელი ფურცელზე წერს სტუდენტების სახელებს, ყრის მათ ყუთში და შემდეგ იღებს ფურცლებს ყუთში ჩაუხედავად.

სულაც არ არის გასაკვირი, რომ ადამიანები ერთმანათისგან განსხვავებიან, ამიტომ საზომები, რომლთაც ჩვენ მათთვის ვიყენებთ, იცვლებიან ერთი ადამიანიდან მეორეზე გადასვლისას. როგორც ვნახეთ, (GSS)-ის კითხვაზე ტელევიზორის ყურებაზე განსხვავებული პასუხები იყო. Exit Poll-ის მაგალითშიც ყველას ერთი და იგივე არჩევანი არ გაუკეთებია. თუ საგნები იქნებოდა ერთნაირი, ჩვენ ამოვირჩევდით მხოლოდ ერთ მათგანს. ამ ცვალებადობას უფრო უკეთ შევისწავლით, რაც უფრო მეტ ადამიანს ავირჩევთ. თუ გვინდა შედეგის პროგნოზირება, უმჯობესია ავირჩიოთ ასი, ვიდრე ერთი, და ჩვენი პროგნოზი იქნება უფრო მეტად სანდო, თუ შევარჩევთ 1000 ამომრჩეველს.

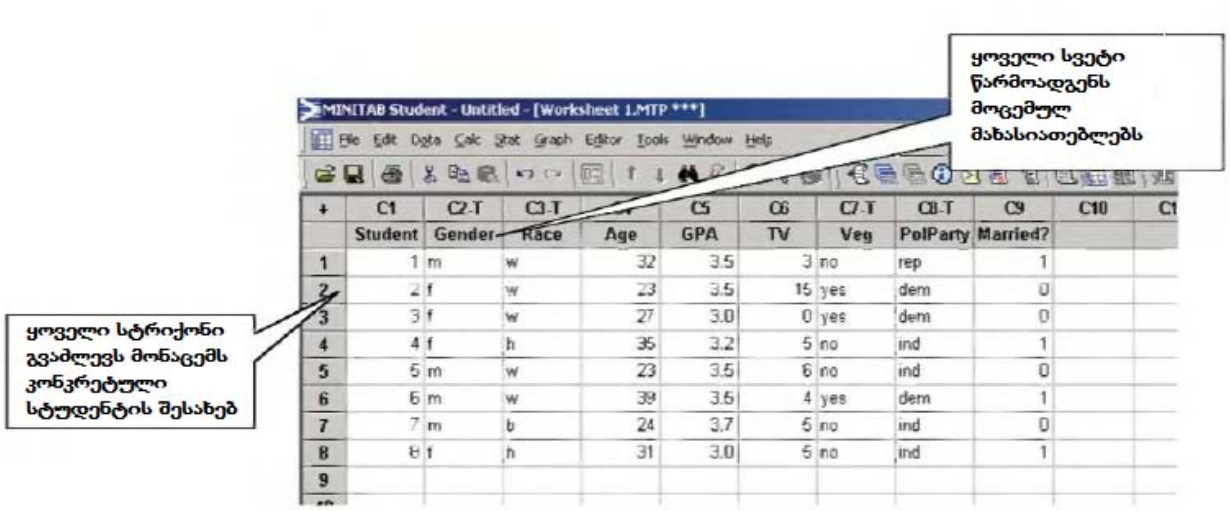
როგორც იცვლება ხალხი, ისე იცვლება შერჩევაც.

დავუშვათ, რომ პროგნოზირებისთვის თქვენ გააკეთეთ Exit Poll-ის გამოკითხვები 1000 ამომრჩეველზე. დავუშვათ, რომ ორგანიზაცია Gallup-იც აკეთებს Exit Poll -ს 1000 სხვა ამომრჩეველზე. შედეგად პროგნოზები იქნება განსხვავებული. შესაძლებელია თქვენი 1000 ამომრჩევლიდან 480-მა ხმა მისცა რესპუბლიკელების კანდიდატს, მაშინ თქვენი პროგნოზი იქნება 48% ამ პიროვნების სასარგებლოდ. შესაძლებელია Gallup-ს 1000 ამომრჩევლიდან 440-მა მისცა ხმა რესპუბლიკელების კანდიდატს, მაშინ მათი პროგნოზი იქნება 44% იგივე პიროვნების სასარგებლოდ. სინამდვილეში შემთხვევითი შერჩევის დროს ცვლილებათა რაოდენობა შერჩევიდან შერჩევამდე რეალურად საკმაოდ პროგნოზირებადია. ორივე თქვენი პროგნოზი 5%-ს სიზუსტით დაემთხვა რესპუბლიკელების ფაქტიურ პროცენტს, თუ შერჩევა იყო შემთხვევითი. მეორე მხრივ, აუცილებლად გასათვალიწინებელია ის ფაქტი, რომ რესპუბლიკელები უფრო ხშირად ვიდრე დემოკრატები, უარს ამბობენ ეგზიტპოლებში მონაწილეობაზე. 2004 წელს აშშ-ს საპრეზიდენტო არჩევნების შემდეგ ბევრი კამათი წარმოიშვა იმის თაობაზე, რომ ჯორჯ ბუშმა არჩევნები მოიგო ზოგიერთ ისეთ შტატში, სადაც ეგზიტპოლები ჯონ კერის მოგებას პროგნოზირებდნენ. ხომ არ არის მოსალოდნელი, რომ გზამ, რომლის მიხედვითაც ჩატარდა ეგზიტპოლები, მიგვიყვანა არასწორ პროგნოზებამდე?

სტატისტიკური პროგრამული უზრუნველყოფისა და კალკულატორების გამოყენება (და ბოროტად გამოყენება). დღევანდელ მკვლევარებს განსაკუთრებით გაუმართლათ წინა თაობებთან შედარებით. მათ აღარ უწევთ დიდი სტასტისტიკური გამოთვლების ხელით ჩატარება. ახლა ადვილად ხელმისაწვდომია კალკულატორები და კომპიუტრები სტატისტიკური ანალიზისთვის. ისინი

შესაძლებელს ხდიან ისეთი დიდი გამოთვლების ჩატარებას, რაც ადრე ძალიან ძნელი იყო და ზოგჯერ შეუძლებელიც.

MINITAB და SPSS ორი პოპულარული სტატისტიკის პროგრამული უზრუნველყოფის პაკეტებია. T l-83+ და Tl-84 გრაფიკული კალკულატორებია, რომლებსაც იმ შემთხვევაში, თუ საქმე გვაქვს მარტივ გრაფიკებთან და გამოთვლებთან, იმავე შედეგებს იძლევიან და უფრო პორტატული ინსტრუმენტია. Microsoft Excel-საც აქვს პროგრამული უზრუნველყოფა ზოგიერთი სტატისტიკური მეთოდისთვის. მაგალითად, შეუძლია მონაცემთა დახარისხება და ნაწილობრივ ანალიზიც, მაგრამ ამ მხრივ მისი შესაძლებლობები შეზღუდულია. ჩვენი ძირითადი აქცენტი არის იმაზე, თუ როგორი ინტერპრეტაცია გავუკეთოთ შედეგებს და არა იმ დეტალებზე, თუ როგორ უნდა ვისარგებლოთ პროგრამით.



	C1	C2-T	C3-T	C4	C5	C6	C7-T	C8-T	C9	C10	C1
	Student	Gender	Race	Age	GPA	TV	Veg	PolParty	Married?		
1	1	m	w	32	3.5	3	no	rep	1		
2	2	f	w	23	3.5	15	yes	dem	0		
3	3	f	w	27	3.0	0	yes	dem	0		
4	4	f	h	35	3.2	5	no	ind	1		
5	5	m	w	23	3.5	6	no	ind	0		
6	6	m	w	39	3.6	4	yes	dem	1		
7	7	m	b	24	3.7	5	no	ind	0		
8	8	f	h	31	3.0	5	no	ind	1		
9											

ნახ. 2.2

მონაცემთა ფაილები. სტატისტიკური ანალიზი ადვილი რომ იყოს, მონაცემთა დიდი კომპლექტი მოვათავსოთ მონაცემთა ფაილში. ამ ფაილს როგორც წესი, აქვს ცხრილის ფორმა. ნახაზზე მოცემულია მონაცემთა ფაილის ნაწილი. იგი გვიჩვენებს, როგორ გამოიყურება მონაცემთა ფაილი MINITAB-ში. ფალი გვიჩვენებს რვა სტუდენტის მონაცემებს შემდეგი მახასიათებლების მიხედვით:

- სქესი (f = ქალი, m = მამაკაცი)
- რასობრივ - ეთნიკური ჯგუფი (b = შავი, h = ესპანური, w = თეთრი)
- ასაკი (წლებში)
- კოლეჯის GPA (მასშტაბი 0-დან 4-მდე)
- ტელევიზორის ყურების საშუალო საათების რაოდენობა კვირაში

- არის თუ არა ვეგეტარიანელი (დიახ, არა)
- პოლიტიკური პარტია (dem = დემოკრატი, rep = რესპუბლიკური, ind = დამოუკიდებელი)
- ოჯახური მდგომარეობა (1 =დაოჯახებული, 0 = დაუოჯახებელი)

ნახ.2.2 გვიჩვენებს მონაცემთა ფაილის აგების ორი ძირითად წესს:

- ყოველი სტრიქონი შეიცავს საზომებს კონკრეტული საგნისთვის (მაგალითად, ჩვენს შემთხვევაში ადამიანისთვის).
- ნებისმიერი სვეტი შეიცავს საზომებს კონკრეტული მახასიათებელისთვის.

აქ ზოგიერთი მახასიათებელი მოცემულია რიცხვებში, ზოგი „რესპ/დემ/ დამ“, ზოგიც - „კი/არა“ ტერმინებში.

მონაცემთა ბაზები. დიზაინის კვლევებიდან ერთერთი უმთავრესია მონაცემთა მოძიება. ამიტომ უპრიანია ვისარგებლოთ უკვე არსებული მონაცემთა ბაზებით, რომლებიც არის დაარქივებული მონაცემთა ბაზების კოლექციაში და ჰქვია **databases.** ბევრი ინტერნეტ მონაცემთა ბაზები ხელმისაწვდომია.

თუ დაინტერესებული ხართ ბეისბოლით ან კალათბურთით ან ფეხბურთით დააწკაპუნეთ სტატისტიკის ბმულებზე: www.mlb.com ან www.nba.com ან www.nfl.com.

გაინტერესებთ გაიგოთ რისი სჯერა ხალხს ყველაზე მეტად, მთელსმსოფლიოში? მოინახულეთ www.globalbarometer.net ან www.europa.eu.int/com/public_opinion ან www.latinobarometro.org.

გაინტერესებთგაიგოთან Gallup გამოკითხვების შედეგები ადამიანების რწმენის შესახებ? იხილეთ: www.galluppoll.com.

დაინტერესებული ხართ მოსახლეობის ზრდის ტემპით, უმუშევრობის დონით, ან გავრცელების სქესობრივი გზით გავრცელებადი გადამდები დაავადებებით? იხილეთ www.google.com/publicdata და სხვა.

Google-ს საძიებო სისტემაში ჩაწერეთ "U.S. Census data", გამოჩნდება ფანჯარა, სადაც იქნება მონაცემთა იმ ბაზების სია, რომლებიც მხარდაჭერილია აშშ-ს მოსახლეობის აღწერის ბიუროს მიერ. "Statistics Canada data" არის კანადის აღწერის მონაცემთა ბაზა. შესაძლებელია ძიება ჩატარდეს კონკრეტული თემის მიხედვით. მაგალითად, თუ გინდათ ნახოთ კანადელების აზრი მარიხუანას ლეგალიზაციაზე მკურნალობის თვალსაზრისით, შეიყვანეთ:***poll legalize marijuana Canada medicine*** ძიების ფანჯარაში.

ყოველთვის შეამოწმეთ წყაროები! მონაცემთა ყველა ბაზა როდი გვაძლევს სანდო ინფორმაციას. ამიტომ სანამ მონაცემებს ან დასკვნებს დავუჯერებდეთ, უპირველეს ყოვლისა უნდა გადამოწმდეს ინფორმაციის წყაროს სანდოობა და რომ წყარო წარმოგვიდგენდეს ინფორმაციას იმაზეც, თუ როგორაა კვლევები ჩატარებული. მაგალითად, ტელევიზიებში რომელიმე გადაცემის მსვლელობისას გაკეთებული ინტერაქტიული გამოკითხვა თითქმის არასდროს არ არის სანდო, რასაც ყოველთვის გრძნობს მთელი მოსახლეობა. ამის მიზეზი არის ის, რომ პირები ვინც გადაცემას გამოეხმაურნენ, არიან პოპულაციის წარმომადგენლობითი შერჩევიდან.

პრაქტიკულინაწილი.

2.1. შერჩევა საჭიროებს სიფრთხილეს. იმ პირებს ვისაც ჰყავთ შვილები და კითხულობენ ენ ლანდერსის საგაზეთო პუბლიკაციებს ბავშვთა მოვლაზე, ჰკითხეს, თუ მათ მოუწევთ აკეთონ ისევ და ისევ ის, რაც საჭიროა ბავშვთა აღსაზრდელად, ისურვებენ თუ არა, კიდევ იყოლიონ ბავშვები? დაახლოებით 10 000 გამოკითხულიდან მხოლოდ 30%-მა უპასუხა, დიახ. რატომ არ არის უსაფრთხო, დავასკვნათ რამე ამ გამოკვლევიდან მოსახლეობის მთელ პროცენტულ შემადგენლობაზე მათთვის ვისაც ისევ უნდა, რომ ჰყავდეს შვილები?

2.2. UW -ს სტუდენტთა გამოკითხვა. ეს გამოკითხვა ჩატარდა ვისკონსინის უნივერსიტეტში (UW) ალკოჰოლის მოხმარებაზე სტუდენტებს შორის. გამოიკითხა 100 სტუდენტი 40 858 სტუდენტიდან. კითხვა იყო ერთი: „გასული კვირის განმავლობაში რამდენი დღე იყო ისეთი, როცა ერთხელ მაინც მოგიწიათ დალევა?“

ა)განსაზღვრეთ პოპულაცია და ამორჩევა.

ბ) UW-ს 40 858 სტუდენტს ვაფასებთ მხოლოდ ერთი მახასიათებლით და გვაინტერესებს, რამდენმა პროცენტმა უპასუხა „ნული“, ანუ არცერთხელ. დავუშვათ, რომ ამრჩეული 100 სტუდენტიდან ასეთი პასუხი დააფიქსირა 29%-მა. ნიშნავს თუ არა ეს იმას, რომ სტუდენტთა მთელი პოპულაციის 29% იმავე პასუხს გაგვცემდა? დაასაბუთეთ!

გ) რიცხითი შეფასება 29% ამორჩევის სტატისტიკაა თუ პოპულაციის პარამეტრი?

2.3. ESP. საერთო სოციალური კვლევა სვამს კითხვას: „რამდენჯერ გიგრძვნიათ, რომ ვიღაც გეხებათ, როცა ეს ვიღაც საკმაოდ შორსაა თქვენგან?“ შერჩევის 3 887 ობიექტიდან, ვისაც ამის შესახებ აზრი გააჩნდა, 1 407- მა თქვა, რომ არასდროს. ხოლო

2 480 თქვა, რომ ასეთ შემთხვევას ერთხელ მაინც ჰქონდა ადგილი. ასე, რომ მათი წილი ვინც თქვა „ერთხელ მაინც“, არის 2840/3887=0.638.

ა) აღწერეთ ინტერესის პოპულაცია.

ბ) ახსენით, როგორაა შერჩევის მონაცემები წარმოდგენილი აღწერითი სტატისტიკის დახმარებით?

გ) პოპულაციის რომელ პარამეტრზე შეგვიძლია დასკვნის გაკეთება?

2.4. პრეზიდენტის პოპულაროობა. ყოველ თვე ორგანიზაცია Gallup ატარებს გამოკითხვებს პრეზიდენტის პოპულაროობის შესახებ (იხ. www.pollingreport.com). 2011 წლის 2-3 თებერვალს გამოიკითხა დაახლოებით 1500 ამერიკელი, რომელთაგან 45%-მა განაცხადა, რომ მხარს უჭერს ობამას კურსს. სტატისტიკის ცდომილებაა პლუს-მინუს 3%. ახსენით მაქსიმუმ რისი ტოლი შეიძლება იყოს სტატისტიკური ცდომილება, რომ წარმოდგენილ იქნეს ასეთი დასკვნითი სტატისტიკური ანალიზი?

2.5. სტრესის შემცირება. თქვენს სკოლას სურს იცოდეს იმ სტუდენტების პროცენტული რაოდენობა, ვისაც უნდა რომ ჰქონდეთ რამდენიმე დღიანი პერიოდი სწავლის დამთავრებასა და გამოცდების დაწყებას შორის. ამით სკოლას სურს, შეამციროს სტუდენტების სტრესის დონე გამოცდების დაწყების წინ. კვლევაში მონაწილეობა მიიღო 100-მა სტუდენტმა.

ა) განსაზღვრეთ პოპულაცია და შერჩევა.

ბ) განსაზღვრეთ კვლევის მიზანი, (1) აღწერითი სტატისტიკისა და (2) დასკვნითი სტატისტიკის მიხედვით.

2.6. გასაკვირია ESP მონაცემები? სავარჯიშო 2.3-ში 3887 შერჩეული სუბიექტიდან, 63.8% ამბობს, რომერთხელ მაინც უგრძვნია შეხება იმ ადამიანთან, ვინც იმ დროს საკმაოდ შორს იყო, ანუ 3887 შერჩეული სუბიექტიდან, ვისაც გააჩნდა რაიმე აზრი ამის თაობაზე, „არასოდეს“ უპასუხა 1407-მა და „ერთხელ მაინც“ 2480-მა. დავუშვათ, რომ მთელი პოპულაციის მხოლოდ 20% გიპასუხებთ, რომ ასეთი განცდა ერთხელ მაინც ქონია. თუ 3887 პიროვნება იყო შემთხვევითი შერჩევა, მაშინ პროპორ-ციულობის შედეგი 0.638 იქნება გასაკვირი? გააკეთეთ ამორჩევის იმიტაცია (სიმულაცია) პოპულაციიდან პროპორციით 0.20. შერჩევის ზომა აიღეთ 4 000, რომელიც ახლოსაა შერჩევის ფაქტობრივ ზომასთან 3887.

2.7. შექმენით მონაცემთა ფაილი. გამოიყენეთ სტატისტიკური პროგრამული უზრუნველყოფა. გაარკვიეთ, თუ როგორ უნდა შეხვიდეთ მონაცემთა ფაილში. შექმენით მონაცემთა ფაილი. ისარგებლეთ ნახ.2.2-ის მონაცემებით.

3

მონაცემთა შესწავლა. მონაცემთა სახეები. მონაცემთა გრაფიკული გამოსახვა. ჰისტოგრამა. წრიული დიაგრამა.

ცვლადი არის რაიმე მახასიათებელი, რომელსაც ვაკვირდებით შესწავლის პროცესში. ტერმინი ცვლადი ნიშნავს, რომ მისი მნიშვნელობები განსხვავდებიან. მაგალითად, ტემპერატურა დღის განმავლობაში. დაკვირვების ყოველი შედეგი შეიძლება იყოს **რიცხვი**, მაგალითად, ნალექის რაოდენობა დღის განმავლობაში, გამოსახული სანტიმეტრებში. ზოგჯერ ცვლადი მიეკუთვნება სპეციალურ **კატეგორიას**, მაგალითად, როგორიცაა პასუხები „დიახ“ ან „არა“ კითხვაზე: „დღეს იწვიმა?“

*ცვლადს ეწოდება **კატეგორიული**, თუ მისი ყოველი დაკვირვება მიეკუთვნება კატეგორიის ერთ რომელიმე სიმრავლეს. ხოლო ცვლადს ეწოდება **რაოდენობრივი**, თუ მასზე დაკვირვებები ღებულობენ რიცხვით მნიშვნელობებს. ისინი წარმოადგენენ ცვლადის განსხვავებულ სიდიდეებს.*

ყოველდღიური ტემპერატურის სიდიდე და ნალექების რაოდენობა რაოდენობრივი სიდიდეებია. რაოდენობრივი ცვლადის სხვა მაგალითებია: ასაკი, და-ძმების რაოდენობა, წლიური შემოსავალი ან განათლებაზე დახარჯული წლების რაოდენობა. თუ საგნები ადამიანებია, მაშინ კატეგორიული ცვლადები შეიძლება იყოს სქესი (კატეგორიებით ქალი და მამაკაცი), რელიგიური კუთვნილება (როგორიცაა, მაგალითად, კათოლიკე, იუდეველი, მუსულმანი, პროტესტანტი, სხვა, არცერთი), საცხოვრებლის ტიპი (სახლი, ბინა, აპარტამენტი, საერთო საცხოვრებელი, სხვა) და გჯერათ თუ არა სიკვდილის შემდეგ ცხოვრების (დიახ, არა).

რაოდენობრივი ცვლადის განსაზღვრისას რატომ ვამბობთ, რომ რიცხვითი მნიშვნელობები წარმოადგენენ *სხვადასხვა სიდიდეებს?* რაოდენობრივი ცვლადის საზომია რიცხვი, რომელიც პასუხობს კითხვას „რამდენი?“ ასეთი ცვლადი ხასიათდება რაოდენობით. რაოდენობრივი ცვლადების შემთხვევაში შესაძლებელია მათი შეფასება არითმეტიკული თვალთახედვით, მაგალითად, როგორიცაა საშუალო. თუმცაღა ზოგიერთი რიცხვითი ცვლადი, მაგლითად, ადგილის კოდი, არ განიხლება როგორც რაოდენობრივი ცვლადი, რადგანაც ისინი არ იცვლებიან რაოდენობრივად. ანუ ბანკის მაგალითზე რომ განვიხილოთ, თუ ბანკი

დაინტერესებულია საშუალო რაოდენობის სესხების გაცემით, მაშინ ადგილის კოდს არავითარი აზრი არა აქვს.

გრაფიკები და რიცხვითი შეფასებები აღწერენ ცვლადის უმთავრეს თავისებურებებს:

- რაოდენობრივი ცვლადებისთვის ძირითადი მოთხოვნაა, რომ აღწეროს გვაქვს თუ არა **ცენტრი** და მონაცემთა **ცვალებადობა** (როგორც ზოგჯერ მოიხსენიებენ **„გავრცელება“**). მაგალითად, რა არის ნალექების ტიპური წლიური რაოდენობა? არის თუ არა დიდი ცვლილებები წლიდან წლამდე?
- **კატეგორიული** ცვლადებისთვის ძირითადი მოთხოვნაა აღწეროს სხვადასხვა კატეგორიები და მათთან დაკავშირებული დაკვირვებათა რაოდენობები. მაგალითად, სტუდენტთა რამდენი პროცენტია ამა თუ იმ კოლეჯში დემოკრატი?

რაოდენობრივი ცვლადები არის ან დისკრეტული ან უწყვეტი. რაოდენობრივი ცვლადების ყოველი მნიშვნელობა არის რიცხვი და ჩვენ ვახდენთ რაოდენობრივი ცვლადების კლასიფიკაციას როგორც დისკრეტული და უწყვეტი ცვლადებისას. *რაოდენობრივი ცვლადი არის **დისკრეტული**, თუ მისი შესაძლო მნიშვნელობები აღებულია განცალებული (იზოლირებული) რიცხვების სიმრავლიდან, როგორიცაა 0, 1, 2, 3, 4, ... ხოლო რაოდენობრივი ცვლადი არის **უწყვეტი**, თუ მისი შესაძლო მნიშვნელობები ადგენენ ინტერვალს.*

დისკრეტული ცვლადის მაგალითებია, შინაური ცხოველების რაოდენობა საოჯახო მეურნეობაში, ბავშვების რაოდენობა ოჯახში, უცხო ენების რაოდენობა, რომელ-ზედაც შეუძლია პიროვნებას საუბარი. მათი მნიშვნელობები *განცალებული* რიცხვებია, როგორიცაა {0, 1, 2, 3, 4, ...}. ცვლადის შედეგი არის რაოდენობა. *ნებისმიერი ცვლადი, რომელსაც გააჩნია სასრული რაოდენობის შესაძლო მნიშვნე-ლობები, არის დისკრეტული.*

უწყვეტი ცვლადის მაგალითებია, სიმაღლე, წონა, ასაკი და დროის ის რაოდენობა, რომელიც საჭიროა რაიმე დავალების შესასრულებლად. უწყვეტი ცვლადის ყველა შესაძლო მნიშვნელობათა სიმრავლე არ შედგება *განცალებული (იზოლირებული)* რიცხვებისგან. ის არის ცვლადის მნიშვნელობათა უსასრულო არე. დროის რაოდენობა რომელიც საჭიროა დავალების შესასრულებლად, შეიძლება იყოს 2.976657... სთ. *უწვეტ ცვლადებს შეუძლიათ მიიღონ უსასრულო, კონტინუუმ რაოდენობის შესაძლო მნიშვნელობები.*

სიხშირეთა ცხრილები.

პირველი ნაბიჯი, რომელიც უნდა გადაიდგას ცვლადების მონაცემთა რიცხობრივად დახასიათების საქმეში არის ის, რომ უნდა დავაკვირდეთ ყველა შესაძლო მნიშვნელობას და დავითვალოთ რამდენად ხშირად გვხვდება თითოეული მათგანი. კატეგორიული ცვლადებისთვის ყოველი დაკვირვება ვარდება ერთერთ კატეგორიაში. კატეგორიას, რომელსაც აქვს ყველაზე მაღალი სიხშირე, ეწოდება **მოდალური კატეგორია.** რაოდენობრივი ცვლადისთვის რიცხვით მნიშვნელობას, რომელიც ყველაზე ხშირად გვხვდება, ეწოდება **მოდა**. ჩვენ შეგვიძლია გამოვიყენოთ პროცენტები და პროპორციები იმისათვის, რომ შევაჯამოთ დაკვირვებათა რაოდენობები სხვადასხვა კატეგორიაში.

იმ დაკვირვებათა **პროპორცია**, რომელიც ეკუთვნის გარკვეულ კატეგორიას, ეწოდება დავირვებათა სიხშირის შეფარდებას დაკვირვებათა მთელ რაოდენობასთან. **პროცენტული მაჩვენებელი** კი არის პროპორცია გამრავლებული 100-ზე. პროპორციას და პროცენტულ მაჩვენებელს სხვანაირად **ფარდობით სიხშირესაც** უწოდებენ და გამოიყენება კატეგორიაში კატეგორიული ცვლადის გაზომვების შესაფასებლად.

სიხშირეთა ცხრილი არის ცვლადის ყველა შესაძლო მნიშვნელობების ჩამონათვალი ისე, რომ თითოეულ მნიშვნელობას მიწერილი უნდა ჰქონდეს დაკვირვებათა რაოდენობა (სიხშირე).

ზვიგენების თავდასხმა (მაგალითი კატეგორიულ ცვლადზე). შეერთებულ შტატებში ზვიგენების ადამიანებზე თადახმის ყველაზე მეტი ფაქტი დაფიქსირებულია ფლორიდის შტატში. 2000 – 2010 წლებში 268 ასეთ შემთხვევას ჰქონდა ადგილი. ფლორიდის გარდა მსოფლიოს კიდევ რომელ რეგიონებში მოხდა ზვიგენების თადასხმის ფაქტები და რამდენი იყო თითოეულ მათგანში? The International Shark Attack File (ISAF) წარმოგვიდგენს ამ მონაცემებს. ცხრილი 3.1 აკეთებს 715 ასეთი შემთხვევის კლასიფიკაციას, რომელთაც ადგილი ჰქონდათ 2000-2010 წლებში. მასში ჩამოთვლილია ის ქვეყნები და ამერიკის ის შტატები, სადაც ყველაზე მეტადაა გახშირებული ეს შემთხვევები. თავდასხმების რიცხვი კონკრეტული რეგიონისთვის არის **სიხშირე**. პროპორცია მიღებულია სიხშირის გაყოფით შემთხვევათა სრულ რაოდენობაზე ანუ 715-ზე. პროცენტული მაჩვენებელი ტოლია პროპორცია გამრავლებული 100-ზე.

კითხვები შესწავლისთვის:

ა) რა არის ცვლადი? კატეგორიულია ის თუ რაოდენობრივი?

ბ) ცხრილის მონაცემების მიხედვით აჩვენეთ, როგორ მიიღება პროპორცია და პროცენტული მაჩვენებელი ფლორიდისთვის.

გ) ამ მონაცემებისთვის განსაზღვრეთ მოდალური კატეგორია.

რეგიონი	სიხშირე	პროპორციულად	პროცენტი
ფლორიდა	268	0.375	37.5
ჰავაი	41	0.057	5.7
კალიფორნია	32	0.045	4.5
სამხრეთ კაროლინა	32	0.045	4.5
ჩრდილოეთ კაროლინა	28	0.039	3.9
ავსტრალია	120	0.168	16.8
სამხრეთ აფრიკა	41	0.057	5.7
ბრაზილია	20	0.028	2.8
ბაჰამის კუნძულები	12	0.017	1.7
სხვები	121	0.169	16.9
საერთო ჯამი	**715**	**1.000**	**100.0**

ცხრილი. 3.1. *ზვიგენების თავდასხმათა სიხშირე სახვადასხვა რეგიონების მიხედვით 2000 – 2010 წლებში.*

წყარო: www.flmnh.ufl.edu/fish/sharks/statistics/statsw.htm.

შემდგომი მოსაზრებები:

ა) თითოეული დაკვირვება (ზვიგენის თავდასხმა) აკეთებს იმ რეგიონის იდენტიფიკაციას, სადაც ეს მოხდა. რეგიონი არის ცვლადი. იგი კატეგორიული ცვლადია ყველა იმათთან ერთად, რომლებიც ჩამოთვლილია ცხრილის პირველ სვეტში.

ბ) დაფიქსირებული 715 შემთხვევიდან 268 იყო ფლორიდაში. ეს არის პროპორცია დაახლოებით 4 შემთხვევა ყოველი 10-დან. პროცენტული თანაფარდობა არის 100(0.375)=37.5%.

გ) ჩამოთვლილი რეგიონებიდან თავდასხმების უდიდესი რიცხვი დაფიქსირდა ფლორიდაში, რამაც მსოფლიო მასშტაბით შეადგინა 37.5%. ფლორიდა არის მოდალური კატეგორია, რადგან მას აქვს ყველაზე დიდი სიხშირე.

წვდომის უნარი (ინტუიცია):

არ აურიოთ სიხშირეები ცვლადის მნიშვნელობებში. ისინი უბრალოდ ჯამებია, თუ რამდენჯერ მოხდა დაკვირვება (ზვიგენის თავდასხმა) თითოეულ კატეგორიაში (რეგიონში). ცვლადი ამ შემთხვევაში არის რეგიონი, რომელშიც ამას ადგილი ჰქონდა. ცხრილებში, რომლებშიც წარმოდგენილია სიხშირეები, *პროპორციათა ჯამი შეადგენს 1.0-ს. ხოლო პროცენტების ჯამი არის 100%.* პრაქტიკაში ცალკეულმა რიცხვებმა შეიძლება მოგვიყვანოს ოდნავ განსხვავებულ რიცხვთან (როგორიცაა 99.9% ან 100.1%), დამრგვალებების გამო.

კატეგორიული ცვლადების გრაფიკები.

ხშირად ხდება ხოლმე, რომ უკეთ შევიგრძნობთ მონაცემთა სიმრავლეს, როცა ვუყურებთ გრაფიკს, ვიდრე მაშინ როცა ვუყურებთ საწყის მონაცემთა რიგს, ან თუნდაც სიხშირეთა ცხრილს. არსებობს კატეგორიული ცვლადის გრაფიკული გამოსახვის ორი უმარტივესი ტიპი: **წრიული დიაგრამა** და **ჰისტოგრამა.**

- **წრიული დიაგრამა** წამოადგენს წრეს სადაც გვაქვს წრიული სექტორები („ნამცხვრის ნაჭრები“) თითოეული კატეგორიისთვის. სექტორის ზომა შეესაბამება კატეგორიაში დაკვირვებათა პროცენტულ მაჩვენებელს.
- **ჰისტოგრამა** წარმოადგენს ვერტიკალურ ზოლს თითოეული კატეგორიისთვის. ზოლის სიმაღლე შეესაბამება დაკვირვებათა პროცენტულ მაჩვენებელს კატეგორიაში.

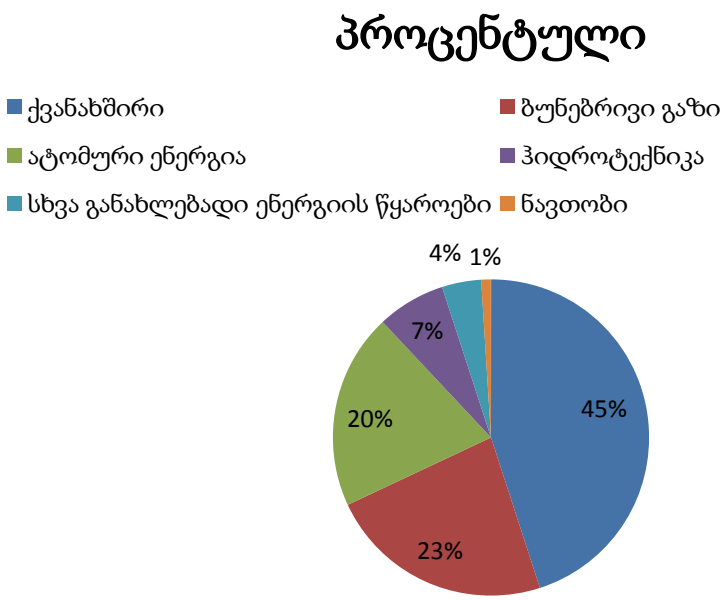
„ნამცხვრის ნაჭრის“ მაგიერ შემდგომში ვიხმაროთ ტერმინი „წრიული სექტორი“ ან უბრალოდ „სექტორი“

შემდეგი ცხრილი გამოსახავს აშშ-სა და კანადაში 2009 წელს გამომუშავებული ელექტროენერგიის წილობრივ წარმოდგენას წყაროების მიხდვით (პროცენტებში).

წყარო	პროცენტი აშშ	პროცენტი კანადა
ქვანახშირი	45	17
ბუნებრივი გაზი	23	7
ატომური ენერგია	20	15
ჰიდროენერგეტიკა	7	59
სხვა განახლებადი ენერგიის წყაროები	4	1
ნავთობი	1	1
საერთო	**100**	**100**

ცხრილი 3.2. *ელექტროენეგიის წყაროები აშშ-სა და კანადაში.*

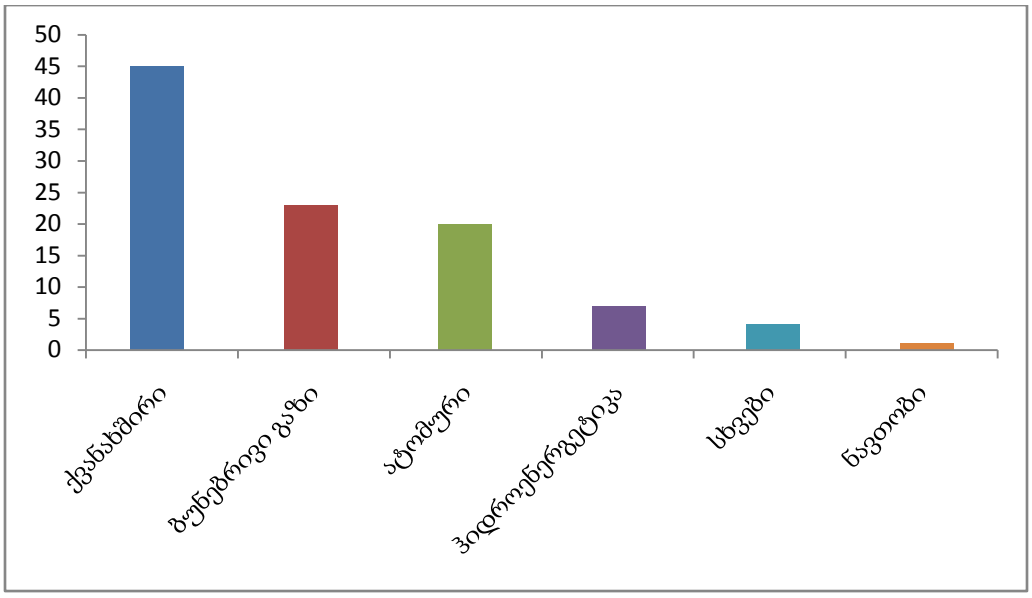
ამ ცხრილის შესაბამის წრიულ დიაგრამას აშშ-სთვის ექნება ასეთი სახე:



ნახაზი 3.3. *ელექტროენეგიის წყაროები აშშ-ში.*

ნაჭერი (ან სექტორი) სადაც მონიშნულია მაგ. ქვანახშირი, გვიჩვენებს რომ ქვანახშირის საშუალებით მიღებული ელექტროენერგია წარმოადგენს მთელი ელექტროენერგიის 45%-ს. ასეთ დიაგრამაზე თვალნათლივ ჩანს თუ ერთი ნაჭერი რადენად დიდია მეორეზე.

შემდეგი გრაფიკი წარმოადგენს ჰისტოგრამას (ამ შემთხვევაშიც აშშ-სთვის). იმ კატეგორიებს, რომელთაც აქვთ მეტი პროცენტულიმაჩვენებელი შეესაბამება მაღალი სვეტი. სვეტების სიგანე უცვლელია ყველა სვეტისთვის.



ნახაზი 3.4. *ჰისტოგრამა.*

ნახაზზე სვეტები წარმოდგენილია დალაგებულად უდიდესიდან უმცირესისკენ. ასეთ დროს ადვილია მაღალპროცენტიანი კატეგორიების ვიზუალური განცალკევება.

პარეტოს დიაგრამები.

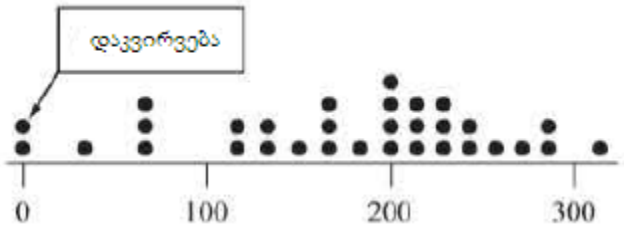
უკანასკნელი ნახაზი წარმოადგენს ჰისტოგრამის განსაკუთრებულ ტიპს, რომელსაც ჰქვია **პარეტოს დიაგრამა.** ეს სახელი უწოდეს იტალიელი ეკონომისტის ვილფრედო პარეტოს (1848-1923) პატივსაცემად. იგი იყენებდა მსგავს ჰისტოგრამებს ისეთი კატეგორიებისთვის, რომლებიც დალაგებულია სიხშირის მიხედვით, ანუ უმაღლესი სვეტიდან უმცირესისკენ. პარეტოს დიაგრამა ხშირად გამოიყენება ბიზნეს აპლიკაციებში ყველაზე უფრო ზოგადი შედეგების წარმოჩენის მიზნით, როგორიცაა მაგალითად, პროდუქტის იდენტიფიცირება გაყიდვების მაღალ მოცულობასთან, ან განისაზღვროს ყველაზე უფრო გავრცელებული პრეტეზიები, რომლებსაც იღებს მომხმარებელთა მომსახურების ცენტრი. დიაგრამა გვეხმარება **პარეტოს პრინციპის** უკეთ წარმოჩენაში, რომელიც მდგომარეობს იმაში, რომ *კატეგორიების მცირე ქვესიმრავლე ხშირად შეიცავს დაკვირვებათა უდიდეს ნაწილს.* მაგალითად, დიაგრამა გვიჩვენებს, რომ სამ კატეგორიაზე (ნახშირი, ატომური და ბუნებრივი გაზი) მოდის აშშ-ს ელექტროენეგიის წყაროების დაახლოებით 88.5%.

გრაფიკები რაოდენობრივი ცვლადებისთვის.

მოდით ახლა განვიხლოთ, როგორ დავხასიათოთ *რაოდენობრივი* ცვლადები გრაფიკულად. შევხედოთ სამი ტიპის გამოსახვას: წერტილოვანს, ღეროვან-ფოთლოვანს და ჰისტოგრამას. საილუსტრაციოდ წარმოვადგინოთ სურსათის (მარცვლოვანები) ყოველდღიური მოხმარების მონაცემების ანალიზის გრაფიკები.

წერტილოვან გრაფიკზე თითოეული წერტილი გამოსახავს დაკვირვებას, რომელიც განთავსებულია უშუალოდ რიხვითი მნიშვნელობის ზუსტად თავზე. თვით ეს მნიშვნელობა კი მონიშნულია ე.წ. დაკვირვების რიცხვით ღერძზე. წერტილოვანი გრაფიკის შედგენა ხდება შემდეგნაირად:

- დავხაზოთ ჰორიზონტალური წრფე. გავუკეთოთ წარწერა ცვლადის სახელის შესახებ და მოვნიშნოთ მასზე ცვლადის რეგულარული მნიშვნელობები.
- ყოველი დაკვირვებისთვის მოვნიშნოთ წერტილი უშუალოდ შესაბამისი მნიშვნელობის თავზე.



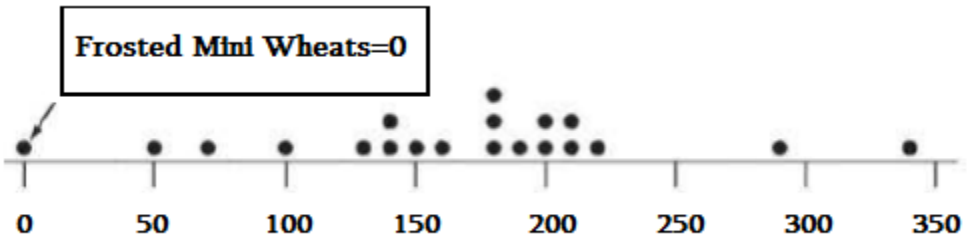
მოდით გამოვიკვლიოთ შაქრისა და ნატრიუმის რაოდენობა საუზმის მარცვლეულში. ნატრიუმიც და შაქარიც უწყვეტი რაოდენობრივი ცვლადებია, ამიტომ მონაცემები დამრგვალებულია.

დიეტოლოგების რჩევით დღიური მოხმარება არ უნდა აღემატებოდეს 2400 მგ ნატრიუმს და 50 გრამ შაქარს (2000 კალორიან დღიურ დიეტაში). A-ტიპი პოპულარულია მოზრდილებში, C-ტიპი ბავშვებში.

Cereal	Sodium (mg)	Sugar (g)	Type
Frosted Mini Wheats	0	11	A
Raisin Bran	340	18	A
All Bran	70	5	A
Apple Jacks	140	14	C
Cap'n Crunch	200	12	C
Cheerios	180	1	C
Cinnamon Toast Crunch	210	10	C
Crackling Oat Bran	150	16	A
Fiber One	100	0	A
Frosted Flakes	130	12	C
Froot Loops	140	14	C
Honey Bunches of Oats	180	7	A
Honey Nut Cheerios	190	9	C
Life	160	6	C
Rice Krispies	290	3	C
Honey Smacks	50	15	A
Special K	220	4	A
Wheaties	180	4	A
Corn Flakes	200	3	A
Honeycomb	210	11	C

ცხრილი 3.5. *ნატრიუმისა და შაქრის რაოდენობა 20 სახეობის მარცვლეულის საუზმეში.*

შემდეგი წერტილოვანი გრაფიკი გვიჩვენებს ცხრილის სურათს ნატრიუმისთვის.



ნახაზი 3.6. *წერტილოვანი დიაგრამა.* ნატრიუმის რაოდენობა წარმოდგენილია წერტილებით შესაბამისი რიცხვის ზემოთ.

კითხვა: რას ნიშნავს, რომ ერთზე მეტი წერტილი ჩანს მნიშვნელობის თავზე? (პასუხი: სიხშირე).

გრაფიკის მეორე ტიპი არის **ღეროვან-ფოთლოვანი** დიაგრამა. იგი მსგავსია წერტილოვანი დიაგრამის, რადგან გამოსახავს ცალკეულ დაკვირვებებს. თითოეული დაკვირვება გამოსახულია ღეროთი და ფოთლით. ზოგადად, ღერო შეიცავს ყველა ციფრს გარდა უკანასკნელისა, რომელიც არის ფოთოლი. მონაცემები დალაგებულია უმცირესიდან უდიდესისკენ. ღერო განლაგებულია ვერტიკალურად და ისევ უმცირესიდან უდიდესისკენ. ღეროს შემდეგ არის ვერტიკალური ხაზი, რის შემდეგაც მოდიან ფოთლები ჰორიზონტალურად და ისევ მზარდი რიგით.

განვიხლოთ მაგალითი. წინა მონაცემებიდან ამოვიღოთ ნაწილი, მხოლოდ ნატრიუმისთვის. ასევე, მონაცემები რომ იყოს კომპაქტური **მოვჭრათ** ბოლო ციფრები (ამ დროს არ ვამრგვალებთ) და მონაცემები წარმოვადგინოთ ასეთი სახით:

Truncated Data	
Frosted Mini Wheats	0
Raisin Bran	34
All Bran	7
Apple Jacks	14
Cap'n Crunch	20
Cheerios	18
Cinnamon Toast Crunch	21
Crackling Oat Bran	15
Fiber One	10
Frosted Flakes	13
Froot Loops	14
Honey Bunches of Oats	18
Honey Nut Cheerios	19
Life	16
Rice Krispies	29
Honey Smacks	5
Special K	22
Wheaties	18
Corn Flakes	20
Honeycomb	21

ცხრილი 3.7.

აქ გვაქვს მონაცემები 0, 34, 7, 20 და ა.შ. ნაცვლად წინა მონაცემების 0, 340, 70, 200 და ა.შ. დავალაგოთ ფოთლები ჰორიზონტალურად ზრდის მიხედვით და მივიღებთ ღეროვან ფოთლოვან დიაგრამას.

0 057
1 0344568889
2 001129
3 4

შევნიშნოთ, რომ ეს დიაგრამა *ზედმეტად* კომპაქტურად გამოიყურება, ამიტომ კარგად არ ჩანს სად ვარდებიან მონაცემები, როგორც ეს იყო მაგალითად, წერტილოვანი გრაფიკების შემთხვევაში.

ასეთ დროს უკეთესია ღეროში მონაცემები ორ-ორჯერ წარმოვადგინოთ და ფოთლები ერთგან იყოს 0-დან 4-მდე, მეორეზე 5-დან 9-მდე. მაშინ ჩვენ მივიღებთ ასეთ სურათს:

0 0
0 57
1 0344
1 568889
2 00112
2 9
3 4

როგორც წერტილოვანი დიაგრამის შემთხვევაში, აქაც იგრძნობა, რომ ხშირ შემთხვევაში ნატრიუმის მნიშვნელობა მაღალია იმ ორი-სამი სახის მარცვლეულთან შედარებით, რომელთაც ნატრიუმის ნაკლები შემცველობა აქვთ.

