Veleučilište u Šibeniku

Poslovna statistika za stručne studije

Patrick R. McMullen

Prijevod:

Kristina Devčić

Ana Perišić

Šibenik, 2017.

Naslov originala:
"UNDERGRADUATE BUSINESS STATISTICS"
Wake Forest University, School of Business
ISBN-10: 1535091630, ISBN-13: 978-1535091633

Recenzent:

doc.dr.sc. Božidar Ivanković, znanstveni suradnik

ISBN: 978-953-7566-39-5

Sadržaj

Uvod

Odluku o pisanju ove knjige donio sam tijekom vremena, a na kraju se pokazala neizbježnom. Primijetio sam kako studenti i njihove obitelji troše previše novca na udžbenike, a i školarine su postale vrlo visoke. Ovaj problem nisam u potpunosti doživio sve dok moja djeca nisu postala studenti te sam tek tada taj problem doživio na vlastitoj koži.

U skladu s tim, počeo sam pisati ovu knjigu u rujnu 2015. godine kako bi studenti dobili dobru literaturu koja bi im pomogla pri učenju početne poslovne statistike izbjegavajući ovisnost o velikim izdavačkim kućama i njihovim pretjeranim cijenama.

Još jedan razlog za pisanje ove knjige je povećanje konzistentnosti između mojih predavanja i sadržaja u knjizi. Jedna od uobičajenih povratnih informacija koje sam dobivao od mojih studenata je da im se literatura ne sviđa. Ne mislim da je problem u kvaliteti literature, nego u načinu na koji sam ju koristio. Uz ovu knjigu, moći ću bolje uskladiti svoja predavanja s literaturom, što bi trebalo koristiti studentima. Time će se povećati povezanost između literature, predavanja i zadataka za domaću zadaću.

Također, zadaci za domaću zadaću dani u knjizi odnosit će se na podatke koje sam pripremio što će također rezultirati boljom povezanošću između predavanja i zadataka.

Programska podrška

Odgovarajući programski paket koji možete koristiti je Microsoft Excel, verzija 2010 i novije. Iako ne smatram da je Microsoft Excel najbolji dostupan statistički program, u konačnici, to je programski paket kojem svi studenti imaju pristup. Usporedit ću sposobnost Microsoft Excela u izračunavanju statističkih pokazatelja sa švicarskim nožem - puno stvari radi dobro, ali nijednu od njih iznimno dobro. Microsoft Excel je fleksibilan i pruža alate za uspješnu analizu podataka u nastavi.

Drugi programski paket koji će se povremeno koristiti je JMP, opći statistički programski paket kojeg je proizveo SAS. Zbog ograničene dostupnosti nakon studiranja JMP se ne koristi tako često kao Microsoft Excel. Ipak, JMP će biti korišten povremeno kako bi se pokazale neke programske sposobnosti koje Excel jednostavno ne može izvesti. To se osobito odnosi na izradu određenih tipova grafikona.

Drugi programski paket koji će se rijetko koristiti je R. R je ono što je poznato kao programski paket otvorenog koda kojeg distribuira grupa korisnika poznata kao *R-projekt*. To znači da se R može besplatno preuzeti, a korisnici mogu, ukoliko to odluče, raditi poboljšanja programa. Program se često ažurira, a neka od novih poboljšanja su uključena u novim verzijama, uz pretpostavku da nova poboljšanja vodstvo R-projekta smatra vrijednima. R je vrlo moćan program te može riješiti bilo koju vrstu problema koju susrećemo u ovom kolegiju. Jedini nedostatak programa R je da je potrebno malo vremena za svladavanje njegovih funkcija. No, nakon što se osnove usvoje, R postaje vrlo koristan.

Podaci se nalaze na CD-u kao prilog tiskanoj verziji udžbenika. Također, dostupni su u elektronskoj verziji na mrežnim stranicama kolegija na kojima se ovaj udžbenik koristi.