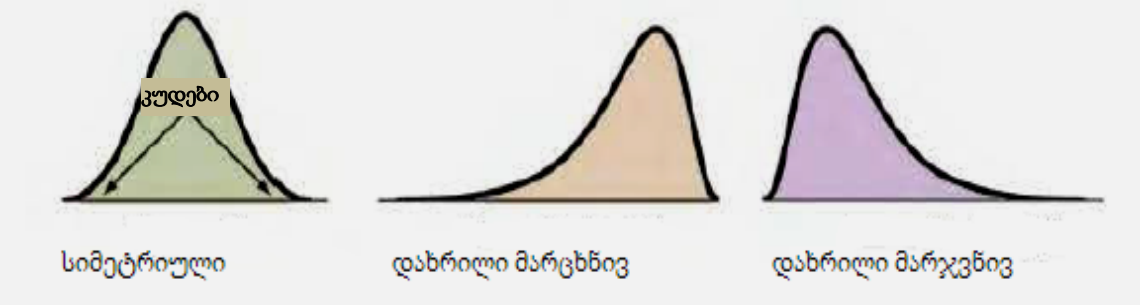
წერტილოვანი და ღეროვან-ფოთლოვანი დიაგრამიდან ადვილად შეიძლება აღვადგინოთ საწყისი მონაცემები იმიტომ, რომ დიაგრამები გამოსახავენ სათითაო დაკვირვებას. მაგრამ ეს ყველაფერი ხდება უსაშველოდ დიდი, როცა დიდია მონაცემთა სიმრავლე. ასეთ დროს უნივერსალური გზა არის სვეტებით სარგებლობა, რომელიც წარმოგვიდგენს სიხშირეებს შეჯამებული სახით. როგორც უკვე ვხვდებით, ეს არის **ჰისტოგრამა**, რომელიც რაოდენობრივი ცვლადებისთვის იყენებს სვეტებს სიხშირის გამოსასახავად.

ყურადღება: ტერმინი ჰისტოგრამა გამოიყენება რაოდენობრივი ცვლადების შემთხვევაში, ხოლო ტერმინი სვეტოვანი გრაფიკი კატეგორიული ცვლადების დროს.

ჰისტოგრამის აგების საფეხურები:

- მონაცემთა დიაპაზონი დავყოთ ტოლ მონაკვეთებად. თუ დისკრეტულ ცვლადს გააჩნია ცოტა რაოდენობის მნიშვნელობები, მაშინ ვიყენებთ ყველა მნიშვნელობას.
- გამოვითვალოთ დაკვირვებათა რაოდენობა (სიხშირე) თითეულ ინტერვალში და შევქმნათ სიხშირეთა ცხრილი.

- ჰორიზონტალურ ღერძზე მოვნიშნოთ ინტერვალების ბოლო წერტილები. დავხაზოთ სვეტი ყოველი მნიშვნელობისთვის ან ინტერვალისთვის, რომლის სიმაღლე ტოლი უნდა იყოს სიხშირის (ან პროცენტული მაჩვენებლის). ეს სიმაღლეები მონიშნული უნდა იყოს ვერტიკალურ ღერძზე.
- *განაწილების ფორმა.* მონაცემთა სიმრავლის გრაფიკი აღწერს მონაცემთა განაწილებას. ეს არის ცვლადის მიერ თითოეული მნიშვნელობის მიღების სიხშირის მაჩვენებელი. განაწილება შეიძლება აღიწეროს განაწილების ცხრილითაც.
- თუ მონაცემებს აქვთ ერთი ბორცვი, მაშინ ასეთ განაწილებას ეწოდება **უნიმოდალური. მოდა** არის ცვლადის ის მნიშვნელობა, სადაც გვაქვს უმაღლესი წერტილი. განაწილებას, რომელსაც აქვს ორი ბორცვი, ეწოდება **ბიმოდალური**. ბიმოდალური შემთხვევა უმეტესად მაშინ გვაქვს, როცა პოპულაცია პოლარიზებულია სადავო საკითხის გამო. განაწილების ფორმა ხშირად არის **სიმეტრიული** ან **ასიმეტრიული**. განაწილება სიმეტრიულია, თუ განაწილების ფერდები ცენტრალური მნიშვნელობის მიმართ სარკისებურად აისახებიან ერთმანეთში. განაწილება არის ასიმეტრიული, მისი ერთი ფერდი მეტადაა გაწელილი ვიდრე მეორე. სიმეტრიული თუ ასიმეტრიული განაწილებების წარმოსადგენად ზოგადად უკეთესია გლუვი წირების გამოყენება. შემდეგ ნახაზებზე შეჯამებულია ჰისტოგრამის ფორმები.



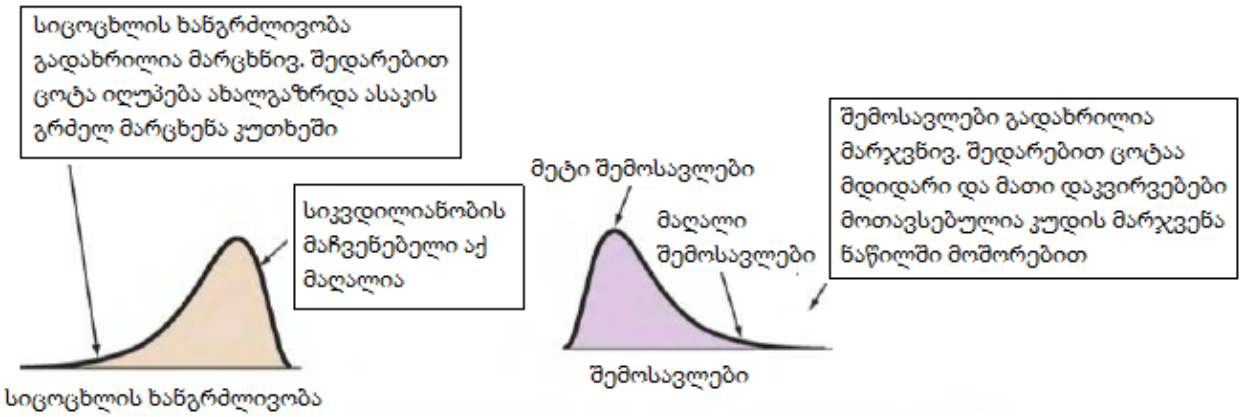
ნახაზი 3.8.

დავუკვირდეთ რა მოხდება, თუ ჰისტოგრამაში ავირჩევთ რაოდენობით უფრო მეტ და სიგრძით უფრო პატარა ინტერვალებს? მონაცემების შეგროვებისას ჰისტოგრამა გახდება თანდათანობით უფრო „გლუვი“. წირის დახრილ ნაწილებს ეწოდებათ განაწილების **კუდები.** განაწილება **დახრილია მარცხნივ,** თუ მარცხენა კუდი გრძელია მარჯვენაზე. ანალოგიურად, **დახრილია მარჯვნივ,** თუ მარჯვენა კუდი გრძელია მარცხენაზე.

როგორი იქნება მაგალითად, IQ-ს შემთხვევაში? ცნობილია, რომ მონაცემები გროვდება 100-ის გარშემო და კლებულობს ორივე მხარეს ერთნაირად, მსგავსად სარკისებური ასახვისა. (იხ. ნახ. 3.9)

ნახაზი 3.9.

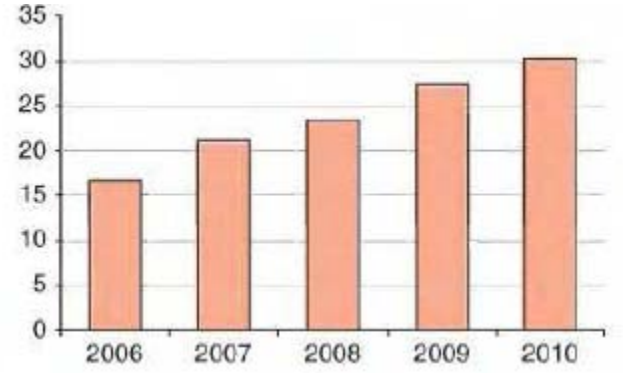
ახლა ვნახოთ როგორ გამოიყურება მრუდი სიცოცხლის საშუალო ხანგრძლივობის მონაცემების განხილვისას. ხალხის უდიდესი ნაწილი განვითარებულ საზოგადოებაში ცოცხლობს სულ მცირე 60 წლამდე, თუმცა ზოგიერთები კვდებიან ძალიან ადრეულ ასაკში. ამიტომ გრაფიკი დახრილია მარცხნიდან. (იხ. ქვედა მარცხენა ნახაზი). როგორ ფორმას უნდა ველოდეთ თუ განვიხილავთ ადამიანების წლიური შემოსავლების განაწილებებს. ამ შემთხვევაში გრძელი იქნება მარჯვენა კუდი. ანუ, ზოგიერთ ადამიანს აქვს გაცილებით დიდი შემოსავალი ვიდრე საზოგადოების უდიდეს ნაწილს. ეს კი მეტყველებს იმაზე, რომ გრაფიკი დახრილი იქნება მაჯვნიდან. (იხ. ქვედა მარჯვენა ნახაზი).



ნახაზი 3.10.

დროითი გრაფიკები: მონაცემების გამოსახვა დროზე დამოკიდებულებით.

ზოგიერთ ცვლადზე დაკვირვება მიმდინარეობს გარკვეული პერიოდის განმავლობაში. მგალითად, ყოველდღიურად ფიქსირდება აქციების ფასის ცვალებადობა, ან ათ წელიწადში ერთხელ ქვეყანაში მიმდინარეობს მოსახლეობის აღწერა და ა.შ. მონაცემთა სიმრავლეს, შეგროვებულს დროის განმავლობაში ეწოდება **დროითი მწკრივი**. შესაძლებელია დროითი მწკრივების გრაფიკული გამოსახვა ე.წ. **დროითი გრაფიკების** საშუალებით. ამ შემთხვევაში დიაგრამის ვერტიკალურ ღერძზე გადაიზომება თითოეული დაკვირვება, ხოლო დრო ჰორიზონტალურ ღერძზე. ზოგადი შაბლონია, დავაკვირდეთ ტენდენციას დროის მიხედვით და მივუთითოთ ეს ტენდენცია დროზე დამოკიდებულებით იზრდება თუ მცირდება. ტენდენციის უკეთ დასანახად ზოგჯერ სასარგებლოა მონაცემთა წერტილები შევაერთოთ და დავაკვირდეთ წირს. დროითი მწკრივების მონაცემთა წარმოდგენის მეორე გზაა, ჰისტოგრამა, ანუ სვეტოვანი დიაგრამა. შემდეგ დიაგრამაზე წარმოდგენილია გაერთიანებულ სამეფოში (დიდ ბრიტანეთში) იმ ადამიანების რაოდენობა, ვინც 2006 - 2010 წლებში იყენებდა ინტერნეტს (მილიონებში).

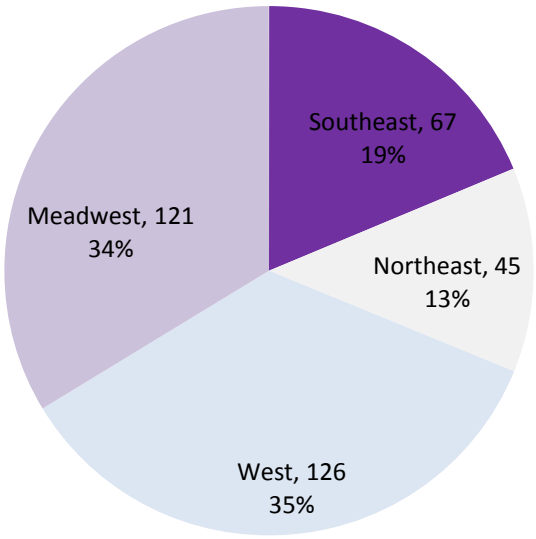


დადგენილია, რომ 2006 წელს 16.5 მილიონი კაცი იყენებდა ინტერნეტს. 2010 წელს კი თითქმის გაორმაგდა და გახდა 30 მილიონი. ამ შემთხვევაში არსებობს დროის მიხედვით ზრდის მკაფიო ტენდენცია. თუმცა პრაქტიკაში ტენდენცია ყოველთვის ასე თვალნათლივად არ ჩანს ხოლმე.

საიტზე www.gapminder.org/ არის ბევრი მაგალითი და ინსტრუმენტი დინამიური დროითი მწკრივებისთვის.

პრაქტიკული ნაწილი.

3.1. მეტეოსადგურები. წრიულ დიაგრამაზე გამოსახულია მეტეოსადგურების რეგიონალური განაწილება აშშ-ში.



ნახაზი 3.11.

ა) წრიული სექტორები (i) ცვლადებია თუ (ii) ცვლადების კატეგორიები?

ბ) ახსენით რას გამოსახავს ის ორ ორი რიცხვი, რომელიც წერია თითოეულ წრიულ სექტორთან.

გ) რიცხვებზე დაკვირვების გარეშე შესაძლებელია თუ არა ამ გრაფიკის ან შესაბამისი ჰისტოგრამის გამოყენებით ადვილად გავარკვიოთ, რომელია მოდალური კატეგორია? რატომ?

3.2. საფრაგეთი არის ყველაზე პოპულარული დასასვენებელი ადგილი. რომელ ქვეყნებს სტუმრობენ ყველაზე ხშირად ტურისტები? ცხრილში მოყვანილია მონაცემები ჟურნალიდან *„მოგზაურობა და დასვენება“*(2005წ).

ყველაზე მეტად მონახულებადი ქვეყნები, 2005წ

ქვეყანა	ვიზიტების რაოდენობა (მილიონი)
საფრანგეთი	77.0
ჩინეთი	53.4
ესპანეთი	51.8
აშშ	41.9
იტალია	39.8
გაერთ. სამეფო	24.2
კანადა	20.1
მექსიკა	19.7

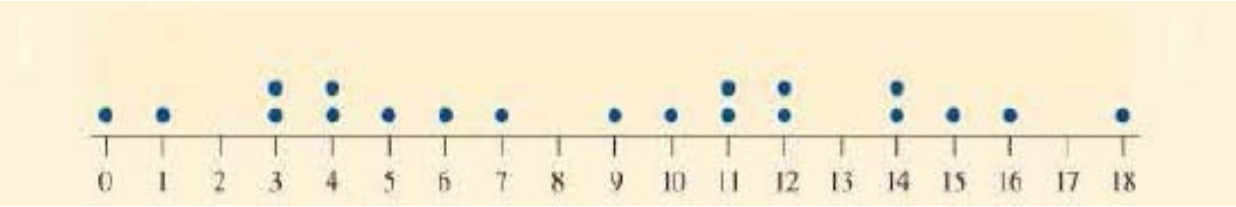
ცხრილი 3.12.

ა) მონახულებული ქვეყნები კატეგორიული თუ რაოდენობრივი ცვლადებია?

ბ) ჰისტოგრამის ასაგებად რომელი უფრო კარგი ვერსიაა: ჩამოვთვალოთ ქვეყნები ანბანის მიხედვით, თუ პარეტოს დიაგრამის სახით? ახსენით.

გ) წერტილოვანი დიაგრამის აგება უფრო აზრიანი იქნება თუ ღეროვან-ფოთლოვანის? ახსენით.

3.3. შაქრის წერტილოვანი დიაგრამა. ზემოთ გვქონდა მოცემული ჯანმრთელ მარცვლოვან საუზმეში შაქრის შემცველობის მონაცემთა ცხრილი. ახლა მსგავსი რამ წარმოდგენილია წერტილოვანი დიაგრამის საშუალებით:



შაქარი (გრ)

ნახაზი 1.13.

ა) განსაზღვრეთ შაქრის მინიმალური და მაქსიმალური რაოდენობა.

ბ) შაქრის რა მნიშვნელობები გვხვდება ყველაზე ხშირად? რა ეწოდება ამ მნიშვნელობებს?

3.4. საგამოცდო ქულების გრაფიკი. ლექტორი უჩვენებს სუდენტებს შუალედური გამოცდის შედეგებს გამოსახულს ღეროვან ფოთლოვანი დიაგრამის სახით:

4	**13**	**0000**
6	**14**	**00**
7	**15**	**0**
8	**16**	**0**
(4)	**17**	**0000**
6	**18**	**00**
4	**19**	**00**
2	**20**	**0**
1	**21**	**0**
1	**22**	
1	**23**	
1	**24**	**0**

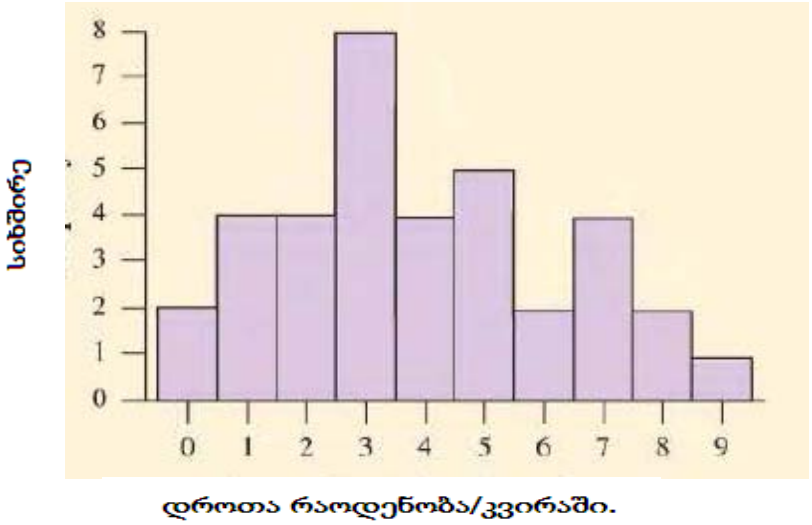
ცხრილი 3.14.

ა) განსაზღვრეთ სტუდენტების რაოდენობა და მათი მაქსიმალური და მინიმალური ქულები.

ბ) წარმოადგინეთ იგივე მონაცემები წერტილოვან დიაგრმაზე.

გ) წარმოადგინეთ იგივე მონაცემები ჰისტოგრამის სახით ოთხი ინტერვალისთვის.

3.5. რამდენად ხშირად კითხულობენ სტუდენტები გაზეთებს? ჰისტოგრამის საშუალებით შეაფასეთ საშუალოდ რა დროს ანდომებენ სტუდენტები ყოველდღიურად გაზეთების კითხვას.



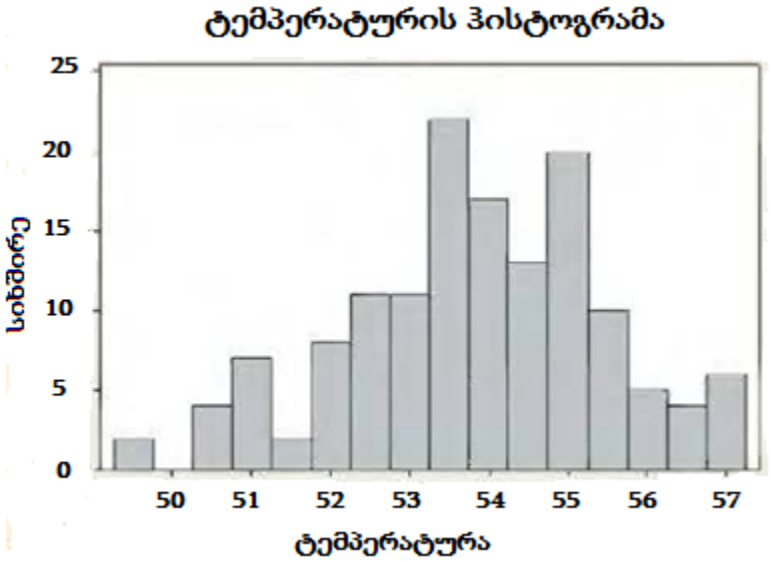
ნახაზი 3.15.

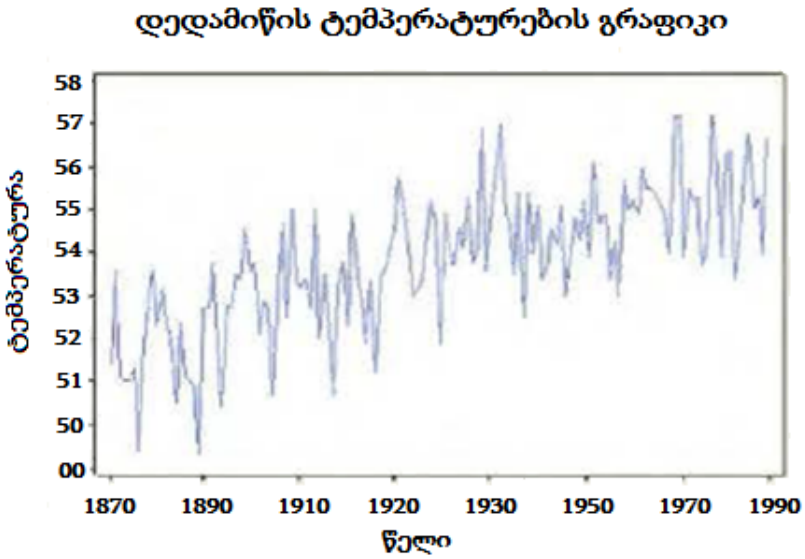
ა) ეს ცვლადი უწყვეტია თუ დიკრეტული? ახსენით.

ზ) მოცემული ჰისტოგრამა იძლევა ამ ცვლადის შედეგს. იგი მიიღებულ იქნა ჯორჯიის უნივერსიტეტის 36 სტუდენტის გამოკითხვის შედეგად. დაასკვენით რა არის: (i) მინიმალური პასუხი, (ii) მაქსიმალური პასუხი, (iii) სტუდენტთა რაოდენობა, ვინც საერთოდ არ კითხულობს გაზეთებს, (iv) მოდა.

გ) აღწერეთ განაწილების ფორმა.

3.6. ჰაერის ტემპერატურა ცენტრალურ პარკში. პირველ ნახაზზე ჰისტოგრამა გვიჩვენებს ნიუ-იორკის ცენტრალურ პარკში საშუალო წლიური ტემპერატურის მონაცემებს 1869 წლიდან 2010 წლამდე.





ა) აღწერეთ რა ფორმისაა განაწილება.

ზ) რა ინფორმაცია შეიძლება მოგვცეს დროითმა გრაფიკმა ისეთი, რასაც ვერ გვაძლევს ჰისტოგრამა?

გ) და პირიქით: რა ინფორმაცია შეიძლება მოგვცეს ჰისტოგრამამ ისეთი, რასაც ვერ გვაძლევს დროითი გრაფიკი?

4

მონაცემთა საშუალო და მედიანა. მათი შედარება.

დისპერსია, სტანდარტული გადახრა.

წინა პარაგრაფებში ვნახეთ, რომ გრაფიკები კარგად ასახავენ პროპორციულ დამოკიდებულებებს კატეგორიული ცვლადებისთვის. რაოდეობრივი ცვლადებისთვის კი გრაფიკი გვიჩვენებს **ფორმის** განსაკუთრებულ მხარეებს. ეს ყოვლთვის კარგი იდეაა იმისთვის, რომ უკეთ „შევიგრძნოთ“ მონაცემები პირველივე შეხედვით. მერე უკვე დავაკვირდეთ ამორჩევების რიცხობრივ რეზიუმეს (**სტატისტიკას**). რაოდენობრივი ცვლადებისთვის შევეცდებით, პასუხი გავცეთ კითხვებს, „რა არის წამომადგენლობითი დაკვირვება“ და „დაკვირვებები ღებულობენ მსგავს მნიშვნელობებს, თუ ოდნავ მაინც განსხვავდებიან?“ პირველ კითხვაზე პასუხის გასაცემად აღვწეროთ განაწილების **ცენტრი.** მეორისთვის კი განაწილების **ცვალებადობა** (გაფანტულობა). რაოდენობრივი ცვლადის განაწილების ყველაზე ცნობილი და ყველაზე ხშირად ხმარებადი ცენტრის მაჩვენებელია საშუალო. იგი მიიღება დაკვირვებათა გასაშუალებით (მაგ. საშუალო არითმეტიკუ-ლის მოძებნით).

საშუალო არის დაკვირვებათა ჯამი შეფარდებული დაკვირვებათა რაოდენობასთან. მას წარმოვიდგენთ, როგორც განაწილების წონასწორობის წერტილს.

მეორე პოპულარული საზომი არის მედიანა. დაკვირვებების ნახევარი მეტია მასზე, ნახევარი კი - ნაკლები.

მედიანა არის დაკვირვებათა საშუალო მაჩვენებელი, როდესაც დაკვირვებები დალაგებულია უმცირესიდან უდიდესისკენ (ან უდიდესიდან უმცირესისკენ).

რეზიუმე: *როგორ განვსაზღვროთ მედიანა?*

- დავალაგოთ n დაკვირვება ზრდადობის რიგით.
- თუ დაკვირვებათა რაოდენობა n არის კენტი, მაშინ მედიანა არის შუა დაკვირვება დალაგებული ამორჩევიდან.
- თუ დაკვირვებათა რაოდენობა n არის ლუწი, მაშინ ვიღებთ შუა ორ დაკვირვებას, და მედიანა არის მათი არითმეტიკული საშუალო.

შევხედოთ ახლოდან საშუალოს და მედიანას. საშუალოს აღნიშვნა გამოიყენება როგორც ფორმულაში, აგრეთვე როგორც შემოკლება ტექსტში.

სიტყვებით: ფორმულა

$$\bar{x} = \frac{\sum x}{n}$$

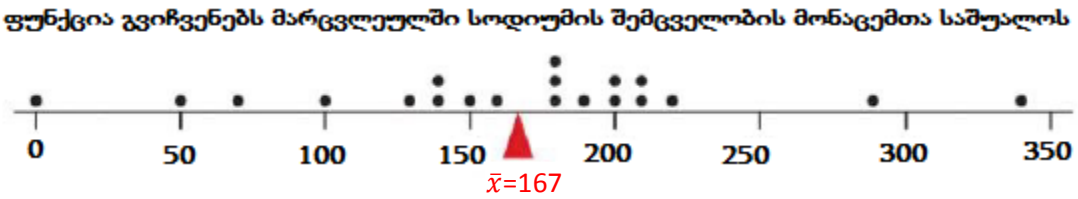
არის შემოკლებული ჩაწერა გამონათქვამის: „ x ცვლადის მნიშვნელობათა ჯამი გაყოფილი შერჩევის ზომაზე“.

იცით თუ არა რომ...

საშუალო აგრეთვე წარმოადგენს „სამართლიანი წილის“ მნიშვნელობას. მაგალითად, როცა ვიხილავთ იმ 20 მარცვლეულს, მათი საშუალო 167მგ შეიძლება წარმოვიდგინოთ, როგორც ნატრიუმის შემცველობა თითოეულ მარცვლეულში, ყველა მარცვლეულს რომ ჰქონოდა ერთი და იგივე შემცვლობა.

საშუალოს ზოგიერთი ძირითადი თვისება:

- საშუალო არის მონაცემთა წონასწორობის წერტილი: თუ ღერძზე, სადაც მონიშნულია დაკვირვებები მოვათავსებთ ერთი იმავე წონებს, მაშინ ღერძი იქნება წონასწორობაში, თუ საყრდენს მოვათავსებთ ღერძის შუაში.



- არათანაბარი განაწილების დროს საშუალო გადაიხრება პიკის მხარეს მედიანასთან მიმართებაში.
- საშუალოზე ძლიერ გავლენას ახდენს მონაცემები, რომლებიც „ამოვარდნილები“, ანუ არიან საკმარისად მცირეები ან საკმარისად დიდები.

ამოვარდნა არის დაკვირვება, რომელიც ვარდება მონაცემთა ძირითადი ბლოკის ძალიან ზევით ან ძალიან ქვევით. ამოვარდნები საჭიროებენ განსაკუთრებულ კვლევას იმისთვის, რომ დავინახოთ ეს მონაცემთა შეცდომის ბრალია, თუ საოცარ და არაჩვეულებრივ მოვლენასთან გვაქვს საქმე.

საშუალოს და მედიანის შედარება. განაწილების ფორმის საშუალებით შეიძლება დავადგინოთ, საშუალო მეტია თუ ნაკლებია მედიანაზე. მაგალითად, თუ ექსტრემალურად დიდი მნიშვნელობა არის კუდის მარჯვენა მხარეს საშუალოსაც წაანაცვლებს მარჯვნივ. ზოგადად, თუ ფორმა არის

- სავსებით სიმეტრიული, მაშინ საშუალო ტოლია მედიანის.
- დახრილი მარჯვნივ, მაშინ საშუალო მეტია მედიანაზე.
- დახრილი მარცხნივ, მაშინ საშუალო ნაკლებია მედიანაზე.

ნახ. 4.1. *ურთიერთდამოკიდებულება მედიანასა და საშუალოს შორის.*

ნახაზები გვიჩვენებენ, რომ საშულო ინაცვლებს გრძელი კუდის მხარეს. გრაფიკის გადახრა მიუთითებს, რომ საშუალო და მედიანა განსხვავდებიან. თუ განაწილება ახლოსაა სიმეტრიულთან, მაშინ კუდები იქნება თითქმის ერთნაირი ზომის, ამიტომ საშუალო და მედიანაც ახლოს იქნებიან ერთმანეთთან. ხოლო რაც მეტია განაწილების დახრა, მით მეტად განსხვავდებიან საშუალო და მედიანა.

რატომ არ მოქმედებს მედიანაზე ამოვარდნა? თუ რამდენად შორსაა ამოვარდნა განაწილების ცენტრიდან, არ აისახება მედიანაზე. ეს იმიტომ ხდება, რომ მედიანა განისაზღვრება მხოლოდ იმით, რომ თანაბარი რაოდენობის მონაცემები უნდა იყოს მის ზევით და ქვევით. რადგანაც საშუალო იყენებს *ყველა* რიცხვით მონაცემს, ამიტომ მედიანისგან განსხვავებით იგი გამოკიდებულია იმაზე, თუ ცენტრიდან რამდენად შორს ვარდება დაკვირვება. ვინადან საშუალო არის წონასწორობის წერტილი, ამიტომაც ექსტრემალური მნიშვნელობის არსებობა ცენტრიდან მარჯვენა მხარეს მარჯვნივ ქაჩავს საშუალოსაც. ხოლო რადგანაც მედიანა არ არის დამოკიდებული ასეთ ცვლილებებზე, ამიტომ ამბობენ, რომ ის **რეზისტენტულია** (მდგრადია) ასეთი ექსტრემალური დაკვირვებების მიმართ.

დასკვნა: მედიანა რეზისტენტულია ამოვარდნების მიმართ.

ვიტყვით, რომ დაკვირვებათა რიცხვითი შეჯამება არის რეზისტენტული, თუ ექსტრემალურ დაკვირვებებს, თუ ასეთები არსებობენ, აქვთ მცირე გავლენა მათ მნიშვნელობებზე.

ე. ი. მედიანა რეზისტენტულია, საშუალო - არა.

ამ თვისებებიდან გამომდინარე, შეიძლება იფიქროთ, რომ ყოველთვის უკეთესია გამოვიყენოთ მედიანა ვიდრე საშუალო. ეს არ იქნებოდა მართებული საშუალოს აქვს სხვა მეტად საჭირო თვისებებიც. მათ ჩვენ ცოტა მოგვიანებით შევისწავლით. ამის გარდა, არსებობს გარკვეული უპირატესობები, რომ საზომებად გამოვიყენოთ ხოლმე *მთლიანი მონაცემების* რიცხვითი მნიშვნელობები. მაგალითად, *დისკრეტული მონაცემებისთვის, რომლებიც ღებულობენ სულ რამდენიმე მნიშვნელობას, მონაცემთა სრულიად განსხვავებული მოდელები (patterns) იძლევიან ერთი და იმავე შედეგს მედიანისთვის.* ეს კი ნიშნავს, იგი *ნამეტანი* მდგრადია. ამის საილუსტრაციოდ მოვიყვანოთ სიტუაციური მაგალითი, რომელიც გვიჩვენებს აგრეთვე, როგორ ვიპოვოთ საშუალო და მედიანა სიხშირეთა ცხრილიდან.

სცენარი 4.2. ქორწინებათა სტატისტიკა. აღწერის ბიუროს აგარიშებში არის მონაცემები, თუ რამდენჯერ იყვნენ აშშ-ს მცხოვრებნი ქორწინებაში, სხვადასხვა ასაკობრივი ჯგუფების მიხედვით.

(წყარო: www.census.gov/hhes/socdemo/marriage/data/sipp/2004/tables.html.)

ცხრილში მოყვანილია მონაცემები 20-24 წლიანი ასაკობრივი ჯგუფისთვის. სიხშირეები სინამდვილეში ათასეულებია, მაგალითად, მამაკაცების, რომლებიც არასდროს ყოფილან ქორწინებაში არის 8 418 000, მაგრამ ეს არ მოქმედებს საშუალოს ან მედიანის გამოთვლაზე.

	სიხშირე	
ქორწინებათა რაოდენობა	ქალები	მამაკაცები
0	7350	8418
1	2587	1594
2	80	10
ჯამი	10,017	10,022

ცხრილი 4.2. *ქორწინებათა რაოდენობა 20-24 წლის ასაკამდე.*

კითხვები შესწავლისთვის:

ა) იპოვეთ მედიანა და საშუალო თითოეული სქესისთვის.

ბ) რატომ არ არის მედიანა მაინცდამაინც დიდად ინფორმაციული?

შემდგომი მოსაზრებები: **ა)** 10 017 ქალისთვის დალაგებული შერჩევა იქნება

0 0 0 0 0 .... 01 1 1 1 1 .... 1 2 2 2 2 2 2

(7350-ჯერ) (2587-ჯერ) (80-ჯერ)

რადგანაც n=10 017, დაკვირვებები გადანომრილი იქნება 1-დან 10 017-მდე. შუა ნომერი იქნება (1+10 017)/2=5009-ე ნომერი. სულ გვაქვს სამი განსხვავებული პასუხი და 5 009-ზე მეტი პასუხი არის 0, ამიტომ მედიანა არის 0. ანალოგიურად, მამაკაცებისთვისაც მედიანა არის 0.

ქალებისთვის მონაცემთა ჯამი გამოითვლება შემდეგნაირად:
$\sum \boldsymbol{x} = 7350(0) + 2587(1) + 80(2) = 2747$

რადგანაც შერჩევის ზომა არის 10 017, საშუალო იქნება 2747/10017=0.274. მსგავსად შეიძლება გამოვთვალოთ მამაკაცებისთვისაც და მათთვის საშუალო იქნება 0.161

ბ) მედიანა ქალებისთვის და კაცებისთვისაც რომ არის 0, არ უნდა ვიფიქროთ, რომ არ ყოფილა განსხვავება მათ მონაცემების შორის. საშუალო იყენებს მონაცემების ყველა რიცხვით მნიშვნელობას, დალაგებას არავითარი მნიშვნელობა არ აქვს. საბოლოოდ ამ ასაკბრივი ჯგუფის მონაცემებზე დაკვირვებით შეიძლება დავასკვნათ, რომ ქალები უფრო ხშირად იყვნენ ქორწინებაში ვიდრე კაცები.

წვდომის უნარი (ინტუიცია): მედიანა იგნორირებას უკეთებს დიდ ინფორმაციას, მაშინ როცა მონაცემები დისკრეტულია და ღებულობენ ცოტა მნიშვნელობებს. ექსრემალურ შემთხვევაში, როდესაც გვაქვს **ბინარული მონაცემები,** რომლებიც იღებენ მხოლოდ ორ მნიშვნელობას, 0 და 1, მედიანა ღებულობს იმ მნიშვნელობას, რომელიც ხშირად მოდის, მაგრამ არ იძლევა არავითარ ინფორმაციას ამ ორი დაკვირვების შედეგების ფარდობითობაზე.

მაგალითად, განვიხილოთ შერჩევა ზომით 5. ცვლადები იყოს ქორწინებათა რაოდენობა. დაკვირვებას (1, 1, 1, 1, 1) და დაკვივებას (0, 0, 1, 1, 1) მედიანა აქვთ 1. საშუალო კი (1, 1, 1, 1, 1) -სთვის არის 1, ხოლო (0, 0, 1, 1, 1) -სთვის არის 3/5.

როდესაც დაკვირვება იღებს მხოლოდ 0 და 1 მნიშვნელობას, მაშინ საშუალო იმ *წილადის* ტოლია, რომლის მრიცხველია იმ დაკვირვებების რაოდენობა, რომლებიც ღებულობენ მნიშვნელობას 1-ს, ხოლო მნიშვნელია მონაცემეთა მთლიანი რაოდენობა.

ეს უფრო მეტად ინფორმაციულია, ვიდრე მედიანა. როდესაც მონაცემები დისკრეტულია, მაგრამ ღებულობენ ორზე მეტ მნიშვნელობებს, მაგალითად, როგორც მოცემულ ცხრილში, უფრო მეტად ინფორმაციულია მივაწოდოთ პროცენტული მაჩვენებელი ან პროპორცია შესაძლო შედეგებზე , ვიდრე ინფორმაცია მედიანაზე ან საშუალოზე.

თუ გვინდა, რომ შევაჯამოთ მონაცემები, სადაც გვაქვს ამოვარდნები, მაშინ მედიანა შეიძლება იყოს მეტად აქტუალური. ამიტომაც ამერიკის შეერთებული შტატებისთვის მეტად მნიშვნელოვანია ჯვარედინი შედეგები. ზოგადად, პრაქტიკაში მატად სასარგებლოა, რომ ვიცოდეთ ორივე, საშუალოც და მედიანაც. თუმცა „ქორწინების“ შემთხვევაში საშუალო უფრო მნიშვნელოვანია ვიდრე მედიანა.

ფორმის გავლენა საშუალოსა და მედიანაზე.

- თუ განაწილება საკმაოდაა გადახრილი, მაშინ მედიანა უფრო მნიშვნელოვანია, ვიდრე საშუალო, რადგანაც ის უფრო უკეთ გამოხატავს რა უფრო ტიპურია.
- თუ განაწილება ახლსაა სიმეტრიულთან ან მცირედაა სახეშეცვლილი, ან ეს არის დისკრეტული განაწილება რამდენიმე განსხვავებული მნიშვნელობით, მაშინ, ზოგადად, საშუალოს უფრო აქვს უპირატესობა, რადგანაც იგი იყენებს მთლიანი დაკვირვების ყველა რიცხვით მონაცემებს.

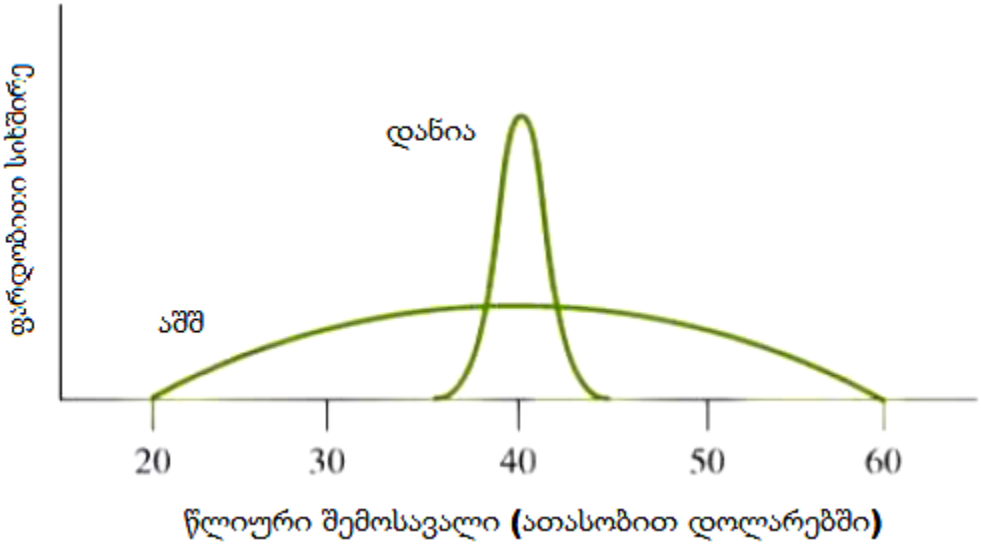
  მოდა.

  ვიცით, რომ **მოდა** არის ის მნიშვნელობა, რომელსაც აქვს ყველაზე დიდი სიხშირე. იგი აღწერს, ტიპიური დაკვირვების თვალსაზრისით ყველაზე გავრცელებულ შედეგს. ანუ ძირითადად გამოიყენება კატეგორიული ცვლადის იმ კატეგორიების აღწერისთვის, რომელთაც აქვთ მაღალი სიხშირე (მოდალური კატეგორია). მოდა უფრო მეტად საჭიროა დისკრეტული რაოდენობრივი ცვლადების შემთხვევაში, როდესაც ცვლადები ღებულობენ განსხვავებულ მნიშვნელობებს მცირე რაოდენობით. მაგალითად, „ქორწინების“ შემთხვევაში ორივე სქესისთვის მოდა არის 0.

  თუმცა მონაცემთა ეს საზომები არ არის საკმარისი რაოდენობრივი ცვლადის განაწილების ადექვატური აღწერისთვის. მაგალითად, იგი არაფერს გვეუბნება

მონაცემთა ცვლილებაზე, განსაკუთრებით მაშინ, როცა მონცემები ძალიან შორს არიან საშუალოსგან. ამ პრობლემებს პასუხი რომ გავცეთ, საჭიროა რიცხვითი მახასითბლები ანუ რიცხვითი რეზიუმეები.

ცვლილების საზომი: რანგი, ანუ დიაპაზონი. განვიხილოთ ნახაზზე წარმოდგენილი სიტუაცია.



გრაფიკი 4.3. *მუსიკის მასწავლებლების შემოსავლების განაწილების გრაფიკი დანიასა და აშშ-ში.*

მასზე წარმოდგენილია საჯარო სკოლების მუსიკის მასწავლებლების ყოველწლიური შემოსავლები დანიასა და აშშ-ში. ორივე განაწილება სიმეტრიულია და აქვთ ერთი და იგივე საშუალო, დაახლოებით $ 40 000. ამასთან, დანიაში წლიური შემოსავლები $ 35 000 -დან $ 45 000 -მდეა, მაშინ როცა აშშ -ში არის $ 20 000-დან $ 60 000 -მდე. შემოსავლები ძალიან ჰგავს ერთმანეთს დანიაში და ძალიანაა განსხვავებული აშშ-ში. ყველაზე მარტივი გზა ამ ცვლილებათა აღწერისთვის არის **რანგი**, ანუ რაც იგივეა **დიაპაზონი**. მართალია, საშუალო ერთი და იგივეა, მაგრამ ამერიკისთვის დიდია გაბნევა საშუალოსგან.

კითხვა. როგორი იქნება რანგი დანიისთვის, რომელიმე ერთ მასწავლებელს რომ ჰქონდეს $ 10 000?

რანგი არის განსხვავება უდიდეს და უმცირეს მნიშვნელობებს შორის.

დანიისთვის იგი არის $ 45 000 - $ 35 000 = $ 10 000, ხოლო ამერიკისთვის კი $ 60 000 - $ 20 000 = $ 40 000. ცხადია, რანგი მით უფრო დიდი იქნება, რაც უფრო დიდია მონაცემთა გაფანტულობა.

რანგი ადვილი გამოსათვლელი და ადვილად გასაგებია, მაგრამ იყენებს მხოლოდ ექსტრემალურ მნიშვნელობებს და იგნორირებას უკეთებს სხვა დანარჩენ მნიშვნელობებს. ამიტომ მასზე სერიოზულად ზემოქმედებენ ამოვარდნები. მაგალითად, თუ დანიის ერთ მასწავლებელს ექნებოდა შემოსავალი $ 100 000, მაშინ რანგი (დიაპაზონი) $ 45 000 - $ 35 000 = $ 10 000 - დან შეიცვლებოდა $ 100 000 - $ 35 000 = $ 65 000 -მდე. რანგი არ არის მდგრადი სტატისტიკა. იგი იზიარებს საშუალოს ყველაზე ცუდ თვისებას, რომ არ არის მდგრადი და მედიანის ყველაზე ცუდ თვისებას, რომ იგნორირებას უკეთებს თითქმის ყველა რიცხვით მნიშვნელობას.

გაზომვის ცვალებადობა. სტანდარტული გადახრა.

პრაქტიკაში ნაკლებად იყენებენ რანგს, რადგან ის იყენებს მხოლოდ უდიდეს და უმცირეს დაკვირვებებს. გაცილებით უკეთესია ცვლილების საზომი, რომელიც გამოიყენებს მთლიან მონაცემებს და აღწერს რამდენად შორს ვარდება მონაცემი საშუალოსთან შედარებით. შეჯამების გზით ამას აკეთებს **გადახრა** საშუალოსგან.

- x დაკვირვებების **გადახრა** $\bar{x}$ საშუალოსგან არის სხვაობა $\left(x-\bar{x}\right)$ დაკვირვებასა ამორჩევის საშუალოს შორის.

კითხვა: როდის იქნება გადახრა დადებითი და როდის უარყოფითი?

- გადახრა დადებითია, როცა დაკვირვებები მეტია საშუალოზე და უარყოფითია, თუ დაკვირვებები ნაკლებია საშუალოზე.
- ყოველ დაკვირვებას აქვს გადახრა საშუალოსგან.
- თუ საშუალოს წარმოვიდგენთ წონასწორობის წერტილად, მაშინ დადებითი გადახრის კონტრბალანსია უარყოფითი გადახრები. ამის გამო, *გადახრების ჯამი ყოველთვის არის 0-ის ტოლი.* ცვალებადობის ზოგად მაჩვენებლებად იყენებენ საშუალოსგან გადახრის კვადრატს ან მის აბსოლუტურ მნიშვნელობას.
- გადახრის კვადრატის საშუალოს ეწოდება **დიპერსია**. იმის გამო, რომ დიპერსია საწყისი მონაცემების საზომად იყენებს კვადრატულ ერთეულებს, ამიტომ დისპერსიიდან კვადრატული ფესვი მოგვცემს უკეთეს ინტერპრეტაციას. მას უწოდებენ **სტანდარტულ გადახრას.**

- სიდიდეს $\sum\left(x-\bar{x}\right)^2$ ეწოდება კვადრატების ჯამი. მის საპოვნელად თითოეული დაკვირვებისთვის უნდა ვიპოვოთ გადახრა, მერე ყველა ავიყვანოთ კვადრატში და შევკრიბოთ.

დაკვირვებათა **სტანდარტული გადახრა s** არის შემდეგი სიდიდე

$$s=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}$$

იგი არის კვადრატული ფესვი **დიპესიიდან** s^2 -დან , რომელიც თავის მხრივ არის გადახრების კვადრატების საშუალო:

$$s^2=\frac{\sum(x-\bar{x})}{n-1}$$

კალკულატორით მარტივადაა შესაძლებელი სტანდარტული გადახრის s -ის გამოთვლა. უხეშად რომ ვთქვათ, სტანდარტული გადახრა s წარმოადგენს ჩვეულებრივ მანძილს ან საშუალო მანძილს დაკვირვებიდან საშუალომდე. მისი ყველაზე მნიშვნელოვანი თვისება კი არის შემდეგი:

- რაც უფრო დიდია სტანდარტული გადახრა, მით მეტია მონაცემთა ცვალებადობა (გაფანტულობა).

მცირე ტექნიკური მომენტი: თქვენ შეიძლება გაგიკვირდეთ, დისპერსიისა და სტანდარტული გადახრის გამოთვლებში რატომ იყენებენ (n-1)-ს, n-ის მაგიერ? როგორც ვთქვით, დიპერსია არის n რაოდენობა გადახრის კვადრატების საშუალო. ამიტომ რატომ არ უნდა გავყოთ n -ზე? ზოგადად ეს ხდება იმიტომ, რომ გადახრა შეიცავს ცვალებადობის შესახებ მხოლოდ n-1 რაოდენობის ინფორმაციის ნაწილს. რადგანაც, როგორც ვნახეთ, გადახრების ჯამი ყოველთვის არის 0. ამიტომ n-1 გადახრის საშუალებით ყოველთვის განისაზღვრება უკანასკნელი გადახრა. ე.ი. სულ გვქონია n-1 დამოუკიდებელი გადახრა. მაგალითად, წარმოვიდგინოთ, რომ გვაქვს n=2 დაკვირვება და პირველი გადახრა არის $(x-\bar{x})=5$. მაშინ მეორე გადახრა იქნება , $(x-\bar{x})=-5$, რადგანაც, გადახრების ჯამი უნდა იყოს 0. იმ შემთხვევაში როცა n=2, გვაქვს მხოლოდ n-1=1 ცალი არაზედმეტი ინფორმაცია. იმ შემთხვევაში, თუ n=1 სტანდარტული გადახრა არ არის განსაზღვრული, ვინადან მხოლოდ ერთი

დაკვირვებით შეუძლებელია, ვიქონიოთ წარმოდგენა, რამდენად განსხვავდებიან მონაცემები, ანუ როგორ ცვალებადობენ.

რეზიუმე: სტანდარტული s გადახრის თვისებები.

- რაც მეტია ცვალებადობა მონაცემთა საშუალოსგან, მით მეტია S-ის მნიშვნელობა.
- S = 0 მხოლოდ მაშინ, როცა ყველა დაკვირვება იღებს ერთი და იმავე მნიშვნელობას. მაგალითად, თუ ვთქვათ, შვიდი ადმიანისთვის ბავშვების ყოლის სასურველი რაოდენობებია 2, 2, 2, 2, 2, 2, 2, მაშინ საშუალო ტოლია 2 -ის, თითოეულის გადახრა არის 0, და ამიტომ S=0. ეს არის ამორჩევაში უმცირესი შესაძლო ცვლილება.
- S -ის სიდიდეზე შეიძლება გავლენა იქონიოს ამოვარდნებმა, რადგანაც იგი იყენებს საშუალოს, რომელზეც მკვეთრად ზემოქმედებენ ამოვარდნები. გარდა ამისა, ამოვარდნებს გააჩნიათ დიდი გადახრები და ამიტომ მათი კვადრატული გადახრები მიისწრაფვიან ექსტრემალურად დიდი რიცხვისაკენ. მათ შეუძლიათ ძალიან გაზარდონ S -ის მნიშვნელობები და მგრძნობიარენი არიან ამოვარდნების მიმართ.

S სიდიდის ინტერპრეტაცია. ემპირიული წესი.

ვიგულისხმოთ, რომ განაწილება უნიმოდალურია და დაახლოებით სიმეტრიულია **ზარის ფორმის.** (იხ. ნახ)

S -ს მნიშვნელობას გააჩნია ზუსტი ინტერპრეტაცია. საშუალოსა და სტანდარტულ გადახრაზე დაყრდნობით შეგვიძლია ავაგოთ ინტერვალები, რომლებიც შეიცავენ მონაცემთა განსაზღვრულ პროცენტებს (მიახლოებით).

სიტყვიერად:

ა) $\bar{x} - s$ აღნიშნავს 1 სტანდარტულ გადახრას საშუალოდან ქვემოთ.

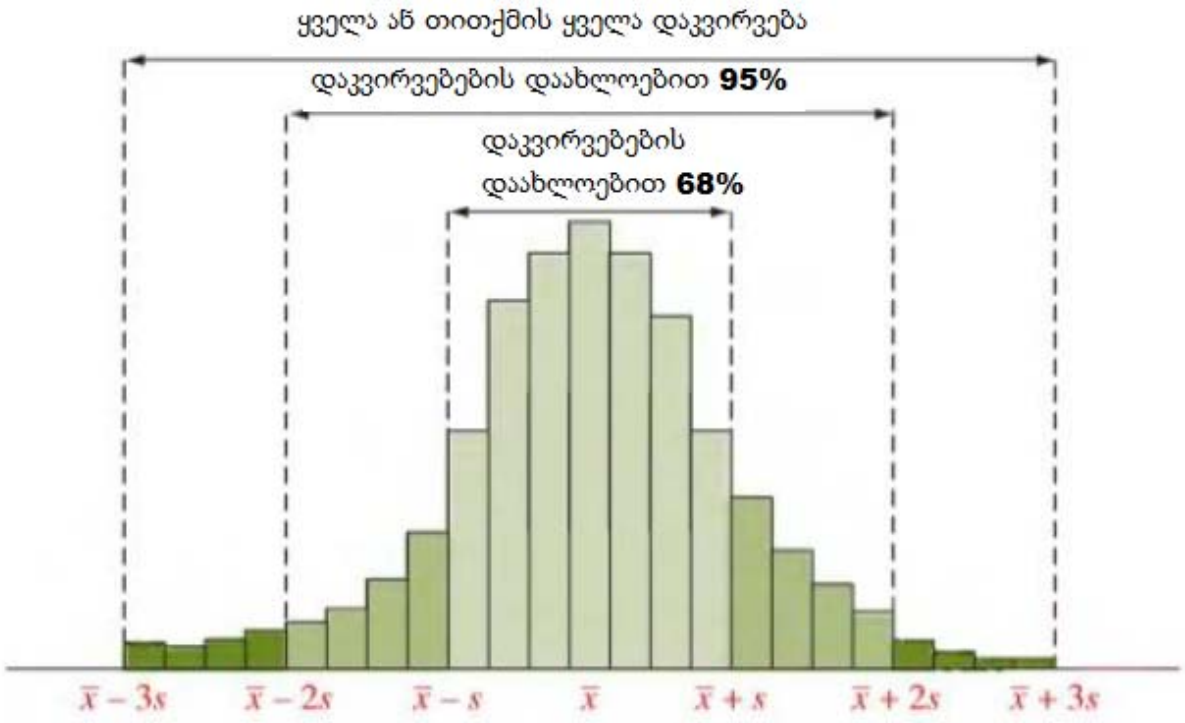
ბ) $\bar{x} + s$ აღნიშნავს 1 სტანდარტულ გადახრას საშუალოდან ზემოთ.

გ) $\bar{x} \pm s$ აღნიშნავს 1 სტანდარტულ გადახრას საშუალოდან ორივე მიმართულებით.

ემპირიული წესი:

თუ მონაცემთა განაწილებას აქვს ზარის ფორმა, მაშინ დაახლოებით

- დაკვირვებათა 68% ვარდება საშუალოდან 1 სტანდარტულ გადახრაში, ანუ მოთავსებულია $\bar{x} - s$-სა და $\bar{x} + s$ მნიშვნელობებს შორის (რომელიც აღინიშნება ასე: $(\bar{x} \pm s)$.
- დაკვირვებათა 95% ვარდება საშუალოდან 2 სტანდარტულ გადახრაში, ანუ მოთავსებულია ($\bar{x} \pm 2s$) შუალედში.
- ყველა ან თითქმის ყველა დაკვირვება ვარდება საშუალოდან 3 სტანდარტულ გადახრაში, ანუ მოთავსებულია $(\bar{x} \pm 3s)$ შუალედში.



დიაგრამა 4.4. *ემპირიული წესი.*

გვახსოვდეს, რომ ემპირიული წესი გამოიყნება მხოლოდ ზარის - ფორმის განაწილებისთვის. შემდგომში ვნახავთ, რატომ „მუშაობს“ ემპირიული წესი.

სპეციალური ფორმის გლუვ ზარის - ფორმის წირთა ოჯახს ეწოდება **ნორმალური განაწილება**. ჩვენ ვნახავთ, როგორ უნდა ვიპოვოთ განაწილების პროცენტული მაჩვენებელი თითოეულ კონკრეტულ რეგიონში, თუ ვიცით მხოლოდ საშუალო და სტანდარტული გადახრა.

ემპირიული წესის გამოყენებით შესაძლებელია მიახლოებით გამოვთვალოთ, ფაქტობრივად, ყველა პროცენტული მაჩვენებელი, რომლებიც ვარდებიან საშუალოდან 1, 2 და 3 სტანდარტულ გადახრებში. ეს წესი ამ მხრივ ცუდია, მონაცემები ძალიან დისკრეტულია და ცვლადები ღებულობენ ძალიან მცირე რაოდენობის მნიშვნელობებს.

შერჩევის სტატისტიკები და პოპულაციის პარამეტრები: გავიხსენოთ, რომ **პოპულაცია** არის ტოტალური ჯგუფი, რომლის შესახებაც გვინდა გავაკეთოთ რაღაც დასკვნები. **შერჩევა** ანუ ნიმუში რეალურად არის პოპულაციის ქვესიმრავლე, რომლის შესახებაც გვაქვს მონაცემები. **პარამეტრი** არის პოპულაციის რიცხვითი მახასიათებელი, ხოლო **სტატისტიკა** არის შერჩევის რიცხვითი მახასიათებელი. რიხვითი მახასიათებლები, საშუალო $\bar{x}$ და სტანდარტული გადახრა s საკმაოდ ხშირად ხმარებადნი არიან პრაქტიკაში. ჩვენ მათ შემდგომშიც გამოვიყენებთ, ოღონდ მონაცემთა შერჩევებისთვის. ანუ ისინი არიან **შერჩევის სტატისტიკები**. ჩვენ განვასხვავებთ შერჩევის სტატისტიკებს და შესაბამის პარემეტრებს პოპულაციებისთვის. პოპულაციის საშუალო არის პოპულაციის ყველა დაკვირვების საშუალო არითმეტიკული, ხოლო პოპულაციის სტანდარტული გადახრა აღწერს მთელი პოპულაციის ცვალებადობას პოპულაციის საშუალოს მიმართ. ზოგადად, ეს სიდიდეები უცნობებია. დასკვნითი სტატისტიკის მეთოდები გვეხმარება მივიღოთ გადაწყვეტილებები ან გავაკეთოთ პროგნოზები პოპულაციის პარამეტრების შესახებ შერჩევის სტატისტიკაზე დაყრდნობით.

შემდგომში, იმისათვის რომ იოლად გავარჩიოთ პოპულაციის პარამეტრები და შერჩევის სტატისტიკები, ჩვენ ვიხმართ განსხვავებულ აღნიშვნებს. ხშირად, პარამეტრების აღსანიშნავად გამოვიყენებთ ბერძნულ ასოებს. მაგალითად, **პოპულაციის საშუალოსთვის** ვიხმართ ასოს $\boldsymbol{\mu}$ („მიუ“), ხოლო პოპულაციის **სტანდარტული გადახრისთვის** ასევე ბერძნულ ასოს $\boldsymbol{\sigma}$ („სიგმა“).

პრაქტიკული ნაწილი.

4.1. მოსალოდნელი სიცოცხლის ხანგრძლივობა. 2006 წელს გაერთიანებული ერების ორგანიზაციის მიერ წარმოდგენილ მოხსენებაში *„ადამიანის განვითარება“,* გამოქვეყნებულია ქვეყნების მიხედვით სიცოცხლის მოსალოდნელი ხანგრძლივობის მონაცემები. დასავლეთ ევროპისთვის მონაცემები ასე განაწილდა: დანია 77, პორტუგალია77, ნიდერლანდები 78, ფინეთი 78, საბერძნეთი 78, ირლანდია 78, გაერთიანებული სამეფო 78, ბელგია 79, საფრანგეთი79, გერმანია 79, ნორვეგია 79, იტალია 80, ესპანეთი 80, შვედეთი 80, შვეიცარია 80.

აფრიკის შემთხვევაში წარმოდგენილი მონაცემები (ბევრ ქვეყანაში ეს მაჩვენებელი შიდსის გავრცელების გამო უფრო დაბალია, ვიდრე იყო მანამდე 5 წლით ადრე) ასეთია: ბოტსვანა 37, ზამბია 37, ზიმბაბვე 37, მალავი 40, ანგოლა 41, ნიგერია 43, რუანდა 44, უგანდა 47, კენია 47, მალი 48, სამხრეთ აფრიკა 49, კონგო 52, მადაგასკარი 55, სენეგალი 56, სუდანი 56, განა 57.

ა) როგორ ფიქრობთ, რომელ ჯგუფს (დასავლეთ ევროპას თუ აფრიკას) აქვს სიცოცხლის ხანგრძლივობის უფრო დიდი სტანდარტული გადახრა? რატომ?

ბ) იპოვეთ სტანდარტული გადახრა ორივე ჯგუფისთვის. შეადარეთ ისინი იმის საილუსტრაციოდ, თუ რომელი ჯგუფისთვის არის s უფრო დიდი და სად გვაქვს უფრო დიდი ცვალებადობა საშუალოსთან შდარებით?

4.2 მოსალოდნელი სიცოცხლის ხანგრძლივობა, მათ შორის რუსეთისაც. გაერთიანებული ერების ორგანიზაციის ცნობით რუსეთისთვის მოსალოდნელი სიცოცხლის ხანგრძლივობა არის 65 წელი, (ამ მაჩვენებელმა ბოლო წლებში დაიწია წინა წლებთან შედარებით. მაშინ იგი შეადგენდა 75 წელს). დავუშვათ, რომ ეს მონაცემი დავუმატოთ წინა სავარჯიშოს დასავლეთ ევროპის მონაცემებს. რას ვარაუდობთ, გაიზრდება თუ შემცირდება სტანდარტული გადახრა იმასთან შედარებით, რაც იყო მხოლოდ დასვლეთ ევროპისთვის ქვეყნებისთვის? რატომ?

4.3. სახლის ფასების ფორმა? ბინათმშენებლობის ნაციონალური ასოციაციის მონაცემებზე დაყრდნობით აშშ-ში 2007 წლის იანვარში გაყიდული ახლი სახლების ფასების მედიანა იყო $ 239 800. შემდეგი მოცემული რიცხვებიდან $ 15 000, $ 1000, $ 60 000 და $ 1 000 000 რომელი მნიშვნელობა მიგაჩნიათ მეტად დამაჯერებლად, რომ იყოს სტანდარტული გადახრა? რატომ? ახსენით რას ხედავთ არარეალურს თითოეულ სხვა მონაცემში?

4.4. სიგარეტზე გადასახადის ფორმა. ბოლო ხანებში ჩატარდა კვლევა და გაკეთდა რეზიუმე, რომელიც შეეხებოდა სიგარეტზე დადებული გადასახადების (ცენტების სიზუსტით) განაწილების ფორმას. იგი ჩატარდა აშშ-ს 50 შტატსა და ვაშინგტონში შედეგად მიღებული იყო, რომ $\bar{x}$ = 73, ხოლო s= 48. როგორ ფიქრობთ, ამ მონაცემებზე დაყრდნობით შეიძლება ითქვას, რომ განაწილება არის ზარის ფორმის? თუ ასეა, რატომ? თუ არა, რატომ? თქვენ რომელ ფორმას ვარაუდობთ?

4.5 ემპირიული წესი და არასათანადო, ძალიან დისკრეტული განაწილება.
მოცემულია მონაცემები ქორწინებათა რაოდენობის შესახებ. დაკვირვებები კაცებზე გვიჩვენებს, რომ საშუალო $\bar{x}$ = 0.16 და s= 0.37.

რამდენჯერ	რაოდენობა	პროცენტული მაჩვენებელი
0	8418	84,0
1	1594	15,9
2	10	0,1
ჯამი	**10022**	**100.0**

ა) იპოვეთ ფაქტობრივი პროცენტული მაჩვენებელი საშუალოდან 1, 2 და 3 სტანდარტული გადახრებისთვის. როგორ შევადაროთ ეს მონაცემები ემპირიული წესით მიღებულ პროგნოზულ პროცენტულ მაჩვენებლებს?

ბ) როგორ ახსნით ა) პუნქტში მიღებულ შედეგებს?

4.6. რამდენი მეგობარი გყავთ? საყოველთაო სოციალურმა კვლევებმა რესპოდენტებს დაუსვა ერთი კითხვა: „რამდენი ახლო მეგობარი გყავთ?“. შერჩევაში იყო 1467 კაცი, საშუალო იყო 7.4 და სტანდარტული გადახრა 11.4. დისტრიბუციას ჰქონდა მედიანა 5 და მოდა 4. (წყარო: მონაცემები აღებულია CSM, UC Berkeley -დან.)

ა) ამ სტატისტიკებზე დაყრდნობით რა განაწილების ფორმას შემოგვთავაზებდით? რატომ?

ბ) გამოსაყენებელია თუ არა ემპირიული წესი ამ მონაცმებში? რატომ? ან რატომ არა?

5

ასოციაცია: შემთხვევითობა, კორელაცია, რეგრესია.

როდესაც ვაანალიზებთ ორი ცვლადის მონაცემებს, პირველი ნაბიჯი არის ის, რომ განვასხვაოთ საპასუხო ცვლადი განმარტებითი ცვლადისგან. მონაცემთა ანალიზი განიხილავს, როგორაა შედეგები დამოკიდებული საპასუხო ცვლადზე ან როგორ აიხსნება ის განმარტებითი ცვლადით. **საპასუხო ცვლადი** არის შედეგის ცვლადი, რომელზეც გაკეთებულია შედარება. როდესაც **განმარტებითი ცვლადი** არის კატეგორიული, მაშინ იგი განსაზღვრავს შესადარებელ ჯგუფებს საპასუხო ცვლადის მნიშვნელობების მიმართ. ხოლო როდესაც განმარტებითი ცვლადი არის რაოდენობრივი ცვლადი, მაშინ ის განსაზღვრავს სხვადასხვა რიცხვითი მნიშვნელობების ცვლილებას საპასუხო ცვლადის მნიშვნელობებთან შედარებით.

ორი ცვლადის მონაცემთა ანალიზის მთავარი მიზანია გამოიკვლიოს, ხომ არ არსებობს რაიმე ასოციაცია (კავშირი) მათ შორის და აღწეროს ამ ასოციაციის ბუნება.

არსებობს **ასოციაცია** ორ ცვლადს შორის, თუ ერთი ცვლადის კონკრეტული მნიშვნელობები მეტად მოსალოდნელია, რომ გამოჩნდნენ მეორე ცვლადის განსაზღვრულ მნიშვნელობებთან ერთად.

ჩვენ განვიხილავთ მეთოდს, რომლის საშუალებითაც დავადგენთ, თუ არსებობს რაიმე ასოციაცია და აღვწერთ რამდენად მკაცრნი არიან ისინი. გამივიკვლევთ ასოციაციებს კატეგორიულ ცვლადებს შორის.

ასოციაცია ორ კატეგორიულ ცვლადს შორის. როგორ უპასუხებდით კითხვაზე: „ყველამ ერთად, შეგიძლიათ თქვათ, რომ ხართ ძალიან ბედნიერნი, საკმარისად ბედნიერნი, არც თუ ისე ბედნიერნი?“ ჩვენ შეგვეძლო გამოგვეთვალა პროცენტული მაჩვენებლები თითოეული ამ სამი შესაძლო პასუხისთვის, რომლებიც არიან კატეგორიული ცვლადები. გამოგვეყენებინა ცხრილები, ჰისტოგრამა, ან წრიული დიაგრამა. თუმცა ჩვენ მაინც გვენდომებოდა გაგვეგო როგორაა ეს პროცენტები დამოკიდებული სხვა ცვლადებზე. მაგალითად, როგორიცაა ადამიანების ოჯახური მდგომარეობა (დაოჯახებული ან დაუოჯახებელი). მათგან ვინც უპასუხა, რომ არის ძალიან ბედნიერი, დაოჯახებულების პროცენტი უფრო მაღალია თუ დაუოჯახებლების?

სცენარი 5.1. პესტიციდები ნატურალურ პროდუქტებში. ჩვენ ვირჩევთ საკვებად ნატურალურ პროდუქტებს, რადგანაც გვჯერა, რომ ისინი არიან პესტიციდების გარეშე და, შესაბამისად, ჯანსაღი პროდუტებია. თუმცა, მცირეა იმ ფართობების რაოდენობა სადაც ნატურალური პროდუქტები მოყავთ (მხოლოდ 2% შეერთებულ შტატებში). როგორც წესი მომხმარებელი ნატურალურ პროდუქტში იხდის გაცილებით მეტს ვიდრე ჩვეულებრივში. ისმის კითხვა: როგორ გავიგოთ, რა არის პროცენტული რაოდენობა იმ (ჩვეულებრივი) პროდუქტებისა, სადაც არის ნარჩენი პესტიციდები, ნატურალურ პროდუქტებთან შედარებით. კალიფორნიის შტატში ამერიკის სოფლის მეურნეობის დეპარტამენტთან (USDA) ერთად მომხარებელთა საბჭომ შერჩევის საფუძველზე ჩაატარა კვლევა. შერჩევა იყო ჩვეულებრივი, რიგითი მონიტორინგის ნაწილი პესტიციდების ნარჩენენებზე. ამ შესწავლის საფუძველზე ქვემოთ მოყვანილი ცხრილი გვიჩვენებს პროდუქტების სიხშირეების ყველა შესაძლო შემთხვევებს კატეგორიების მიხედვით. გვაქვს ორი ცვლადი: საკვების სახეობა და პესტიციდების მდგომარეობა.

	პესტიციდების მდგომარეობა		
საკვების სახეობა	**აღმოჩნდა**	**არ აღმოჩნდა**	**ჯამი**
ნატურალური	29	98	127
ჩვეულებრივი	19,485	7,086	26,571
ჯამი	**19,514**	**7,184**	**26,698**

ცხრილი 5.2. ჯამური სტრიქონი და ჯამური სვეტი არის კატეგორიის სიხშირეები თითოეული ცვლადისთვის. ცხრილის შიგნით მონაცემები იძლევა ინფორმაციას ასოციაციაზე.

ა) რომელი არის საპასუხო ცვლადი და რომელი განმარტებითი?

ბ) როგორ შევადაროთ პესტიციდებიანი პროდუქტების პროპორციები ამ ორი ტიპის პროდუქტისთვის?

გ) მთელი შერჩევის რა ნაწილი შეიცავს პესტიციდების ნარჩენებს?

<u>მსჯელობა</u>: ა) პესტიციციდების სტატუსი, რადგანაც მათი ნარჩენები არსებობს, განსაზღვრავს ინტერესს შედეგის მიმართ. აქედან გამომდინარე, პესტიციდების

სტატუსი არის საპასუხო ცვლადი, ხოლო სურსათის ტიპი არის განმარტებითი ცვლადი.

ბ) ცხრილიდან ჩანს, რომ 127 ნატუალური პროდუქტიდან 29 შეიცავს პესტიციდების ნარჩენებს, ამიტომ პროპორცია პესტიციდებისთვის არის 29/127 = 0.228.მაშინ, როცა 19 485 ჩვეულებრივი პროდუქტი 26 571 -დან შეიცავს პესტიციდებს. მისთვის პროპორცია ასეთია 19 485/26 571 = 0.733, რომელიც გაცილებით მაღალია, ვიდრე ნატურალურ შემთხვევაში. შეფარდებიდან 0.733/0.228 = 3.2 შეიძლება დავასკვნათ, რომ პესტიციდების რაოდენობა ჩვეულებრივ პროდუქტში დაახლოებით სამჯერ მეტია ვიდრე ნატურალურში.

გ) შერჩევის პროდუქტების საერთო პროპორცია პესტიციდების ნარჩენებისთვის ტოლია პესტიციდებიანი პროდუქტების მთელი რაოდენობა შეფარდებული პროდუქტების მთელ რაოდენობასთან, ანუ (29 + 19 485) / (127 + 26 571) = 19 514 / 26 698 = 0.731. შეიძლებოდა ეს შედეგი მიგვეღო სვეტების ჯამებისგანაც.

<u>გააზრებისთვის:</u> საერთო პროპორციის მნიშვნელობა 0.731 ძალიან ახლოსაა პროპორციასთან 0.733, რომელიც მივიღეთ ჩვეულებრივი პროდუქტებისთვის. ამის მიზეზი არის ის, რომ ჩვეულებრივ პროდუქტებს დიდი წილი აქვთ შერჩევაში. (ჯამური სვეტის მიხედვით შესაბამისი პროპორცია მხოლოდ ჩვეულებრივი პროდუქტისთვის არის 26 571 / 26 698 = 0.995), ე.ი. პესტიციდების ნარჩენები შერჩევის საგანთა 73% -ში. ამასთან ეს მაჩვენებელი სამჯერ უფრო მაღალია ჩვეულებრივ პროდუქტებში, ვიდრე ნატურალურში.

შემთხვევითობის ცხრილი. წინა ცხრილში ჩვენ გვქონდა ორი კატეგორიული ცვლადი სურსათის ტიპი და პესტიციდების სტატუსი. ჩვენ შეგვიძლია, ჩავატაროთ ანალიზი კატეგორიული ცვლადებისთვის ცალ-ცალკე. ჯამური სვეტების მიხედვით პესტიციდების სტატუსისთვის და ჯამური სტრიქონების მიხედვით სურსათის ტიპებისთვის. ეს ჯამურები არის დათვლის კატეგორია განცალკევებული ცვლადებისთვის, მაგალითად, (19 514, 7 184) კატეგორიებისთვის (არის, არ არის) პესტიციდების სტატუსისთვის. ჩვენ აგრეთვე შეგვიძლია შევისწავლოთ ასოსიაცია მათშორის, როგორც ეს ახლახანს გავაკეთეთ კატეგორიების კომბინაციების დათვლითა და პროპორციების გამოთვლით. ზემოთ მოყვანილი ცხრილი წარმოადგენს *შემთხვევითობის ცხრილის* მაგალითს.

შემთხვევითობის ცხრილი არის ორი კატეგორიული ცვლადის გამოსახვა. მისი სტრიქონები არის კატეგორიების სია პირველი ცვლადისთვის და სვეტები არის

კატეგორიების სია მეორე ცვლადისთვის. თითოეული ჩანაწერი ცხრილში არის დაკვირვებათა რაოდენობა შერჩევაში, კატეგორიის კონკრეტული კომბინაციისთვის. ეს ხდება ორი კატეგორიული ცვლადისთვის.

თითოეული სვეტისა და სტრიქონის კომბინაციას შემთხვევითობის ცხრილში ეწოდება **უჯრა**. მაგალითად, პირველი უჯრა მეორე სტრიქონში არის სიხშირე 19 485 ანუ დაკვირვებათა რაოდენობა ჩვეულებრივი პროდუქტის კატეგორიაზე პესტიციდების არსებობაზე. მონაცემთა აღების პროცესს და სიხშირეთა მიძებნას შემთხვევათა ცხრილის თითოეული უჯრისთვის ეწოდება მონაცემთა შეუღლება ან **კროს-ტაბულაცია**.

გაფანტულობის გრაფიკი. ორი რაოდენობრივი ცვლადის შემთხვევაში ხშირად საჭიროა ორივე ცვლადის ერთდროული გრაფიკული გამოსახვა. მაგალითად, საპასუხო ცვლადი აღვნიშნოთ y-ით, ხოლო განმარტებითი ცვლად x-ით. ამ აღნიშვნებს გამოვიყენებთ იმისთვის, რომ მონაცემები გამოვსახოთ გრაფიკულად ასოციაციის შესასწავლად. ასეთი დამოკიდებულების მონიშვნას ჰქვია გაფანტულობის წერტილოვანი გრაფიკი. x და y სიდიდეები გამოსახულია წერტილებით ორი ღერძის მიმართ. დაკვირვება n საგანზე არის n წერტილი გაფანტულობის გრაფიკზე.

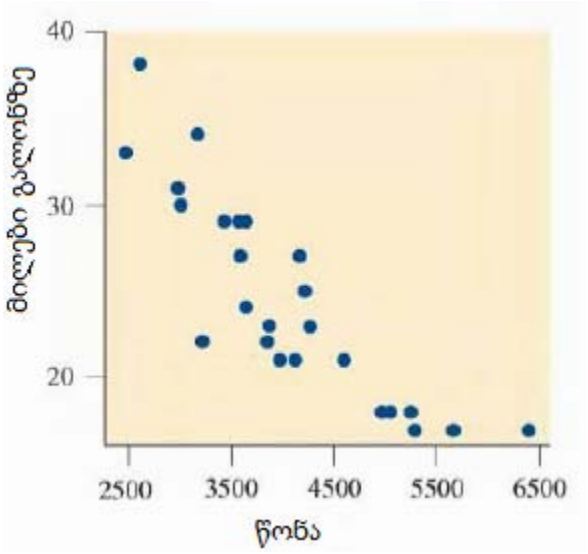
ეს გრაფიკი გამოვიყენოთ ასოციაციის შესასწავლად. როგორ იცვლებიან საპასუხო ცვლადების მნიშვნელობები განმარტებით ცვლადების მნიშვნელობებზე დამოკიდებულებით? კერძოდ, რა ტენდენციაა დიაგრამაზე? დადებითია ასოციაცია თუ უარყოფითი?

ვიტყვით, რომ ორი რაოდენობრივს ცვლადს გააჩნია დადებითი ასოციაცია, თუ x ცვლადის მაღალი მნიშვნელობები გამოჩნდებიან y-ის მაღალ მნიშვნელობებთან ერთად და x -ის მცირე მნიშვნელობების დროს გვაქვს y-ის მცირე მნიშვნელობები. **დადებითი ასოციაცია:** იზრდება x ცვლადი, იზრდება y ცვლადიც.

ანალოგიურად, ორ რაოდენორივ ცვლადს გააჩნია უარყოფითი ასოციაცია, როდესაც ერთი ცვლადის მაღალ მნიშვნელობას შეესაბამება მეორე ცვლადის დაბალი მნიშვნელობა და პირიქითაც, ერთის დაბალ მნიშვნელობებთან გვხვდებიან მეორის მაღალი მნიშვნელობები. **უარყოფითი ასოციაცია:** თუ x მიდის ზევით (ანუ იზრდება), მაშინ y მოდის ქვევით (ანუ მცირდება).

თუ გვინდა, რომ შევიწავლოთ ასოციაცია მანქანის x = წონასა და y = საწვავის რაოდენობას შორის (მილების რაოდენობა თითოეულ გალონზე), ცხადია ველო-

დებით, რომ ასოციაცია უნდა იყოს უარყოფითი: ვინადან რაც უფრო მძიმეა ავტომობილი, მით ნაკლებ მანძილს გადის ერთი გალონით.



ეს ნახაზი აღწერს ამ დამოკიდებულებას. (მონაცემები აღებულია მონაცემთა ფაილებიდან).

რამდენიმე კითხვა იმისთვის, რომ უკეთ დავუკვირდეთ გაფანტულობის დიაგრამას (გრაფიკს).

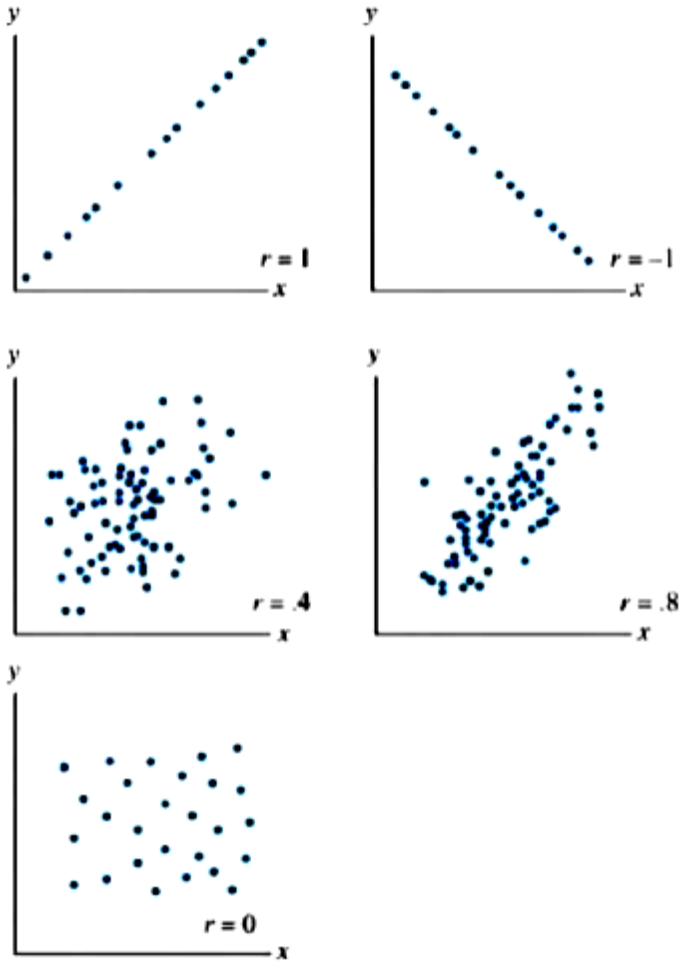
- თქვენ აქ ხედავთ დადებით ასოციაციას, უარყოფით ასოციაციას თუ საერთოდ არ არის ასოციაცია.
- შეიძლება თუ არა, რომ მოვახდინოთ მოცემული წერტილების დამოკიდებულების საკმარისად კარგი აპროქსიმაცია წრფის საშუალებით? ასეთ შემთხვევაში მონაცემთა წერტილები უნდა მოხვდნენ წრფესთან ახლოს ან იყვნენ გადახრილები მისგან საკმაოდ მცირედ.
- არის ზოგიერთი დაკვირვება არაბუნებრივი? რას გვეუბნებიან არაჩვეულებრივი წერტილები?

ასოციაციის სიმტკიცე: კორელაცია. თუ დაკვირვების წერტილები რაიმე წრფის უშუალო მახლობლობაშია, მაშინ ამბობენ, რომ ცვლადებს აქვთ მიახლოებით წრფივი დამოკიდებულება. ზოგიერთ შემთხვევაში ხდება, რომ მონაცემთა წერტილები ვარდებიან წრფის მახლობლობაში, მაგრამ მცირე გადახრებით წრივი დამოკიდებულებისგან. ასეთი დამოკიდებულების დასკვნით მაჩვენებელს უწოდებენ კორელაციას და იგი მიუთითებს წრფივი ასოციასიის სიმტკიცეზე.

კორელაცია განსაზღვრავს ორ რაოდენობრივ ცვლადს შორის ასოციაციის მიმართულებას და მათი წრფივი დამოკიდებულებისადმი სიმტკიცეს. იგი აღინიშნება r ასოთი და იღებს მნიშვნელობებს -1 -სა და 1 -ს შორის.

- r-ის დადებითი მნიშვნელობა მიუთითებს დადებით ასოციაციაზე, ხოლო უარყოფითი მნიშვნელობა - შესაბამისად უარყოფითზე.
- რაც უფრო ახლოსაა r ერთთან ან მინუს ერთთან, მით ახლოს არიან მონაცემთა წერტილები წრფესთან და მით უფრო ძლიერია ასოციაცია. რაც უფრო ახლოსაა ნულთან მით უფრო სუსტია ასოციაცია.

მოდით, ვნახოთ როგორ გამოიყურება კორელაცია, როდესაც მოცემული გვაქვს გაფანტულობის გრაფიკი.



ამ გრაფიკებზე მოცემულია გაფანტულობები და მათი კორელაციები. როცა კორელაცია ახლოსაა ± 1 -თან, მაშინ, როგორც ვთქვით, მონაცემები დალაგებულია თითქმის წრფეზე.

კითხვა: რატომ ითვლება, რომ თუ მონაცემთა წერტილები დალაგებულია წრფესთან ძალიან ახლოს, მაშინ იგი უნდა განვიხილოთ როგორც ძლიერი ასოციაცია?

კორელაცია r ღებულობს ექსტრემალურ +1 ან -1 მნიშვნელობებს მხოლოდ მაშინ, როცა წერტილები *სრულყოფილად* მიუყვებიან *ფორმას* (ნიმუშს), როგორც ეს გამოსახულია პიველ ორ ნახაზზე. როცა r=+1, წრფის დახრილობა არის დადებითი. ასეთ დროს ასოციაციაც არის დადებთი, რადგანაც x-ის მეტ მნიშვნელობებს შეესაბამება y-ის მეტი მნიშვნელობები. თუ r=-1, მაშინ წრფის დახრილობა უარყოფითია და შესაბამისად გვაქვს უარყოფითი ასოციაცია.

პრაქტიკაში არ უნდა ველოდოთ, რომ წერტილები განლაგდებიან ზუსტად წრფეზე. თუმცა რაც უფრო ახლოს არიან იდეალურ წრფესთან, მით უფრო ახლოსაა კორელაცია +1 -თან ან -1 -თან. მაგალითად, წერტილოვანი გრაფიკი ნახაზზე, სადაც კორელაცია r = 0.8 მიუთითებს უფრო ძლიერ ასოციაციაზე, ვიდრე იმ ნახაზზე, სადაც r = 0.4, რომლის წერტილები ვარდებიან უფრო მოშორებით წრფიდან.

კორელაციის თვისებები:

- **კორელაცია r ყოველთვის *არის -1 -სა და 1 -ს შორის.*** რაც უფრო მეტადაა მათთან ახლოს, მით უფრო ძლიერია **წრფივი ასოციაცია** და შესაბამისად მონაცემთა წერტილები არიან თითქმის წრფეზე.
- **დადებითი კორელაცია** მიგვითითებს **დადებით ასოციაციაზე,** ხოლო **უარყოფითი - უარყოფითზე.**
- კორელაციის სიდიდე **არ არის დამოკიდებული ცვლადების სიდიდეებზე.** მაგალითად, ვიგულისხმოთ, რომ ერთ-ერთი ცვლადია პიროვნების შემოსავალი დოლარებში. თუ ჩვენ შევცვლით დოლარს ევროთი ან შემოსავალს გამოვსახავთ ათასებში, შედეგად მივიღებთ იგივე კორელაციას.
- **ორ ცვლადს აქვს იგივე კორელაცია, იმისგან დამოუკიდებლად რომელს ჩავთვლით საპასუხო ცვლადად და რომელს - განმარტებით ცვლადად.**

კორელაციის გამოსათვლელი ფორმულა. ჯერ განვმარტოთ ე.წ. Z-შეფასება (ანუ Z-ქულა). იგი გვიჩვენებს სტანდარტული გადახრის რაოდენობას საშუალოდან. დადებითი Z-შეფასება (ქულა) ნიშნავს, რომ დაკვირვება მეტია საშუალოზე, ხოლო უარყოფითობა - ნაკლებია საშუალოზე. იგი ასე გამოითვლება:

Z- შეფასება = (გადახრა - საშუალო)/(სტანდარტული გადახრა)

მიუხედავად იმისა, რომ შესაძლებელია კომპიუტერზე უშუალოდ გამოვთვალოთ კორელაცია, საკითხის უკეთ გაგებისთვის ძალიან კარგი იქნება, თუ შევხედავთ და კარგად დავაკვირდებით მის გამოსათვლელ ფორმულას. x-ით აღვნიშნოთ განმარტებითი ცვლადი. ხოლო Z_x იყოს Z-შეფასება, რომელიც აღნიშნავს სტანდარტული გადახრების რაოდენობას და მიმართულებას, რომ x -ები ვარდებიან x -ის საშუალოსგან მოშორებით. ეს არის

$$Z_x = \frac{\textit{დაკირვებული სიდიდე– საშუალო}}{\textit{სტანდარტული გადახრა}} = \frac{(x-\bar{x})}{s_x},$$

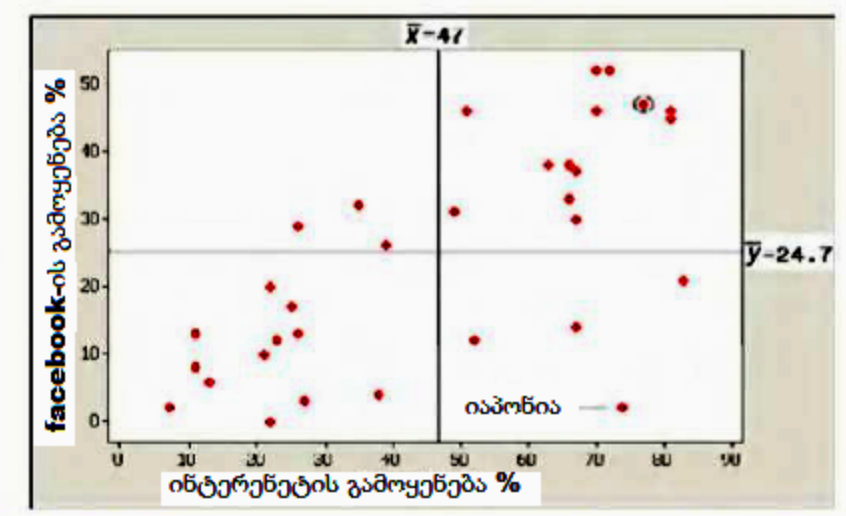
სადაც s_x აღნიშნავს x სიდიდის სტანდარტულ გადახრას. ანალოგიურად იქნება საპასუხო y ცვლადისთვისაც Z_y. r -ის საპოვნელად ვიპოვოთ $Z_x Z_y$ ნამრავლი ყოველი დაკვირვებისთვის, მერე გამოვთვალოთ მათი მახასიათებელი სიდიდე, მათი ნამრავლების საშუალო.

r -კორელაციის გამოთვლა

$$\boldsymbol{r} = \frac{1}{\boldsymbol{n}-1}\sum \boldsymbol{z}_x \boldsymbol{z}_y = \frac{1}{\boldsymbol{n}-1}\sum \left(\frac{\boldsymbol{x}-\bar{\boldsymbol{x}}}{\boldsymbol{s}_y}\right)\left(\frac{\boldsymbol{y}-\bar{\boldsymbol{y}}}{\boldsymbol{s}_y}\right)$$

სადაც n არის წერტილების რაოდენობა, $\bar{x}$ და $\bar{y}$ არიან საშუალოები და s_x და s_y სტანდარტული გადახრებია x -ის და y -ის. ჯამი აიღება ყველა n დაკვირვების მიმართ.

საილუსტრაციოდ მოვიყვანოთ კვლევა, რომელიც ჩატარდა 33 ქვეყანაში ინტერნეტისა და Facebook-ის გამოყენების შესახებ. მონაცემები მოცემულია გაფანტულობის გრაფიკის საშუალებით.



ვერტიკალური ხაზი გავლებულია x -ის საშუალოზე, ჰორიზონტალური კი y -ის საშუალოზე. ისინი არეს ყოფენ ოთხ კვადრანტად. სტატისტიკები ასეთია:

$\bar{x}$ = 47.0 $\bar{y}$ = 24.7 s_x = 24.4 s_y = 16.5

იაპონიის შესაბამის წერტილს (x = 74 , y = 2) აქვს Z -შეფასებები, Z_x = 1.11 და Z_y = - 1.3 . ეს წერტილი მონიშნულია ნახაზზე, რომელიც მოხვდა ქვედა მარჯვენა კვადრანტში, რაც იაპონიას აქცევს ოდნავ არატიპურ ქვეყნად, რადგან იგი მიეკუთვნება იმ 7 ქვეყნის ჯგუფს 33 -დან, რომელთა შესაბამისი წერტილები მოხვდა ქვედა მარჯვენა ან ზედა მარცხენა კვადრანტში, დანარჩენი 26 ქვეყნის შესაბმისი წერტილები კი განლაგებულია ზედა მარჯვენა და ქვედა მარცხენა კვადრანტებში.

კითხვა: ამ კვადრანტების წერტილებმა კორელაციის მნიშვნელობის გამოთვლაში რა წვლილი შეიტანა - დადებითი თუ უარყოფითი? (*მითითება:* Z-შეფასების ნამრავლები ამ წერტილებისთვის დადებითია თუ უარყოფითი?).

რეზიუმე: Z-შეფასებების ნამრავლი და კორელაცია.

- მარჯვენა ზედა კვადრანტში Z-შეფასებების ნამრავლი დადებითია ყოველი წერტილისთვის. ყოველი წერტილისთვის ნამრავლი აგრეთვე დადებითია ქვედა მარცხენა კვადრანტშიც. ასეთი წერტილები განაპირობებენ დადებით კორელაციას.
- მარცხენა ზედა და მარჯვენა ქვედა კვადრანტებში Z-შეფასებების ნამრავლი უარყოფითია ყოველი წერტილისთვის. ასეთი წერტილები განაპირობებენ უარყოფით კორელაციას.

ე. ი. თუ ყველა წერტილი ვარდება ზედა მარჯვენა ან ქვედა მარცხენა კვადრანტში, კორელაცია უნდა იყოს დადებითი.

შედეგის პროგნოზირება. ჩვენ ვნახეთ, როგორ უნდა გამოვიკვლიოთ დამოკიდებულება ორ რაოდენობრივ ცვლადს შორის გაფანტულობის გრაფიკის საშუალებით. როდესაც დამოკიდებულება წრფივია, მაშინ კორელაცია აღწერს მას რიცხობრივად. შეგვიძლია, ჯერ ჩავატაროთ ანალიზი მონაცემებზე და შემდეგ დავწეროთ იმ წრფის განტოლება, რომელიც ყველაზე უკეთ აღწერს სქემას (ნიმუშს). ეს წრფე პროგნოზირებს საპასუხო ცვლადის შედეგებს, განმარტებითი (დამოუკიდებელი) ცვლადის საშუალებით. როგორც გვახსოვს, კორელაცია არ მოითხოვს მხოლოდ ერთ ცვლადს, იგი ერთმანეთთან აკავშირებს საპასუხო და განმარტებით ცვლადებს.

რეგრესიის წრფე: პროგნოზის განტოლება საპასუხო ცვლადის შედეგზე.

რეგრესიის წრფე პროგნოზირებს y საპასუხო ცვლადის მნიშვნელობას, როგორც განმარტებით x ცვლადზე დამოკიდებულ წრფივ ფუნქციას. $\hat{y}$-ით აღვნიშნოთ y-ის **პროგნოზირებადი მნიშვნელობა**. რეგრესიის წრფის განტოლებას აქვს შემდეგი სახე

$$\hat{y}=a + bx$$

სადაც a აღნიშნავს წრფის თანაკვეთას y -ღერძთან, ხოლო b არის დახრილობა.

რეგრესიის განტოლების პოვნა.