Potrebno predznanje

Za gradivo obuhvaćeno ovim kolegijem potrebno je temeljno razumijevanje srednjoškolske ili više matematike. Osim toga, povremeno je potrebna određena "matematička zrelost". U statistici, često zbrajamo veličine i pritom koristimo indeksni zapis. Razlog tome nije da budemo pretenciozni, nego je potrebno objasniti nešto što je kraće moguće.

Čitajući knjigu, a osobito kada se pokušava razumjeti formule, važno je osvijestiti da se ne radi o literaturi za čitanje prije spavanja. Određeno vrijeme je potrebno da se shvati objašnjeni sadržaj. Zato su dani primjeri kad god se to smatralo potrebnim.

Druga pretpostavka je da je važno poznavati osnove Microsoft Excela. Pod osnovama se očekuje da student može napisati formulu, kopirati, zalijepiti i kreirati osnovne vrste grafikona.

Novosti ovog izdanja

U ovo izdanje ugrađeno je nekoliko poboljšanja u odnosu na prethodno. Dodano je više zadataka za domaću zadaću, a ispravljene su i gramatičke greške koje su prije prošle neopaženo. Gramatičke nedosljednosti su riješene te je dodano novo poglavlje o pogreškama tipa II i snazi testa.

Zahvale

Ironično, moram se zahvaliti izdavačkim kućama i sveučilišnim knjižarama što su mi dali motivaciju za pisanje ove knjige. Zbog njihovog partnerstva i činjenice da naplaćuju studentima previše novca, smatram da je sada, više nego ikad, potrebno imati pristupačan udžbenik.

Želio bih se zahvaliti mom prijatelju i kolegi Jonu Pinderu što me na ovo nagovorio. Jon je počeo raditi isto prije nekoliko godina s namjerom da studentima uštedi novac.

Također, želio bih zahvaliti Mikeu DiCellou na njegovoj pomoći oko fotografija, Karen Combs i Mariu Rodriguezu Neda na njihovoj pomoći u uređivanju i Vickie Whapham na administrativnoj pomoći. Također, važno mi je zahvaliti se Kevinu Benderu, Patu Peacocku, Carol Oliff, Lynn Zimmerman, Sharon Payne i Chasu Mansfieldu na njihovoj beskrajnoj pomoći u pomaganju studentima da bolje razumiju važnost procesa učenja.

Najvažnije, želio bih se zahvaliti profesoru Larryju Richardsu sa Sveučilišta u Oregonu. Dok sam bio tamo, održao sam nekoliko predavanja iz statistike s Larryjem. Pod njegovim mentorstvom sam

shvatio da je statistika najvažnije predavanje iz matematike koje postoji. Tada sam shvatio da bi bilo zabavno podučavati statistiku jednog dana. Larryjevo mentorstvo mi je dalo samopouzdanje da podučavam statistiku. Stati pred studente koji imaju strah od statistike nije jednostavan zadatak. Larry mi je dao samopouzdanje da to učinim.

Patrick R. McMullen
Winston-Salem, Sjeverna Karolina

1. Zašto statistika?

Statistika je predmet straha u poslovnim školama - kako za studente preddiplomskih, tako i diplomskih studija. Autor osobno vjeruje u ovu teoriju s obzirom da je dobio ocjenu dovoljan kad je prvi put polagao statistiku kao student preddiplomskog studija tehničkog fakulteta. Pod pretpostavkom da niste bacili svoju knjigu nakon navedenog priznanja, nastavit ćemo dalje. ☺

Prije prvog odslušanog predavanja iz statistike dio studenata je uzbuđen. Razlog je taj što misle da znaju sve o statistici. Kao školarci, prijatelji razmjenjuju sličice kako bi skupili kompletne albume. U SAD-u se uglavnom skupljaju sličice baseball igrača. Tako, ako netko ima dvije sličice Petea Rosea iz 1974. godine, a nedostaje mu sličica Henryja Aarona iz 1974. godine, mijenja jednu od svojih sličica Petea Rosea za sličicu Henryja Aarona.