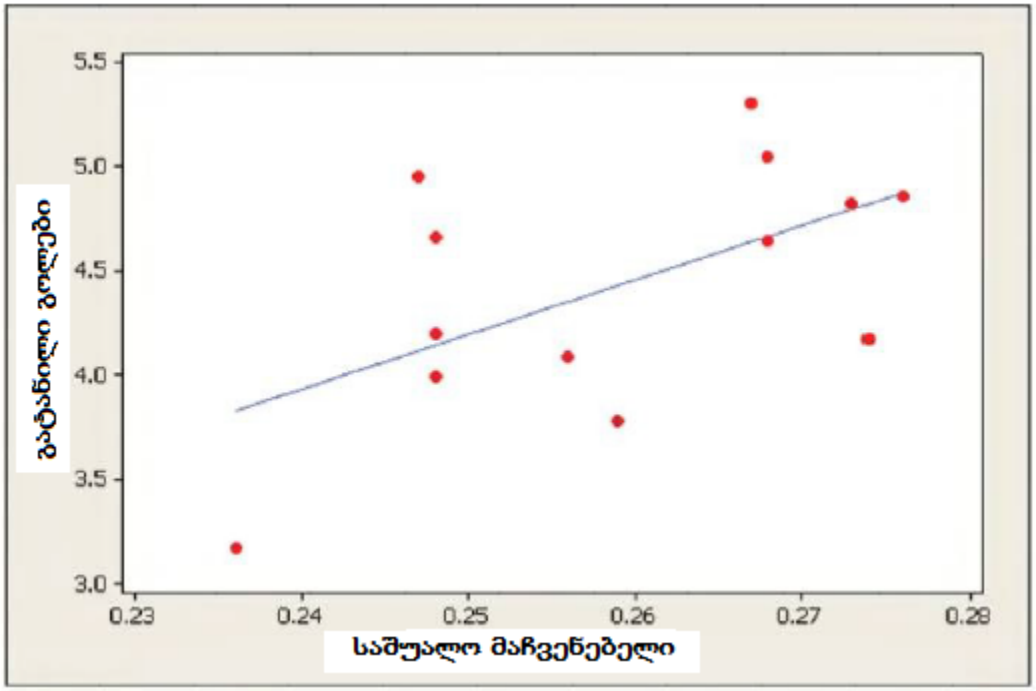
როგორ ვისარგებლოთ მონაცემებით იმისთვის, რომ ვიპოვოთ რეგრესიის განტოლება? პირველ რიგში უნდა ავაგოთ გაფანტულობის გრაფიკი, რათა დავრწმუნდეთ, რომ დამოკიდებულებას აქვს დაახლოებით წრფის ტენდენდენცია. თუ ეს მართლაც ასეა, პროგრამულ უზრუნველყოფას და კალკულატორებს ადვილად შეუძლიათ იპოვონ იმ წრფის განტოლება, რომელიც ყველაზე უკეთ მოერგება მონაცემებს.

სცენარი 5.6. ბეისბოლის ანგარიშის პროგნოზირება.

ბეისბოლში არსებობს გუნდის შეფასების ორი ტიპის რეზიუმე, პროპორციულად რამდენ ხანს ფლობენ ინიციატივას და გუნდის შედეგიანობა (ანუ საშუალოდ გატანილი გოლების რაოდენობა ერთ მატჩში). ქვემოთ მოყვანილი ცხრილი გვიჩვენებს ამერიკის ლიგის გუნდების 2010 წლის სტატისტიკას. გატანილი გოლები არის საპასუხო ცვლადი y, ხოლო განმარტებითი (დამოუკიდებელი) ცვლადი x არის გუნდის საშუალო დონე.

გუნდი	საშუალო მაჩვენებელი	გატანილი გოლები
NY Yankees	0.267	5.30
Boston	0.268	5.05
Tampa Bay	0.247	4.95
Texas	0.276	4.86
Minnesota	0.273	4.82
Toronto	0.248	4.66
Chicago Sox	0.268	4.64
Detroit	0.268	4.64
LA Angels	0.248	4.20
Kansas City	0.234	4.17
Oakland	0.256	3.09
Cleveland	0.248	3.99
Baltimore	0.259	3.78
Seattle	0.236	3.17

არსებობს დადებითი კორელაციის ტენდენცია, r = 0.568 .



კითხვა: როგორ შეაფასებთ პროგნოზის ცდომილებას, რომელიც წარმოიშობა იმის გამო, რომ ვიღებთ რეგრესიის განტოლებას გატანილი გოლების პროგნოზირებისთვის.

კითხვები შესწავლისთვის:

ა) პროგრამული უზრუნველყოფის მიხედვით რა ითვლება რეგრესიის განტოლებად?

ბ) თუ გუნდს აქვს საშუალო დონე 0.275, შემდგომ წელს როგორია პროგნოზი მატჩში საშუალო შედეგიანობის შესახებ.

გ) როგორი დახრა ექნება ამ წრფეს ?

გააზრებისთვის:

ა) როდესაც MIMITAB–ის დახმარებით სტატისტიკის მენიუში, რეგრესიის ნაწილში ჩვენ ავარჩევთ რეგრესიის განტოლებას, პასუხი იქნება ასეთი: რეგრესიის განტოლებაა:

გატანილი გოლები = - 2.32 + 26.1საშუალო დონე

Y -ღერძთან თანაკვეთა არის a = -2.32, ხოლო დახრილობა განმარტებით ცვლადთან განისაზღვრება BAT_A VG მონაცემთა ფაილით. იგი ტოლია b = 26.1 და საბოლოოდ რეგრესიის განტოლებას აქვს ასეთი სახე:

$$\hat{y}=a + bx = - 2.3 + 26.1x.$$

ბ) ჩვენი პროგნოზით ამერიკის ლიგის გუნდი საშუალო 0.275 მაჩვენებლით, საშუალოდ ერთ თამაშში გაიტანს $\hat{y}$ = - 2.3 + 26.1(0.275) = 4.86 გოლს.

გ) რადგან დახრილობა b არის დადებითი, ამიტომ ასოციაციაც არის დადებითი. გუნდის პროგნოზირებადი ქულები იზრდება, თუ იზრდება საშუალო მაჩვენებელი. რადგანაც x = საშუალო დონეს, რომელიც არის პროპორცია. გრაფიკზე გუნდი საშუალო მონაცემებით ვარდება შუალედში 0.23-სა და 0.28-ს შორის ანუ დიაპაზონია 0.05. x-ის ზრდა 0.5-ით იწვევს პროგნოზირებადი გატანილი გოლების ზრდას (0.05)26.1=1.3-ით. ცხადია, ძლიერი გუნდებისთვის ეს მაჩვენებლი მაღალია, ვიდრე სუსტებისთვის.

პრაქტიკული ნაწილი.

5.1. რომელია საპასუხო /განმარტებითი ცვლადი? ცვლადების შემდეგი წყვილისთვის რომელია მეტად ბუნებრივად საპასუხო და განმარტებითი ცვლადი?

ა) შვილების რაოდენობა და დედების რელიგიურობა.

ბ) ბედნიერება (არცთუ მაინც და მაინც ბედნიერი, საკმაოდ ბედნიერი, ძალიან ბედნიერი) და დაქორწინებული ხართ? (დიახ, არა).

5.2. ნახმარი ავტომობილები და ასოციაციის მიმართულება. 100 ავტომობილიდან, თუ მათ განვიხილავთ ექსპლოატაციაში ყოფნის ვადების მიხედვით, რას უფრო ელოდებით, დადებით ასოციაციას, უარყოფით ასოციაციას თუ არანაირ ასოციაციას შემდეგ ცვლადებს შორის? ახსენით რატომ.

ა) ავტომობილის ასაკი და ოდომეტრზე გამოსახული მილების რაოდენობა.

ბ) ავტომობილის ასაკი და გასაყიდი ფასი.

გ) ავტომობილის ასაკი და მის რემონტებზე გაწეული სრული დანახარჯი.

დ) ავტომობილის წონა და მილების რაოდენობა, რომლის გავლაც მას შეუძლია ერთი გალონი საწვავით.

5.3. შუალედურ-ფინალური გამოცდების კორელაცია. მინესოტას ერთ-ერთი კოლეჯის სტუდენტებისთვის, რომელთაც ჩასაბარებლებლად აირჩიეს სტატისტიკა, შუალედური და ფინალური გამოცდების ქულების საშუალო ტოლია 70-ის, ხოლო სტანდარტული გადახრა ტოლია 10-ის. პროფესორი იკვლევს ფინალური გამოცდის პროგნოზს შუალედური გამოცდის შედეგების (ქულების) მიხედვით. რეგრესიის განტოლებას, რომელიც ფინალური გამოცდის y ქულებს აკავშირებს შუალედური გამოცდის x ქულებთან, აქვს შემდეგი სახე: $\hat{y}= 30+0.60x$.

ა) იპოვეთ პროგნოზირებული ფინალური საგამოცდო ქულა იმ სტუდენტისთვის, ვისთვისაც

(**i**) შუალედურის ქულა 100,

(**ii**) შუალედურის ქულაა 50.

ყურადღება მიაქციეთ, რომ თითოეულ შემთხვევაში პროგნოზიერბული საფინალო გამოცდის ქულა რეგრესირებს საშუალოსთან ანუ 75-თან. (ეს არის რეგრესიის განტოლების თვისება).

ბ) აჩვენეთ, რომ კორელაცია ტოლია 0.60-ის და გამოსახეთ ის. (მითითება: გამოიყენეთ დამოკიდებულება დახრილობასა და კორელაციას შორის.)

5.4. ფინალური გამოცდის შედეგის პროგნოზირება შუალედური გამოცდების შედეგების საშუალებით. შუალედური გამოცდის საშუალო აღვნიშნოთ x -ით, ხოლო ფინალური გამოცდის საშუალო კი y -ით. ორივეს საშუალოა 80 და სტანდარტული გადახრა 10. კორელაცია საგამოცდო ქულებს შორის არის 0.70.

ა) იპოვეთ რეგრესიის განტოლება.

ბ) გააკეთეთ საფინალო გამოცდის შედეგის პროგნოზირება სტუდენტისთვის, ვისაც აქვს შუალედური შეფასება 80.

6

მონაცემთა შეგროვება.

ექსპერიმენტი და დაკვირვება. კარგი და ცუდი ამორჩევა. კარგი და ცუდი ექსპერიმენტები.

სცნარი 6.1. ფიჭური ტელეფონები და თქვენი ჯანმრთელობა.

რამდენად უსაფრთხოა ფიჭური ტელეფონების ხმარება? ფიჭურ ტელეფონებს გააჩნიათ ელექრომაგნიტური გამოსხივება, და მისი ანტენა არის ამ ენერგიის ძირითადი წყარო. რაც უფრო ახლოსაა ანტენა მომხმარებლის თავთან, მით უფრო მეტია გამოსხივების ზემოქმედება. ფიჭური ტელეფონებისადმი გაზრდილ მოთხოვნასთან ერთად გაიზარდა შეშფოთება ჯანმრთელობის პოტენციურ რისკებთან დაკავშირებით. რამდენიმე გამოკვლევამ შეისწავლა ამ რისკების შესაძლებლობა.

კვლევა 1. გერმანული კვლევამ (*Stang et al., 2001)* შეადარა 118 პაციენტი თვალის კიბოს იშვიათი დაავადებით, რომელსაც ჰქვია უვეალური მელანომა, 475 ჯანმრთელ პაციენტს, რომელთაც არა აქვთ თვალის კიბოს აღნიშნული დაავადება. პაციენტების მიერ მობილური ტელეფონებით სარგებლობა გაზომეს კითხვარების საშუალებით. აღმოჩნდა, რომ თვალის ონკოლოგიური დაავადების მქონე პაციენტები უფრო ხშირად სარგებლობენ მობილური ტელეფინებით, ვიდრე სხვები.

კვლევა 2. ბრიტანულმა კვლევამ *(Hepworth et al., 2006*) შეადარა ტვინის კიბოთი დაავადებული 966 პაციენტი 1716 პაციენტს, რომელთაც არა აქვთ ეს დაავადება. პაციენტების მიერ მობილური ტელეფონებით სარგებლობა გაზომეს კითხვარების საშუალებით. მობილური ტელეფონებით სარგებლობის რაოდენობა ორივე ჯგუფში იყო თითქმის ერთი და იგივე.

კვლევა 3. კვლევა, რომელიც გამოქვეყნებულია ჟურნალში „*American Medical Association"* (Volkow et. al., 2011) ასაბუთებს, რომ მობილური ტელეფონების გამოყენება აჩქარებს ტვინის აქტიურობას. შემთხვევითი ჯვარედინი კვლევის ჩარჩოებში 47 მონაწილეს მოარგეს ფიჭური ტელეფონები თითოეულ ყურზე და შემდეგ გაიარეს პოზიტრონულ ემისიური ტომოგრაფია, რომელიც გაზომვისთვის სკანირებას უკეთებს ტვინის გარკვეულ ტიპის აქტივობებს. ერთი ციკლის

განმავლობაში ორივე ფიჭური ტელეფონი გამორთეს. შემდეგი სკანირების დროს 50 წუთიანი სიჩუმის შემდეგ შევიდა ზარი მარჯვენა ყურში. ზარების მიღების მიმდევრობა (პირველი და მეორე სკანირების დროს) შემთხვევითია. სკანირებების შედარებამ გვიჩვენა ტვინის იმ ნაწილის გააქტიურების მნიშვნელოვანი ზრდა, რომელიც უფრო ახლოს იყო ანტენასთან ზარის გაშვების დროს.

პირველ და მესამე კვლევაში ჩანს ფსიქილოგიური რეაგირება მობილური ტელეფონის ზემოქმედებაზე, მეორე კვლევაში კი არ შეინიშნება.

კითხვები კვლევისთვის

- რატომ არის ხვადასხვა სამედიცინო კვლევების შედეგები ზოგჯერ განსხვავებული?
- რომელია საუკეთესო კვლევა მონაცემთა შეგროვებისთვის, გამოვიკვლიოთ თუ როგორ ასოცირდება ფიჭური ტელეფონის ზემოქმედება ადამიანის ორგნიზმის სხვადასხვა ფიზიოლოგიურ აქტიურობებთან? ან იქნებ, შეუძლია ასეთ კვლევას დაადგინოს უშუალო მიზეზ-შედეგობრივი კავშირი ფიჭური ტელეფონების გამოყენებასა და ჯანმრთელობის რისკებს შორის? და საერთოდ, რამდენად ეთიკურია ასეთი კვლევების ჩატარება?

გააზრებისთვის

მონაცემთა შეგროვებისთვის სხვადასხვა კვლევების ცოდნა გვეხმარება გავიგოთ, მეცნიერულ კვლევებზე დაყრდნობით რამდენად წინააღმდეგობრივი შედეგები შეიძლება დადგეს და განვსაზღვროთ, რომელი კვლევები იმსახურებენ ყველაზე მეტად ჩვენს სანდოობას. ამ თავში ვნახავთ, თუ როგორ შეიძლება, რომ კვლევამ დიდი გავლენა იქონიოს შედეგზე, ან კვლევა კაგად იყოს განსაზღვრული, ან შედეგი იყოს უაზრო ან უფრო ცუდი, შევყავდეთ დაბნეულობაში.

ექსპერიმენტული კვლევები და დაკვირვებები. ჩვენ სტატისტიკას ვიყენებთ პოპულაციის შესასწავლად. გავიხსენოთ, რომ პოპულაცია შედგება ყველა იმ საგნებისგან, რომლებიც წარმოადგენენ ჩვენს ინტერსს, ხოლო ამორჩევა (ნიმუში) კი არის პოპულაციის ქვესიმრავლე, რომელზედაც კვლევა აგროვებს მონაცემებს.წინა მაგალითში გერმანულ და ბრიტანულ კვლევებში განიხილავდნენ საგნებს (ობიექტებს), რომლებიც იყენებდნენ ფიჭურ ტელეფონებს. კვლევა წარმოებს იმაზე, აქვთ თუ არა აღნიშნულ საგნებს ანუ ობიექტებს თვალის ან თავის ტვინის კიბო. რადგან ასეთი კვლევების ჩატარება ძალიან ძვირია და საჭიროებს ხანგრძლივ დროს იმისთვის, რომ მთელი გერმანიის და ბრიტანეთის მასშტაბით ვეძებოთ

წარმომადგენლები მონაცემთა შესაგროვებლად, ამიტომ მონაცემთა შესაგროვებლად იყენებენ ამორჩევას. ამ კვლევასაც, ისევე როგორც სხვა ძალიან ბევრს, გააჩნია ორი ცვლადი; პირველადი ინტერესის - საპასუხო ცვლადი და განმარტებითი ცვლადი. საპასუხო ცვლადი ზომავს და აინტერესებს გამოსავალი. კვლევას აინტერესებს საკითხი, როგორაა შედეგი დამოკიდებული განმარტებით ცვლადზე. ზემოთ მოტანილ კვლევებში საპასუხო ცვლადია, აქვს თუ არა მოცემულ საგანს კიბო (თვალის ან ტვინის), ხოლო განმარტებითი ცვლადია მათი რაოდენობა, ვინც იყენებს ფიჭურ (მობილურ) ტელეფონს.

ზოგიერთი კვლევა, როგორიცაა მაგალითად წინა ამოცანაში კვლევა 3, მოითხოვს **ექსპერიმენტის** ჩატარებას. იგი მოითხოვს, რომ მობილურები იყოს გამორთული და ჩაირთოს საჭირო დროს, რადგან გვინდა, რომ შევადაროთ საპასუხო ცვლადის შედეგები. ექსპერიმემენტის ამ პირობებს ეწოდება **პროცედურები**. ისინი შეესაბამებიან განმარტებითი ცვლადის სხვადასხვა მნიშვნელობებს. ზოგჯერ არის პირიქითაც, მრავალი კვლევა მოითხოვს მხოლოდ *დავაკვირდეთ* საპასუხო ცვლადის მნიშვნელობებს და განმარტებითი ცვლადის მნიშვნელობებს მხოლოდ შერჩევის საგნებისთვის ისე, რომ არანაირად არ ვზემოქმედებთ მათზე. ასეთ კვლევას ეწოდება **დაკვირვებადი**, იგი არაექსპერიმენტულია.

ექსპერიმენტის უპირატესობა დაკვირვებასთან შედარებით. დაკვირვებადი კვლევების დროს დაფარულმა ცვლადებმა შეიძლება გავლენა მოახდინონ შედეგზე. **დაფარული ცვლადი** ისეთი ცვლადია, რომელზედაც ვერ ხდება დაკვირვება. აქედან გამომდინარე, იგი გალენას ახდენს ცვლადებს შორის ასოციაციაზე.

მაგალითად, კვლევა 1-ში ნაპოვნია ასოციაცია მობილური ტელეფონები მოხმარებასა და თვალის კიბოს შორის. თუმცაღა შესაძლებელია იყოს მათ შორის განსხვავება ცხოვრების წესის მიხედვით, გენეტიკით, განსხვავებული ჯანმრთელობა იმ საგნებს შორის, რომელთაც აქვთ თვალის კიბო და ვისაც არა აქვს. დაფარული ცვლადი გავლენას მოახდენს ასოციაციაზე. მაგალითად, დაფარული ცვლადი შეიძლება იყოს ცხოვრების წესი, ვთქვათ, კომპიუტერის მოხმარება. შესაძლებელია, ვინც ხშირად იყენებს მობილურ ტელეფონს, ასევე ხშირად ხმარობდეს კომპიუტერსაც. შესაძლებელია, კომპიუტერის ეკრანის ზემოქმედებაც ზრდიდეს თვალის კიბოს რისკს. ამ შემთხვევაში თვალის კიბოს მაღალი გავრცელებადობა შესაძლებელია გამოწვეული იყოს კომპიუტერის ხშირი მოხმარებით და არა ფიჭური ტელეფონის გამო.

ამისგან განსხვავებით, ექსპერიმენტი საგრძნობლად ამცირებს დაფარული ცვლადების გავლენას შედეგზე. რატომ? იმიტომ რომ საგნებს ვირჩევთ შემთხვევით და გვინდა, რომ მოვახდინოთ მათი რაიმენაირი ფორმირება და გამოვიყვანოთ შედეგები საგანთა ჯგუფისთვის, რომელთაც აქვთ ერთნაირი განაწილება, მაგრამ სხვა ცვლადების მიმართ.

მიზეზ-შედეგობრივი კავშირის დადგენა მთავარია მეცნიერებაში. მაგრამ ყოველთვის ვერაა შესაძლებელი დადგენა მიზეზსა და ეფექტურად განსაზღვრული შედეგისა, დაკვირვებადი კვლევების საშუალებით. აქ ყოველთვისაა შანსი, რომ დაფარულმა ცვლადმა გავლენა იქონიოს ასოციაციაზე. თუ იმ ხალხს, ვინც ხშირად ხმარობს მობილურ ტელეფონს აქვთ თვალის კიბოს მაღალი მაჩვენებელი, შესაძლებლია ეს იყოს იმის გამო, რომ რომელიღაც ცვლადმა, რომელიც ჩვენ გამოგვეპარა და ვერ გავზომეთ, იმოქმედა შედეგზე. მაგალითად, როგორიცაა კომპიუტერის გამოყენება. როგორც ადრეც შევნიშნეთ, ასოციაციის არსებობა არ იწვევს მიზეზობრიობის დადგენას.

ამიტომ ადვილია, დაფარული ცვლადის გავლენა შევამციროთ ექსპერიმენტში, ვიდრე ეს ხდებოდა დაკვირვების დროს. ექსპერიმენტის დროს მკვლევარი აკონტროლებს სიტუაციას და საგნებს. ამიტომ, თუ ჩვენ აღმოვაჩენთ ასოციაციას საპასუხო ცვლადსა და განმარტებით ცვლადს შორის, იგი იქნება მეტად სანდო, ვიდრე დაკვირვებითი კვლევის დროს. აქედან გამომდინარე, ექსპერიმენტული მეთოდი გაცილებით უკეთესია ვიდრე დაკვირვების მეთოდი.

კვლევის ტიპის გასაზღვრა.

თუ ექსპერიმეტული მეთოდი უკეთესია, მაშინ რატომღა ვიყენებთ დაკვირვების მეთოდს? იმის გამო, რომ ზოგჯერ ექსპერიმენტი საჭიროებს ძალიან დიდ დროს, ვიყენებთ დაკვირვების მეთოდს. მაგალითად, ფიჭური ტელფონების შემთხვევაში, სიტუაცია შეიძლება იყოს შემდეგი: ავირჩიოთ უნივესიტეტის სტუდენტების ნახევარი და ვუთხრათ მათ, რომ შემდეგი 50 წლის განმავლობაში ყოველდღიურად გამოიყენონ მობილური ტელეფონები, ხოლო მეორე ნახევარს კი - ვუთხრათ, რომ ამ ხნის განმავლობაში მათ არასდროს გამოიყენონ იგი. ჩვენ კი აქედან 50 წლის შემდეგ ვთქვათ, რომ კიბოს მეტი შემთხვევა გვქონდა მათთან, ვინც იყენებდა მობილურ ტელეფონებს.

მოვიყვანოთ თვალსაჩინო მიზეზები, რომლებიც ართულებენ ექპერიმენტული მეთოდის გამოყენებას:

- არაეთიკურია, რომ კვლევის საგანი დიდი ხნის განმავლობაში ვამყოფოთ დაკვირვების ქვეშ. როგორც ეს იყო ზემოთ აღწერილ შემთხვევაში.
- პრაქტიკაში სირთულეა ისიც, რომ დარწმუნებული უნდა ვიყოთ, რომ საგანი ყოველთვის დაგვიჯერებს. როგორ შევამოწმოთ ჩვენ, რომ საგნები 50 საექსპერიმენტო დროის განმავლობაში არ შეიცვლიან დამოკიდებულებას?
- ვის უნდა, რომ 50 წელი ელოდოს პასუხს?

გამოყენებული მონაცემთა გაზა უკვე ხელმისაწვდომია.

ცხადია, რომ თქვენ არ ჩაატარებთ ყოველ ჯერზე ახალ კვლევებს, თუ გვინდა რაიმე კითხვაზე პასუხის გაცემა. ადამიანის ბუნება ისეთია, რომ განმეორებით ჩატარების მაგიერ დაეყრდნოს უკვე არსებულ მონაცემებს. ყველაზე იოლად ხელმისაწვდომია ის მონაცემები, რომელებიც თქვენი კვლევის შედეგადაა მიღებული. ვთქვათ, თქვენს ნაცნობს, რომელიც იყო ფიჭური ტელეფონის ხშირი მომხმარებელი, დაუდგინეს ტვინის კიბო. არის თუ არა ეს იმის დასაბუთება, რომ მობილური ტელეფონის ხშირი მოხმარება ზრდის ტვინის კიბოს შანსს? სამწუხაროდ, არ არსებობს არანაირი საშუალება დავადგინოთ, რომ თუ ამორჩევა რეპრეზენატიულია (წარმომადგენლობითია), მაშინ შედეგი გავცელდება მთელი პოპულაციისთვის. ზოგჯერ გვსმენია ანეკდოტური მტკიცებულება იმისათვის, რომ გააქარწყლონ მიზეზ-შედეგობრივი დამოკიდებულებები. მაგ. „ ბიძაჩემი ჯეფრი, რომელიც არის 85 წლის, მთელი თავისი ზრდასრული ცხოვრების მაძილზე ყოველ დღე ეწეოდა თითო კოლოფ სიგარეტს, მაგრამ დღესაც ცხენივით ჯანმრთელია“. არ არის სავალდებულო, რომ ასოციაცია იყოს სრულყოფილი, იგი უნდა იყოს მიზეზ - შედეგობრივი. ყველა ვინც დღეში ეწევა ერთ კოლოფ სიგარეტს, არ დაავადდება ფილტვის კიბოთი, მაგრამ გასაკუთრებით დიდი წილი ამ დაავადების მოდის მათზე ვინც ამას აკეთებს, ვიდრე არამწევლებზე. შესაძლებელია ბიძა ჯეფრის აქვს გამორჩეული ჯანმრთელობა, მაგრამ ამან არ უნდა გვიბიძგოს რეგულარული მწევლობისკენ.

იმის მაგიერ, რომ გამოვიყენოთ არაოფიციალური მონაცემები, უკეთესია დავეყრდნოთ ავტორიტეტულ მონაცემთა ბაზებს. თქვენ შეგიძლიათ მოიპოვოთ კვლევების შედეგები თემების მიხედვით, რისთვისაც საჭიროა შეიყვანოთ საძიებო სისტემაში საკვანძო სიტყვები. ეს ძიება მიგიყვანთ გამოქვეყნებულ კვლევის შედეგებამდე. ყველა შემთხვევაში ეს უფრო მეტად სანდო იქნება, ვიდრე არაოფიციალური მონაცემები.

აღწერები და სხვა შერჩევების გამოკვლევები.

საერთო სოციალური გამოკითხვა (GSS) არის შერჩევის გამოკითხვის (კვლევის) მაგალითი. იგი აგროვებს ინფორმაციას ზრდასრული ამერიკელებისგან, რომ წარმოადგინოს სრული სურათი პოპულაციაზე. შერჩევის გამოკითხვა არის არაექსპერიმენტული კვლევის ერთ-ერთი ტიპი. ქვეყნების უმეტეს ნაწილში რეგულარულად ტარდება მოსახლეობის აღწერები. ეს არის კვლევა, რომელიც ცდილობს დაადგინოს მოსახლეობის რაოდენობა და გაზომოს მათ შესახებ ზოგიერთი მახასიათებელი. იგი განსხვავდება ყველა სხვა კვლევებისგან, რადგან ისინი გამოიყენებენ მხოლოდ მცირე შერჩევას მთელი პოპულაციიდან.

ამერიკის შეერთებული შტატების კონსტიტუციაში წერია, რომ ქვეყანაში მოსახლეობის საყოველთაო აღწერა უნდა ჩატარდეს ყოველ 10 წელიწადში. პირველი ასეთი აღწერა ჩატარდა 1790 წელს. მაშინ აშშ-ს მოსახლეობა იყო 3.9 მილიონი ადამიანი, ხოლო დღესთვის აშშ-სმოსახლეობა გაზრდილია 300 მილიონზე მეტი ადამიანით და იგი შეადგენს 308,745,538 ადამიანს (2010 აშშ-ს აღწერის მონაცემებით). გარდა დათვლისა არსებობს აშშ-ში მოსახლეობის აღწერის ჩატარების კიდევ სამი ძირითადი მიზეზი:

- კონსტიტუციით მანდატების რაოდენობა წარმომადგენლობით პალატაში განისაზღვრება შტატის მოსახლეობის წილის მიხედვით მთელი ქვეყნის მოსახლეობაში, რომელიც ეფუძნება მოსახლეობის აღწერას. როდესაც 2010 წლის აღწერა დასრულდა, კონგრესმა ადგილების რაოდენობა 435-ით განსაზღვრა. ყოველი აღწერის შემდეგ ცალკეულმა შტატმა შეიძლება დაკარგოს ან მიიღოს დამატებითი ადგილები კონგრესში იმისდა მიხედვით, თუ როგორია მისი მოსახლეობის რაოდენობა სხვა შტატების მოსახლეობასთან შედარებით.
- აღწერის მონაცემები გამოიყენება აგრეთვე ამომრჩეველთა ოლქების საზღვრების გადანაწილებაშიც.
- აღწერის მონაცემები გამოიყენება ადგილობრივი კომუნიკაციებისთვის ფედერალური ბიუჯეტის გადანაწილებაში.

იმისათვის რომ შერჩევის გამოკვლევა იყოს ინფორმაციული, მნიშვნელოვანია, რომ შერჩევა კარგად ასახავდეს პოპულაციას. აღმოჩნდა, რომ თურმე შემთხვევით შერჩევაში თუ განვსაზღვრავთ საკვანძო საგნებს, მივიღებთ კარგ შერჩევას.

კარგი და ცუდი გზები შერჩევების მისაღებად. პირველი ნაბიჯი შერჩევის კვლევაში არის იმ პოპულაციის განსაზღვრა, რომელიც წარმოადგენს კვლევის მიზანს. მეორე ნაბიჯია, შევადგინოთ საგანთა ნუსხა, რომლებისგანაც შეგვეძლება შერჩევის შედგენა. ამ სიას ეწოდება **შერჩევის ჩარჩო.** იდეალურ შემთხვევაში, შერჩევის ჩარჩოში ჩამოთვლილია მთელი პოპულაციის ინტერესი. თუმცა, რა თქმა უნდა ძნელია, პრაქტიკაში მოვახდინოთ პოპულაციის ყველა საგნის იდენტიფიცირება. მაგრამ თუ შევადგენთ შერჩევას მხოლოდ კომფორტული საგნებისგან, შედეგები ვერ იქნება რეპრეზენტატიული (წარმომადგენლობითი) მთელი პოპულციისთვის.

უკეთესია შეირჩეს რეპრეზენტატიული, ვიდრე შერჩევა კომფორტის მიხედვით. ანუ ყველა საგანს უნდა ჰქონდეს ერთნაირი შანსი შერჩევაში მოხვედრის. იგი აგრეთვე უნდა იძლეოდეს ანალიზის შანსს. ანუ განვსაზღვროთ რამდენად ალბათურია, რომ აღწერითი სტატისტიკა (მაგ. შერჩევის საშუალო) ჩავარდება რეალურ მნიშვნელობასთან საკმაოდ ახლოს, ვინაიდან ჩვენს მიზანს შეადგენს, გამოვიტანოთ დასკვნები მთელი პოპულაციის შესახებ. ამის გამო უნდა გამოყენებულ იქნეს შემთხვევითი შერჩევა.

ამორჩევის შემთხვევით ტიპს ეწოდება მარტივი შემთხვევითი შერჩევა. უხეში მიახლოებით, n საგნიანი მარტივი შემთხვევითი შერჩევის ცდომილების შეფასებაა:

$$\text{ცდომილების მიახლოვებითი მნიშვნელობა} = \frac{1}{\sqrt{n}}100\%.$$

კარგი და ცუდი ექსპერიმენტები. განვიხილოთ მაგალითი, სადაც ვნახავთ, როგორ შეიძლება ექსპერიმენტი გავხადოთ კარგი. საყოვლთაოდ ცნობილია, რომ სიგარეტზე თავის დანებება საკმაოდ რთულია. ზოგიერთი მეცნიერის აზრით მწევლებს ექნებათ ნაკლები შანსი რეციდივზე, თუ ისინი მოწევაზე თავის დანებების შემდეგ მიიღებენ ანტიდეპრესანტებს. როგორ დააყენებთ ექპერიმენტს შესწავლაზე, ეხმარება თუ არა ანტიდეპრენსანტი მწეველს სიგარეტზე თავის დანებებაში?

ასეთი ტიპის კვლევებში, ისე როგორც სამედიცინო ექსპერიმენტების უდიდეს ნაწილში, შეუძლებელია გამოვიყენოთ შემთხვევითი შერჩევა (ყველა მწეველი, ვისაც უნდა სიგარეტის გადაგდება). ასე რომ საჭიროა კომფორტული შერჩევა. მაგალითად, სამედიცინო ცენტრმა შეიძლება გააკეთოს რეკლამა და მიიწვიოს მოხალისენი, ვისაც უნდა მოწევისთვის თავის დანებება.

საკონტროლო შედარების ჯგუფი. დავუშვათ გყავთ 400 მოხალისე. თქვენ ეკითხებით მათ, ვის უნდა თავის დანებება დღესვე. შემდეგ თითოეულს უნდა შესთავაზოთ

ანტიდეპრენსანტი და ერთიწლის შემდეგ შეამოწმოთ, რამდენი რეციდივი იყო. ვთქვათ, იყო 42%. ეს არ არის საკმარისი ინფორმაცია. ეს შედეგი უნდა შეადაროთ რეციდივების იმ პროცენტულ მაჩვენებელს, რომელიც გვექნებოდა, თუ არ მიიღებდნენ ანტიდეპრესანტებს. ამიტომ მოხალისეები უნდა გავყოთ ორ ჯგუფად. ერთი ჯგუფი იღებს ატიდეპრენსანტებს, მეორე კი - არა. თქვენ შეგიძლიათ მეორე ჯგუფს მისცეთ ტაბლეტები, რომლებიც ზუსტად ისე გამოიყურებიან როგორც ანტიდეპრესანტები, მაგრამ არ შეიცავენ არანაირ აქტიურ ნივთიერებას. ამ მეორე ჯგუფს ვუწოდოთ საკონტროლო ჯგუფი. გარკვეული დროის შემდეგ რეციდივები უნდა შედარდეს.

შემთხვევითობა. თუ თქვენ თვითონ გადაწყვეტთ, რომელმა საგანმა (პაციენტმა) მიიღოს ანტიდეპრენსანტი და რომელმა არა, შეიძლება ამან მიგიყვანოთ განსხვავებულ შედეგებამდე და შეგნებულად თუ შეუგნებლად გააკეთოთ მცდარი დასკვნები.

აქედან გამომდინარე, უკეთესია გამოვიყენოთ შემთხვევითი შერჩევა. შევარჩიოთ შემთხვევით 200 ვინც უნდა მიიღოს ანტიდეპრენსანტი და 200 საკონტროლო ჯგუფისთვის. შემთხვევითობა გვეხმარება თავიდან ავიცილოთ გაურკვეველობები ან აგვერიოს ერთი ჯგუფის ტენდენციები მეორეში, როგორიცაა, მაგალითად, უკეთესი ჯანმრთელობა ან ახალგაზრდობა. ასეთი შერჩევის დროს შემცირებულია დაფარული ცვლადების გავლენაც.

პრაქტიკული ნაწილი.

6.1. ექსპერიმენტი თუ დაკვირვება? ახსენით ექსპერიმენტია თუ დაკვირვებადი კვლევა? უკეთესი იქნებოდა გვეთქვა: გამოიკვლიეთ შემდეგი სიტუაცია.

ა) მოწევა მოქმედებს თუ არა გულის დაავადებებზე.

ბ) SAT-ის ქულები დადებითად ასოცირდება თუ არა უმაღლეს კოლეჯში სწავლასთან?

გ) კატალოგის გარეთა მხარეს მიკრული სპეციალური კუპონი განაპირობებს თუ არა იმას, რომ მიმღები სავარაუდოდ შეუკვეთავს პროდუქციას ამანათების კომპანიიდან?

6.2. უსაფრთხოების ღვედი. ენდის აქვს მოსმენილი, რომ ავტოკატასტროფის მსხვერპლი გარდაიცვალა იმის გამო, რომ მას ეკეთა უსაფრთხოების ღვედი, რომელმაც არ მისცა მოძრაობის საშუალება და დაზიანდა მანქანის ნამტვრევებით.

ამის შედეგად ენდი აღარ იკეთებს მართვის დროს უსაფრთხოების ღვედს. როგორ ახსნით ენდის მცდარ დასკვნას? ეყრდნობა იგი არაოფიციალურ მონაცემებს?

6.3. იარაღის კონტროლი. ამერიკელების 75% -ზე მეტმა უპასუხა „დიახ“ კითხვაზე: „მხარს უჭერთ თუ არა კონტროლის გამკაცრებას იარაღის არალეგალურ გაყიდვაზე? მაგრამ აგრეთვე 75% -ზე მეტმა თქვა „არა“ კითხვაზე: „მხარს უჭერთ თუ არა კანონს, რომლის მიხედვითაც პოლიციამ უნდა გადაწყვიტოს, ვის ექნება უფლება ჰქონდეს ცეცხლსასროლი იარაღი. ახსენით რა არის არასწორი ან ფორმულირებებში.

6.4. მოწევა გავლენას ახდენს ფილტვის კიბოზე? ხომ არ გამოიკვლევდით საკითხს: მწევლებს უფრო მეტად ემუქრებათ თუ არა ფილტვის კიბო, ვიდრე არამწევლებს? სტუდენტებიდან შემთხვევით შეარჩიეთ ნახევარი ისეთი სტუდენტების, ვინც ეწევა ერთ კოლოფ სიგარეტს დღეში. ნახევარი კი მათგან ვისაც არასდროს მოუწევია. 50 წლის შემდეგ თქვენ გააანალიზეთ სად უფრო მეტია ფილტვის კიბოს შემთხვევა, ვინც ეწეოდნენ, თუ ვინც არ ეწეოდნენ?

ა) ეს ექსპერიმენტია თუ დაკვირვება? რატომ?

ბ) ჩამოთვალეთ ასეთი დაგეგმილი კვლევის განხორციელებისთვის სამი მაინც პრაქტიკული სირთულე.

7

ალბათობის თეორია.

ძირითადი ცნებები:

ხდომილება, ხდომილებათა სახეები, ელემენტარულ ხდომილებათა სივრცე. ხდომილების ალბათობა. ხდომილებათა ჯამისა და ნამრავლის ალბათობები.

ყოველდღიურ ცხოვრებაში ხშირად გვიწევს ისეთი გადაწყვეტილებების მიღება, როდესაც არ ვართ დარწმუნებული შედეგში. გინდათ ჩადოთ ფული საფონდო ბაზარში? გინდათ დააზღვიოთ ავტომობილი? გინდათ დაიწყოთ ახალი ბიზნესი, მაგალითად, გახსნათ პიცერია უნივერსიტეტის კამპუსის წინ? ან უფრო მშვიდობიანი გადაწყვეტილება, იწვიმებს თუ არა, რომ წაიღოთ ქოლგა? ამისთვის უნდა შემოვიტანოთ ალბათობის ცნება - ანუ როგორ გამოვსახოთ რიცხვობრივად „გაურკვევლობა“. ვნახავთ, როგორ გამოვთვალოთ შემთხვევითი მოვლენების შესაძლო შედეგების მოსვლის შანსები - ყოველდღიური ვარიანტები, რომელთა მოსვლის შედეგები არ არის ზუსტად ცნობილი. მაგალითად, ალბათობის დახმარებით თქვენ შეგეძლებათ გამოთვალოთ ლატარიის გათამაშებაში თქვენი მოგების შანსი. შეგეძლებათ აგრთვე გამოთვალოთ ალბათობა იმისა, რომ დამქირავებლის მიერ ნარკოლოგიური შემოწმების ტესტი რა ალბათობით იძლევა სწორ დასკვნას ნარკოტიკების მოხმარების ან არ მოხმარების შესახებ. ალბათობის დახმარებით შესაძლებელია გაიზომოს უზუსტობა, რომელიც შეიძლება გამოჩნდეს შემთხვევითად შერჩეული ექსპერიმენტული კვლევის დროს. ჩვენი მიზანია შევიწსავლოთ ალბათობის საფუძვლები, რათა გავაკეთოთ სწორი დასკვნები მონაცემთა ბაზაზე დაყრდნობით.

ამ გზაზე პირველი ნაბიჯი არის ყველა შესაძლო შედეგის ჩამოთვლა. ყველა შესაძლო შემთხვევითი ელემენტარული შედეგების სიმრავლეს ეწოდება **ელემენტარულ ხდომილებათა სივრცე.** მაგალითად, თუ კამათელს გავაგორებთ ერთხელ, მაშინ ელემენტარულ ხდომილებათა სივრცე შედგება ექვსი შესაძლო შედეგისგან: { 1, 2, 3, 4, 5, 6 }. თუ მონეტას ავაგდებთ ორჯერ, მაშინ სივრცე ოთხელემენტიანია

{*გგ, გს, სგ, სს*}, სადაც მაგალითად, *სგ* ნიშნავს პირველად საფასურის მოსვლას, შემდეგ გერბის.

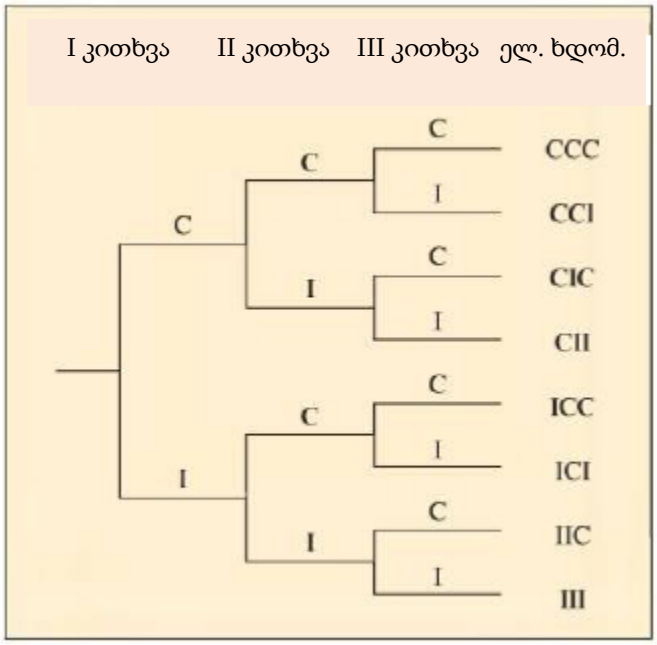
სცენარი 7.1. პოპ ვიქტორინის მრავალჯერადი არჩევანი. დავუშვათ, უნივერსიტეტში ტარდება ვიქტორინა, სამ ეტაპად. თითოეულ კითხვას აქვს ხუთი შესაძლო პასუხი. ვიქტორინაში მონაწილე სტუდენტების პასუხები ფასდება მხოლოდ ორი შეფასებით: სწორი - " C " (correct) და არასწორი - " I " (incorrect). თუ სტუდენტმა სწორად უპასუხა პირველ და მეორე კითხვას და მესამეს შეცდომით, მაშინ მისი შედაგია CCI.

კითხვა შესწავლისთვის:

რა არის სტუდენტთა პასუხების ელემენტარულ ხდომილებათა სივრცე?

გააზრებისთვის:

ელემენტარულ ხდომილებათა სივრცის შედგენის ერთერთი მეთოდია დიაგრამა-ხის გამოყენება, რომლის ტოტები გვიჩვენებენ თუ რა შეიძლება მოხდეს ყოველ ცდაზე. სტუდენტის ეფექტურობა სამი კითხვის შემდეგ გამოისახება შემდეგი ხით და ტოტებით.



დიაგრამა 7.2.*ხის დიაგრამა სიტუაციისთვის: სტუდენტმა უპასუხა სამ კითხვას.* პირველი სიმრავლიდან გამოდის სამი ტოტი, მესამე სიმრავლეში კი უკვე რვა ტოტია, რომლებიც განსაზღვრავენ ყველა შედეგს ელემენტარულ ხდომილებათა სივრცეში.

დიგრამის მიხედვით ამ შემთხვევაში გვაქვს რვა შესაძლო შემთხვევა:

$$\{\ CCC,\ CCI,\ CIC,\ CII,\ ICC,\ ICI,\ IIC,\ III\ \}\ .$$

კითხვა: რამდენი შესაძლო შედეგი იქნებოდა ოთხი კითხვის შემთხვევაში.

ხდომილებები.

ხშირად საჭიროა შევადგინოთ შედეგების კონკრეტული სიმრავლე, რომელიც ცხადია იქნება ელემენტარულ ხდომილებათა სივრცის ქვესიმრავლე. ვთქვათ ასეთი: იმ შედეგების ქვესიმრავლე, როდესაც სტუდენტი პასუხობს სწორად ორ შეკითხვას მაინც სამიდან.

ელემენტარულ ხდომილებათა სივრცის ნებისმიერ ქვესიმრავლეს ეწოდება **ხდომილება.**

ხდომილებები, ზოგადად, აღინიშნება დიდი ლათინური ასოებით A, B, C,... მაგალითად, A = სტუდენტმა უპასუხა სწორად სამივე შეკითხვას = $\{CCC\}$, B = სტუდენტმა სწორად უპასუხა ორ შეკითხვას მაინც = $\{CCI,\ CIC,\ ICC,\ CCC\}$.

ყოველ შედეგს ელემენტარულ ხდომილებათა სივრციდან გააჩნია ალბათობა. აგრეთვე ყოველ ხდომილებას გააჩნია ალბათობა. რომ ვიპოვოთ ეს ალბათობა, უნდა ამოვწეროთ ელემენტარულ ხდომილებათა სივრცე და გამოვთქვათ დამაჯერებელი ვარაუდები შედეგებზე. ზოგჯერ შეიძლება დავუშვათ, რომ შედეგები ტოლალბათურია. ალბათობები უნდა ემორჩილებოდნენ ორ ძირითად წესს:

- ნებისმიერი ხდომილების ალბათობა არის 0 -სა და 1 -ს შორის.
- ყველა ელემენტარულ ხდომილებათა ალბათობების ჯამი უნდა იყოს 1 -ის ტოლი.

ხდომილების ალბათობა. A ხდომილების ალაბათობა აღინიშნება P(A) სიმბოლოთი და ტოლია ხდომილებაში შემავალი ყველა ელემენტარულ ხდომილებათა ალბათობების ჯამის.