Baseball sličice obiluju statističkim podacima o igri prikazanog igrača u cijeloj njegovoj karijeri. Henry Aaron je 1974. godine srušio rekord svih vremena u optrčavanju kojeg je davno postavila baseball legenda Babe Ruth. Kad je Hank Aaron uspio u tom naizgled nemogućem podvigu, zauvijek se upisao u legendu baseball-a. Tablica 1.1 prikazuje odabrane statističke podatke o karijeri Henryja Aarona.

Karijera	OU	BU	B	U	O	RBI	PBU
23 godine	3298	12364	2174	3771	755	2297	0.305

Tablica 1.1. Odabrani statistički pokazatelji o karijeri Henryja Aarona (OU – broj odigranih utakmica, BU – broj prilika za udarac, B – broj bodova, U – broj udaraca, O – broj optrčavanja, RBI, PBU – prosječan broj udaraca)

Iz pregleda tablice 1.1 kao i pregleda drugih detaljnih statističkih pokazatelja o bacanjima u karijeri postaje jasno da je Henry Aaron nedvojbeno bio jedan od najboljih bacača u povijesti ove igre. Konkretno, on ima više RBI (2297) nego itko u povijesti, a 755 optrčavanja u karijeri je bilo najviše ikad dok Barry Bonds nije oborio rekord 2007. godine sa 762 optrčavanja. Nažalost za Barryja Bondsa, oko njegovog rekorda će uvijek biti nedoumica zbog njegovog korištenja lijekova za poboljšavanje izvedbe[1]. Zbog kontroverzi oko Bondsa mnogi još uvijek smatraju Henryja Aarona kraljem optrčavanja.

Da se vratim na temu, to je u biti bilo ono što sam ja smatrao statistikom - proučavanje mnogo brojeva - nešto slično proučavanju bejzbolskih sličica kako bih saznao nešto o uspjehu igrača. Nisam mogao biti u većoj zabludi.

Proučavanje mnogo brojeva, poput brojeva na baseball sličicama, ipak je dio statistike, i to vrlo mali dio statistike.

Statistika je, u svom najopćenitijem obliku, učenje o populaciji. Učenje o populaciji zahtijeva proučavanje populacije što je prilično teško s obzirom na veličinu populacije. Problem je moguće riješiti prikupljanjem podataka iz dijela populacije ili *uzorka* kako bi pomogli donijeti određene zaključke o populaciji.

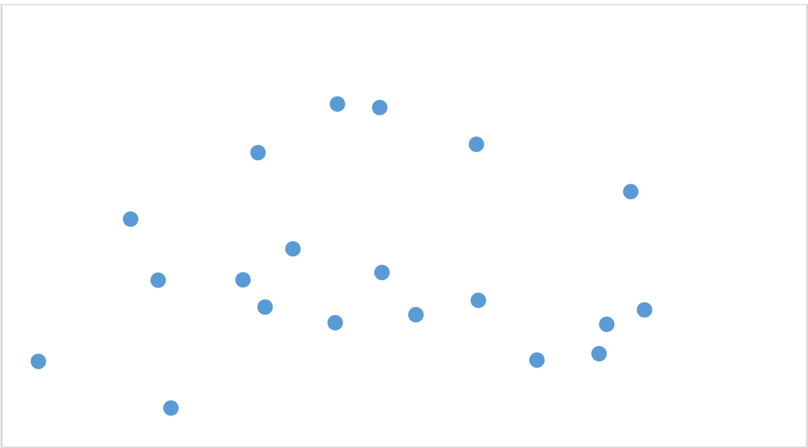
Iako to može zvučati prilično lako, treba biti oprezan prilikom prikupljanja podataka. Važno je da su podaci nasumično odabrani kako bi se izbjegla bilo kakva pristranost.

Slijedi primjer pristranog uzorka. Pretpostavimo da je zadatak procjene prosječne cijene kuća u SAD-u zadan studentu iz Brentwooda[2] u Kaliforniji. Jednostavan način bio bi pronaći nedavne transakcije u prometu nekretninama u susjedstvu (Brentwood), izračunati prosječnu vrijednost tih transakcija i uzeti ju kao prosječnu cijenu kuće u SAD-u.

Problem u navedenom slučaju je prilično jasan. Uzorak nije dobar reprezentant populacije kuća američkog stanovništva. Brentwood u Kaliforniji je vrlo bogato područje i prosječna cijena kuća je mnogo viša nego što bi bila za cijeli SAD. Umjesto toga, trebalo je *nasumce* ili

[1] Bonds je priznao da je koristio lijek za poboljšavanje izvedbe, ali je tvrdio da nikad nije bio svjestan činjenice da se radi o takvom lijeku.
[2] vrlo bogato područje u Kaliforniji

slučajno odabrati uzorak kuća iz cijelog SAD-a i provesti analizu takvog uzorka. Kada je za analizu nekog obilježja populacije nasumce i nepristrano odabran uzorak, tada se zaključci mogu primijeniti na cijelu populaciju. Grafikon 1.1 prikazuje slučajan odabir uzorka ili uzorkovanje gdje su točkice nasumično odabrani elementi statističkog skupa ili šire populacije koju predstavlja pravokutnik.



Grafikon 1.1. Uzorak odabran iz populacije

Nakon prikupljanja slučajnog uzorka iz populacije, provodi se odgovarajuća analiza. Namjera je iz podataka dobiti rezultate koji se mogu javno objaviti. Nakon objave rezultata, očekuje se dobro definiran skup alata i semantičkih pravila za interpretaciju dobivenih rezultata. Prije objavljivanja rezultata provodi se strukturirani ili formalni test koji pokazuje je li u provedenim analizama primijenjen pravilan protokol.

U znanstveno-istraživačkoj zajednici se bilo kakva bitna tvrdnja ne uzima ozbiljno ako nije popraćena adekvatnim statističkim testiranjem. Poslovna zajednica možda nije tako stroga kao znanstveno-istraživačka zajednica, ali provođenje odgovarajuće analize i testiranja svakako je potrebno u većini situacija.

Zamislimo tvrtku koja će upravo započeti novu reklamnu kampanju za svoj proizvod. Prije nego potroši milijune dolara kako bi pokrenula kampanju, morat će najprije provesti testno oglašavanje na fokus grupama kako bi se utvrdilo reagiraju li fokus grupe pozitivno na novu reklamnu kampanju. Da bi se to utvrdilo, moraju se provesti formalne statističke analize koje će uvjeriti viši menadžment da je ulaganje u kampanju isplativo.

Drugi primjer se odnosi na Agenciju za hranu i lijekove (FDA) te njihovo moguće odobrenje novog lijeka. Njihov je posao osigurati da je određeni lijek učinkovit i bez opasnih nuspojava. Kako bi se novi lijek odobrio, FDA mora imati *pravi statistički dokaz* o učinkovitosti lijeka bez opasnih nuspojava.

Formalni statistički alati koji se koriste u takvim postupcima i odlukama su ukratko opisani u sljedećem potpoglavlju.