- როდესაც ყველა შესაძლო ხდომილება ტოლალბათურია,

$$P(A)=\frac{A \text{ ხდომილების ყველა ელემენტის რაოდენობა}}{\text{ელემენტარულ ხდომილებათა სივრცის ელემენტების რაოდენობა}}$$

ხდომილებათა წყვილისთვის ალბათობების მოძებნის ძირითადი წესები: ზოგიერთი ხდომილება გამოსახულია როგორც შედეგი, რომელიც

(ა) არის ერთ ხდომილებაში **და არ არის** რომელიმე სხვაში,

(ბ) არის ერთ ხდომილებაში **და** არის სხვა ხდომილებაშიც,

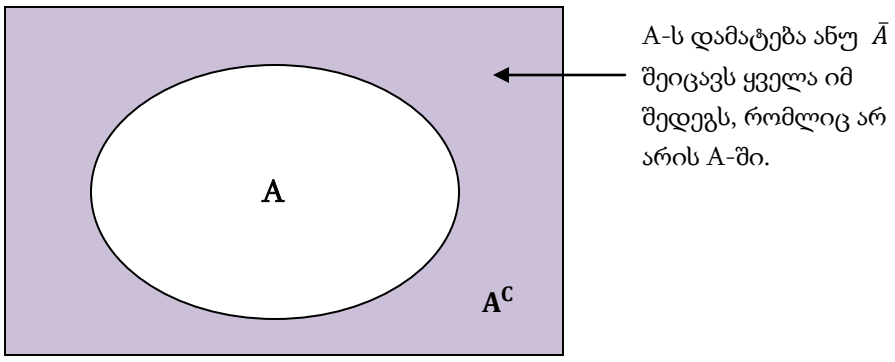
(გ) არის ერთ ხდომილებაში **ან** არის სხვა $\bar{A}$ ხდომილებაში.

ვნახოთ როგორ გამოითვლება ალბათობები ამ სამ შემთხვევაში.

A ხდომილების **დამატება** აღინიშნება $\bar{A}$ –ით და არის იმ ელემენტების სიმრავლე ელემენტარულ ხდომილებათა სივრციდან, რომლებიც არ არიან A-ში. აქედან გამომდინარე A და $\bar{A}$ ხდომილებათა ალბათობების ჯამი ერთის ტოლია. ამიტომ

$$P(\bar{A}) = 1 - P(A).$$

შემდეგ ნახაზზე A ხდომილება არის ფიგურა მართკუთხედში, ხოლო $\bar{A}$ არის მართკუთხედის დარჩენილი დაშტრიხული ნაწილი. A და $\bar{A}$ ხდომლებები ერთად ფარავენ მთელ ელემენტარულ ხდომილებათა სივრცეს.



ნახაზი 7.3. *A ხდომილებისა და მისი დამატების $\bar{A}$ -ის ილუსტრაცია.*

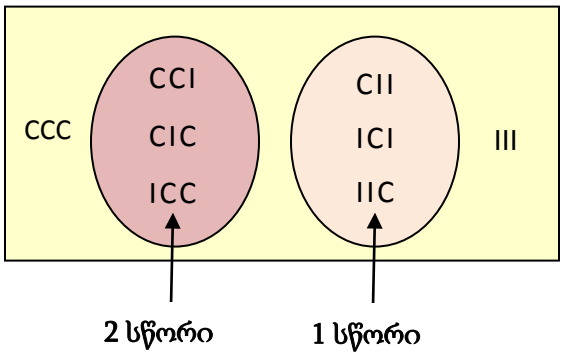
ასეთი ტიპის დიაგრამებს ეწოდებათ **ვენის დიაგრამები.**

კითხვა: შეგიძლიათ, დახაზოთ A და B ხდომილობების ისეთი დიაგრამა, რომ მათ ჰქონდეთ საერთო შედეგები (ელემენტები), მაგრამ ზოგიერთი ელემემენტი იყოს მხოლოდ A -ში ან მხოლოდ B -ში.

ზოგჯერ რაიმე ხდომილობის ალბათობის საპოვნელად უკეთესია, ვიპოვოთ მისი დამატების ალბათობა და შემდეგ 1 -ს გამოვაკლოთ ეს ალბათობა.

არათავსებადი ხდომილებები ეწოდება ისეთ ხდომილობებს, რომელთაც საერთო შედეგები არ გააჩნიათ ანუ არიან არათანამკვეთი სიმრავლეები.

შემდეგ დიაგრამაზე მოცემულია არათავსებადი ხდომლებების ერთი წყვილი სცენარი 7.1 -დან.



ნახაზი 7.4. *ვენის დიაგრამა არათავსებადი ხდომილებებისთვის.*

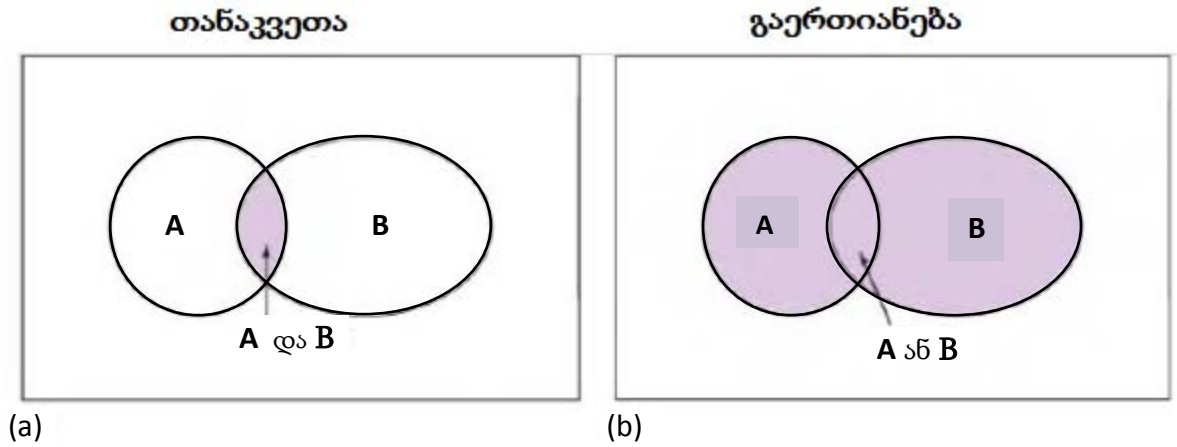
ხდომილება - სტუდენტმა სწორად უპასუხა ზუსტად ერთ შეკითხვას არათავსებადია ხდომიებასთან - სტუდენტმა სწორად უპასუხა ზუსტად ორ შეკითხვას.

კითხვა: ამ დიაგრამაზე გამოვყოთ ხდომილება: სტუდენტმა სწორად უპასუხა პირველ შეკითხვას. აქვს თუ არა ამ ხდომილებას საერთო ვენის დიაგრამაზე გამოსახულ რომელიმე მოცემულ ხდომილებასთან?

განმარტებიდან გამომდინარე A და $\bar{A}$ არათავსებადი ხდომილებებია.

ზოგიერთი ხდომილება შეგენილია სხვა ხდომილებებისგან ან რაც იგივეა, ხდომილებებისგან შიძლება შევადგინოთ ახალი ხდომილებები. მაგალითად, ხდომილებას, რომელსაც ადგილი აქვს, მაშინ როცა ერთდროულად სრულდება A და B ხდომილებები, ეწოდება A და B **ხდომილებათა თანაკვეთა.** ხოლო ხდომილებას, რომელსაც ადგილი აქვს მაშინ, როცა სრულდება A ან B ხდომილება, ეწოდება **ხდომილებათა გაერთიანება.** ეს არის უდიდესი სიმრავლე, რომელიც შეიცავს A-ს და B-ს თანაკვეთას, აგრეთვე ხდომილებებს A -დან, რომლებიც არ არიან B -ში, და ხდომილებებს B-დან, რომლებიც არ არიან A-ში. შემდეგი დიაგრამა გვიჩვენებს თანაკვეთისა და გაერთიანების ვენის დიაგრამებს. თანაკვეთა გვიჩვენებს, რომ როცა

ხდება A ხდომილება, ხდება B -ც. აღინიშნება „ A და B”. გაერთიანება კი ნიშნავს, რომ ხდება A ან ხდება B ან ორივე. აღინიშნება „ A ან B”.



ნახაზი 7.5. *ორი ხდომილების თანაკვეთა და გაერთიანება.*

კითხვა: შეგიძლიათ გამოსახოთ P(A ან B), თუ იცით P(A), P(B) და P(A და B)?

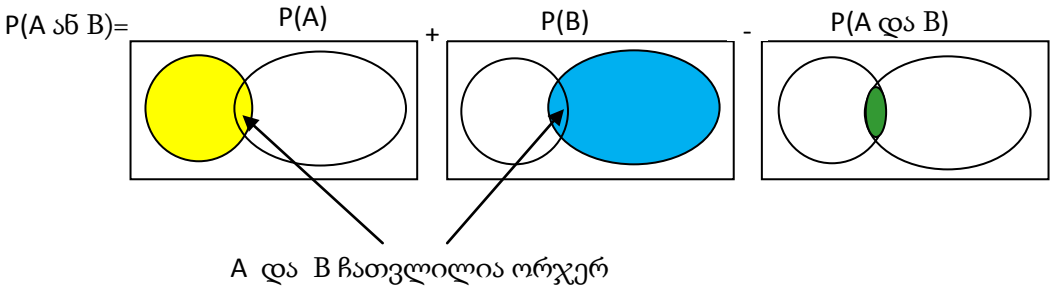
მაგალითად, პოპ ვიქტორინის შემთხვევაში განვიხილოთ ხდომილებები:

A = სტუდენტმა სწორად უპასუხა პირველ კითხვას = $\{CCC,\ CCI,\ CIC,\ CII\}$,

B = სტუდენტმა სწორად უპასუხა ორ კითხვას = $\{\ CCI,\ CIC,\ ICC\}$.

მაშინ A-ს და B-ს თანაკვეთა იქნება $\{CCI,\ CIC\}$, ხოლო გაერთიანება $\{CCC,\ CCI,\ CIC,\ CII,\ ICC\}$.

როგორ გამოვთვალოთ ჯამის ალბათობა? დავაკვირდეთ ქვემოთ მოცემულ ვენის დიაგრამებს



ნახაზი 7.6. *გაერთიანების ალბათობა.*

ჩვენ შეგვიძლია შევკრიბოთ P(A) და P(B), მაგრამ რადგანაც თანაკვეთა შედის ორივეში, ამიტომ მათ ჯამს უნდა გამოვაკლოთ თანაკვეთის ალბათობა. თუ მათი თანაკვეთა ცარიელია, მაშინ ჯამს გამოვაკლოთ 0, ანუ გექნება უბრალოდ ალბათობების ჯამი. ამიტომ, ზოგადად,

$$P(\text{A ან B}) = P(\text{A}) + P(\text{B}) - P(\text{A და B}) .$$

კითხვა: როდის ექნება ადგილი ტოლობას: $P(\text{A ან B}) = P(\text{A}) + P(\text{B})$?

პასუხი: როცა A და B არათავსებადი ხდომილებებია, ანუ როცა $P(\text{A და B}) = 0$, მაშინ $P(\text{A ან B}) = P(\text{A}) + P(\text{B})$. ეს ჩანაწერი იგივეა, რაც $P(A+B) = P(A) + P(B)$.

თუ გვინდა, რომ ვიპოვოთ ნამრავლის ალბათობა, შეგვიძლია გავამრავლოთ მათი ალბათობები ერთმანეთზე, მაგრამ აუცილებლად უნდა გვახოვდეს, ასეთ დროს ხდომილებები უნდა იყვნენ **დამოუკიდებლები**, ანუ ერთის მოხდენა ან არ მოხდენა გავლენას არ უნდა ახდენდეს მეორეზე, ე.ი.

$$P(\text{A და B}) = P(\text{A}) \times P(\text{B}) , \qquad \text{ანუ} \qquad P(AB) = P(A)P(B).$$

სცენარი 7.7. პოპ ვიქტორინის პასუხების გამოცნობა. ვიქტორინის კითხვებზე პაუხის გასაცემად სტუდენტი სრულიად მოუმზადებელია და შეთხვევით ირჩევს 5 სავარაუდო პასუხიდან ერთს. მაშინ ალბათობა იმისა, რომ სწორად უპასუხებს შეკითხვას, ცხადია, ტოლია 1/5 -ის, ანუ 0.20-ის. წინა კითხვაზე გაცემული პასუხის შედეგი არ მოქმედებს შემდეგ შედეგზე.

კითხვები შესწავლისთვის:

ა) გამოთვალეთ სტუდენტის პასუხების ყველა შესაძლო შემთხვევების ალბათობები. გამოსახეთ სწორი პასუხი (C) და არასწორი (I) აღნიშვნებში.

ბ) გამოთვალეთ ალბათობა იმისა, რომ სტუდენტი სწორად უპასუხებს ორ შეკითხვას მაინც.

გაანზრება:

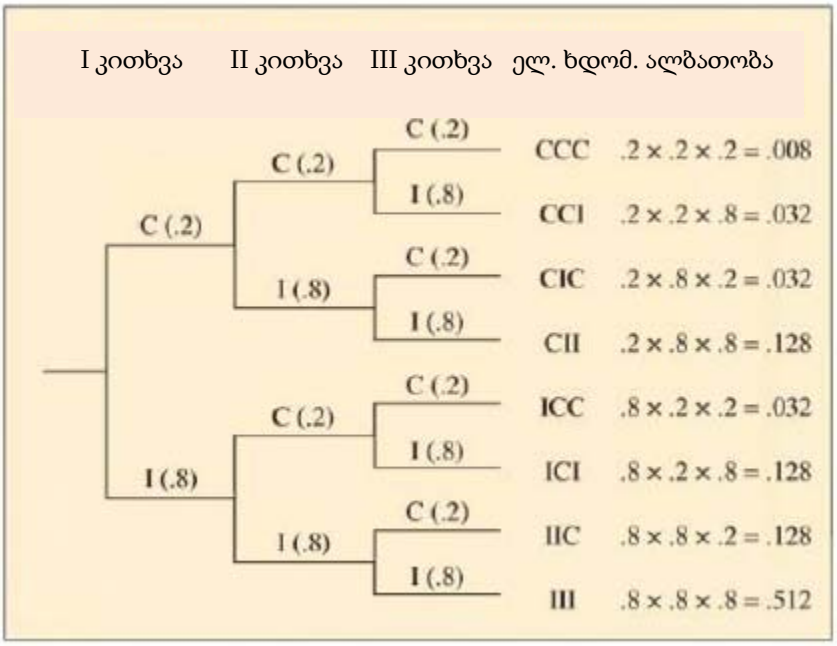
ა) ყოველი პასუხისთვის $P(C) = 0.20$ და ყოველი $P(I) = 1 - 0.20 = 0.80$. ალბათობა იმისა, რომ სტუდენტი სწორად უპასუხებს ყველა შეკითხვას არის

$$P(CCC) = P(C) \times P(C) \times P(C) = 0.20 \times 0.20 \times 0.20 \times 0.20 = 0.008 ,$$

რაც ერთი შეხედვით არაბუნებრივია. ამის მსგავსად, ალბათობა იმისა, რომ ორი პასუხი სწორია და მესამე არასწორი, არის:

$$P(\text{CCI}) = P(\text{C}) \times P(\text{C}) \times P(I) = 0.20 \times 0.20 \times 0.20 \times 0.80 = 0.032 .$$

ასევე იქნება, P(CIC) -ის და P(ICC) -სთვისაც, ანუ სხვა თანმიმდევრობით მიღებული ორი სწორი პასუხის შემთხვევაში. ქვემოთ მოცმული ხისებრი დიაგრამა გვიჩვენებს, როგორ უნდა გადავამრვლოთ ალბათობები, რომ მივიღოთ ალბათობები ყველა რვავე შემთხვევისთვის.



დიაგრამა 7.8. *ხის დიაგრამა სამი კითხვის გამოცნობაზე.* ყოველი გზის გასწვრივ ალბათობების ნარავლები იძლევია შესაბამისი ხდომილებების ალბათობებს, თუ ისინი დამოუკიდებლებია. აქ იგულისხმება, რომ ხდომილებები დამოუკიდებელნი არიან.

კითხვა: თქვენ მოელით, რომ ხდომილებები იქნება დამოუკიდებელი თუ სტუდენტი მხოლოდ ალალბედზე არ უპასუხებს ყველა კითხვას? რატომ კი, ან რატომ არა.

ბ) ალბათობა იმისა, რომ ორი სწორი პასუხი მაინც გვაქვს, არის

$$P(\text{CCC}) + P(\text{CCI}) + P(\text{CIC}) + P(\text{ICC}) = 0.008 + 3 \times 0.032 = 0.104 .$$

საბოლოოდ მივიღეთ, რომ თუ სტუდენტი შემთხვევითად პაუხობს კითხვებს, მაშინ მისი წარმატების შანსი დაახლოებით არის მხოლოდ 10%. შემოწმების მიზნით,

თქვენ შეგიძლიათ ნახოთ, რომ რვავე ხდომილების ალბათობების ჯამი არის 1. ალბათობა გვიჩვენებს, რომ შემთხვევითი გამოცნობა არ არის სტუდენტის საუკეთესო არჩევანი.

ალბათობათა გამოთვლის წესები:

- ყოველი კონკრეტული ხდომილების ალბათობა, მოთავსებულია 0 -სა და 1 -ს შორის. ყველა ელემენტარულ ხდომილებათა ალბათობების ჯამი ტოლია 1 -ის. ხდომილების ალბათობა ტოლია, ამ ხდომილებაში შემავალ ყველა ელემენტარულ ხდომილებათა ალბათობების ჯამის.
- A და მისი $\bar{A}$ დამატებისთვის სამართლიანია ტოლობა: $P(\bar{A}) = 1 - P(A)$.
- ორი ხდომილების გაერთიანებისთვის გვაქვს: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- თუ A და B დამოუკიდებელი ხდომილებებია, მაშინ ორი ხდომილების თანაკვეთის ალბათობა ტოლია: $P(A \cap B) = P(A) \times P(B)$, ანუ $P(A \cdot B) = P(A) \times P(B)$.
- თუ A და B არათავსებადი ხდომილებებია, ანუ მათ არ გააჩნიათ საერთო ელემენტები, მაშინ $P(A \cap B) = 0$ და $P(A \cup B) = P(A) + P(B)$, ანუ $P(A \cdot B) = 0$ და $P(A+B) = P(A) + P(B)$.

პრაქტიკული ნაწილი.

7.1. არასწორი ელემენტარულ ხდომილებათა სივრცე. წყვილი გეგმავს, რომ იყოლიონ ოთხი ბავშვი. მამა ამბობს, რომ ელემენტარულ ხდომილებათა სივრცე გოგონებისთვის არის 0, 1, 2, 3 და 4. იგი აგრძელებს, რომ ეს სივრცე შედგება 5 შედეგისგან, თანაც გოგოს და ბიჭის დაბადება თითქმის ტოლალბათურია. ამიტომ ალბათობა იმისა, რომ ოთხივე ბავშვი გოგო იქნება არის 1/5. ახსენით რაში მდგომარეობს ამ მსჯელობის ხარვეზი.

7.2. დაზღვევა. ყოველ წელს სადაზღვევო ინდუსტრია მნიშვნელოვან რესურსებს ხარჯავს რისკის ალბათობის შეფასებაზე. რომ დააგროვოთ ერთი სიკვდილის რისკები ერთი მილიონიდან, უნდა იაროთ მანქანით 100 მილი, გადაუფრინოთ ქვეყანას თვითმფრინავით, იმუშაოთ პოლიციაში 10 საათი, იმუშაოთ ქვანახშირის მაღაროში 12 საათი, მოწიოთ 2 ღერი სიგარეტი და თუ არამწეველი ხართ იცხოვროთ 2 კვირის განმავლობაში მწევლებთან, დალიოთ 70 პინტი ლუდი ყოველწლიურად (Wilson and Crouch, 2001, pp. 208-209). აჩევნეთ, რომ მილიონ სიკვდილში ერთის ალბათობა დაახლოვებით იგვეა, რაც წესიერი მონეტის აგდებებისას 20-ჯერ ზედიზედ საფასურის მოსვლა.

7.3. გლობალური დათბობა და ხეები. გამოკითხვა ეკითხება საგნებს, სჯერათ თუ არა, რომ გლობალური დათბობა ნამდვილად ხდება (შესაძლო პასუხებია: დიახ ან არა) და რა რაოდენობის საწვავის დახარჯვას გეგმავენ წლიურად მომავალში, გასულ წლებთან შედარებით (შესაძლო პასუხებია: ნაკლები, დაახლოებით იგივე, მეტი).

ა) გამოსახეთ ყველა შესაძლო შედეგების ელემენტარულ ხდომილებათა სივრცე ხის დიაგრამის საშუალებით, სადაც პირველი პასუხი გლობალურ დათბობას ეხება, მეორე კი - საწვავის გამოყენებას.

ბ) ვთქვათ, A იყოს ხდომილება „დიახ“ (პასუხი გლობალური დათბობაზე) და B იყოს ხდომილება „ნაკლები“ (პასუხი მომავალში საწვავის ხარჯზე). ვიგულისხმოთ, რომ P(A და B) > P(A) P(B). არის თუ არა A და B დამოუკიდებელი ხდომილებები? ახსენით არატექნიკური ტერმინებით.

7.4. ნახატებისა და ხელნაკეთობების გაყიდვები. ქალაქის ცენტრში არის ნახატებისა და ხელნაკეთობების მაღაზია. დაკვირვებები გაყიდვებზე გვიჩვენებს, რომ მაღაზიაში შემოსული ხალხის 20% რაღაცას ყიდულობს. სამი პოტენციური მყიდველი შევიდა მაღაზიაში.

ა) რამდენი შესაძლებლობა არსებობს, რომ გამყიდველი მათ მიჰყიდის რამეს (ელემენტარულ ხდომილებათა სივრცე). ააგეთ ხის დიაგრამა და აჩვენეთ შესაძლო შედეგები. (ვთქვათ, Y = გაყიდვა, N = არ გაყიდვა)

ბ) იპოვეთ ალბათობა იმისა, რომ სამი მყიდველიდან ერთი მაინც იყიდის.

გ) რა დაშვებას აკეთებთ ბ) ნაწილში გამოთვლების ჩატარების დროს? აღწერეთ სიტუაცია, რომელიც ამ დაშვებას არარეალურს გახდიდა. (მითითება: ხდომილებათა ურთიერთდამოკიდებულების თვალსაზრისით).

7.5. უსაფრთხოების ღვედების გამოყენება და ავტოავარიები. უკანასკნელ წელს ავტოავარიების შესახებ ჩანაწერებზე დაყრდნობით ფლორიდის სატრანსპორტო საშუალებებისა და საგზაო მოძრაობის უსაფრთხოების დეპარტამენტმა წარმოადგინა მონაცემები. ამ მონაცემებში (S) -ით აღნიშნულია, რომ პიროვნება გადარჩა, ხოლო(D) ნიშნავს, რომ გარდაიცვალა. შესაბამისად, ეკეთა თუ არა მას ღვედი (Y = დიახ, N = არა). მონაცემები წარმოდგენილია შემდეგი ცხრილის სახით:

ავტო ავარიის შედეგების უსაფრთხოების ღვედების ტარებაზე დამოკიდებულება			
ეკეთა თუ არა უსაფრთხ. ღვედი	გადარჩა (S)	დაიღუპა (D)	სულ
კი (Y)	412,368	510	**412,878**
არა (N)	162,527	1,601	**164,128**
სულ	**574,895**	**2,111**	**577,006**

ა) რა არის ყველა შესაძლო შედეგების ელემენტარულ ხდომილებათა სივრცე შემთხვევით შერჩეული პიროვნებისთვის, რომელიც მოყვა ავტოავარიაში. (მითითება: ერთერთი შესაძლო შედეგია YS).

ბ) მონაცემების გამოყენებით შეაფასეთ: i) P(D) და ii) P(N).

გ) შეაფასეთ ალბათობა იმისა, რომ პიროვნებას არ ეკეთა უსაფრთხოების ღვედი და გარდაიცვალა.

დ) ა) პუნქტზე დაყრდნობით რომ ვუპასუხოთ გ) პუნქტს, ყოფილან თუ არა N და D ხდომილებები დამოუკიდებლები? მონაცემების კონტექსტიდან გამომდინარე, გაანალიზეთ, რას ნიშნავს კითხვა: "არიან თუ არ არიან ისინი დამოუკიდებლები?"

8

ხდომილობათა დამოკიდებულება და დამოუკიდებლობა. პირობითი ალბათობა. მიღებული ფორმულების გამოყენებები.

ხშირად ხდომილებები არ არიან დამოუკიდებელი. პრაქტიკაში არაა აუცილებელი, რომ ხდომილებები იყვნენ დამოუკიდებელი. მაგალითად, ვიქტორინაზე მხოლოდ ორი სწორი პასუხით, ინსტრუქტორმა იპოვა შემდეგი პროპორცია თავისი სტუდენტების ფაქტობრივ პასუხებში (C = სწორი, I = არასწორი):

შედეგი:	II	IC	CI	CC
ალბათობა:	0.26	0.11	0.05	0.58

ვთქვათ, A აღნიშნავს - { პირველი პასუხი სწორია }, და B აღნიშნავს - { მეორე პასუხი სწორია }. ამ ალბათობებზე დაყრდნობით

P(A) = P({CI, CC}) = 0.05 + 0.58 = 0.63,

P(B) = P({IC, CC}) = 0.11 + 0.58 = 0.69,

P(A და B) = P({CC}) = 0.58.

A და B დამოუკიდებლები რომ ყოფილიყვნენ, უნდა გვქონოდა

P(A და B) =P(A) × P(B) = 0.63 × 0.69 = 0.43.

რადგანაც P(A და B) ნამდვილად არის 0.58, რომელიც განსხვავებულია 0.43-სგან, ამიტომ A და B არ არიან დამოუკიდებლები.

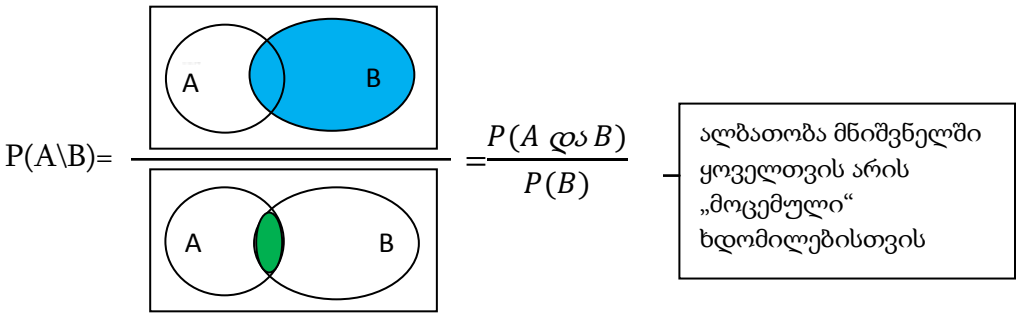
აქ ჩვენ უნდა შემოვიტანოთ **პირობითი ალბათობის** ცნება, რომელიც შეისწავლის ისეთი ხდომილებების ალბათობებს, როდესაც ვიცით, რომ მიღებული შედეგი იყო ელემენტარულ ხდომილებათა სივრცის ერთი კერძო ნაწილი. ყველაზე ხშირად იგი გამოიყენება ერთი ცვლადის კატეგორიის ალბათობის გამოსათვლელად (მაგალითად, პიროვნება არის ნარკოტიკების მომხმარებელი), როდესაც ვიცით პირველი ცვლადის შედეგი (მაგალითად, ვიცით ტესტის შედეგი ნარკიტოკების მოხმარებაზე). ამრიგად, პირობითი ალბათობის არსია შემდეგი: მოცემულია ხდომილება B, მაშინ A ხდომილების ალბათობა არის პროპორცია A და B ხდომილებათა თანაკვეთისა B ხდომილებასთან, რაც ფორმულით ასე გამოისახება:

P(A და B)/P(B). ეს ფარდობა არის A ხდომილების ალბათობა, როდესაც მოცემულია B ხდომილება.

სხვანაირად რომ ვთქვათ, A და B ხდომილებებისთვის A ხდომილების პირობითი ალბათობა, რომ B ხდომილება უკვე მოხდა, გამოითვლება შემდეგნაირად:

$$P(A|B) = \frac{P(A \text{ და } B)}{P(B)}$$

P(A|B) ასე იკითხება: „ A ხდომილების ალბათობა B პირობით“. საილუსტრაციოდ განვიხილოთ შემდეგი ნახაზი.



ნახაზი. 8.1. *ვენის დიაგრამა A ხდომილების პირობითი ალბათობისთვის B პირობით.*

ყველა იმ შემთხვევებიდან როცა B მოხდა, P(A|B) აღნიშნავს იმ პროპორციას, როცა A ხდომილებაც მოხდა.

კითხვა: დახაზეთ P(B|A) -ს წარმოდგენა. P(A|B) აუცილებლად უდრის P(B|A) -ს?

სცენარი 8.2. სამმაგი სისხლის ტესტი დაუნის სინდრომზე: *(პირობითი ალბათობა).* დიაგნოსტიკური ტესტი ამბობს, რომ პასუხი **დადებითია**, თუ ის მდგომარეობა, რაზეც მიმდინარეობს კვლევა, სახეზეა. ხოლო **უარყოფითია**, თუ იგივე მდგომარეობა არ აღმოჩნდა. რამდენად ზუსტია დიაგნოსტიკური ტესტი? სიზუსტის შეფასების ერთერთი მეთოდი არის შესაძლო შეცდომების ალბათობების გაზომვა.

ცრუ დადებითი: ტესტი ასაბუთებს, რომ მდგომარეობა სახეზეა, მაგრამ სინამდვილეში ის არ არსებობს.

ცრუ უარყოფითი: ტესტი ასაბუთებს, რომ მდგომარეობა არ აღმოჩნდა, თუმცა სინამდვილეში მდგომარეობა სახეზეა.

სისხლის სამმაგი ტესტი იკვლევს ფეხმძიმე ქალს და იძლევა რისკის შეფასებას იმაზე, რომ ბავშვი, რომელიც უნდა დაიბადოს, იქნება დაუნის სინდრომის გენეტიკური დარღვევით. ეს სინდრომი, რომელიც გვხვდება დაახლოვებით 800 ახალშობილიდან ერთში, წარმოიშობა უჯრედის არასწორად გაყოფის შედეგად, რაც გამოიხატება ნაყოფში 21-ე ქრომოსომის დამატებით ასლის გაჩენაში. ეს არის გონებრივი ჩამორჩენილობის ყველაზე უფრო გავრცელებული გენეტიკური მიზეზი. ბავშვის დაუნის სინდრომით დაბადების შანსი იზრდება, თუ ქალის ასაკი აღემატება 35 წელს. გამოკვლეული იქნა 35 წელზე მეტი ხნის 5282 ქალბატონი სამმაგი სისხლის ტესტზე, რათა შეესწავლათ თვით ტესტის სიზუსტე. შედეგი ასეთია: 5282 ქალიდან ფაქტს ადგილი ჰქონდა 54-ჯერ. მოხსენებაში წერია: „54 შემთხვევიდან 48-ს დაუნის სინდრომი დაუდგინდა ტესტის შედეგად და იმ ფეხმძიმეთაგან, ვინც საერთოდ შეუხებელი დარჩა, 25%-ს განესაზღვრათ განსაკუთრებულად მაღალი რისკი (ეს არის ცრუ დადებითი)".

კითხვები შესწავლისთვის:

ა) შეადგინეთ შემთხვევების ცხრილი, რომელიც გვიჩვენებს სისხლის ტესტის შესაძლო შედეგების რაოდენობას, აქვს თუ არა ნაყოფს დაუნის სინდრომი.

ბ) დავუშვათ, რომ პოპულაციიდან აღებული შერჩევა არის წარმომადგენლობითი. შეაფასეთ დადებითი ტესტის ალბათობა შემთხვევითად შერჩეული ფეხმძიმე ქალისთვის, რომელიც არის 35 წლის ან მეტის.

გ) მოცემულია, რომ დიაგნისტიკური ტესტის შედეგი დადებითია. შეაფასეთ ალბათობა იმისა, რომ დაუნის სინდრომი ნამდვილად არის.

გაანზრება:

ა) ჩვენ გამოვიყენებთ ორი ცვლადის შესაძლო შედეგებისთვის შემდეგ აღნიშვნებს:

დაუნის სინდრომის სტატუსი: D = დაუნის სინდრომი სახეზეა, D* = არ აღმოაჩნდა. სისხლის ტესტის შედეგი: POS = დადებითი, NEG = უარყოფითი.

ქვემოთ მოყვანილი ცხრილი გვიჩენებს შედეგების ოთხ შესაძლო კომბინაციას.

	სისხლის ტესტი		
დაუნის სინდრომი	**დადებითი**	**უარყოფითი**	**სულ**
D(დაუნი)	48	6	**54**
D*(არ აღმოაჩნდა)	1307	3921	**5228**
სულ	**1365**	**3927**	**5282**

ცხრილი 8.3. *შემთხვევების ცხრილისამმაგი სისხლის ტესტისთვის დაუნის სინდრომზე.*

ციტატაში ნათქვამია, დაუნის სინდრომით იყო 54 შემთხვევა. ეს არის ჯამის პირველი სტრიქონი. აქედან 48 ტესტირებულია დადებითად, ანუ 54 - 48 = 6 ტესტირებულია უარყოფითად. ეს რაოდენობებიც მოცემულია პირველ სტრიქონში. დაუნი აღირიცხა 54 შემთხვევაში n = 5282 -დან, ანუ 5282 – 54 = 5228 შემთხვევაში არ აღმოჩნდა, ეს არის ხდომილება D*. ეს ჯამის მეორე სტრიქონია. ამ 5228 -დან 25% -ს, ანუ 5228 × 0.25 = 1307 -ს ჰქონდათ დადებითი ტესტი. დარჩენილ 5228 – 1307 = 3921 კი უარყოფითი ტესტი. ეს რაოდენობები მოცემულია მეორე სტრიქონში.

ბ) ცხრილიდან შევაფასებთ (ანუ გამოვთვლით) დადებითი ტესტის ალბათობას: P(POS) = 1355/5282 = 0.257.

გ) დაუნის სინდრომის ალბათობა იმ პირობით, რომ ტესტი დადებითია, არის პირობითი ალბათობა, P(D|POS). კონცენტრირება დადებით ტესტზე ნიშნავს, რომ ჩვენ ვიხილავთ მხოლოდ პირველი სვეტის შემთხვევებს. ანუ 1355 -დან, ვინც ტესტირებული იყო დადებითად, 48 -ს მართლაც აღმოაჩნდა დაუნის სინდრომი, ანუ P(D|POS) = 48/1355 = 0.035. ვნახოთ, როგორ შეიძლება მივიღოთ იგივე, პირობითი ალბათობის განმარტებიდან,

$$P(D|POS) = \frac{P(D \text{ და } POS)}{P(POS)} .$$

ვინაიდან, P(POS) = 0.257 და P(D და POS) = 48/5282 = 0.0091, ამიტომ P(D|POS) = 0.0091/0.257 = 0.035.

საბოლოოდ, დაუნის სინდრომზე დადებითი ტესტის მქონე პაციენტებიდან, სინამდვილეში დაუნის სინდრომიანი ნაყოფი აღმოაჩნდა 4% -ზე ნაკლებს. ეს რაღაც დამამშვიდებელი ამბავია იმ ქალბატონებისთვის, ვისაც ტესტზე აქვს დადებითი შედეგი.

წვდომის უნარი: რატომ უნდა გაიაროს ქალმა ეს ტესტი, როდესაც უმეტეს შემთხვევაში დადებითი პასუხი არის ცრუ დადებითი? ცხრილიდან ჩანს, რომ P(D) = 54/5282 = 0.0102, საიდანაც შეფასება იმისა, რომ 35 წლის და მეტი ასაკის ქალბატონებში დაუნის სინდრომის შანსი თითქმის 1% -ა. ცხრილიდან აგრეთვე ჩანს, რომ P(D|NEG) = 6/3927 = 0.0015, ოდნავ მეტია ვიდრე 1 ყოველი 1000-დან. ქალებს, ვისაც აქვთ უარყოფითი პასუხები, შეუძლიათ ნაკლებად ინერვიულონ დაუნის სინდრომზე, რადგანაც ამის შანსი ოდნავ მეტია ვიდრე ათასში ერთი. შეადარეთ საერთო შედეგს ერთი ასიდან.

2011 წელს მკვლევარებმა დაანონსეს დაუნის სინდრომის სისხლის ახალი ტესტის შესახებ დნმ-ის გამოყენებით. მათ აღნიშნეს, რომ აუცილებელია მეტი მუშაობა იმისთვის, რომ გაუმჯობესდეს ტესტის სიზუსტე და რომ მცირე მასშტაბიანი გამოკვლევები უნდა გაფართოვდეს მოსახლეობის უფრო ფართო გამოკვლევებამდე. (www.technologyreview.com/biomedicine/18139/pagel/).

გამრავლების წესი P(A და B) -ს საპოვნელად. როდესაც კითხულობთ, ან ისმენთ რაიმე ინფორმაციას, რომელიც იყენებს ალბათურ განაცხადებს ან დასაბუთებებს, იყავით ფრთხილად, ხომ არ არის ეს პირობითი ალბათობა, განსხვავებული რეალურისგან? როგორც წესი განცხადებების უდიდესი ნაწილი პირობითია რაღაც ხდომილებაზე და უნდათ, რომ წარმოგვიდგინონ ამ კონტექსტში. მაგალითად, საზოგადოებრივი აზრის გამოკითხვები ხშირადაა პირობითობის შემცველი, რადგან შერჩეულია სქესის, რასის ან ასაკობრივი ჯგუფების მიხდვით.

როგორც ზემოთ ვნახეთ, თუ A და B დამოუკიდებელი ხდომილებებია, მაშინ

$$P(\text{A და B}) = P(A) \times P(B).$$

პირობითი ალბათობის ცნების შემოტანამ შესაძლებელი გახადა უფრო ზოგადი ფორმულის მიღება P(A და B) -სთვის, რომელიც გამოდგება იმ შემთხვევებშიც, როცა A და B დამოუკიდებელი ხდომილებებია. გადავწეროთ თავიდან პირობითი ხდომილების განმარტება: P(A|B) = P(A და B)/P(B); გავამრავლოთ ორივე მხარე P(B) -ზე, მივიღებთ:

$$P(B) \times P(A|B) = P(B) \times P(\text{A და B})/P(B) = P(\text{A და B}),$$

ანუ

$$P(\text{A და B}) = P(B) \times P(A|B).$$

თუ A და B ხდომილებეს აქვთ მოსვლის ერთნაირი შანსი, მაშინ ერთდროულად სამართლიანია ორივე ფორმულა:

$$P(\text{A და B}) = P(B) \times P(A|B)$$

და

$$P(\text{A და B}) = P(A) \times P(B|A).$$

ამორჩევა დაბრუნებით და დაბრუნების გარეშე. შერჩევის პროცესებში ვხვდებით შემთხვევებს, როცა შერჩევა, რომელიც აღებულია პოპულაციიდან, აღარ შეიძლება, რომ უკან დაბრუნდეს და თავიდან იქნეს არჩეული. ასეთ შერჩევას ეწოდება **შერჩევა დაბრუნების გარშე.** ამ პროცესის ნებისმიერ ეტაპზე რაიმე შედეგის მოსვლის ალბათობა დამოკიდებულია წინა შედეგზე. ასეთ დროს გამოიყენება პირობითი ალბათობა. როდესაც პოპულაციის ზომა მცირეა, მაშინ პირობითი ალბათობები საგრძნობლად იცვლება, რადგან პოპულაციის ერთი ერთეულით შემცირება შესამჩნევია, მაშინ როცა დიდი მოცულობის პოპულაციისთვის ეს პროცესი თითქმის შეუმჩნეველი რჩება. ასეთ დროს შერჩევა დაბრუნებით და დაბრუნების გარეშე იძლევა თითქმის ერთი და იმავე შედეგს.

ორ A და B ხდომილებებს ეწოდება **დამოუკიდებელი**, თუ ერთი ხდომილების ალბათობა არ არის დამოკიდებული მეორე ხდომილების მოხდენაზე ან არ მოხდენაზე. ასეთ დროს მართებულია შემდეგი ფორმალური დამოკიდებულებები

$$P(A|B) = P(A),$$

ან მისი ეკვივალენტური

$$P(B|A) = P(B).$$

მაგალითად, განვიხილოთ შემთხვევა: ოჯახი ორი ბავშვით, F = გოგონა, M = ბიჭი. ელემენტარულ ხდომილებათა სივრცეს აქვს სახე: { FF, FM, MF, MM } . დავუშვათ, გვაქვს ასეთი ხდომილებები: A = პირველი ბავშვი გოგონაა, B = მეორე ბავშვი გოგონაა. მაშინ P(A) = 1/2. იგივენაირად, P(B) = 1/2. აგრეთვე, P(A და B) = 1/4. პირობითი ალბათობის განმარტებიდან გამომდინარე,

$$P(B|A) = P(\text{A და B})/P(A) = (1/4)/(1/2) = 1/2.$$

ასე რომ P(B|A) = 1/2 = P(B), ანუ A და B დამოუკიდებელი ხდომილებებია.

სცენარი 8.4. ორი ხდომილება დიაგნოსტიკური ტესტირებიდან: *(დამოუკიდებლობის შემოწმება).*

ზემოთ მოცემულ ცხრილში სისხლის ტესტის შესახებ იყო მონაცემები ნაყოფის დაუნის სინდრომით დაავადების შესახებ. თითოეული მონაცემის შეფასება ალბათურად, რომელიც დაფუძნებულია სიხშირეებზე, მოცემულია შემდეგ ცხრილში:

	სისხლის ტესტი		
სტატუსი	**დადებითი**	**უარყოფითი**	**სულ**
D	0,009	0,001	**0,010**
D*	0,247	0,742	**0,990**
სულ	**0,257**	**0,743**	**1,00**

ცხრილი 8.5. *სისხლის ტესტი.*

კითხვები შესწავლისთვის:

ა) ხდომილებები POS და D დამოუკიდებლებია თუ დამოკიდებულები?

ბ) ხდომილებები POS და D* დამოუკიდებლებია თუ დამოკიდებულები?

გაანგრება:

ა) დადებითი ტესტის ალბათობა არის 0.257. თუმცა ალბათობა იმისა რომ ტესტი დადებითია იმ პირობით, რომ ადგილი აქვს დაუნის სინდრომს, ტოლია შემდეგი სიდიდის:

$$P(POS|D) = P(POS \text{ და } D)/P(D) = 0.009/0.010 = 0.90.$$

რადგანაც $P(POS|D) = 0.90$ განსხვავდება $P(POS) = 0.257$-გან, ამიტომ ხდომილებები POS და D დამოკიდებული ხდომილებებია.

ბ) ანალოგიურად, $P(POS|D^*) = P(POS \text{ და } D^*)/P(D^*) = 0.247/0.990 = 0.250$. ეს მცირედ განსხვავდება $P(POS) = 0.257$-სგან და POS და D*აგრეთვე დამოკიდებული ხდომილებებია.

წვდომის უნარი: როგორც ველოდით, დადებითი პასუხის ალბათობა დამოკიდებულია იმაზე, აქვს თუ არა ნაყოფს დაუნის სინდრომი, და ეს არის გაცილებით მაღალი თუ ნაყოფს ეს გააჩნია. დიაგნოსტიკური ტესტი იქნებოდა უსარგებლო, თუ დაავადების სტატუსი და ტესტის შედეგი იქნებოდნენ დამოუკიდებლები.

ზოგადად, თუ A და B დამოკიდებული ხდომილებებია, მაშინ ასევეა A და B*, ასევეა A* და B, და ასევეა A* და B*. მაგალითად, თუ A დამოკიდებულია იმაზე, რომ მოხდა B, მაშინ A-ს მოხდენა აგრეთვე იქნება დამოკიდებული B-ს არ მოხდენაზე. ამიტომ თუ დავადგინეთ, რომ POS და D ხდომილებები დამოკიდებული ხდომილებებია, მაშინ POS და D* ხდომილებებიც აგრეთვე დამოკიდებულია.

ახლა უკვე შეგვიძლია კიდევ ერთხელ დავრწმუნდეთ ადრე ნაჩვენები ფორმულების მართებულობაში. დამოუკიდბელი ხდომილებებისთვის P(A და B) = P(A) × P(B). ეს არის გამრავლების წასის კერძო შემთხვევა, ანუ

$$P(A \text{ და } B) = P(A) \times P(B|A).$$

თუ A და B დამოუკიდებლები არიან მაშინ, P(B|A) = P(B) და ისევ გამრავლების წესით:

$$P(A \text{ და } B) = P(A) \times P(B).$$

რეზიუმე: დამოუკიდებლობის შემოწმება.

ამგვარად გვაქვს დამოუკიდებლობის შემოწმების სამი ხერხი:

- P(A|B) = P(A)
- P(B|A) = P(B)
- P(A და B) = P(A) × P(B).

თუ რომელიმე ერთი მათგანი ჭეშმარითია, მაშინ ჭეშმარიტია დანარჩენი ორიც და A და B დამოუკიდებელი ხდომილებებია.

სტუდენტებს ხშიარდ უჭირთ გაარჩიონ ერთმანეთისგან არათავსებადი და დამოუკიდებელი ხდომილებები. სინამდვილეში, თუ შევადარებთ მათი ხდომილებების ზუსტ მნიშვნელობებს ელემენტარულ ხდომილებათა სივრცეში, ვნახავთ, რომ ისინი სრულიად განსხვავებულნი არიან. საილუსტრაციოდ მოვიყვანოთ შემდეგი მაგალითი.

სცენარი 8.5. განსხვავება არათავსებად და დამოუკიდებელ ხდომილებებს შორის: (დამოუკიდელობის შემოწმება).

განვიხილოთ სამი ხდომილება: სიარული, საღეჭი რეზინის ღეჭვა და ფეხსაცმლის შეკვრა: W = სიარული, C = სარეჭი რეზინის ღეჭვა, T = ფეხსაცმლის შეკვრა. წარმოვიდგინოთ, რომ როცა ჯო მუშაობს კამპუსის კაფეტერიაში, ალბათობა იმისა რომ ის მოცემულ მომენტში დადის, არის 0.1; ალბათობა იმისა, რომ ის ღეჭავს საღეჭ რეზინს, არის 0.3, და ალბათობა იმისა, რომ იკრავს ფეხსაცმელს, არის 0.001. წარმოვიდგინოთ აგრეთვე, რომ W და C დამოუკიდებელი ხდომილებებია.

კითხვები შესწავლისთვის:

ა) რას ნიშნავს კონტექსტში, რომ W და C დამოუკიდებელი ხდომილებებია?

ბ) რას W და C ხდომილებათა თანაკვეთის ალბათობა? არიან თუ არა W და C არა-თავსებადი ხდომილებები?

გ) W, C და T ხდომილებებიდან რომელ წყვილს ასრულებს ჯო ერთდროულად ყველაზე ნაკლებად? რა არის ამ ორი ხდომილების თანაკვეთის ალბათობა? არიან თუ არა ისინი არათავსებადი ხდომილებები?

გააზრება:

ა) W და C ხდომილებათა დამოუკიდებლობა ნიშნავს, რომ ჯო ღეჭავს თუ არა საღეჭ რეზინს, არ არის დამოკიდებული იმაზე, დადის თუ არა იგი ამ დროს. მაგალითად, ჯო შეიძლება უბრალოდ ღეჭავს საღეჭ რეზინს, მაშინ როდესაც დადის და მაშინაც როცა არ დადის.