1.1 Statistički alati

S obzirom da je ova knjiga zamišljena kao uvod u statistiku, dan je prikaz obrađenih tema. One su sljedeće:

- ***Deskriptivna statistika*** je temelj svih daljnjih analiza. Ovdje će se opisati skupove podataka na najosnovniji mogući način, brojčano i grafički.
- ***Vjerojatnost*** je mjera izvjesnosti određenog slučajnog događaja u uvjetima nesigurnosti.
- ***Slučajne varijable*** pomoću kojih se proučava struktura određenih vrsta distribucije podataka.
- ***Procjenjivanje*** je postupak kojim se na temelju uzorka izabranog iz (veće) populacije računaju određene populacijske veličine te se određuje interval u kojem se nalazi promatrana populacijska veličina s određenom vjerojatnosnom pouzdanošću.
- ***Testiranje hipoteza*** je najvažnije područje koje pokriva ova knjiga. Kada se postavi određena tvrdnja, tada se to mora popratiti formalnim testom. Testiranje hipoteza omogućuje točno definiranu strukturu koju je potrebno koristiti da bi se donio zaključak o odbacivanju ili neodbacivanju neke tvrdnje.

- ***Jednosmjerna analiza varijance*** se koristi za usporedbu više populacija, tj. kako bi se utvrdila podudarnost populacija s obzirom na neko obilježje.
- ***Hi kvadrat test*** je postupak kojim se (među ostalim) uspoređuje dana distribucija s nekom teoretskom distribucijom.
- ***Jednostavna linearna regresija*** je još jedna od važnijih tema u ovoj knjizi. U jednostavnoj linearnoj regresiji se istražuje moguća veza između dvaju numeričkih obilježja. Ako postoji smislena veza, možemo ju iskoristiti za procjenu i/ili predviđanje jedne veličine uz poznatu vrijednost druge veličine.
- ***Analiza odlučivanja*** je postupak u kojem se odlučuje između mogućih alternativa u uvjetima neizvjesnosti u budućnosti. Odluke se donose pomoću različitih strategija na temelju procjene razine neizvjesnosti.

1.2 Zaključci

Čini se razumnim razmišljati o statistici kao o alatu pomoću kojeg se može bolje razumjeti okolinu na temelju skupa podataka. Ovi alati pomažu nam u proučavanju okoline. Kad bolje razumijemo svoju okolinu, možemo poboljšati i unaprijediti položaj naše organizacije, bez obzira na svrhu te pojedine organizacije.

1.3 Zadaci

1. Bi li Boston u Massachusettsu bio dobro mjesto za odabir uzorka na temelju kojeg bi se procijenilo prosječne bodove ostvarene na završnim ispitima (npr. državna matura/SAT) za sve srednjoškolce? Zašto da ili zašto ne?
2. Bi li država Ohio bila dobro mjesto za odabir uzorka za procjenu je li se kupcima svidio novi prehrambeni proizvod? Zašto da ili zašto ne?
3. Zašto bi upotreba statistike u Agenciji za hranu i lijekove bila važnija nego u bilo kojoj organizaciji?

2. Opisivanje podataka

Statistika se bavi analizom podataka i interpretacijom dobivenih rezultata. Prije provođenja formalnih statističkih testova, potrebno je opisati podatke u njihovom najosnovnijem obliku. Postoje tri osnovna pristupa u opisu podataka: analiza kvalitativnih podataka, analiza kvantitativnih podataka i grafičko prikazivanje podataka.

2.1 Analiza kvalitativnih podataka

Opisivanje kvalitativnih podataka je uglavnom jednostavno, kao vježba zdravog razuma. Podaci kvalitativne prirode su podaci koji nisu dani u brojčanom, nego u opisnom obliku. Podaci se svrstavaju u kategorije ili razrede (engl. *bin*). Ovakva vrsta podataka najčešće se organizira u tablicu i/ili prikazuje jednostavnim grafikonom.

Primjerice, pretpostavimo da se tvrtka, koja se bavi proizvodnjom i prodajom odjeće, želi usredotočiti na populaciju koja često pohađa rock koncerte. Zainteresirani su za dobivanje informacija o boji hlača koju ta populacija nosi. Zapošljavaju ljude koji odlaze na koncerte i broje osobe koje nose određenu boju traperica. Odlučili su se za četiri kategorije: crne, plave, bež i ostalo. Osim po boji traperica, posjetitelje koncerta su kategorizirali i prema spolu. Nakon prebrojavanja, svoje rezultate su predstavili u tablici 2.1.a u kojoj su dani podaci o svakoj kategoriji boje razvrstani prema spolu.

	Crne	Plave	Bež	Ostalo	Ukupno
Muškarci	236	326	73	22	657
Žene	158	75	42	31	306

Tablica 2.1a. Broj traperica prema spolu

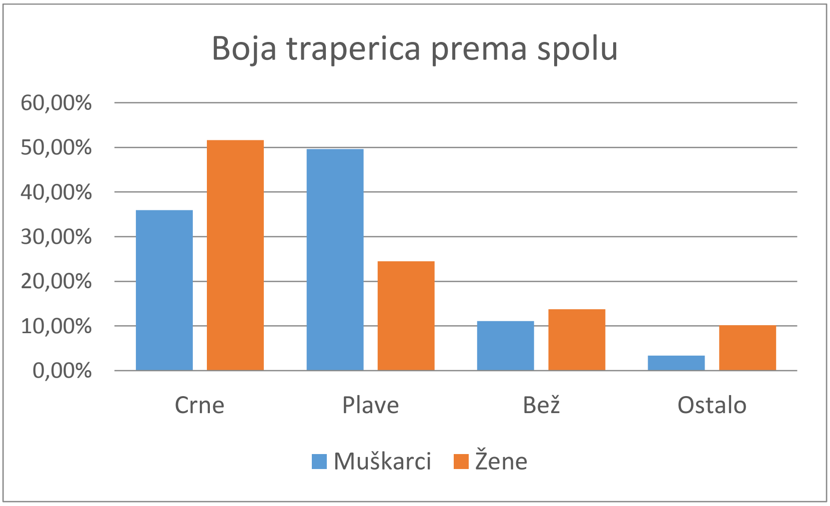
Analizu se može provesti i korak dalje tako da se izračunaju postoci o bojama traperica prema spolu. Da bi se to učinilo za muškarce, svaki element prvog retka treba podijeliti sa 657, primjerice 236 treba podijeliti sa 657 kako bi se dobio postotak muškaraca koji nosi traperice crne boje, 326 treba podijeliti sa 657 kako bi se dobio postotak muškaraca koji nosi traperice plave boje i tako dalje. Kada se učini isto za žene (podijeli frekvencija drugog retka svake boje s 306), dobiju se rezultati prikazani tablicom 2.1b.

	Crne	Plave	Bež	Ostalo	Ukupno
Muškarci	35.92%	49.62%	11.11%	3.35%	100%
Žene	51.63%	24.51%	13.73%	10.13%	100%

Tablica 2.1b. Postoci za boje traperica prema spolu

Tablica 2.1b je bolji način prikazivanja dobivenih podataka budući da je prikaz standardiziran u odnosu na početne podatke.

Standardizirani podaci se mogu prikazati stupčastim grafikonom (ili histogramom) gdje visina svakog stupca odgovara postotku odabrane boje traperica, posebno za muškarce i za žene. Ovaj grafikon je nastao korištenjem "Stupčastog grafikona" (engl. "*Column Chart*") u programu Microsoft Excel. Grafikon u obliku torte (engl. „*Pie Chart*") je još jedna mogućnost prikazivanja ovih podataka gdje veličina svake kriške predstavlja postotak određene boje. Ipak, u posljednjih nekoliko godina, statističari tvrde da ova vrsta grafikona iskrivljuje perspektivu. Kao takav, stupčasti grafikon je prikazan kao preferirani grafički alat za podatke dobivene prebrojavanjem.



Grafikon 2.1. Stupčasti grafikon boja traperica prema spolu

2.2 Analiza kvantitativnih podataka

Ovdje počinje prava analiza podataka. Najprije se deskriptivnom statistikom dobiva određeni skup opaženih podataka, zatim se za njih izračunavaju brojni statistički pokazatelji iz kojih se interpretiraju dobiveni rezultati na smislen način. Dva pokazatelja su najvažnija: očekivanje i disperzija.