ბ) რადგანაც W და C დამოუკიდებელი ხდომილებებია, P(W და C) = P(W) × P(C) = 0.1×0.3 = 0.003. რადგანაც P(W და C) ≠ 0, ამიტომ ხდომილებები C და W მართალია დამოუკიდებლები არიან, მაგრამ არ არიან არათავსებადი.

გ) ამრიგად, ჩვენ ვნახეთ, რომ ჯომ შესაძლებელია თან იაროს და თან ღეჭოს საღეჭი რეზინი. ასევე შესაძლებელია შეიკრას ფეხსაცმელი და თან ღეჭოს საღეჭი რეზინი. ამიტომ P(T და C) უბრალოდ მეტი იქნება ნულზე. მაშინ როცა უბრალოდ შეუძლებელია ერთდროულად თან იაროს და თან ფეხსაცმელი შეიკრას. (თუ არ გჯერათ, სცადეთ!) ამიტომ P(W და T) = 0. ხდომილებები სიარული და ფეხსაცმლის შეკვრა დამოუკიდებელი ხდომილებებია.

ჩვდომის უნარი: სავსებით შესაძლებელა, რომ სხვა პიროვნებისთვის W და C ხდომილებები არ არიან დამოუკიდებლები. მაგალითად, წარმოვიდგინოთ, რომ ჯენიფერი ღეჭავს საღეჭ რეზინს მხოლოდ მაშინ, როცა მიდის კლასში ან ბრუნდება უკან. რეზინის ღეჭვა დამოკიდებულია თუ არა სიარულზე? ფაქტია, რომ თუ ის არ მიდის, მაშინ არ ღეჭავს რეზინს. ჯოსთვის (ან ნებისმიერი სხვისთვის), როდესაც ის მიდის ან არ მიდის, მასზე მკაცრადაა დამოკიდებული ხდომილება იკრავს თუ არა ფეხსაცმელს. თუ იკრავს ფეხსაცმელს, მაშინ მას არ შეუძლია სიარული. აგრეთვე, თუ ის დადის, მაშინ მას არ შეუძლია შეიკრას ფეხსაცმელი. ესენი არათავსებადი ხდომილებებია (მათი ერთდროულად მოხდენა შეუძლებელია). *ეს და ნებისმიერი სხვა წყვილი არათავსებადი ხდომილებებისა არ შეიძლება იყოს დამოუკიდებელი.*

მიღებული ფორმულებისგამოყენებები.

მოვლენების დამთხვევა ნამდვილად უჩვეულო ხდომილებაა?

მოვლენები, რომლებიც ხშირად გვეჩვენება შემთხვევითად, არც თუ ისე უჩვეულონი არიან, თუ მათ განვიხილავთ ისეთ კონტექსტში, როგორც ყველა შესაძლო ხდომილებები მთელ დროში. ბევრმა საკითხმა შეიძლება გამოიწვიოს შეგრძნება უჩვეული დამთხვევისა. მგალითად, ბევრ კითხვას აჩენს, თუ შეხვდით ადამიანს ვისაც აქვს თქვენი გვარი, ან დაამთავრა იგივე სკოლა, აქვს იგივე პრიფესია, იგივე დაბადების დღე და ა.შ. ეს ნამდვილად არც ისე გასაოცარია, როგორც მოგზაურობის დროს ზოგჯერ რომ ხვდებით თქვენს მეგობარს, ან ვინმეს ვისაც იცნობთ.

როდესაც გვაქვს მოსახლეობის და დროის დიდი ზომის შერჩევა, ეს მართლაც გასაოცრად გამოიყურება, თუმცა დარწმნებულები ვართ, რომ ეს მოხდება. მოვლენები, რომლებიც იშვიათია კონკრეტული ადამიანიასთვის, საკმაოდ ხშირად ხდება ადამიანთა დიდი რაოდენობის შემთხვევაში. თუ კონკრეტული შემთხვევა შეემთხვა მილიონდან ერთ ადამიანს ყოველდღიურად, მაშინ შეერთებულ შტატებში ჩვენ უნდა ველოდეთ 300 ასეთ მოვლენას დღეში და 100 000-ზე მეტს წელიწადში. მილიონიდან ერთი შანსი ყოველთვის ხდება და ჩვენ ვრჩებით გაოცებულები, თუ იგი ჩვენ შეგვემთხვევა.

თუ ახლა აიღებთ მონეტას, ააგდებთ მას 10-ჯერ, სავარუდოდ თქვენ დარჩებით გაკვირვებული, ათივეჯერ რომ დაჯდეს საფასური. მაგრამ თუ თქვენ მას აგდებთ ხანგრძლივი დროის განმავლობაში, მაშინაც იქნებით გაოცებული რაიმე მომენტში რომ მოგივიდეთ ათი საფასური ზედიზედ? შეიძლება კი, მაგრამ არ უნდა იყოთ. მაგალითად, თუ აგდებთ წესიერ მონეტას 2000-ჯერ, მაშინ თქვენ ელოდებით, რომ ყელაზე გრძელი რიგი საფასურის უნდა იყოს 10?. თუ ასეთი უჩვეულო მოვლენა

თქვენ შეგემთხვევათ, ფიქრობთ, რომ ეს გამოიყურება, ისე თითქოს 10-ჯერ მოვიდა საფასური?. ან უფრო მეტიც, ვხედავთ ზედიზედ 10 საფასურს აგდებათა გრძელ სერიაში. ეს ნამდვილად არ არის უჩვეულო მოვლენა.

იმის საილუსტრაციოდ, რომ ზოგიერთი ხდომილება მოგვეჩვენოს მოვლენად (იშვიათ დამთხვევად), რაშიც სინამდვილეში გასაოცარი არაფერია, მოდით ვუპასუხოთ ასეთ კითხვას: „რა არის ალბათობა იმისა, რომ თქვენს კლასში ორ პიროვნებას მაინც ერთ დღეს ექნება დაბადების დღე?“

სცენარი 8.6. დაბადების დღეების დამთხვევა. დავუშვათ, რომ კლასში არის 25 სტუდენტი. თუ ვიგულისხმებთ, რომ წელიწადში არის 365 დღე, ინტუიცია გვკარნახობს, რომ დაბადების დღეების დამთხვევის ალბათობა იქნება მცირე. დავუშვათ, რომ დაბადების დღეების ალბათობა ყველა სტუდენტისთვის ერთი და იგივეა და შეადგენს 1/365 (უგულებელვყოფთ ნაკიან წელს), ამასთან ყველა ხდომილება დამოუკიდებელია (გამოვრიცხავთ ტყუპების არსებობას კლასში).

კითხვები შესწავლისთვის:

ა) რა არის ალბათობა იმისა, რომ 25 სტუდენტიდან ორს მაინც ერთ დღეს აქვს დაბადების დღე?

ბ) სწორია თქვენი ინტუიცია იმის შესახებ, რომ ალბათობა მცირეა?

გააზრება:

ერთ დამთხვევაში მაინც იგულისხმება, რომ ადგილი აქვს ერთ დამთხვევას, ორ დამთხვევას და მეტს. რომ ვიპოვოთ შესაბამისი ალბათობა, ადვილია ვიპოვოთ დამატების ალბათობა, ანუ არ დამთხვევის ალბათობა. მაშინ

P(ერთი მანც დაემთხვევა) = 1 – P(არ დაემთხვევა).

დასაწყისისთვის ვიგულისხმოთ, რომ გვყავს ორი სტუდენტი. პირველის დაბადების დღე შეიძლება იყოს ნებისმიერი 365-დან. მაშინ მეორის რომ იყოს განსხვავებული, ალბათობა იქნება 364/365, ხოლო დამთხვევის ალბათობა გამოვა 1 - 364/365 = 1/365.

დავუშვათ, ახლა გვყავს სამი სტუდენტი. ალბათობა იმისა, რომ მათ აქვთ განსხვავებული დაბადების დღეები, არის

P(არ ემთხვევა) = P(სტუდენტი 1 და 2 და 3 აქვთ განსხვავებული დაბადების დღეები) = P(სტუდენტი 1 და 2 განსხვავებულია) × P(სტუდენტი 3 განსხვავებულია | სტუდენტი1 და 2 განსხვავებულია) = (364/365)×(363/365).

ლოგიკურად გავაგრძელებთ 25 სტუდენტისთვის. ასე მივალთ 25-ე სტუდენტამდე, რომლისთვისაც დაბადების დღე განსხვავდება 24 დანარჩენისგან, ანუ დარჩება 341 დღე 365 -დან. ასე, რომ

P(არ ემთხვევა) =

= P(სტუდენტი 1 და 2 და 3... და 25 აქვთ განსხვავებული დაბადების დღეები) =

= (364/365)×(363/365) × (362/365)× ... × (341/365).

ეს ნამრვალი ტოლია 0.43. თუ გამოვიყენებთ დამატების ალბათობას, მივიღებთ

P(ერთი მაინც დაემთხვევა) = 1 – P(არ დაემთხვევა) = 1 – 0.43 = 0.57.

ალბათობა უკვე აჭარბებს 1/2 -ს 25 სტუდენტში.

წვდომის უნარი: ელოდით ასეთ მაღალ ალბათობას? ეს არ უნდა იყოს გასაკვირი, თუ წარმოვიდგენთ, რომ 25 სტუდენტი ქმნის *300 წყვილს,* ვისი დაბადების დღეებიც შეიძლება რომ დაემთხვას. გახსოვთ, რა მოხდა ბევრი შესაძლებლობების შემთხვევაში? ასეთ დროს თუ ვიღაცას რაღაც შეემთხვა, ასეთი დამთხევევა არც თუ ისე გასაკვირია.

გყავთ კლასში სტუდენტი ვისი დაბადების დღეც თქვენსას ემთხვევა? არის ეს იშვიათი მოვლენა? თუ სტუდენტების რაოდენობა თქვენს კლასში 23 მაინც არის, მაშინ ალბათობა იმისა, რომ ერთი თანხვედრა მაინც იქნება, მეტია 1/2 -ზე. თუ კლასში 50 სტუდენტია, მაშინ ალბათობა უკვე 0.97-ის ტოლია. 100 სტუდენტისთვის კი შეადგენს 0.9999997 -ს (აქ არის *4950 განსხვავებული წყვილი*). მოვიყვანოთ სხვა ფაქტები დაბადების დღეების დამთხვევაზე, რამაც შეიძლება გაგაკვირვოთ:

- 88 ადამიანში ალბათობა იმისა, რომ სამს მაინც აქვთ საერთო დაბადების დღე, არის1/2.
- 14 ადამიანში ალბათობა იმისა, რომ ორს მაინც აქვთ დაბადების დღეები ერთი დღის დაშორებით, არის 1/2.

9

შემთხვევითი სიდიდე. დისკრეტული შემთხვევითი სიდიდის განაწილების კანონი. უწყვეტი ტიპის შემთხვევითი სიდიდის განაწილების კანონი. შემთხვევითი სიდიდის რიცხვითი მახასიათებლები.

შემთხვევითი ცვლადი (ან **შემთხვევითი სიდიდე**) არის შემთხვევითი მოვლენის რიცხვითი საზომი. ცვლადების გამოსახვისთვის ჩვენ ვისარგებლებთ ანბანის ბოლო ასოებით, როგორიცაა მაგალითად x. აგრეთვე ვისარგებლებთ x ცვლადით შემთხვევითი სიდიდის (ცვლადის) შესაძლო მნიშვნელობის აღსანიშნავად. როდესაც ვსაუბრობთ შემთხვევით სიდიდეზე თავისთავად და არა მის რაიმე კონკრეტულ მნიშვნელობაზე, მაშინ ვიხმართ დიდ ასოებს, როგორიცაა ისევ ვთქვათ X. მაგალითად, X = მონეტის სამი აგდების დროს საფასურის მოსვლათა რაოდენობა. ეს არის შემთხვევითი სიდიდე. ხოლო x=2 არის მისი ერთერთი შესაძლო მნიშვნელობა.

რადგანაც შემთხვევითი სიდიდე აღწერს შემთხვევით მოვლენებს, ამიტომ თითოეულ შედეგს აქვს თავისი მოსვლის ალბათობა. შემთხვევითი სიდიდის **ალბათობის განაწილება** (ანუ შემთხვევითი სიდიდის განაწილების კანონი) განსაზღვრავს შესაძლო მნიშვნელობებს და მათ შესაბამის ალბათობებს. უმჯობესია, რომ ცვლადი იყოს შემთხვევითი, რადგან ასეთ დროს შესაძლებელი ხდება ალბათობების განსაზღვრა. შემთხვევითობის გარეშე ჩვენ არ შეგვეძლება, შორეულ პერსპექტივაში პროგნოზირება გავაკეთოთ შესაძლო შედეგის ალბათობაზე.

დისკრეტული შემთხევითი სიდიდის განაწილების კანონი

როდესაც შემთხვევითი ცვლადი ღებულობს იზოლირებულ მნიშვნელობებს, როგორც იყო მაგალითად, მონეტის აგდების შემთხვევაში 0, 1, 2, 3 საფასურის მოსვლათა რაოდენობა, ეწოდება **დისკრეტული შემთხვევითი სიდიდე**. ასეთ დროს განაწილების კანონი დიკრეტული სიდიდის ყოველ მნიშვნელობას შეუსაბამებს მის ალბათობას. ყოველი ალბათობა ვარდება 0 და 1 მნიშვნელობებს შორის და ყველა შესაძლო მნიშვნელობათა ალბათობების ჯამი ტოლი უნდა იყოს 1-ის. P(x)-ით აღვნიშნოთ ის ალბათობა, რომელიც შეესაბამება x-ის შესაძლო მნიშვნელობას. მაგალითად, P(2) აღნიშნავს ალბათობას, როცა შემთხვევითი სიდიდე ღებულობს მნიშვნელობას 2-ს. მაგალითად, ვთქვათ, X არის კამათლის გაგორებისას მოსული

შედეგები. მისი შესაძლო მნიშვნელობები იქნება x = 1, 2, 3, 4, 5, 6 და P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.

შემთხვევითი სიდიდე შეიძლება იყოს **უწყვეტი**, თუ მისი შემთხვევითი მნიშვნელობები წარმოადგენს ინტერვალს და არა ცალკეული რიცხვების ჯგუფს.

შემთხვევითი სიდიდის საშუალო

გავიხსნოთ, რომ **პარამეტრი** არის პოპულაციის რიცხვითი მახასიათებელი, როგორიცაა, მაგალითად, საშუალო (ანუ მათემატიკური ლოდინი), მედიანა, დისპერსია, სტანდარტული გადახრა და სხვა. განაწილების კანონის აღწერისთვის ჩვენ ვისარგებლებთ ნებისმიერი რიცხვითი მახასიათებლით. მათ შორის ყველაზე გავრცელებულია *საშუალო*, რომელიც ნიშნავს ცენტრის პოვნას და *სტანდარტული გადახრა,* რომელიც აღწერს ცვალებადობას.

წარმოვიდგინოთ, რომ ჩვენ ხშირად ვაკვირდებთ შემთხვევითი სიდიდის მნიშვნელობებს, როგორიცაა, ვთქვათ, ვაკვირდებით შედეგებს, როცა ვაგორებთ კამათელს. შემთხვევითი სიდიდის განაწილების საშუალო არის მნიშვნელობა, რომელსაც მივიღებთ მნიშვნელობათა საშუალოს გამოთვლით.

დისკრეტული **შემთხვევითი სიდიდის საშუალო** (მათემატიკური ლოდინი) არის:

$$\mu = \sum xP(x),$$

სადაც ჯამი აიღება x-ის ყველა შესაძლო მნიშვნელობების მიმართ.

განვიხილოთ კერძო შემთხვევა, როდესაც ყველა შედეგი ტოლალბათურია. მაგალითად, n არის სხვადასხვა შედეგების რაოდენობა და თითოეულის ალბათობა არის 1/n. თუ n = 6, როგორც ეს არის კამათლის გაგორების დროს, მაშინ თითოეული ალბათობა ტოლია 1/6 -ს და მაშინ,

$$\mu = \sum xP(x) = \sum x \times \frac{1}{n} = (\sum x)/n\ .$$

ფორმულა $\sum x \times \frac{1}{n}$ იგივეა რაც ჩვეულებრივად შერჩევის საშუალო. მაშინ ფორმულა $\mu = \sum xP(x)$ არის წინა ფორმულის განზოგადება იმისთვის, რომ სხვადასხვა შედეგები, რომელთაც აქვთ სხვადასხვა ალბათობები, გამოვიყენოთ ალბათობის განაწილება და შერჩევის მონაცემები.

შემთხვევითი X ცვლადის ალბათობის განაწილების საშუალოს ეწოდება X-ის მოსალოდნელი სიდიდე (ან მათემატიკური ლოდინი). იგი გამოსახავს არა იმას, რასაც ვაკვირდებით ერჯერადად, არამედ იმას, თუ რას ველოდებით საშუალოდ გრძელვადიან პერსპექტივაში.

სცენარი 9.1. რისკების საპასუხოდ.

ხართ თქვენ რისკებისადმი მიდრეკილი პიროვნება, ვინც უპირატესობას აძლევს, რომ იყოს დარწმუნებული უკეთეს შედეგში? თუ ხართ ისეთი რისკის მოყვარული, რომ მზად ხართ ითამაშოთ უკეთესი შედეგის მიღების იმედით?

კითხვები შესწავლისათვის:

ა) თქვენ მოცემული გაქვთ $1000 ინვესტიცია. თქვენ უნდა აირჩიოთ ერთერთი

(i) რომ ნაღდი მოგება $500, ან

(ii) 0.5 არის შანსი რომ მოიგოთ $1000 და 0.5, რომ ვერაფერი ვერ მოიგოთ.

რა არის მოსალოდნელი მოგება თითოეულ სტრატეგიაში? რომელს ანიჭებთ უპირატესობას?

ბ) თქვენ მოცემული გაქვთ $2000 ინვესტიცია. თქვენ უნდა აირჩიოთ ერთერთი (i) დარწმუნებული ხართ, რომ დაკარგავთ $500, ან

(ii) 0.5 ალბათობით კარგავთ $1000-ს და 0.5 ალბათობით არაფერს არ კარგავთ.

რა არის მოსალოდნელი წაგება თითოეულ სტრატეგიაში? რომელს ანიჭებთ უპირატესობას?

გაანგრებისთვის:

ა) მოსალოდნელი მოგება დარწმუნებული სტრატეგიით არის $500 (i), რადგან ამ სტრატეგიას აქვს $500 მოგება 1.0 ალბათობით. სტრატეგიაში მოგება-რისკით (ii) ალბათობის განაწილება მოცემულია ცხრილით:

მოგება x	***P(x)***
0	0,5
1000	0,5

ცხრილი 9.2. *რისკის გაწევის ალბათობის განაწილება.*

მოსალოდნელი მოგება $\mu = \sum xP(x) = \$0 \times 0.5 + \$1000 \times 0.5 = \$500$.

ბ) მოსალოდნელი წაგება დარწმუნებული სტრატეგიით არის $500 (i). სტრატეგიაში მოგება-რისკით (ii) მოსალოდნელი წაგება $\mu = \sum xP(x) = \$0 \times 0.5 + \$1000 \times 0.5 = \$500$.

ჩვდომის უნარი: თითოეულ შემთხვევაში მოსალოდნელი მნიშვნელობა არის ერთი და იგივე თითოეული სტრატეგიისთვის. თუმცა ადამიანების უდიდესი ნაწილი უპირატესობას ანიჭებს დარწმუნებულ სტრატეგიას (i) **ა)** შემთხვევაში, და რისკის შემცველ სტრატეგიას (ii)-ს **ბ)** შემთხვევაში. ისინი რისკს თავს არიდებენ **ა)** შემთხვევაში, მაგრამ რისკავენ **ბ)** შემთხვევაში. ეს უპირატესობა შესწავლილ იქნა კვლევაში Daniel Kahneman და AmosTversky -ის მიერ,რომელისთვისაც Kahneman -მა მიიღო ნობელისპრემია 2002 წელს.

ალბათობის განაწილების ცვალებადობის რეზიუმე

ისევე როგორც მონაცემთა შერჩევის განაწილების შემთხვევაში, სასარგებლოა, ალბათობის განაწილებისთვისაც დავაკვირდეთ ცვალებადობას და ცენტრს. წარმოვიდგინოთ, რომ ორ სხვადასხვა სანვესტიციო სტრატეგიას აქვს ერთი და იგივე მოსალოდნელი მოგება. აქვს თუ არა ერთერთ სტრატეგიას ამონაგების მეტი ცვალებადობა?

ალბათობის განაწილების **სტანდარტული გადახრა** აღინიშნება σ ასოთი და იგი არის საშუალოდან გადახრის საზომი. σ-ს დიდი მნიშვნელობა შეესაბამება დიდ ცვალებადობას. უხეშად რომ ვთქვათ, σ აღწერს საშუალოსგან რამდენად შორს ვარდებიან შემთხვევითი სიდიდეები. ჩვენ ჯერ არ გვინდა მოვიყვანოთ σ-ს გამოსათვლელი ფორმულები. ეს საქმე ცოტა ხნით შემდეგისთვის გადავდოთ.

სცენარი 9.3. რისკის გაწევა იწვევს მეტ ცვალებადობას. დავუბრუნდეთ წინა სცენარს. მოცემული გაქვთ $1000 ინვესტიცია და თქვენ უნდა აირჩიოთ ერთერთი (i) ნაღდი მოგება $500, ან (ii) 0.5 არის შანსი რომ მოიგოთ $1000 და 0.5, რომ ვერაფერი ვერ მოიგოთ. ქვემოთ მოყვანილ ცხრილში ნაჩვენებია X -ის მოგების ალბათობის განაწილების ორივე სტრატეგიისთვის.

დარწმუნებული სტრატეგია		*დარწმუნებული სტრატეგია*	
x	P(X)	x	P(X)
500	1.0	0	0.5
		1000	0.5

ცხრილი 9.4. *X = მოგების ალბათობის განაწილება.*

კითხვები შესწავლისათვის:

როგორც ვნახეთ, ორივე სტრატეგიას აქვს ერთი და იგივე საშუალო, კერძოდ $500. ამ ორი ალბათობის განაწილებიდან რომელ მათგანს ექნება უფრო დიდი სტანდარტული გადახრა?

გააზრებისთვის:

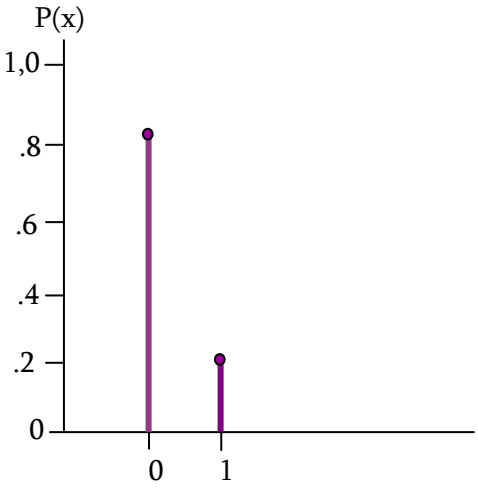
დარწმუნებული სტრატეგიისთვის არა გვაქვს არანაირი ცვალებადობა. სტანდარტული გადახრა ამ განაწილებისთვის არის 0, უმცირესი მნიშვნელობა. სარისკო სტრატეგიისთვის კი არ არის დამოკიდებული შედეგზე ($0 იქნება თუ $1000) შედეგი იქნება $500 გადახრა საშუალო $500-დან. მართლაც, სარისკო სტრატეგიის ალბათობის განაწილების სტანდარტული გადახრა არის $500. ამ სტრატეგიაში არის ძალიან დიდი ცვალებადობა, რაც კი შეიძლება მომხდარიყო.

ჩვდომის უნარი: პრაქტიკაში სხვადასხვა საინვესტიციო სტრატეგიებს ხშირად ადარებენ ხოლმე მათი ცვალებადობებით. ინვესტორი დაინტერესებულია არა მხოლოდ ინვესტიციების მოსალოდნელი დაბრუნებით, არამედ სარგებლის თანმიმდევრულობითაც წლიდან წლამდე.

ალბათობის განაწილება კატეგორიული ცვლადებისთვის

აქამდე განხილულ მაგალითებში ცვლადები იყვნენ უფრო მეტად რაოდენობრივი, ვიდრე კატეგორიული. ნაწილობრივ ეს ხდება იმიტომ, რომ შემთხვევითი სიდიდე არის *რიცხვითი* საზომი შემთხვევითი მოვლენებისთვის. თუმცა, ჩვენვ ხედავთ, რომ კატეგორიული ცვლადებისთვის, გვაქვს მხოლოდ ორი კატეგორია, ამიტომ სასარგებლოა ეს ორი შესაძლო შედეგი წარმოვადგინოთ რიცხვითი სიდიდეებით, დავუშვათ, 0 -ით და 1 -ით.

მაგალითად, წარმოვიდგინოთ, რომ თქვენ ჩაატარეთ მარკეტინგული კვლევა ბევრ პოტენციურ მომხმარებელთან იმისთვის, რომ შეაფასოთ ალბათობა იმისა, იყიდის თუ არა მომხმარებელი ახალ პროდუქციას, რომელიც თქვენ აწარმოეთ და აპირებთ გაყიდვას. თქვენმა კვლევამ მოგცათ შეფასება, რომ ალბათობა არის 0.2. თუ თქვენ შესაძლო შედეგებს (წარმატება, მარცხი) აღნიშნავთ (1, 0)-ით, ანუ მომხმარებელმა იყიდა თუ არ იყიდა პროდუცია, მაშინ X -ის შედეგების ალბათობის განაწილება არის ქვემოთ მოცემული ცხრილის სახით. ამ მონაცემების გამოსახვა შესაძლებელია გრაფიკულადაც.

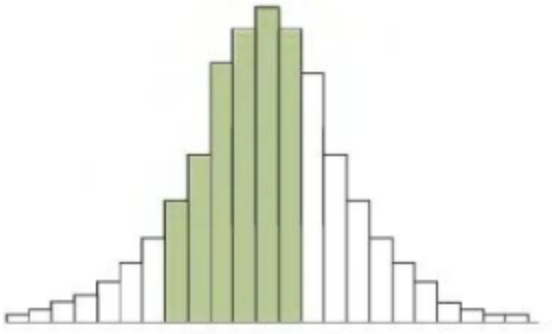


x	P(x)
0	0,8
1	0,2

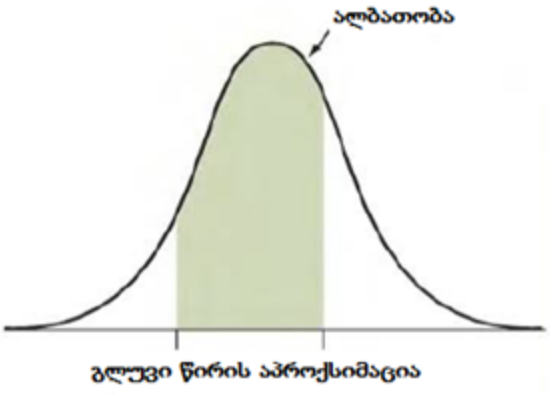
ალბათობის განაწილების საშუალო არის:

$$\mu = \sum xP(x) = 0 \times 0.8 + 1 \times 0.2 = 0.2.$$

საშუალო არის წარმატების ალბათობა. *შემთხვევითი ცვლადისთვის, რომლის შესაძლო მნიშვნელობებია 0 და 1, საშუალო არის იმ შედეგის ალბათობა, რომელიც აღნიშნულია 1 -ით.*



ბევრი ინტერვალი



გლუვი წირის აპროქსიმაცია

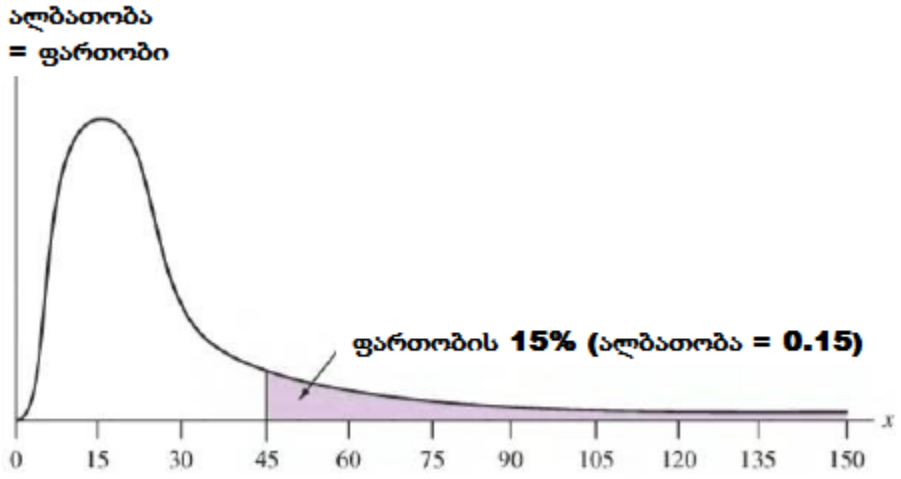
უწყვეტი შემთხვევითი ცვლადის ალბათობის განაწილება

შემთხვევით სიდიდეს ეწოდება **უწვეტი**, თუ მისი შესაძლო მნიშვნელობები აიღება ინტერვალიდან. მაგალითად, ბოლო დროის კვლევების მიხედვით, რომელიც ჩაატარა აშშ მოსახლეობის აღწერის ბიურომ, ანალიზი ჩაატარა, თუ რა დროს ანდომებს ხალხი სამუშაო ადგილამდე მისვლას. გადაადგილების დრო შეიძლება გაიზომოს ნამდვილი რიცხვებით 0-დან 150 წუთამდე.

უწყვეტი შემთხვევითი სიდიდის ალბათობის განაწილება ნიშნავს, მივანიჭოთ ყველა შესაძლო მნიშვნელობათა ინტერვალებს შესაბამისი ალბათობები. მაგალითად, გადაადგილების დროზე დახარჯული ალბათობის განაწილება უზრუნველყოფს ალბათობას იმისას, რომ გადადგილებაზე დახარჯული დრო არ აღემატება 15 წუთს ან მგზავრობის დრო არის 30-სა და 60 წუთს შორის. ალბათობა იმისა, რომ შემთხვევითი ცვლადი ვარდება კონკრეტულ ინტერვალში არის 0-სა და 1-ს შორის, და ალბათობა ინტერვალისთვის, რომელიც შეიცავს ყველა შესაძლო ინტერვალს = 1.

თუ შემთხვევითი სიდიდე უწვეტია, მაშინ მნიშვნელობათა ინტერვალები ჰისტოგრამისთვის შეიძლება შერჩეულ იქნას სურვილის მიხედვით. მაგალითად, ერთი შესაძლო გადაადგილების დრო არის {0-დან 30-მდე, 30-დან 60-მდე, 60-დან 90-მდე, 90-დან 120-მდე, 120-დან 150-მდე }, რაც საკმაოდ დიდი ინტერვალებია. ამისგან განსხვავებით შეიძლება ავიღოთ საკმაოდ მცირე ინტერვალები { 0-დან 1-მდე, 1-დან 2-მდე, … ,149-დან 150-მდე }. თუ ინტერვალების რაოდენობას თანდათან გავზრდით, მაშინ მათი ზომები დაწვრილდება და ჰისტოგრამის ფორმა თანდათან მიუახლოვდება გლუვ წირს. ჩვენ ვისარგებლებთ ასეთი წირებით უწყვეტი შემთხვევითი სიდიდის ალბათობის განაწილების გამოსასახავად.

ქვემოთ მოცემული გრაფიკი გვიჩვნებს ალბათობის განაწილებას უწყვეტი X ცვლადისთვის. X = აშშ-ში ადამიანების სამსახურში მისასვლელად დახარჯული დრო. ისტორიულად გამოკითხვებისას ხალხი თანახმა იყო სამსახურში მისასვლელად დაეხარჯა მაქსიმუმ 45 წუთი. თუმცა, 2000 წლის ამერიკის აღწერის ბიუროს ანგარიშ შინათქვამია, რომ აშშ-ს იმ მშრომელი მოსახლეობის 97% -დან, რომლებიც არ მუშაობენ სახლში, 15% -ს სამსახურამდე მისასვლელად სჭირდება 45 წუთზე მეტი.



გრაფიკი 9.5. *გადაადგილების დროის ალბათობის განაწილება.* გრაფიკის ქვეშ მოთავსებული ფიგურის ფართობი იმ მნიშვნელობებისთვის, რომლებიც მეტია 45 წუთზე, არის 0.15.

კითხვა: მიუთითეთ გრაფიკის ქვეშ მოთავსებული ფიგურის ნაწილი, რომელიც შეესაბამება 15 წუთზე ნაკლებ დროს და რომელიც ტოლია 0.29 .

ფიგურაზე დაშტრიხული ფიგურის ფართობი შეესაბამება იმ მნიშვნელობათა ალბათობას, რომლებიც მეტია 45-ზე. ეს ფართობი წირის ქვეშ მოთავსებული მთელი ფართობის 15%-ის ტოლია. ამიტომ ალბათობა უდრის 0.15-ს. მოხსენებაში ნათქვამია, რომ ალბათობის განაწილების საშუალოა 25.5. (მას შემდეგ რაც გამოქვეყნდა 2010 წლის მონაცემები, შედარებისას შესაძლოა, პროცენტული მაჩვენებელი შეიცვალოს).

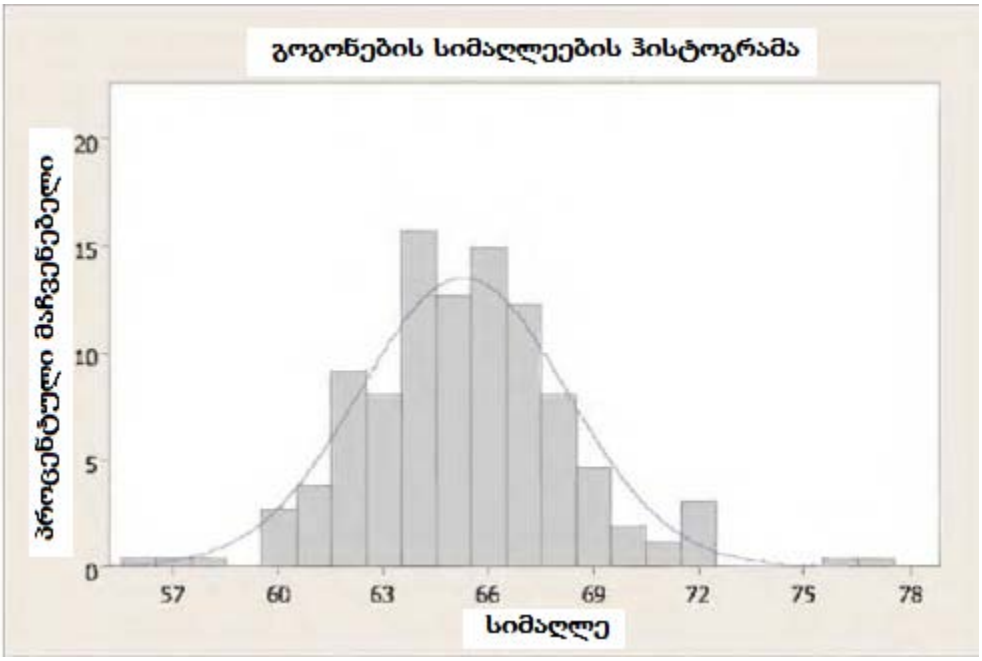
უწყვეტი შემთხვევითი სიდიდის შემთხვევაში უნდა დავამრგვალოთ გაზომვები. ალბათობები მოცემულია ინტერვალებისთვის და არა ცალკეული მნიშვნელობები-სთვის. გადადგილებისთვის საჭირო დროის შეფასებისას აშშ მოსახლეობის აღწერის ბიურომ დასვა კითხვა: „ზოგადად რამდენი წუთი დაგჭირდათ სახლიდან სამსახურში მისვლამდე გასულ კვირას?“ და მიიღო შესაძლო პასუხები 0, 1, 2, 3, ... ამ შემთხვევაში, მაგალითად, 24 წუთი სინამდვილეში აღნიშნავს ნამდვილ რიცხვებს, რომლებიც მთელამდე დამრგვალებისას იძლევა 24-ს. ეს არის წირის ქვემოთ მოთავსებული ფიგურის ფართობი 23.5-სა და 24.5 მნიშვნელობებს შორის. *პრაქტიკულად, ალბათობა იმისა, რომ გადაადგილების დროის მნიშვნელობა ვარდება რაღაც ინტერვალში, როგორიცაა 45 წუთის ზემოთ, ან 30-სა 60 წუთს შორის, უფრო მეტად საინტერესოა, ვიდრე ალბათობა რომელიმე კონკრეტული მნიშვნელობისთვის.*

უწყვეტი შემთხვევითი სიდიდის შემთხვევაში, ფართობი რომელიმე კონკრეტული მნიშვნელობის ზემოთ ტოლია ნულის. მაგალითად, ალბათობა იმისა, რომ მისვლის დრო ტოლია 23.7693611045... არის 0, მაგრამ ალბათობა იმისა, რომ ეს დრო მოთვასებულია 23.5 და 24.5 წუთებს შორის არის დადებითი.

პრაქტიკაში უწყვეტი ცვლადები იზომებიან დისკრეტულის მსგავსად, დამრგვალების გამო. დამრგვალებებისას უწყვეტმა შემთხვევითმა სიდიდემ შეიძლება მიიღოს ბევრი ცალკეული მნიშვნელობა. უწყვეტი შემთხვევითი სიდიდის ალბათობის განაწილება გამოიყენება შესაძლო დამრგვალებული მნიშვნელობების აპროქსიმაციისთვის (მიახლოებისთვის).

სცენარი 9.6. სიმაღლე.*(ალბათობის განაწილება).*

შემდეგ გრაფიკზე ჰისტოგრამა გვიჩვენებს ჯორჯიის უნივერსიტეტში სტუდენტი გოგონების სიმაღლეებს, რომელსაც ფარავს გლუვი წირი. (*წყარო:* ჯორჯიის უნივერსიტეტის მონაცემები.)



გრაფიკი 9.7.*სტუდენტი გოგონების ჰისტოგრამა დაფარულია ზარის-ფორმის წირით.*

სიმაღლე არის უწყვეტი, მაგრამ გაზომილია როგორც დისკრეტული სიდიდე, რადგანაც დამრგვალებულია უახლოეს დუიმებამდე. გლუვი წირი არის აპროქსიმაცია - სიმაღლის ალბათობის განაწილება და განიხილავს მას როგორც უწყვეტ შემთხვევით სიდიდეს.

კითხვა: როგორ აღწერთ განაწილების ცვალებადობის ფორმას და ცენტრს?

კითხვა შესწავლისათვის: რას წარმოადგენს გლუვი წირი?

გააზრებისთვის: თეორიაში სიმაღლე არის უწყვეტი შემთხვევითი სიდიდე, პრაქტიკაში, ის იზომება უახლოეს დუიმებამდე დამრგვალებით. მაგალითად, ჰისტოგრამის სვეტი 64-ს ზემოთ (თავზე) წარმოადგენს სიმაღლეებს 63.5-დან 64.5-მდე, რომლებიც დამრგვალებულია 64 დუიმამდე. ჰისტოგრამა გვაძლევს მონაცემთა განაწილებას. ფართობი გლუვი წირის ქვეშ ორ წერტილს შორის არის მიახლოებითი ალბათობა იმისა, რომ სიმაღლის მნიშვნელობა ვარდება ამ ორ წერტილს შორის.

ჩვდომის უნარი: ჰისტოგრამა არის შერჩევის გრაფიკული წარმოდგენა. გლუვი წირი არის პოპულაციის გრაფიკული წარმოდგენა. როგორც ვხდავთ, მას ისევ ზარის ფორმა აქვს. შემდეგ ნაწილში შევისწავლით, ალბათობის განაწილებებს ასეთი ფორმებისთვის და ვნახავთ, თუ როგორ უნდა გამოვთვალოთ მათთვის ალბათობები. ამისთვის ზოგჯერ გამოვიყენებთ ელემენტარულ ხდომილებათა სივრცეს, ზოგჯერ ეს შეიძლება მოხერხდეს ფორმულების გამოყენებით, ზოგჯერ ცხრილების ან გრაფიკების დახმარებით. ზოგჯერ კი აუცილებელი ხდება განაწილების აპროქსიმაციის მოდელირება.

10

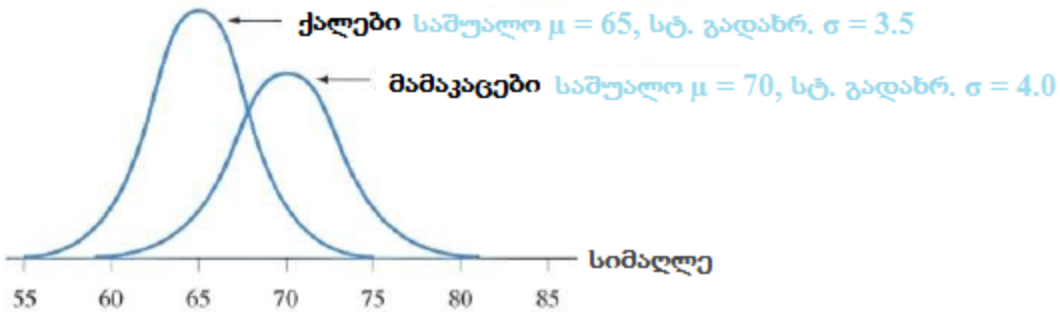
ნორმალური განაწილება. ალბათობების გამოთვლა ნორმალური განაწილებისთვის. ალბათობების გამოთვლა სტანდარტული გადახრის საშუალებით.

ზოგიერთი ალბათობის განაწილება იმსახურებს განსაკუთრებულ ყურადღებას, რადგანაც ისინი მეტად სასარგებლონი არიან ძალიან ბევრ გამოყენებებში. მათთვის გვაქვს ფორმულები, ცხრილები, რომლებიც განსაზღვრავენ ყველა შესაძლო შედეგების ალბათობებს. ჩვენ შევისწავლით ე.წ. **ნორნალურ განაწილებას**, რომელიც ზოგადად გამოიყენება უწყვეტი ტიპის შემთხვევითი სიდიდეების გაანალიზებისას. იგი ხასიათდება კონკრეტული სიმეტრიული ზარის-ფორმის წირით, რომელიც დამოკიდებულია ორ პარამეტრზე μ საშუალოსა (მათემატიკურ ლოდინსა) და σ სტანდარტულ გადახრაზე.

ნორმალური განაწილება:

ალბათობის განაწილება ზარის-ფორმის წირის საშუალებით.

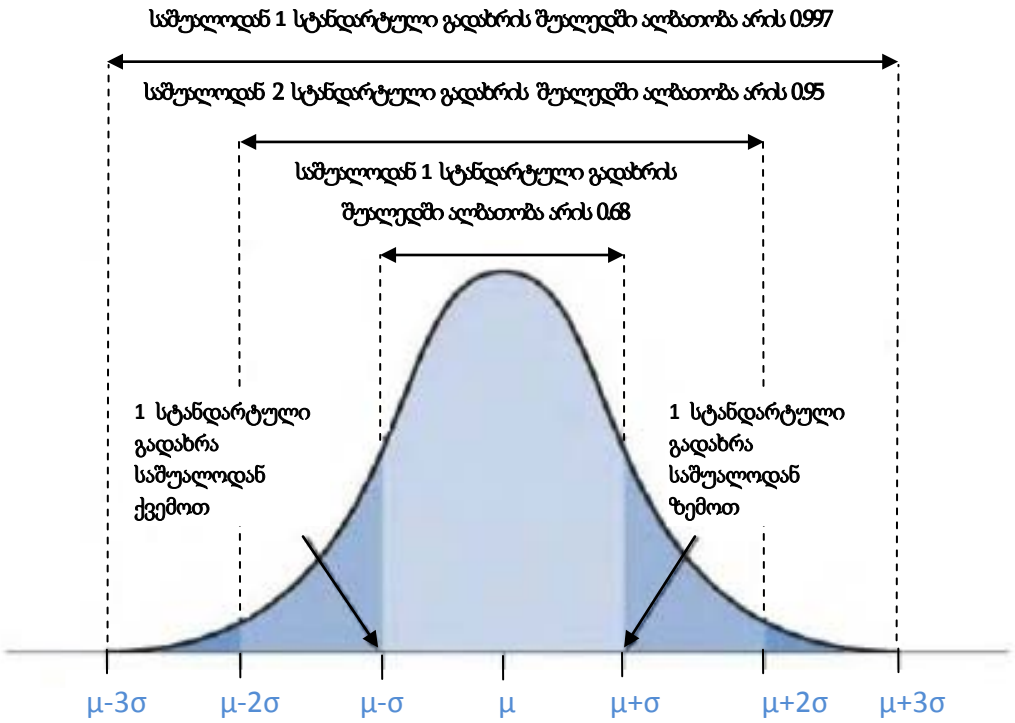
წინა ლექციის ბოლოს ჩვენ ვნახეთ, რომ ჯორჯიის უნივერსიტეტის სტუდენტი გოგონების სიმაღლეთა განაწილებას ჰქონდა დაახლოებით ზარის-ფორმის განაწილება. აპროქსიმაციის წირი აღწერს ალბათობის განაწილებას, რომლის საშუალოა 65.0 დუიმი, ხოლო სტანდარტული გადახრა - 3.5 დუიმი. ფაქტობრივად, მოზრდილი ქალბატონების სიმაღლეებს ჩრდილოეთ ამერიკაში დაახლოებით აქვს ნორმალური განაწილება $\mu = 65.0$ დუიმი, $\sigma = 3.5$ დუიმი სტანდარტული გადახრით. მოზრდილი მამაკაცების სიმაღლეებსაც აქვთ დაახლოებით ნორმალური განაწილება $\mu = 70.0$ დუიმი, $\sigma = 4.0$ დუიმი სტანდარტული გადახრით. როგორც წესი, ზრდასრული მამაკაცები საშუალოდ უფრო მაღლები არიან (რადგანაც $70 > 65$) და სიმაღლის ცვალებადობაც საშუალოსთან შედარებით ოდნავ მეტია ვიდრე ქალებში (რადგანაც $4.0 > 3$).



გრაფიკი 10.1 *ქალების და კაცების სიმაღლეთა ნორმალური განაწილებები.* გრაფიკები მოცემულია μ და σ პარამეტრების და სხვადასხვა მნიშვნელობებისთვის.

კითხვა: მოცემულია, რომ μ = 70 და σ = 4; რომელ ინტერვალში ვარდება მამაკაცების სიმაღლეთა თითქმის ყველა მნიშვნელობა?

ნორმალური განაწილება არის სიმეტრიული, ზარის-ფორმის და ხასიათდება მისი μ საშუალოთი და σ სტანდარტული გადახრით. μ საშუალოდან σ-ს რაიმე კონკრეტულ რიცხვჯერ გადანაცვლებულ შუალედში ალბათობა არის ერთი და იგივე ყველა ნორმალური განაწილებისთვის. ეს ალბათობა ტოლია 0.68-ის ერთი სტანდარტული გადახრის შუალედში, 0.95 ორი სტანდარტული გადახრის შუალედში და 0.997 სამი სტანდარტული გადახრისთვის. უკეთ წარმოდგენისთვის დავაკვირდეთ 10.2 გრაფიკს.



გრაფიკი 10.2. *ნორმალური განაწილება.*

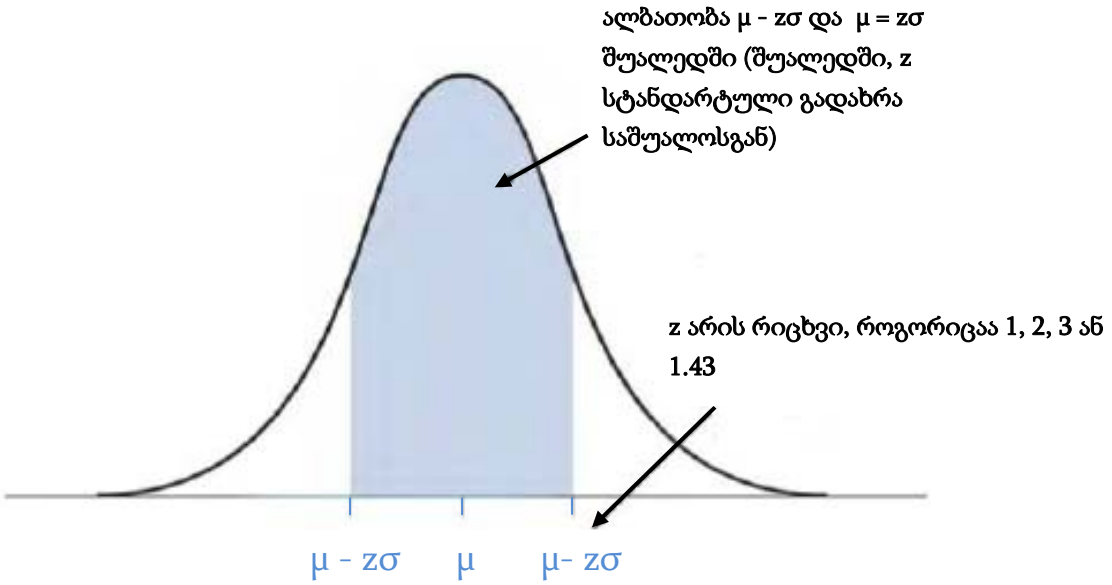
კითხვა: როგორაა ეს ალბათობები დაკავშირებული ემპირიულ წესთან?

ქალბატონების სიმაღლეებისთვის $\mu = 65.0$ და $\sigma = 3.5$ დუიმს. ამიტომ

$$\mu - 2\sigma = 65.0 - 2 \times 3.5 = 58.0 \quad \text{და} \quad \mu + 2\sigma = 65.0 + 2 \times 3.5 = 72.0,$$

ე. ი. ქალბატონების სიმაღლეთა დაახლოებით 95% ვარდება 58 დიუმსა და 72 დიუმს (6 ფუტი) შორის. მამაკაცთა სიმაღლეებისთვის $\mu = 70.0$ და $\sigma = 4.0$ დიუმი, დაახლოებით 95% ვარდება $\mu - 2\sigma = 70.0 - 2 \times 4.0 = 62$ დუიმი და $\mu + 2\sigma = 70.0 + 2 \times 4.0 = 78$ დუიმი (6.5 ფუტი).

1, 2 და 3-ის ჯერადად სტანდარტული გადახრები საშუალოსგან ზოგადად განისაზღვრებიან z სიმბოლოს გამოყენებით. მაგალითად, 2 სტანდარტული გადახრისთვის $z = 2$. z-ის ყოველი ფიქსირებული მნიშვნელობისთვის, z სტანდარტული გადახრა საშუალოსგან არის წირის ქვემოთ მოთავსებული ფიგურის ფართობი $(\mu - z\sigma, \mu + z\sigma)$ შუალედში, როგორც ეს ნაჩვენებია 10.3 ნახაზზე. ყოველი ნორმალური განაწილებისთვის, როცა $z = 1$, ალბათობა არის 68%. ასევე, $z = 2$ -სთვის ალბათობა არის 95% და ალბათობა არის დაახლოებით 1, როცა $z = 3$, რაც შეესაბამება $(\mu-3\sigma, \mu+3\sigma)$ შუალედს. მთლიანად ალბათობა კი ნებისმიერი ნორმალური განაწილებისთვის, ცხადია, ტოლია 1-ის.



გრაფიკი 10.3. *ალბათობა $\mu - \sigma$ და $\mu + \sigma$ შუალედში.* ეს ფართობი მონიშნულია ნახაზზე. იგი საერთოა ყველა ნორმალური განაწილებისთვის და დამოკიდებულია მხოლოდ z-ის მნიშვნელობაზე. z = 1, 2 და 3 მნიშვნელობებისთვის ნაჩვენებია ფრაფიკი 10.2-ზე. შევნიშნოთ, რომ z არ არის მხოლოდ მთელი რიცხვი, იგი შეიძლება იყოს ნებისმიერი ნამდვილი რიცხვი.

ნორმალური განაწილება არის ყველაზე მნიშვნელოვანი განაწილება სტატისტიკაში, რადგანაც ბევრ ცვლადს მიახლოებით ნორმალური განაწილება აქვს. ნორმალური განაწილება კიდევ იმითაა მნიშვნელოვანი, რომ იგი არის კარგი აპროქსიმაცია ბევრი დიკრეტული განაწილებისთვის, როდესაც შესაძლო შედეგების რაოდენობა არის საკმარისად ბევრი. ძირითადი მიზეზი ნორმალური განაწილების პოპულარობისა არის ის, რომ ბევრი სტატისტიკური მეთოდი იყენებს მას მაშინაც კი, როდესაც მონაცემებს არა აქვთ ზარის ფორმა. თუ რატომ, ამას მალე ვნახავთ.

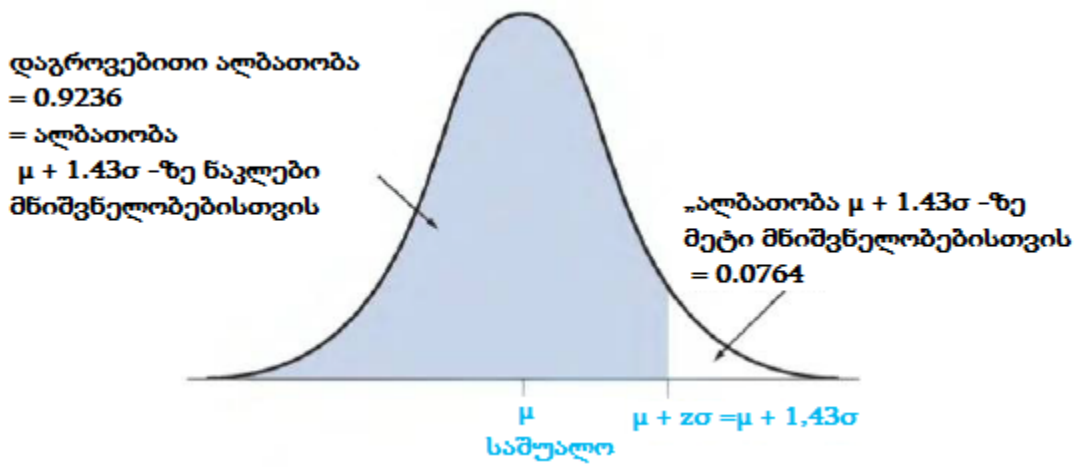
ნორმალური განაწილების ალბათობების გამოთვლა.

როგორც ვნახავთ, ალბათობები 0.68, 0.95 და 0.997 შუალედებში 1, 2 და 3 სტანდანდარტული გადახრებისთვის საშუალოდან არ არის გასაკვირი, მაგრამ რა მოხდება, თუ გვინდა, რომ გამოვთვალოთ ალბათობა ვთქვათ, 1.43 სტანდარტული გადახრისთვის?

შემდეგი A ცხრილი საშუალებას გვაძლევს გამოვთვალოთ ნორმალური განაწილების ალბათობები. იგი ითვლის დაგროვებით ალბათობას, ალბათობას იმისას, რომ მნიშვნელობა ჩავარდება $\mu + z\sigma$-ის ქვემოთ (ანუ იქნება მასზე ნაკლები). A ცხრილის უკიდურეს მარცხენა სვეტში მოცემულია z -ის მნიშვნელობები მეათედის სიზუსტით, მეასედები კი მითითებულია ჩამოთვლილი სვეტების თავში. მოყვანილი ცხრილი არის პატარა ფრაგმენტი A ცხრილის. ალბათობა z = 1.43 მნიშვნელობისთვის უნდა მოვძებნოთ 1.4-ის შემცველი სტრიქონისა და 0.03 სვეტის თანაკვეთაზე. იგი ტოლია 0.9236-ის. ყველა ნორმალური განაწილებისთვის ალბათობა იმისა, რომ მნიშვნელობა ნაკლები იქნება $\mu + 1.43\sigma$-ზე უდრის 0.9236-ს.

მძიმის შემდეგ ორი ათწილადის ნიშანი z თვის										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	0.9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
…										
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9139	.9147	.9062	.9177
1.4	.9192	.9207	.9222	.9236	.9252	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9394	.9394	.9406	.9418	.9429	.9441

ცხრილი 10.4. *A ცხრილის ნაწილი (მარცხენა - ზედა)* ნორმალური დაგროვებითი განაწილებისთვის.



გრაფიკი 10.5. *ნორმალური დაგროვებითი ალბათობა იმ მნიშვნელობებისთვის, რომლებიც ნაკლებია μ + zσ -ზე.* მონიშნული ფართობი შეესაბამება z = 1.43 მნიშვნელობას.

როგორც გრაფიკიდან ჩანს, (μ+zσ)-ზე მეტი მნიშვნელობის გამოსათვლელად შეგვიძლია ერთს გამოვაკლოთ ნაპოვნი ალბათობა, ანუ 1 – 0.9236 = 0.0764. უარყოფითი z-თვის მნიშვნელობებს მოძებნით A ცხრილის სხვა ნაწილში.

ალბათობების გამოთვლა z = სტანდარტული გადახრების რაოდენობის გამოყენებით.

ჩვენ z სიმბოლოს ვიყენებთ საშუალოდან სტანდარტული გადახრის აღსანიშნავად. თუ გვაქვს შემთხვევითი მნიშვნელობის სიდიდე x, სად ვარდება მნიშვნელობა? სხვაობა x -სა და μ -ს შორის არის x - μ, ხოლო z -ქულა გამოსახავს მათ განსხვავებას, როგორც სტანდარტული გადახრების რიცხვს (რაოდენობას).

ე.ი. z-ქულა შემთხვევითი სიდიდის მნიშვნელობებისთვის გამოითვლება შემდეგნაირად:

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

ეს ფორმულა სასარგებლოა იმ შემთხვევისთვის, როდესაც მოცემული გვაქვს რაიმე ნორმალური შემთხვევითი x სიდიდის მნიშვნელობა და გვინდა გამოვთვალოთ ამ სიდიდეზე დამოკიდებული ალბათობა. ამისათვის x-ს გარდავქმნით z-ქულად და შემდეგ გამოვიყენებთ ნორმალური განაწილების ცხრილს შესატყვისი ალბათობის გამოსათვლელად.

სცენარი 10.6. თქვენი ფარდობითი მდგომარეობა SAT ტესტირების შემდეგ. კოლეჯში მისაღები გამოცდების SAT (Scholastic Aptitude Test) ტესტს აქვს სამი კომპონენტი: კრიტიკული კითხვა, მათემატიკა და წერა. თითოეული კომპონენტის მიხედვით ქულები დაახლოებით განაწილებულია ნორმალურად μ = 500 საშუალოთი და σ = 100 სტანდარტული გადახრით. ქულების დიაპაზონი თითოეულ კომპონენტში არის 200 -დან 800 -მდე.

კითხვები შესწავლისათვის:

ა) თუ თქვენი SAT ქულა რომელიმე კომპონენტში იყო x = 650, რამდენი სტანდარტული გადახრა იყო ეს საშუალოსგან?

ბ) SAT ქულების რამდენი % იყო თქვენს ქულებზე მაღალი?

გააზრებისთვის:

ა) SAT-ის 650 ქულას შეესაბამება z = 1.5 z -ქულა, რადგანაც 650 არის 1.5 სტანდარტული გადახრით ზემოთ ვიდრე საშუალო. სხვა სიტყვებით რომ ვთქვათ, x = 650 =

$\mu + z\sigma = 500 + z \times 100$, სადაც $z = 1.5$. ჩვენ ეს შეგვიძლია ვიპოვოთ უშუალოდ ზემოთ ამოწერილი ფორმულიდან

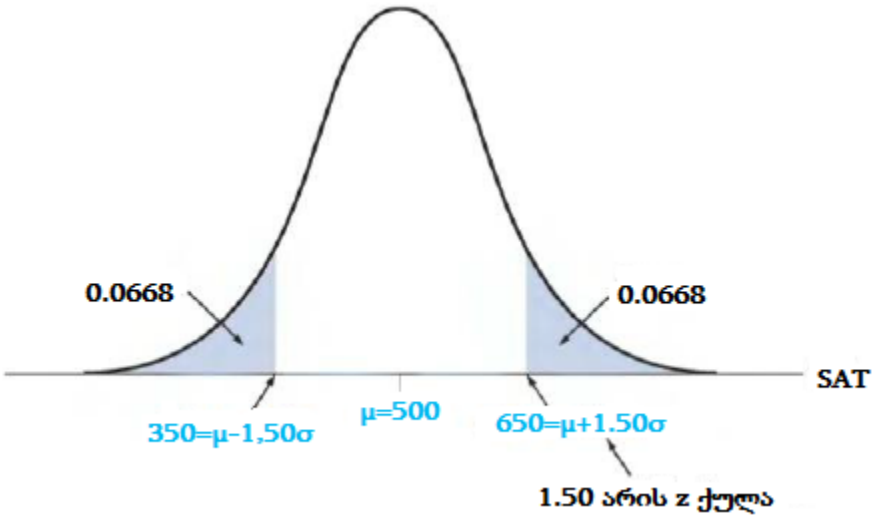
$$z = \frac{x-\mu}{\sigma} = \frac{650-500}{100} = 1.5$$

ბ) უნდა ვიპოვოთ 650 ქულაზე მეტი ქულების პროცენტული მაჩვენებელი ანუ მარჯვენა-კუდის ალბათობა, 650-ის ზემოთ, ნორმალური შემთხვევითი სიდიდისთვის, რომლის საშუალოა $\mu = 500$ და სტანდარტული გადახრა $\sigma = 100$. A ცხრილიდან ჩანს, რომ 1.5-ის შესაბამის z-ქულას აქვს დაგროვებითი ალბათობა 0.9332. ეს არის ალბათობა 650-ის *ქვემოთ*; ამიტომ მაჯვენა-კუდის ალბათობა იქნება $1 - 0.9332 = 0.0668$ (იხ. გრაფიკი 10.7). აქედან ჩანს, რომ SAT ქულების დაახლოებით 7% არის 650-ზე მეტი. საბოლოოდ, აღმოჩნდა, რომ 650 ქულა საკმაოდ მაღალია საშუალოსთან შედარებით, იმ გაგებით, რომ მხოლოდ მცირე რაოდენობას აქვს თქვენზე მეტი ქულა.

წვდომის უნარი: დადებითი z-ქულები ჩნდება მაშინ, როცა x ვარდება საშუალოზე ზემოთ, ხოლო უარყოფითი z -ქულები ჩნდებიან მაშინ როცა x ვარდება საშუალოზე ქვემოთ. მაგალითად, SAT-ის 350 ქულას შეესაბამება ასეთი z-ქულა,

$$z = \frac{x-\mu}{\sigma} = \frac{350-500}{100} = -1.5$$

ანუ გადახრა არის საშუალოს ქვემოთ. ალბათობა იმისა, რომ SAT-ის ქულების მნიშვნელობები იქნება 350-ზე ნაკლებები, არის აგრეთვე 0.0668 (იხ. გრაფიკი 10.7).



გრაფიკი 10.7. *ნორმალური განაწილება SAT-თვის.* SAT ქულებს 650-ს და 350-ს აქვთ z-ქულები 1.5 და -1.5, ამიტომ ისინი ვარდებიან 1.5 სტანდარტული გადახრით საშუალოს ზემოთ და ქვემოთ.

კითხვა: რომელ SAT ქულას აქვს z = 3.0 და z = -3.0?

რეზიუმე: *z - ქულების გამოყენება ნორმალური ალბათობის ან შემთხვევითი სიდიდის მნიშვნელობის საპოვნელად:*

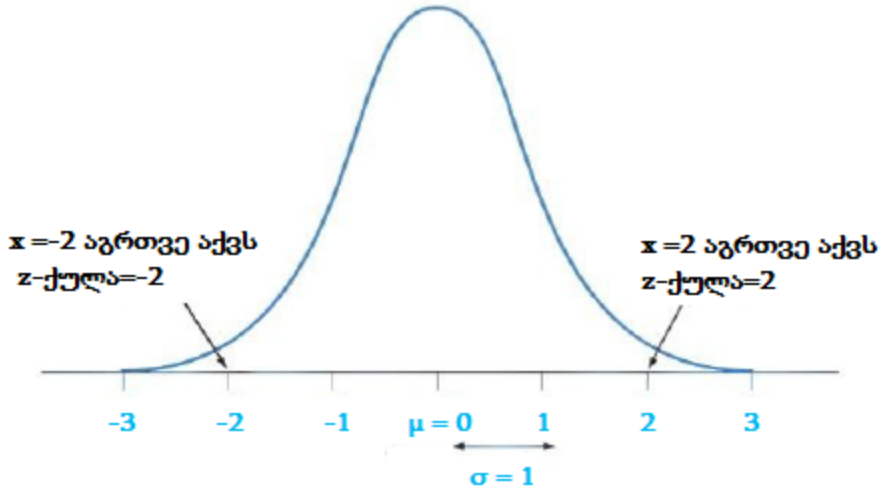
- თუ მოცემული გვაქვს x სიდიდე და გვინდა ვიპოვოთ ალბათობა, ამისათვის გადავიყვანოთ x სიდიდე z-ქულაში ფორულით $z = \frac{x-\mu}{\sigma}$, გამოვიყენოთ ნორმალური ალბათობების ცხრილი (ან პროგრამული უზრუნველყოფა, ან კალკულატორი), რომ ვიპოვთ დაგროვებითი ალბათობა და შემდეგ გარდავქმნათ ჩვენი ალბათური ინტერესების მიხედვით.
- თუ მოცემული გვაქვს ალბათობა და გვინდა ვიპოვოთ x-ის მნიშვნელობა, გარდავქმნათ ალბათობა მასთან დაკავშირებულ დაგროვებით ალბათობად, ვიპოვოთ z-ქულა ნორმალური ცხრილის საშუალებით (ან პროგრამული უზრუნველყოფით, ან კალკულატორით) და შემდეგ გამოვთვალოთ $x = \mu + z\sigma$.

სტანდარტული ნორმალური განაწილების საშუალო = 0 და
სტანდარტული გადახრა = 1.

ბევრი სტატისტიკური მეთოდი ეყრდნობა კონკრეტულ ნორმალურ განაწილებას, რომელსაც ეწოდება **სტანდარტული ნორმალური განაწილება.** იგი არის ისეთი ნორმალური განაწილება, რომლის საშუალო $\mu = 0$ და სტანდარტული გადახრა $\sigma = 1$.

ამ განაწილებისთვის z სტადარტული გადახრების მოხვედრათა რაოდნობა საშუალოს ზევით ტოლია შემდეგი სიდიდის $\mu + z\sigma = 0 + z \times 1 = z$, ანუ ტოლია თვით z-ქულის. მაგალითად, მნიშვნელობა 2-სთვის არის ორი სტანდარტული გადახრა საშუალოდან ზემოთ და მნიშვნელობა -1.3 არის 1.3 სტანდარტული გადახრა საშუალოდან ქვემოთ. როგორც შემდეგი გრაფიკი გვიჩვენებს საწყისი მნიშვნელობები იგივეა, რაც z-ქულები, რადგან

$$z = \frac{x-\mu}{\sigma} = \frac{x-0}{1} = x.$$



გრაფიკი 10.8. *სტანდარტული ნორმალური განაწილება.* საშუალო = 0, სტანდარტული გადახრა = 1. შემთხვევითი ცვლადის მნიშვნელობა x კი იგივეა, რაც z - ქულა.

კითხვა: რომელია ის საზღვრები, რომელთა შორისაც მოთავსებულია თითქმის ყველა მნიშვნელობა?

წინა სცენარში (SAT ქულების შესახებ) გვქონდა μ = 500 და σ = 100. დავუშვათ, რომ გარდავქმენით SAT ყველა x ქულა z-ქულაში ფორმულით $z = \frac{x-\mu}{\sigma} = \frac{x-500}{100}$; მაშინ x = 650 გადადის z = 1.5 -ში და x = 350 გადადის z = -1.5-ში. როდესაც ნორმალური განაწილების მნიშვნელობები გარდაქმინილია z - ქულებში, მაშინ ამ z - ქულებს აქვთ საშუალო 0 და სტანდარტული გადახრა 1. ეს კი ნიშნავს, რომ z - ქულების მთელ სიმრავლეს აქვს სტანდარტული ნორმალური განაწილება.

z - ქულები და სტანდარტული ნორმალური განაწილება. როდესაც შემთხვევით სიდიდეს აქვს სტანდარტული ნორმალური განაწილება, და მისი მნიშვნელობები გადაყვანლია z-ქულებში, ანუ გამოკლებულია საშუალო და გაყოფილია სტანდარტულ გადახრაზე, მაშინ z-ქულებს აქვთ სტანდარტული ნორმალური განაწილება (საშუალო = 0 და სტანდარტული გადახრა = 1). ეს შედეგი მეტად მნიშვნელოვანია სტატისტიკური დასკვნების გასაკეთებლად.

11

ბინომიური განაწილება. ბინომური განაწილების საშუალო და სტანდარტული გადახრა.

ალბათობები, როდესაც თითოეულ დაკვირვებას აქვს ორი შესაძლო შედეგი.

ჩვენ უნდა შევისწავლოთ დისკრეტული შემთხვევითი სიდიდეების ყველაზე მნიშვნელოვანი ალბათობის განაწილება. მისი ცოდნა დაგვეხმარება პასუხები გავცეთ იმ კითხვებს, რომლებიც დასაწყისში იყო დასმული. მაგალითად, კითხვა შეიძლება ყოფილიყო ასეთი: არსებობს თუ არა ქალთა მიმართ დიკრიმინაციის ძლიერი მტკიცებულება თანამშრომლების შერჩევისას ტრენინგების მენეჯმენტში?

ბინომური განაწილება: ალბათობები ორშედეგიან მონაცემთა დასათვლელად.

ძალიან ბევრ გამოყენებებში თითოეული დაკვირვება არის ხოლმე ორშედეგიანი ანუ **ბინარული**: მას აქვს ერთი შედეგი ორი შესაძლოდან. მაგალითად, ადამიანს შეუძლია,

- მიიღოს ან არ მიიღოს შეთავაზება ბანკისგან საკრედიტო ბარათის აღების თაობაზე,
- ჰქონდეს ან არ ჰქონდეს ჯანმრთელობის დაზღვევა,
- რეფერენდუმზე უპასუხოს „დიახ“ ან „არა“ გუბერნატორის გაწვევას.

შერჩევიდან შეგვიძლია შევაჯამოთ ასეთი ცვლადები მათი რაოდენობის უშუალო დათვლით ან პროპორციული შედარებით იმ შედეგების მიხედვით, რომლებიც გვაინტერესებს. მაგალითად, n = 5 ზომის შერჩევაში, დავუშვათ, რომ შემთხვევითი ცვლადი X აღნიშნავს იმ ხალხის რაოდენობას, რომელთა პასუხიც რეფერენდუმის რაღაც კითხვაზე არის „დიახ“. X-ის შესაძლო მნიშვნელობებია 1, 2, 3, 4 და 5. გარკვეულ პირობებში შემთხვევითი სიდიდე, რომელიც ითვლის კონკრეტული ტიპის დაკვირვების რაოდენობას, აქვს ალბათობის განაწილება, რომელსაც ეწოდება **ბინომური.**

განვიხილოთ n შემთხვევა და ვუწოდოთ მათ **ცდები**, სადაც ვაკვირდებით ბინარულ შემთხვევით სიდიდეს. ეს *ფიქსირებული რიცხვია*, როგორიცაა, მაგალითად, n = 5

ხუთ ამომრჩევლიანი შერჩევისათვის. X რიცხვს (ცდები, რომლებშიც ჩვენთვის სასურველ შედეგს აქვს ადგილი) შეუძლია მიიღოს ნებისმიერი ერთი მნიშვნელობა შემდეგი მთელი რიცხვებიდან 0, 1, 2, ... , n. ბინომური განაწილება იძლევა ალბათობებს X-ის ამ შესაძლო მნიშვნელობებისთვის, როდესაც ძალაშია შემდეგი სამი პირობა:

ბინომური განაწილების პირობები

- ყოველ ცდას უნდა ჰქონდეს ორი შესაძლო შედეგი. შედეგს, რომელიც ჩვენ გვაინტერესებს, ვუწოდებთ წარმატებას და მეორე შედეგს ვუწოდებთ წარუმატებლობას.
- თითოეულ ცდას უნდა ჰქონდეს წარმატების ერთი და იგივე ალბათობა. ამ ალბათობას აღვნიშნავთ p სიმბოლოთი, ხოლო წარუმატებლობის ალბათობას კი (1 – p)-თი.
- ეს n ცდა დამოუკიდებელია. ეს იმას ნიშნავს, რომ ყოველი ცალკეული ცდის შედეგი არაა დამოკიდებული სხვა ცდების შედეგებზე.

ბინომური შემთხვევითი სიდიდე X არის წარმატებათა რაოდენობა n ცდაში.

მონეტის n-ჯერ აგდება, სადაც n განსაზღვრულია წინასწარ, არის ბინომური განაწილების პროტოტიპი:

- თითოეული ცდა არის მონეტის აგდება. გვაქვს ორი შესაძლო შედეგი თითოეული ცდისთვის: გერბი და საფასური. დავუშვათ, რომ წარმატება არის საფასურის მოსვლა (შეიძლება ავირჩიოთ ნებისმიერად: გერბი ან საფასური).
- საფასურის მოსვლის ალბათობაა p, რომელიც ტოლია 0.5-ის. იგივე ალბათობა აქვს გერბსაც.
- თითოეული აგდება დამოუკიდებელია, რაც იმას ნიშნავს, რომ კონკრეტული აგდების შედეგი არ არის დამოკიდებული წინა შედეგებზე.

ბინომური შემთხვევითი სიდიდე X ითვლის საფასურების მოსვლის რაოდენობას მონეტის n-ჯერ აგდებისას. თუ n = 3 არის აგდებათა რაოდენობა, მაშინ X = საფასურების რაოდენობა შეიძლება იყოს 0, 1, 2, ან 3.

სცენარი 11.1. ESP ექსპერიმენტი (ბინომური ალბათობები).

ჯონ დოუ ამტკიცებს, რომ იგი ფლობს ექსტრასენსორულ აღქმადობას (ESP - Extrasensory Perception). ჩატარდა ექსპერიმენტი, რომლის დროსაც ერთ ოთახში

პიროვნება შემთხვევით ირჩევს რიცხვს {1, 2, 3, 4, 5} სიმრავლიდან და კონცენტრირდება მასზე ერთი წუთის განმავლობაში. სხვა ოთახში ჯონ დოუ ასახელებს რიცხვს, რომელიც მისი აზრით იქნა შერჩეული. ექსპერიმენტი შედგებოდა სამი ცდისგან. მესამე ცდის შემდეგ შემთხვევითი რიცხვები შედარდა ჯონ დოუს პროგნოზებს. დოუმ ორ შემთხვევაში სწორად ამოიცნო დასახელებული რიცხვები.

კითხვა შესწავლისთვის:

თუ ჯონ დოუს სინამდვილეში არ გააჩნია ESP უნარი და ის უბრალოდ იცნობს რიცხვებს, მაშინ რა არის ალბათობა იმისა, რომ სამიდან ორ ცდაში ის სწორად გამოიცნობს?

გააზრებისთვის:

ვთქვათ, X = სწორად ნათქვამი პასუხები n = 3 ცდაში. მაშინ X = 0, 1, 2 ან 3. ვთქვათ, p არის სწორი პასუხის ალბათობა მოცემულ ცდაში. ერთ ცდაში დოუს მიერ სწორად გამოცნობის ალბათობაა p = 0.2, რადგან კონკრეტულ ცდაში მან უნდა ამოიცნოს ხუთელემენტიანი სიმრავლიდან ერთი შერჩეული რიცხვი. ხოლო 1 − p = 0.8 იქნება არასწორი პასუხის ალბათობა მოცემულ ცდაში. ცდის შედეგი აღვნიშნოთ S-ით ან F-ით, რომელიც აღნიშნავს წარმატებას ან წარუმატებლობას, იმისდა მიხედვით უპასუხებს დოუ სწორად თუ არა. შემდეგი ცხრილი გვიჩვენებს ამ ცდის შესაბამისი ელემენტარულ ხდომილებათა სივრცის ყველა რვა შედეგს. მაგალითად, FSS წარმოადგენს სწორ პასუხს მეორე და მესამე ცდებში. ცხრილი აგრეთვე გვიჩვენებს დამოუკიდებელ ხდომილებათა ალბათობებს გამრავლების წესის გამოყენებით. აქვე გავიხსენოთ, რომ დამოუკიდებელი ხდომილებებისთვის P(A და B) = P(A)P(B). ამიტომ P(FSS) = P(F)P(S)P(S) = $0.8 \times 0.2 \times 0.2 = 0.032$.

შედეგი	ალბათობა	შედეგი	ალბათობა
SSS	$0.2 \times 0.2 \times 0.2 = (0.2)^3$	**SFF**	$0.2 \times 0.8 \times 0.8 = (0.2)^1(0.8)^2$
SSF	$0.2 \times 0.2 \times 0.8 = (0.2)^2(0.8)^1$	**FSF**	$0.8 \times 0.2 \times 0.8 = (0.2)^1(0.8)^2$
SFS	$0.2 \times 0.8 \times 0.2 = (0.2)^2(0.8)^1$	**FFS**	$0.8 \times 0.8 \times 0.2 = (0.2)^1(0.8)^2$
FSS	$0.8 \times 0.2 \times 0.2 = (0.2)^2(0.8)^1$	**FFF**	$0.8 \times 0.8 \times 0.8 = (0.8)^3$

ცხრილი 11.2 *ელემენტარულ ხდომილებათა სივრცე და ალბათობები სამი ცდის შემთხვევაში.* ალბათობები სწორ პასუხზე არის 0.2 ყოველ ცდაზე, თუ ჯონ დოუს არა აქვს ESP უნარი.

არსებობს სამი შესაძლებლობა, რომ ჯონ დოუ სამიდან ორში აკეთებს სწორ არჩევანს. ესენია: SSF, SFS და FSS. თითოეულ მათგანს აქვს ალბათობა $0.2 \times 0.2 \times 0.8 = 0.032$. სრული ალბათობა ორი სწორი პსუხის შემთხვევაში იქნება:

$$3 \times 0.2 \times 0.2 \times 0.8 = 3 \times 0.032 = 0.096.$$

პირველი მამრავლი 3 მიუთითებს იმაზე, რომ ორი სწორი ამოცნობის სამი შესაძლებლელი ვარიანტი არსებობს (SSF, SFS და FSS).

წვდომის უნარი: ალბათობის ტერმინებში, როცა კონკრეტულ ცდაში სწორი პასუხის შემთხვევაში ალბათობაა $p = 0.2$, შეგვიძლია წინას მსგავსად გამოვთვალოთ სხვა შემთხვევებიც. მაგალითად, როცა ერთი სწორად ამოცნობა იქნება დაფიქსირებული კვლავ სამი ცდის შემთხვევაში, მივიღებთ, რომ გვაქვს კვლავ სამი შესაძლო ვარიანტი: SFF, FSF და FFS. ამიტომ ამ შემთხვევაში შესაბამისი ალბათობა იქნება

$$3 \times 0.2 \times 0.8 \times 0.8 = 0.384.$$

ბინომური ალბათობის ფორმულა

როდესაც ცდათა რიცხვი n დიდია, ელემენტარულ ხდომიებათა სივრცის ყველა შედეგის ჩამოწერა დამღლელია. ამიტომ მოსახეხებელია ვისარგებლოთ ფორმულით, რომლითაც შესაძლებელი იქნება გამოვითვალოთ ბინომური ალბათობა ნებისმიერი n -თვის.

გავიხსენოთ, რომ **ბინომური კოეფიციენტი** გამოითვლება ფორმულით:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

სადაც x არის წარმატებათა რაოდენობა n ცდაში. მაგალითად, თუ გვაქვს $x = 2$ წარმატება $n = 3$ ცდაში, გვექნება

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!\,1!}.$$

წარმატებების ალბათობა აღვნიშნოთ p-თი. n დამოუკიდებელი ცდის შემთხვევაში x რაოდენობის წარმატებების მიღწევის ალბათობა ტოლია:

$$P(x) = \frac{n!}{x!\ (n-x)!}\ p^{x}\ (1-p)^{n-x}, \quad x = 0, 1, 2, ..., n$$

სიმბოლო n! იკითხება „n ფაქტორიალი“ და არის ყველა ნატურალური რიცხვის ნამრავლი 1-დან n –მდე ანუ $n! = 1 \times 2 \times 3 \times ... \times n$. მაგალითად, $3! = 1 \times 2 \times 3 = 6$, $4! = 1 \times 2 \times 3 \times 4 = 24$ და ა.შ. შევიშნოთ, რომ 0! განიმარტება, როგორც 1, ე.ი. $0! = 1$. მოცემული p და n რიცხებისათის შეგვიძლია ვიპოვოთ შესაძლო ხდომილების ალბათობა x-ის შესაბამისი მნიშვნელობების ჩასმით ბინომურ ფორმულაში. გამოვიყენოთ ეს ფორმულა ESP სცენარში პასუხის დასადგენად.

- შემთხევითი ცვლადი X წარმოადგენს სწორად ამოცნობილ ვერსიათა რაოდენობას (წარმატებათა რაოდენობა) ESP ექპერიმენტის $n = 3$ ცდაში.
- სწორი პასუხის გამოცნობის ალბათობა ყოველ კონკრეტულ ცდაში არის $p = 0.2$.
- ზუსტად ორი სწორი პასუხის გაცემის ბინომური ალბათობა სამცდიან ექსპერიმენტში ანუ როდესაც $n = 3$ (ცდას) და $x = 2$ (სწორ პასუხს), გამოითვლება ბინომური ალბათობის ფორმულით:

$$P(2) = \frac{n!}{x!\ (n-x)!}\ p^{x}(1-p)^{n-x} = \frac{3!}{2!1!}(0.2)^{2}(0,8)^{1} = 3(0,04)(0,8) = 0.096$$

  რა არის განსხვავებული თანამამრავლების როლი ბინომის ფორმულაში?
- ფაქტორიალის ნაწილი გვიჩვენებს შესაძლებელი შედეგების რაოდენობას, რომელსაც აქვს წარმატება $x = 2$. აქ $[3!/(2!1!)] = (3 \times 2 \times 1)/(2 \times 1) \times 1 = 3$ გვეუბნება, რომ არის მხოლოდ სამი შესაძლებლობა ორი სწორი პასუხის გამოცნობის, ესენია SSF, SFS და FSS.
- წევრი $0.2 \times 0.2 \times 0.8 = 0.032$ არის ალბათობა ორწარმატებიანი ერთი სამეულისა. სრული ალბათობა კი ტოლი იქნება $3 \times 0.032 = 0.096$.

სცადეთ გამოთვალოთ P(1). ამისთვის ბინომურ ფორმულაში ჩასვით $x = 1$, $n = 3$ და $p = 0.2$. თქვენ მიიღებთ 0.384. როგორც უკვე აღვნიშნეთ ზემოთ, აქაც ისევ არის სამი შესაძლო შემთხვევა და თითოეულის ალბათობა არის $0.2 \times 0.8 \times 0.8 = 0.128$. სრული ალბათობა კი ტოლი იქნება $3 \times 0.128 = 0.384$.

ქვემოთ მოყვანილი ცხრილი აჯამებს გამოთვლებს x-ის შესაძლო ყველა ოთხივე მნიშვნელობისთვის. თქვენ აგრეთვე შეგიძლიათ იპოვოთ ბინომური ალბათობები სტატისტიკის პროგრამული უზრუნველყოფის დახმარებით, როგორიცაა MINITAB ან კალკულატორი სტასტისტიკური ფუნქციებით.

x	P(x) = [n! / (x! (n-x)!)] $p^x(1-p)^{n-x}$
0	0.512 = [3! / (0!3!)] $(0.2)^0(0.8)^3$
1	0.384 = [3! / (1!2!)] $(0.2)^1(0.8)^2$
2	0.096 = [3! / (2!1!)] $(0.2)^2(0.8)^1$
3	0.008 = [3! / (3!0!)] $(0.2)^3(0.8)^0$

ცხრილი 11.3. *ბინომური განაწილება, როცა n = 3, p = 0.2.* შემთხვევითმა X სიდიდემ შეიძლება მიღოს ნებისმიერი მთელი მნიშვნელობა 0-სა და 3-ს შორის მათი ჩათვლით. ამ დროს სრული ალბათობა ტოლია 1-ის.

სცენარი 11.4. ტესტირება გენდერულ დისკრიმინაციაზე სამსახურებრივი დაწინაურების სფეროში (ბინომური განაწილება)

მიუხედავად იმისა, რომ ბინომური განაწილება შეიძლება გამოვიყენოთ ისეთი მარტივი ამოცანების ალბათობების გამოსათვლელად, როგორიცაა მონეტის აგდება ან ESP უნარების შემოწმება, იგი არის აგრეთვე მნიშვნელოვანი იარაღი იმისთვის, რომ ჩავწვდეთ უფრო სერიოზულ პრობლემებს. მაგალითად, როცა ადგილი აქვს ქალთა შესაძლო დისკრიმინინაციას სამსახურებში. ქალ თანამშრომელთა ერთი ჯგუფი აცხადებს, რომ ნაკლებად სავარაუდოა სამსახურში დააწინაურონ ქალი, ვიდრე იმავე კვალიფიკაციის მამაკაცი.

კითხვა გააზრებისთვის:

დავუშვათ, რომ არის დიდი ჯგუფი, რომელიც გადის მენეჯმენტში ტრენინგებს და სადაც ნახევარი მანდილოსნებია და ნახევარი მამაკაცები. ცოტა ხნის წინ შეარჩიეს 10 პიროვნება დასაწინაურებლად, რომელთაგან არც ერთი არ იყო მანდილოსანი. რა

იქნება ალბათობა იმისა, რომ 0 მანდილოსანია 10 შერჩეულ პიროვნებაში, თუ აქ მართლა არ არის გენდერული დისკრიმინაცია?

გააზრებისთვის:

წინა შემთხვევების მსგავსად, აქაც ყოველ ჯერზე ალბათობა იმისა, რომ შერჩეული იქნება მანდილოსანი, ტოლია 0.5-ის და ალბათობა იმისა, რომ შერჩეული იქნება მამაკაცი, კვლავ არის 0.5. აღვნიშნოთ X-ით მანდილოსნების რაოდენობა, ვინც შეირჩა დასაწინაურებლად 10 ადამიანის შემთხვევითი შერჩევით. მაშინ X-ის შესაძლო მნიშვნელობებია: 0, 1, ... , 10 და X-ს აქვს ბინომური განაწილება შემდეგი მონაცემებით: n = 10, p = 0.5. ყოველი x-თვის 0-დან 10 -მდე ჩვენ შეგვიძლია ვიპოვოთ ალბათობა იმისა, რომ 10 შერჩეული პიროვნებიდან მანდილოსნების რაოდენობა x-ის ტოლია. გამოვიყენოთ ბინომური ფორმულა, როცა n = 10 და p = 0.5. კერძოდ,

$$P(x) = \frac{n!}{x!\,(n-x)!}\; p^{x}(1-p)^{n-x} = \frac{10!}{x!\,(10-x)!}(0.50)^{x}(0.50)^{10-x}, \quad x = 0, 1, 2, \ldots, 10$$

ალბათობა იმისა, რომ დასაწინაურებლად მანდილოსნები არ არიან შერჩეული (x = 0), ტოლია:

$$P(0) = \frac{10!}{0!10!}(0.50)^{0}(0.50)^{10} = (0.50)^{10} = 0.001\,.$$

როგორც ვიცით, ნებისმიერი (არანულოვანი) რიცხვი 0 ხარისხში არის 1. აგრეთვე, 0! =1 და შეკვეცის შემდეგ დარჩება, რომ $P(0) = (0.5)^{10}$, ე.ი. თანამშრომლები შემთხვევით რომ ყოფილიყვნენ შერჩეულები, იქნებოდა ძალიან მცირე შანსი (ერთი შეფარდებული ათასთან), რომ დაწინაურებითვის შერჩეული 10 ადამიანიდან არცერთი იყოს მანდილოსანი.

წვდომის უნარი: იმის გამო, რომ ალბათობა ასეთი მცირეა, ის ფაქტი, რომ 10 დასაწინაურებელ კანდიდატში არც ერთი იყო მანდილოსანი, ეჭვს აჩენს იმაზე, რომ შერჩევა იყო შემთხვევითი გენდერობის თვალსაზრისით.

დარწმუნდით, რომ შეძლება ბინომური პირობების გამოყენება.

სანამ გამოიყენებდეთ ბინომურ ფორმულას, ჯერ დარწმუნდით, რომ ის სამი პირობა შესრულებულია, რომელზეც ზემოთ ვისაუბრეთ ანუ

(1) მონაცემები ბინარულია (წარმატება და წარუმატებლობა),

(2) ყოველ ცდაში წარმატების ალბათობა არის ერთი და იგივე (აღინიშნება p -თი), და

(3) ფიქსირებულია დამოუკიდებელ ცდათა რაოდენობა n. ამის განსჯისას კითხეთ თქვენს თავს, ეს დაკვირვებები გაგონებენ თუ არა მონეტის აგდებას, ბინომის მარტივ პროტოტიპს. მაგალითად, გამოიყურება დამაჯერებლად, თუ მას გამოვიყენებთ წინა მაგალითის შემთხვევაში გენდერულ თანასწორობაზე. ამ შემთხვევაში

- მონაცემები ბინარულია, რადგან (ქალი, მამაკაცი) თამაშობენ იმავე როლს, რასაც (საფასური, გერბი).
- თუ თანამშრომლები შემთხვევითობითაა შერჩეული, მაშინ ქალების შერჩევის p ალბათობა მოცემულ ცდაში არის 0.5.
- 10 ადამიანის შემთხვევით შერჩევაში ერთი ცდის შედეგი არ არის დამოკიდებული სხვა ცდის შედეგზე.

ბინომური განაწილების საშუალო და სტანდარტული გადახრა.

როგორც ყველა დისკრეტული განაწილების შემთხვევაში, აქაც საშუალოს საპოვნელად შეიძლება გამოვიყენოთ ფორმულა

$$\mu = \sum xP(x).$$

თუმცა, ბინომური განაწილებისთვის μ საშუალოს და σ სტანდარტული გადახრის პოვნა უფრო იოლადაა შესაძლებელი. არსებობს სპეციალური ფორმულები, რომლებიც დაფუძნებულია მხოლოდ n ცდათა რაოდენობასა და თითოეულ ცდაში წარმატებათა p ალბათობაზე. ესენია:

$$\mu = np \quad \text{და} \quad \sigma = \sqrt{np(1-p)}.$$

საშუალოს ფრმულას აქვს ასეთი აზრი: თუ წარმატების ალბათობა მოცემულ ცდაში არის p, მაშინ ველით, რომ პროპორცია n ცდაში მთლიანად უნდა იყოს np. თუ შევარჩევთ n = 10 ადამიანს პოპულაციიდან სადაც ნახევარი მანდილოსანია, მაშინ ველოდებით, რომ დაახლოებით $np = 10 \times 0.5 = 5$ მანდილოსანი უნდა იყოს ამ შერჩევაში.

როდესაც ცდათა რიცხვი n დიდია, მაშინ ძალიან დამღლელი იქნება ყველა შესაძლო შედეგის ბინომური ალბათობის გამოთვლა. ასეთ დროს გაცილებით მოსახერხებელია გამოვიყენოთ საშუალო და სტანდარტული გადახრა, სადაც ჩავარდება ალბათობათა უდიდესი ნაწილი. როდესაც ცდათა რიცხვი n დიდია, ბინომურ განაწლებას აქვს ზარის ფორმა, რომელიც ზემოთ გვქონდა აღწერილი. ამიტომ ჩვენ შეგვიძლია

გამოვიყენოთ ნორმალური განაწილება ბინომური განაწილების აპროქსიმაციისთვის და დავასკვნათ, რომ თითქმის ყველა ალბათობა ვარდება [μ - 3σ ; μ + 3σ] შუალედში.

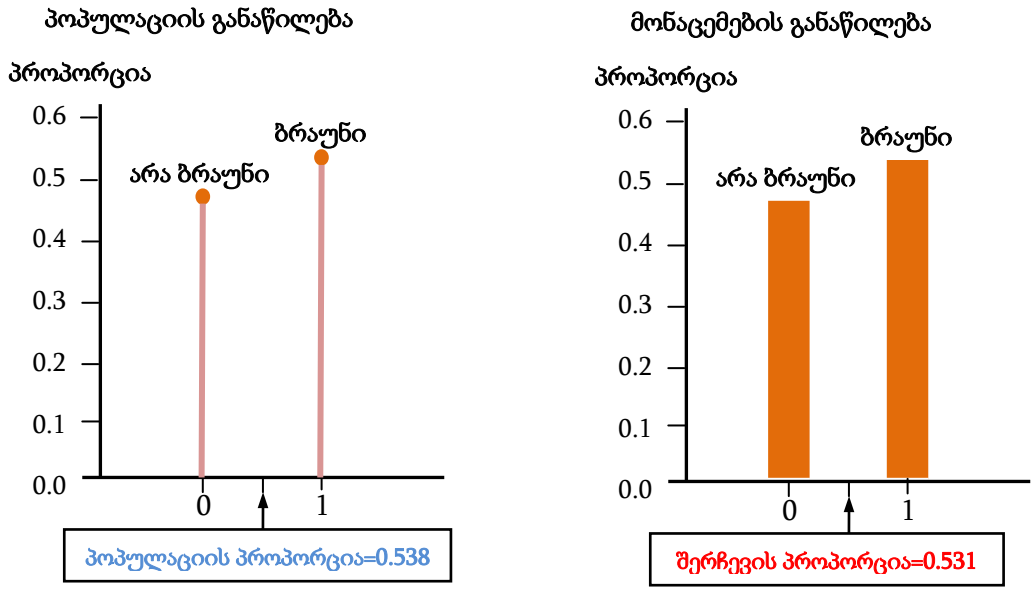
12

ამორჩევის განაწილება. მისი საშუალო და სტანდარტული გადახრა. დიდ რიცხვთა კანონი და ცეტრალური ზღვარითი თეორემა. კავშირი ბინომიალურ და შერჩევის განაწილებებს შორის.

შერჩევის განაწილება გვიჩვენებს, როგორ ცვალებადობს შერჩევის პროპორცია. იგი გვეხმარება განვსაზღვროთ, რამდენად ახლოს მოხვდება შერჩევის სტატისტიკა პოპულაციის პარამეტრთან. ეგზიტპოლის დროს შემთხვევითად შერჩეული ამომრჩევლების მიერ რომელიმე პიროვნებაზე ხმის მიცემა არის კატეგორული ცვლადი, რადგანაც შედეგი (ხმის მიცემა) ცვალებადობს ამომრჩევლიდან ამომრჩევლამდე და ვარდება კატეგორიაში (კანდიდატი). კალიფორნიის გუბერნატორის არჩევნების დროს ჩატარებული ეგზიტპოლების შედეგები ასე გამოიყურება. აღვნიშნოთ X = ხმის მიცემის შედეგი, x = 1 იყოს, ხმა მისცა ჯერი ბრაუნს და x = 0 ყველა სხვა დანარჩენ შემთხვევაში (მეგ ვაითმანს, სხვა კანდიდატს და არ უპასუხია). ასეთ შემთხვევაში ვამბობთ, რომ შედეგი არის ბინარული. ანუ ამომრჩეველს აქვს ორი ვარიანტი: ხმა მისცა ბრაუნს ან ხმა არ მისცა ბრაუნის სასარგებლოდ. შერჩევის პროპორცია შეიძლება განხილიულ იყოს როგორც შემთხვევითი სიდიდე, რადგან იგი, როგორც ვთქვით, იცვლება შერჩევიდან შერჩევამდე. ეგზიტპოლის წინ მოხდა შერჩევა, სადაც პროპორცია, თუ ვინ მისცა ხმა ბრაუნის სასარგებლოდ, იყო უცნობი. ეს იყო შემთხვევითი სიდიდე. პროპორცია ამომრჩევლების, ვინც მას მისცა ხმა სავარაუდოდ განსვხვავდება სხვა ეგზიტპოლის შერჩევისგან. თუ რამდენჯერ გამოჩნდებიან შემთხვევითი სიდიდის შესაძლო მნიშვნელობები (0 და 1) იძლევიან **მონაცემთა განაწილებას** (0.469 და 0.531) ამ ერთი შერჩევისთვის. ამომრჩევლების ყოველი შემთხვევითი შერჩევისთვის გვექნება განსხვავებული მონაცემთა განაწილება.

ოფიციალური მონაცემებით, საბოლოო შედეგით ბრაუნმა მიიღო ხმების 0.538 ნაწილი, ვაითმანმა - 0.409, დანარჩენებმა 0.053. ესენი არის პოპულაციის პროპორციები. ასე რომ 0.462 და 0.538 პროპორციები გვაძლევენ **პოპულაციის განაწილებას**. შევნიშნოთ, რომ x = 0 მნიშვნელობა მოიცავს ვათმანს, დანარჩენებს და

მათაც ვინც არ უპასუხა. ქვემოთ მოცემულია პოპულაციის გრაფიკული წარმოდგენა. მასთან ერთადაა ერთერთი ეგზიტპოლის პროპორცია, რომელიც ძალიან ჰგავს პოპულაციის პროპორციას.



გრაფიკი 12.1. *პოპულაციის (9.5 მილიონი ამომრჩეველი) და მონაცემთა (n = 3889) განაწილებები (0 = არა ბრაუნი, 1 = ბრაუნი).*

კითხვა: რატომ გამოიყურებიან ისინი ასე მსგავსად?

ბრაუნის პროპორციები სხვა ეგზიტპოლებშიც მსგავსია პოპულაციის პროპორციების. იქნებიან ისინი მგავსი, თუ ეს შეიძლება იყოს 0.54 ერთში, 0.57 მეორეში, 0.51 მესამეში? ან იქნებ ძალიან განსხვავებულები არიან, როგორიცაა 0.40 ერთში, 0.59 მეორეში და 0.65 მესამეში? ამისთვის საჭიროა ვიცოდეთ ალბათობის განაწილება, რომელიც იძლევა ამორჩევის პროპორციის ალბათობებს. ალბათობის განაწილებას სტატისტიკისთვის, როგორიცაა მაგალითად შერჩევის პროპორცია, ეწოდება **შერჩევის განაწილება.**

ეგზიტპოლის შერჩევის 3889 ამომრჩევლისთვის, წარმოვიდგინოთ, რომ მივიღეთ განსხვავებული შერჩევები. თითოეულ შერჩევას აქვს შერჩევის პროპორციის ის მნიშვნელობა, რომელიც უჩვენებს ბრაუნის მომხრეთა ხმების წილს. წარმოვიდგინოთ, რომ შეგვიძლია შევხედოთ ყველა შერჩევას და ვიპოვოთ მათი შესაბამისი პროპორციები. შემდეგ შევადგინოთ სიხშირეთა ცხრილი შერჩევათა პროპორციებისთვის (მაგალითად, გრაფიკული ან დავხაზოთ ჰისტოგრამა, სადაც

პროპორციები იქნება რიცხვითი მნიშვნელობები). ეს იქნება შერჩევის პროპორციის შერჩევის განაწილება. ეს არის უბრალოდ ალბათობის განაწილების ერთერთი ტიპი.

შერჩევის განაწილებაში საშუალო და სტანდარტული გადახრა დამოკიდებულია შერჩევის n ზომასა და პოპულაციის პროპორციაზე p-ზე. ანუ შემთხვევითი შერჩევისთვის, რომლის ზომაა n და პროპორცია p, მაშინ შერჩევითი განაწილების

$$\textit{საშუალო} = \mathrm{p} \quad \text{და} \quad \textit{სტანდარტული გადახრა} = \sqrt{\frac{p(1-P)}{n}} \ , \qquad (12.1)$$

შერჩევის განაწილების ყოფაქცევა.

მაშინაც კი, როდესაც პოპულაციის განაწილება არ არის ზარის - ფორმის, შერჩევის განაწილებას, რომლის საშუალოა $\bar{x}$, შეიძლება ჰქონდეს ზარის - ფორმა. შევნიშნოთ აგრეთვე, რომ ამორჩევის განაწილების საშუალო, როგორც ჩანს, არის იგივე, რაც პოპულაციის საშუალო μ , ხოლო შერჩევის სტანდარტული გადახრა კი $\frac{\sigma}{\sqrt{n}}$. ეს ზარის ფორმა შედეგია **ცენტრალური ზღვარითი თეორემის (CLT).** იგი ამბობს, რომ შერჩევის განაწილება, რომლის საშუალოა $\bar{x}$, ხშირად მიახლოებით აქვს ნორმალური განაწილების სახე. ეს შედეგი არ იცვლება იმიდა მიხედვით, თუ როგორი ფორმა აქვს პოპულაციის განაწილებებს, საიდანაც აიღება შერჩევა. შედარებით დიდი ზომის შერჩევისთვის შერჩევის განაწილება არის ზარის ფორმის მაშინაც კი, როცა განაწილება ძალიან დისკრეტული და ძალიან დამახინჯებულია.
ცენტრალური ზღვარითი თეორემა (CLT) აღწერს მოსალოდნელ ფორმას შერჩევის განაწილებისთვის, რომლის საშუალოა $\bar{x}$.

შემთხვევითი შერჩევისთვის, რომლის ზომაა n, პოპულაციის საშუალო μ და სტანდარტული გადახრა σ, თუ n იზრდება, მაშინ შერჩევის განაწილება, რომლის საშუალოა $\bar{x}$, მიისწრაფვის მიახლოებით ნორმალური განაწილებისკენ.

ცენტრალურ ზღვარით თეორემაში გასაოცარია ის ფაქტი, რომ მიუხდავად იმისა, თუ რა ფორმისაა პოპულაციის განაწილება, შერჩევის განაწილება მაინც მიისწრაფვის ნორმალური განაწილებისკენ. მართლაც, უმრავლესი პოპულაციის განაწილება ძალიან სწრაფად უახლოვდება ზარის ფორმას, როცა ზომა n იზრდება. ამიტომ თუ ჩვენი შერჩევის ზომა საკმაოდ დიდია, არაა აუცილებელი ვიფიქროთ იმაზე, რა ფორმისაა პოპოულაციის განაწილება, იმისთვის, რომ ვიმუშაოთ შერჩევის განაწილებებზე.

n-ის მოქმედების ეფექტი შერჩევის განაწილების სტანდარტულ გადახრაზე.

კვლავ განვიხილოთ ფორმულა $\frac{\sigma}{\sqrt{n}}$. ეს არის სტანდარტული გადახრა შერჩევის საშუალოსგან. შევნიშნოთ, რომ როცა შერჩევის *ზომა* n იზრდება, მაშინ სტანდარტული გადახრა შერჩევის საშუალოსგან მცირდება. ამ დამოკიდებულებას აქვს მეტად მნიშვნელოვანი პრაქტიკული შედეგი: **როდესაც შერჩევა დიდია, შერჩევის საშუალო მიისწრაფვის ჩავარდეს რაც შეიძლება ახლოს პოპულაციის საშუალოსთან.**

იმის გამო, რომ შერჩევის პროპორცია არის ბინომური შეთხვევითი სიდიდე გაყოფილი შერჩევის ზომაზე, ამიტომ მისი საშუალოსა და სტანდარტული გადახრის ფორმულები არიან წარმატებათა რაოდენობის საშუალოსა და სტანდარტული გადახრის ფორმულები გაყოფილი n-ზე. ანუ, პროპორციის შერჩევის განაწილების საშუალო და სტანდარტული გადახრა გამოითვლებიან (12.1) ფორმულით. შედეგად მივიღეთ, რომ ბინომური განაწილება არის შერჩევის განაწილება (თუმცა არა აუცილებლად).

13

პოპულაციის პარამეტრების წერტილოვანი და ინტერვალური შეფასებები. პოპულაციის საშუალოსთვის ნდობის ინტერვალის აგება.

ნდობის ინტერვალი.

როგორც უკვე ვიცით, მეტად მიშვნელოვანია მონაცემებზე დაყრდნობით *სტატისტიკური დასკვნების* გაკეთება. პარამეტრების ყველაზე მეტად ინფორმაციული შეფასება ხდება ე.წ. ნდობის ინტერვალის საშუალებით. ეს ისეთი ინტერვალია, სადაც უნდა ჩავარდეს უცნობი პარამეტრი დიდი დამაჯერებლობით.

პოპულაციის პარამეტრებითვის არსებობს ორი ტიპის შეფასება: წერტილოვანი და ინტერვალური. **წერტილოვანი შეფასება** არის შეფასება ერთი რიცხვით და ეს არის საუკეთესო გზა იმისა, რომ გამოვიცნოთ უცნობი პარამეტრი. **ინტერვალური შეფასება** არის რიცხვითი ინტერვალი, რომელშიც უნდა ჩავარდეს უცნობი პარამეტრი დიდი ალბათობით. **ნდობის ინტერვალი** არის ისეთი ინტერვალი, რომელიც შეიცავს პარამეტრის ყველაზე მეტად დამაჯერებელ მნიშვნელობებს. ალბათობას იმისა, რომ ეს მეთოდი წარმოგვიდგენს ინტერვალს, რომელიც შეიცავს პარამეტრს, ეწოდება **ნდობის დონე**. ეს რიცხვი უნდა იქნეს შერჩეული რაც შეიძლება ახლოს 1-თან; ყველაზე ხშირად აიღება 0.95. მაშინ ეს იქნება 95%-ანი ნდობის ინტერვალი. მარგინალური ცდომილება იზომება შერჩევის განაწილების სტანდარტული გადახრის ნამრავლით რაღაც რიცხვჯერ. როგორიცაა მაგალითად,

$$1.96 \times (\text{სტანდარტული გადახრა}),$$

როცა შერჩევის განაწილება არის ნორმალური განაწილება.

მას შემდეგ რაც შერჩევა უკვე გაკეთებულია, თუ შერჩევის პროპორცია ვარდება 1.96 სტანდარტულ გადახრებს შორის, მაშინ ინტერვალი

$$[\ \text{შერჩევის პროპორცია} - 1.96 \times (\text{სტანდარტულ გადახრა})\ ] \text{-დან}$$

$$[\ \text{შერჩევის პროპორცია} + 1.96 \times (\text{სტანდარტულ გადახრა})\ ] \text{-მდე}$$

შეიცავს პოპულაციის პროპორციას. სხვა სიტყვებით რომ ვთქვათ, ალბათობით 0.95 შერჩევის პროპორციის მნიშვნელობები ჩნდებიან ისე, რომ ინტერვალი

შერჩევის პროპორცია ± 196 × (სტანდარტულ გადახრა)

შეიცავს პოპულაციის უცნობ პროპორციას. რიცხვების ეს ინტერვალი არის პოპულაციის განაწილების დაახლობით 95% -იანი ნდობის ინტერვალი.

პოპულაციის საშუალოსთვის ნდობის ინტერვალის აგება t განაწილების გამოყენებით.

სცენარი 13.1. ტელევიზორის ყურებაში დახარჯული საათები.

(პოპულაციის საშუალოსთვის ნდობის ინტერვალის აგება).

რამდენ დღე-ღამეს ატარებს ჩვეულებრივი ადამიანი ტელევიზორის ყურებაში? ტელევიზორის გადაჭარბებული ყურება დასახელდა ერთ-ერთ ფაქტორად, კვების რაციონთან ერთად, ამერიკელების სიმსუქნის მზარდი პროპორციის მიზეზად. რამდენიმე ხნის წინ ზოგადმა სოციალურმა გამოკვლევამ ჰკითხა რესპოდენტებს: „პირადად თქვენ, საშუალოდ, დღეში რამდენ საათს უყურებთ ტელევიზორს?“ შეჯამების შედეგად მიღებულია ასეთი მონაცემები:

N	საშუალო	სტანდ. გადახრა	სტანდ. შეცდ. საშ.	95% ნდ.ინტერ
1324	2.98	2.66	0.0731	(2.8366; 3.1234)

კითხვები შესწავლისთვის:

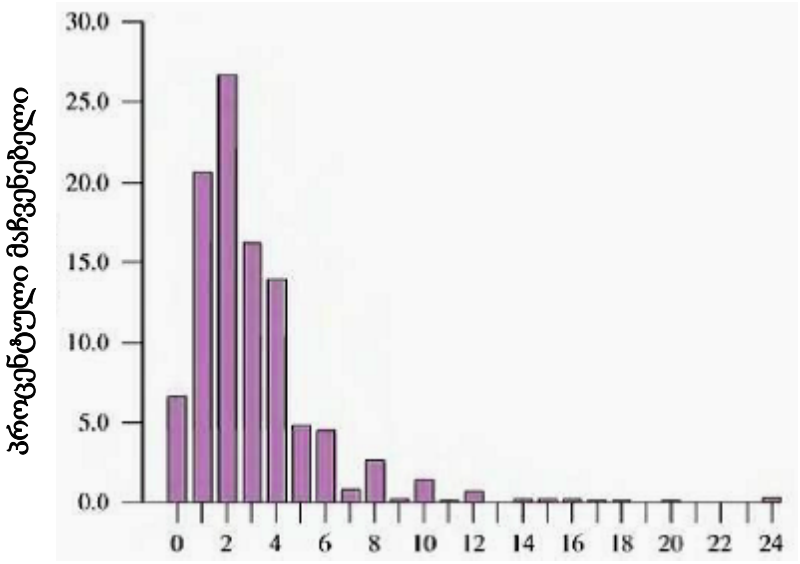
ა) შერჩევის საშუალო და სტადარტული გადახრა სავარაუდოდ განაწილების რა ფორმას გამოსახავენ?

ბ) პროგრამული უზრუნველყოფით როგორ იქნა მიღებული სტანდარტული შეცდომა? რას ნიშნავს ეს?

გ) მოახდინეთ პროგრამული უზრუნველყოფის მიერ წარმოდგენილი ნდობის ინტერვალის ინტერპრეტაცია.

გააზრებისთვის:

ა) შერჩევის საშუალოსთვის $\bar{x}$ = 2.98 და სტანდარტული გადახრისთვის s = 2.66, უმცირესი შესაძლო მნიშვნელობა 0 ვარდება ოდნავ მეტი ვიდრე 1 სტანდარტული გადახრა საშუალოსგან ქვემოთ. ამ ინფორმაციით შეიძლება ვივარაუდოთ, რომ ტელევიზორის ყურების განაწილება შესაძლოა დახრილია მარჯვნივ. იგივეს გვიჩვენებს მონაცემთა ჰისტოგრამა.



გრაფიკი 13.2. *ყოველდღიურად ტელევიზორის ყურების საათების რაოდენობის ჰისტოგრამა.*

კითხვა: დახრილობის ეფექტი მოქმედებს პოპულაციის საშუალოს ნდობის ინტერვალზე?

ბ) თუ შერჩევის სტანდარტული გადახრაა s = 2.66 და შერჩევის ზომა n = 1324, მაშინ შერჩევის საშუალოს სტანდარტული შეცდომა

$se = s/\sqrt{n} = 2.66/\sqrt{1324} = 0.0731$ საათი.

თუ ბევრი კვლევა იქნება ჩატარებული ტელევიზოტის ყურებაზე, როცა თითო-ეულისთვის n = 1324, შერჩევის საშუალო ძალიან არ შეიცვლება მთელი ამ კვლევების დროს.

გ) 95% -იანი ნდობის ინტერვალი ტელევიზორის ყურების პოპულაციის μ საშუალოსთვის აშშ-ში არის (2.84; 3.12) საათი. ჩვენ შეიძლება 95%-ით ვიყოთ დარწმუნებულები, რომ ტელევიზორის ყურების საათების რაოდენობის საშუალო მნიშვნელობა ამერიკელებისთვის დღეში არის 2.84 -სა და 3.12 შორის.

წვდომის უნარი: იმის გამო, რომ შერჩევის ზომა იყო დიდი, შეფასება არის ზუსტი და ნდობის ინტერვალი არის საკმარისად მცირე. რაც მეტია შერჩევის ზომა, მით ნაკლებია სტანდარტული შეცდომა და შემდგომი ცდომილებები.

რეზიუმე: პოპულაციის μ საშუალოსთვის 95% -იანი ნდობის ინტერვალი არის

$$\bar{x} \pm t_{.025} \frac{s}{\sqrt{n}} ,$$

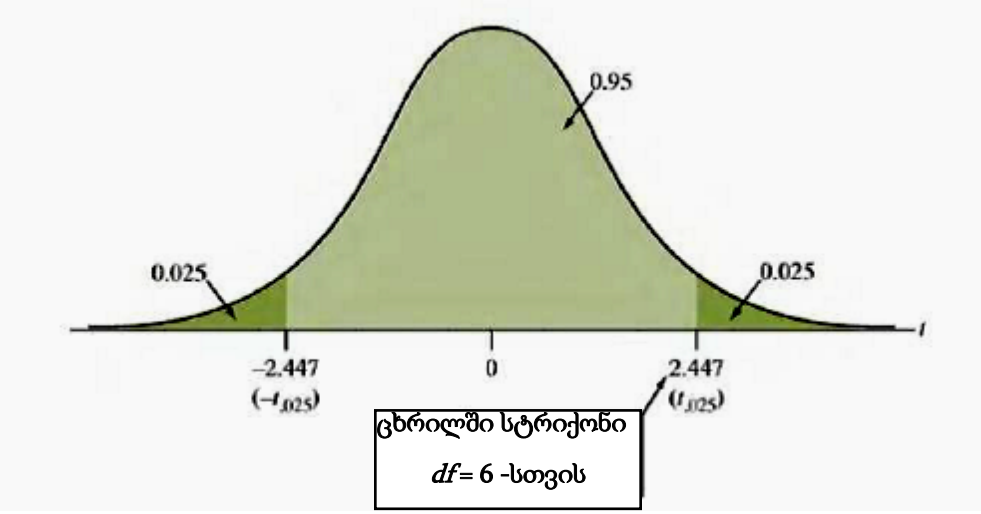
სადაც $t_{.025}$ არის t-ქულის მნიშვნელობა მარჯვენა კუდის 0.025 ალბათობის მნიშვნელობისთვის (სრული ალბათობა ორივე კუდის შემთხვევაში არის 0.05 და 0.95 არის $-t_{.025}$ და $t_{.025}$ მნიშვნელობებს შორის). ამ მეთოდის გამოსაყენებლად ჩვენ გვჭიდება რომ გვქონდეს:

- პოპულაციის დაახლოებით ნორმალური განაწილება,
- შემთხვევითი შერჩევა და
- t - განაწილების ცხრილი, რომელიც ქვემოთაა მოცემული.

	ნდობის დონე					
	80%	**90%**	**95%**	**98%**	**99%**	**99,8%**
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.3
. . .						
6	1.440	1.943	2.477	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785

ცხრილი 13.1. *B -ცხრილის ნაწილი სადაც ნაჩვენებია t -ქულები.* ამ ცხრილში

***df*= n-1** ანუ თუ n = 7, *df* = 6.



გრაფიკი 13. 2. *t -განაწილება df = 6-სთვის.* განაწილების 95% არის -2.447 -სა და 2.447 -ს შორის. ეს t -ქულები გამოიყენება 95% -იანი ნდობის ინტერვალისთვის, როცა n = 7.

კითხვა: df = 6 -სთვის t-ქულის რა მნიშვნელობა შეესაბამება 99%-იან ნდობის ინტერვალს?

14

სტატისტიკურ ჰიპოთეზთა შეფასება.

ძირითადად სტატისტიკური კვლევების მთავარი მიზანია შამოწმოს, მონაცემები უწყობს თუ პირიქით, არ უწყობს ხელს რაიმე დასკვნის გაკეთებას ან პროგნოზს. ამ დასკვნებს ეწოდებათ **ჰიპოთეზები** პოპულაციის შესახებ. ზოგადად, ისინი გამოისახებიან პოპულაციის პარამეტრების ტერმინებში იმ ცვლადების მიმართ, რომლებიც გაზომილ იქნა შესწავლის პროცესში. ე.ი. ჰიპოთეზა, როგორც წესი, ამბობს, რომ პარამეტრი ღებულობს რაიმე კონკრეტულ მნიშვნელობას ან მისი მნიშვნელობა ვარდება მნიშვნელობათა რაღაც დიაპაზონში. ეს პარამეტრები შეიძლება იყოს, მაგალითად, პოპულაციის პროპორციები ან ალბათობები.

მნიშვნელობის ტესტი შედგება ხუთი საფეხურისგან:

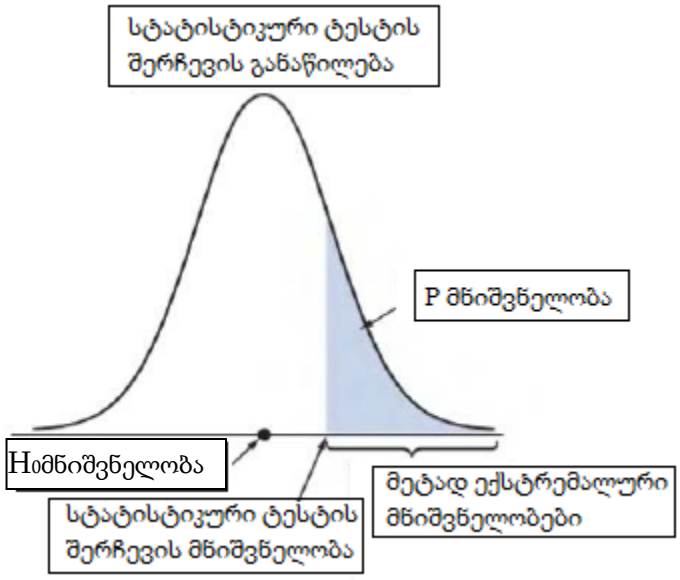
საფეხური 1: **დაშვებები.** ყოველ კრიტერიუმს გააჩნია თავისი გარკვეული დაშვებები, რა პირობებშიც შესაძლებელია მისი გამოყენება. უპირველეს ყოვლისა, ტესტი გულისხმობს, რომ მონაცემები უნდა იყოს შემთხვევითი. სხვა დაშვებები შეიძლება შეეხებოდეს შერჩევის ზომას ან პოპულაციის განაწილებას.

საფეხური 2: **ჰიპოთეზები.** მნიშვნელობის ყოველ ტესტს გააჩნია ორი ჰიპოთეზა პარამეტრების შესახებ: ნულ ჰიპითეზა და ალტერნატიული ჰიპოთეზა. **ნულ ჰიპოთეზა** არის დებულება იმის შესახებ, რომ პარამეტრი ღებულობს რაიმე კონკრეტულ მნიშვნელობას. **ალტერნატიული ჰიპოთეზა** ამბობს, რომ პარამეტრის მნიშვნელობა ვარდება მნიშვნელობათა რაიმე ალტერნატიულ დიაპაზონში.

ნულ ჰიპოთეზის რაიმე მნიშვნელობა, როგორც წესი, არ წარმოადგენს *არანაირ ეფექტს*. ალტერნატიული მნიშვნელობის ჰიპოთეზის მნიშვნელობა კი იწვევს ამა თუ იმ ტიპის ეფექტს. $\boldsymbol{H_0}$ სიმბოლოთი აღინიშნება ნულ ჰიპოტეზა , ხოლო $\boldsymbol{H_a}$ აღნიშნავს ალტერნატიულ ჰიპოთეზას.

საფეხური 3: **სტატისტიკური ტესტი.** პარამეტრს, რომელსაც შეეხება ტესტი, გააჩნია წერტილოვანი შეფასება. სტატისტიკური ტესტი გვიჩვენებს, რამდენად შორს ვარდება წერტილოვანი შეფასება ნულოვანი ჰიპოთეზით გასაზღვული პარამეტრის მნიშვნელობიდან. ცხადია, ეს მნიშვნელობები იზომება სტანდარტული გადახრის ტერმინებში წერტილოვან გადახრასა და პარამეტრს შორის.

საფეხური 4: **P-მნიშვნელობა.** სტატისტიკური ტესტის მნიშვნელობის ინტერპრეტაციისთვის ჩვენ ვიყენებთ ალბათობის შედეგებს ნულ ჰიპოთეზის წინააღმდეგ. ამას ასე ვაკეთებთ: დავუშვათ, რომ H_0 მართალია, რადგანაც მტკიცებულების ტვირთი აწევს H_a-ს. ამის შემდეგ დავალაგოთ მნიშვნელობები, რომელსაც ველოდებით ამ ტესტსიგან შერჩევის განაწილების შესაბამისად, იმ დაშვებით, რომ H_0 მართალია. თუ ამორჩევის სტატისტიკური ტესტი კარგად ვარდება შერჩევის განაწილების კუდში, მაშინ ეს ძალიან შორსაა H_0-ის პროგნოზისაგან. თუ H_0 მართალია, მაშინ ასეთი მნიშვნელობა იქნება უჩვეულო. თუ გვაქვს დიდი რაოდენობით შესაძლო შედეგები, რომელიმე ერთი მნიშვნელობა შეიძლება იყოს ნაკლებად ალბათური. შემდეგ ვადგენთ, რამდენად შორს შეიძლება ჩავარდეს სტატისტიკური ტესტის მნიშვნელობები და უფრო მეტად ექსტრემალური მნიშვნელობები (იგულისხმება კიდევ უფრო შორს ვიდრე ამას პროგნოზირებს H_0). იხილეთ გრაფიკი 14.1.



გრაფიკი 14.1. *დავუშვათ, რომ H_0 მართალია. P-მნიშვნელობა არის სტატისტიკური ტესტის ალბათობა, რომელიც გამოჩნდა ერთხელ ან არის მეტად ექსტრემალური.* ეს არის დაშტრიხული არე.

კითხვა: რომელია უფრო მკაცრი მტკიცებულება ნულ ჰიპითეზის წინააღმდეგ, P-მნიშვნელობა 0.2 თუ 0.01?

ამ ალბათობას ეწოდება **P-მნიშვნელობა.** რაც უფრო მცირეა P-მნიშვნელობა, მით მეტია მტკიცებულება H_0 ჰიპოთეზის წინააღმდეგ.

საფეხური 5: **დასკვნა.** კეთდება დასკვნა ჰიპოთეზის შეფასების შესახებ, რომელიც გვამცნიობს P-მნიშვნელობას და აკეთებს იმ კითხვების ინტერპრეტაციას, რის გამოც იყო მოტივირებული ეს ტესტი. ხანდახან იგი მოიცავს გადაწყვეტილებას H_0 ჰიპოთეზის სამართლიანობის შესახებ. მაგალითად, P-მნიშვნელობის საფუძველზე ჩვენ შეიძლება უკუვაგდოთ H_0 ჰიპოთეზა და დავასკვნათ, რომ ასტროლოგების პროგნოზები უკეთესია, ვიდრე შემთხვევითი გამოცნობა? შეიძლება განვიხილოთ ისიც, რომ გადავაგდოთ H_0 ჰიპოთეზა H_a–ს მხარდასაჭერად მხოლოდ მაშინ, როცა P-მნიშვნელობა არის ძალიან მცირე, როგორიცაა მაგალითად, 0.05 ან უფრო ნაკლები.

სცენარი 14.2. 40-საათიანი სამუშაო კვირა.

(მნიშვნელობის ტესტი პოპულაციის საშუალოს შესახებ).

1938 წელს მიღებული კანონით, სამართლიანი შრომის სტანდარტების შესახებ, აშშ-ში განისაზღვრა 40-საათიანი სტანდარტული სამუშაო კვირა. უკანასკნელ წლებში სამუშაო კვირის სტანდარტმა დასავლეთ ევროპასა და ავსტრალიაში დაიწია 40 საათის ქვემოთ. თუმცა ბევრს სჯერა, რომ შრომისმოყვარე კულტურის მქონე ამერიკაში ზეწოლა მოახდინეს, რომ ყოფილიყო 40-საათიან სტანდარტზე მეტი სამუშაო საათები. ისეთ წარმოებაში, როგორიცაა მაგალითად, საინვესტიციო- საბანკო მომსახურება, 40 - საათიანი კვირა განიხილება როგორც „უსაქმურობა“ და შეიძლება მიგვიყვანოს სამუშაოს დაკარგვამდე.

კითხვა შესწავლისთვის:

როგორ შეგვიძლია ჩამოვაყალიბოთ სამუშაო საათების კვლევა მნიშვნელობათა ტესტის გამოყენებით, რომელიც განსაზღვრავს, ამერიკის მშრომელი მოსახლეობის (პოპულაციის) საშუალო სამუშაო კვირა ტოლია 40 საათის, თუ განსხვავდება 40-საათიანი სტანდარტისგან? ჩამოვაყალიბოთ ნულოვანი და ალტერნატიული ჰიპოთეზები ამ ტესტისთვის.

გააზრებისთვის:

საპასუხო ცვლადი, სამუშაო კვირის ხანგრძლივობა, არის რაოდენობრივი ცვლადი. ჰიპოთეზა შეეხება აშშ-ს შრომისუნარიანი მოსახლეობის სამუშაო კვირის საშუალო ხანგრძლივობას და გვაინტერესებს, საშუალო ხანგრძლივობა μ არის მართლა 40 საათი თუ განსხვავდება 40 საათიანი სტანდარტისგან? რომ გამოვიყენოთ მნიშვნელობათა ტესტი, საჭიროა ჩატარდეს ანალიზი, თუ რამდენად ძლიერია მტკიცებულებები მოცემულ საკითხზე და ჩატარდეს ტესტირება H_0: μ = 40 ჰიპოთეზა H_a: μ ≠ 40 ჰიპოთეზის წინააღმდეგ.

ჩვდომის უნარი: მათ ვინც მუშაობდა 2008 წელს, სოციალური კვლევის კითხვაზე: „რამდენი საათი იმუშავეთ გასულ კვირაში?“ 636 მამაკაცის პასუხებიდან საშუალო იყო 45.5 და სტანდარტული გადახრა 15.16. გამოკითხული 567 ქალბატონისთვის საშუალო აღმოჩნდა 38.08, სტანდარტული გადახრა 12.57. მნიშვნელობათა ტესტის ხუთივე საფეხურის გავლის შემდეგ შეიძლება ჩვენება, რომ ქალბატონებისთვის 95% -იანი ნდობის ინტერვალია (37.040; 39.115).

15

ორი ჯგუფის შედარება.

რამდენად კარგად აფასებს ორი შერჩევის საშუალოებს შორის სხვაობა ($\overline{x_1}$ - $\overline{x_2}$), ორი პოპულაციის საშუალოებს შორის სხვაობას. ეს აღიწერება ე.წ. შერჩევის განაწილების სტანდარტული შეცდომით. დამოუკიდებელი შერჩევებისთვის ($\overline{x_1}$ - $\overline{x_2}$) -ის სტანდარტული შეცდომა არის:

$$\text{se} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

სცენარი 15.1. ნიკოტინზე დამოკიდებულება.

(საშუალოთა შედარება და სტანდარტული შეცდომის პოვნა).

ჩატარდა კვლევა სწავლის პერიოდში ახალგაზრდების ნიკოტინზე დამოკიდებულების შესახებ. გამოიკითხა 679 ბიჭი და 332 გოგონა. დამოკიდებულების ხარისხი შეფასდა 0-დან 10 ქულამდე. მეტი ქულა ნიშნავდა მეტ დამოკიდებულებას. მოზარდების ნიკოტინზე დამოკიდებულების კვლევის ერთერთი განმარტებითი ცვლადი იყო პასუხი კითხვაზე: ვინც ეწეოდა სწავლის პერიოდში, დარჩა თუ არა მწევლად მას მერე, რაც სწავლა დაასრულა? კლევის შედეგად სწავლის დასასრულისთვის აღმოჩნდა, რომ 75 კვლავ მწეველია და 257 ექს-მწეველი. ამასთან, მწეველებად დარჩენილთა საშუალო არის 5.9 (s = 3.3) და 1.0 (s = 2.3) ექს-მწევლებისთვის.

კითხვები შესწავლისთვის:

ა) შეადარეთ ერთმანეთს მწევლების და ექს-მწევლების საშუალო მაჩვენებლები.

ბ) რა არის შეჩევების საშუალოთა სხვაობის სტანდარტული შეცდომა? როგორ გესმით სტანდარტული შეცდომა (se)?

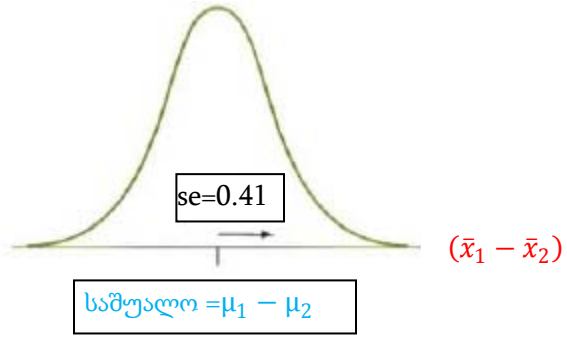
გააზრებითვის:

ა) დავუშვათ, რომ მწევლები არიან ჯგუფი 1 და ექს-მწევლები ჯგუფი 2. მაშინ, $\overline{x_1}$ = 5.9 და $\overline{x_2}$ = 1.0. რადგანაც ($\overline{x_1}$ - $\overline{x_2}$) = 5.9 − 1.0 = 4.9, ეს ნიშნავს , რომ 10 კითხვიდან საშუალოდ მწევლებმა უპასუხეს „დიახ“ ხუთით მეტ კითხვას, ვიდრე ექს-მწევლებმა. ეს ითვლება დიდ სხვაობად.

ბ) თუ ამ მონაცემების პირობებში გამოვიყენებთ se-ს ფორმულას ($\overline{x_1}$ - $\overline{x_2}$) -სთვის, მივიღებთ, რომ

$$\text{se} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(3.3)^2}{75} + \frac{(2.3)^2}{257}} = 0.41.$$

ეს აღწერს($\overline{x_1}$ - $\overline{x_2}$)-სთვის შერჩევის განაწილების ცვალებადობას, როგორც ეს ნაჩვენებია გრაფიკზე



თუ იმავე ზომის შემთხვევით შერჩევებს ავიღებდით შესასწავლი პოპულაციიდან, განსხვავებები შერჩევის საშუალოებს შორის ცვალებადობს კვლევიდან კვლევამდე. ($\overline{x_1}$ - $\overline{x_2}$) -ის მნიშვნელობათა სტანდარტული გადახრა დაახლოებით ტოლია 0.41.

წვდომის უნარი: სტანდარტული შეცდომა გამოიყენება ნდობის ინტერვალებში და მნიშვნელობის ტესტებში საშუალოთა შესადარებლად. მაგალითად, განვიხილოთ,

ნდობის ინტერვალი პოპულაციის საშუალოებს შორის სხვაობისთვის.

ვთქვათ, გვაქვს ორი შერჩევა, რომელთა ზომებია n_1 და n_2, სტანდარტული გადახრებით s_1 და s_2. გვსურს, ავაგოთ 95% -იანი ნდობის ინტერვალი პოპულაციის საშუალოების $(\mu_1 - \mu_2)$ სხვაობისთვის. მას აქვს ასეთი სახე:

$$(\overline{x_1} - \overline{x_2}) \pm t_{.025} \times (\text{se}), \text{ სადაც } \text{se} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

პროგრამული უზრუნველყოფა წარმოგვიდგენს $t_{.025}$-ს, t -ქულის მარჯვენა-კუდის ალბათობით 0.025 (მაშინ სრული ალბათობა = 0.95, რომელიც მოთავსებულია $-t_{.025}$-სა და $t_{.025}$-ს შორის).

ეს მეთოდი გულისხმობს:

- დამოუკიდებელი შემთხვევითი შერჩევები უნდა იყოს ორი ჯგუფიდან, ან შემთხვევითად შერჩეული ან შემთხვევითი ექსპერიმენტით მიღებული.
- პოპულაციის დაახლოებით ნორმალური განაწილებას თითოეული ჯგუფისთვის. (ეს ძირითადად მნიშვნელოვანია მცირე ზომის შერჩევებისთვის, და მაშინაც კი მეთოდი არის საკმაოდ მტკიცე მოთხოვნების დარღვევათა მიმართ).

სცენარი 15.2. ნიკოტინისადმი მიდრეკილება.

(ნდობის ინტერვალი).

მოდით იმავე კვლევაში - ახალგაზრდების ნიკოტინზე დამოკიდებულების შესახებ, დასკვნის სახით შევადაროთ ნიკოტინზე დამოკიდებულების საშუალო მწევლებისა და ექს-მწევლებისთვის. წინა მაგალითში გვქონდა, რომ $\overline{x_1} = 5.9$ და $s_1 = 3.3$, $n_1 = 75$ მწევლისთვის, და $\overline{x_2} = 1.0$ და $s_2 = 2.3$, $n_2 = 257$ ექს-მწევლისთვის.

კითხვები შესწავლისთვის:

ა) მწევლებისთვის და ექს-მწევლებისთვის შერჩევის განაწილებები არიან თუ არა მიახლოებით ნორმალური განაწილებები? როგორ მოქმედებს ეს შედეგზე?

ბ) პროგრამული უზრუნველყოფის ანგარიშებში ნათქვამია, რომ მწევლებისა და ექს-მწევლებისთვის პოპულაციის ქულათა საშუალოებს შორის სხვაობის, 95% -ანი ნდობის ინტერვალი არის (4.1; 5.7). აჩვენეთ, როგორაა ეს ინტერვალი მიღებული და გააკეთეთ შესაბამისი ინტერპრეტაცია.

გააზრებითვის:

ა) ექს-მწევლებისთვის (ჯგუფი 2) შემოთავაზებული სიდიდეები $\overline{x_2} = 1.0$ და $s_2 = 2.3$ შორსაა ზარის-ფორმისგან, რადგან უმცირესი მნიშვნელობა 0 არის უფრო მცირე, ვიდრე 1 სტანდარტული გადახრა საშუალოსგან ქემოთ. ეს არაა პრობლემური. დიდი შერჩევებისთვის ნდობის ინტერვალის აგების მეთოდი არ მოითხოვს პოპულაციის ნორმალურ განაწილებას, რადგან ცენტრალური ზღვარითი თეორემიდან გამომდინარეობს შერჩევის განაწილების ნორმალურობა. (გავიხსნოთ, რომ თუნდაც მცირე შერჩევებისთვის მეთოდი არის მდგრადი მაშინაც კი, როცა პოპულაციის განაწილება არ არის ნორმალური).

ბ) შერჩევის ამ ზომებისა და s_1 და s_2 მნიშვნელობებისთვის პროგრამული უზრუნველყოფის ანგარიშებში ნათქვამია, რომ $df = 95$. t -ქულის მნიშვნელობა 95% -იანი ნდობის ინტერვალისთვის არის $t_{.025} = 1.985$ (მიახლოებით, B ცხრილის მიხედვით $t_{.025} = 1.984$, როცა $df = 100$. $n_1 - 1$ -ზე და n_2 -1 -ზე კიდევ უფრო ნაკლებია 74, ამიტომ თუ თქვენ არ გაქვთ პროგრამული უზრუნველყოფა df -ის საპოვნელად, შეგიძლიათ გამოიყენით t-ქულა $df = 74$ -სთვის).

პოპულაციის საშუალო აღვნიშნოთ μ_1-ით მწევლებისთვის, და μ_2-ით ექს-მწევლებისთვის. წინა სცენარიდან ($\overline{x_1} - \overline{x_2}$) = 4.9 და აქვს სტანდარტული შეცდომა se = 0.41, მაშინ ($\mu_1 - \mu_2$) -სთვის 95% -იანი ნდობის ინტერვალი არის:

$$(\overline{x_1} - \overline{x_2}) \pm 1.985 \times (\text{se}), \text{ ან } 4.9 \pm 1.985 \times 0.41,$$

რომელიც ტოლია, 4.9 ± 0.8, ანუ (4.1; 5.7).

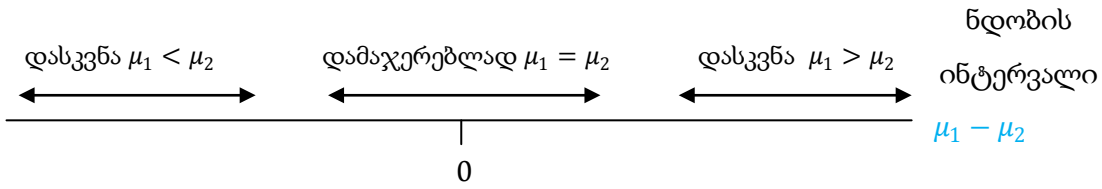
ჩვენ შეგვიძლია 95%-ით ვიყოთ დარწმუნებული, რომ პოპულაციის ქულების საშუალოებს შორის სხვაობა მწევლებისა და ექს-მწევლებისთვის ვარდება შუალედში (4.1; 5.7). შეგვიძლია დავასკვნათ, რომ მწევლებისთვის პოპულაციის საშუალო არის 4.1 -ით მეტიდან 5.7 -ით მეტამდე ვიდრე ექს-მწევლებისთვის.

ჭვდომის უნარი: ჩვენ 95%-ით გვჯერა, რომ მწევლებმა 10 კითხვიდან უპასუხეს „დიახ“, 4-დან 6-მდე მეტ კითხვას, ვიდრე ექს-მწევლებმა. 10 ბალიან სისტემაში ეს არსებითი განსხვავებაა.

საშუალოთა სხვაობის ნდობის ინტერვალის ინტერპრეტაცია.

გარდა ნდობის ინტერვალის ინტერპრეტაციისა, კარგი იქნება, თუ შევადარებთ საშუალოებს იმ კონტექსტში, თუ რამდენად დამაჯერებელია განსხვავება საშუალოებს შორის. თქვენ შეგიძლიათ იმსჯელოთ იმაზეც, რა შედეგებს მივღებთ, თუ გამოვიყენებთ ერთი და იმავე მეთოდებს აქაც და პროპორციების შედარებისთვისაც.

- *შევამოწმოთ, 0 ვარდება თუ არა ინტერვალში.* როდის არის 0 დამაჯერებელი მნიშვნელობა ($\mu_1 - \mu_2$) -სთვის? ანუ რაც იმას ნიშნავს, შესაძლებებელია, რომ $\mu_1 = \mu_2$? ქვემოთ მოტანილი ნახაზი გვიჩვენებს შემდეგს: დავუშვათ, რომ ($\mu_1 - \mu_2$) -სთვის ნდობის ინტერვალია (4.1; 5.7). მაშინ მწევლებისთვის პოპულაციის საშუალო ექს-მწევლებთან შედარებით შეიძლება იყოს სულ მცირე 4.1 -ით მეტი, ან მაქსიმუმ 5.7 -ით მეტი.



ნახაზი 15.1. *ორი საშუალოს საშუალოების ნდობის სამი ინტერვალი.* როდესაც ინტერვალი შეიცავს 0-ს, მაშინ ეს არის დამაჯერებელი, რადგან პოპულაციის საშუალოები შეიძლება ტოლები იყვნენ. წინააღმდეგ შემთხვევაში, ჩვენ შეგვიძლია ვივარაუდოთ, რომელი უფრო დიდია.

კითხვა: თუ ($\mu_1 - \mu_2$)-ის ნდობის ინტერვალი შეიცავს დადებით რიცხვებს, რატომ ვარაუდობენ, რომ $\mu_1 > \mu_2$?

- *თუ ($\mu_1 - \mu_2$) -ის ნდობის ინტერვალი შეიცავს მხოლოდ დადებით რიცხვებს, ვარაუდობენ, რომ ($\mu_1 - \mu_2$) არის დადებითი.* მაშინ ვასკვნით, რომ μ_1 მეტია μ_2 -ზე. 95% -იანი ნდობის ინტერვალია (4.1; 5.7). რადგანაც *($\mu_1 - \mu_2$)* არის სხვაობა მწევლებისა და ექს-მწევლების საშუალოებს შორის, ვასკვნით რომ პოპულაციის საშუალო მაღალია მწევლებისთვის.
- *თუ ($\mu_1 - \mu_2$)-ის ნდობის ინტერვალი შეიცავს მხოლოდ უარყოფით რიცხვებს, ვარაუდობენ, რომ ($\mu_1 - \mu_2$) არის უარყოფითი.* მაშინ ვასკვნით, რომ μ_1 ნაკლებია μ_2 -ზე.
- *რომელი ჯგუფია 1 და რომელია 2, ამის დასახელება ნებისმიერია.* თუ თქვენ გაუცვლით სახელებს, მაშინ ნდობის ინტერვალს ექნება იგივე ბოლო წერტილები, მაგრამ შებრუნებული ნიშნით. მაგალითად, (4.1; 5.7) შეიცვლება (- 5.7; - 4.1) -ით.