2.2.1 Očekivanje

Očekivanje je mjera centralne tendencije – jedinstvena vrijednost koja se koristi za opisivanje sredine danog skupa podataka. Primjerice, kada profesor ocjenjuje ispite, često navodi prosječan uspjeh na ispitu za cijeli razred. Profesor zbroji sve rezultate i dobiveni rezultat podijeli brojem studenata. Profesor u biti interpretira prosječan uspjeh studenata na ispitu. Na temelju konkretnog rezultata, svaki student može usporediti svoj uspjeh s prosječnim. U svijetu statistike, prosječna vrijednost neke veličine se često naziva **aritmetičkom sredinom**.

Matematički, aritmetička sredina uzorka ($\bar{x}$) definirana je formulom:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2-1}$$

gdje je x_i i-to opažanje, a n je ukupan broj opažanja u uzorku.

Prosjek nije jedina mjera koja se može koristiti za opisivanje centralne tendencije. Zamislite da smo dobili zadatak procijeniti očekivanu vrijednost kuće u King Countyju u Washingtonu. Odlučili smo da ćemo za procjenu s tog područja nasumično odabrati 10 kuća u uzorak. Ispostavi se da je jedna od deset odabranih kuća dom osnivača Microsofta Billa Gatesa. Kuća gospodina Gatesa vrijedi nekoliko milijuna dolara - znatno više od ostalih kuća u skupu podataka. To će znatno iskriviti naš prosjek – napuhat će prosjek i lažno prikazati očekivanu vrijednost kuća u King Countyju u Washingtonu. Umjesto izračunavanja prosjeka, možemo poredati naše podatke od najmanje do najveće (ili od najveće do najmanje) vrijednosti i kao središnju vrijednost odabrati onu koja uređen niz dijeli na dva jednakobrojna dijela. Ovdje ćemo za neparan broj podataka kao središnju vrijednost odabrati onaj član niza koji uređen niz dijeli na dva jednakobrojna dijela, dok za paran broj podataka kao središnju vrijednost odabrati prosječnu vrijednost dviju središnjih vrijednosti. Ova vrijednost se naziva **medijanom** i često se koristi kod društveno-ekonomskih podataka kako bi se uklonile bilo kakve pristranosti uzrokovane ekstremnim vrijednostima.

Još jedna mjera centralne tendencije je **mod** – najčešća opažena vrijednost u skupu podataka. Ponekad, ovisno o prirodi podataka, mod ne postoji. S druge strane, u nekim slučajevima postoji više modova. Primjerice, ako skup podataka sadrži dva moda, kažemo da su podaci bimodalni.

Ugrađene funkcije za izračunavanje mjera centralne tendencije u Microsoft Excelu su prikazane tablicom 2.2. U tom kontekstu kroz cijelu knjigu se izraz "raspon podataka" odnosi na stvarne podatke u Excel tablici.

Statistički pokazatelj	Formula u Excelu
Srednja vrijednost	*=AVERAGE(raspon podataka)*
Medijan	*=MEDIAN(raspon podataka)*
Mod	*=MODE(raspon podataka)*

Tablica 2.2. Formule u Excelu za mjere centralne tendencije[3]

2.2.2 Disperzija

Disperzija je nedovoljno cijenjen i ignoriran dio deskriptivne statistike, ali vrlo važan. U stvari, jednako je važna kao i mjere centralne tendencije. Razlog što nije dovoljno cijenjena nalazi se u nedovoljnom razumijevanju iste. U stvari, kad nisam uspio položiti ispit iz statistike prvi put, moje nerazumijevanje disperzije bio je glavni razlog neuspjeha.

Disperzija je raspršenost promatranog skupa podataka.

[3] U novijim verzijama Excela (2010 i novije) koriste se funkcije *=mode.sngl* i *=mode.mult*

Najjednostavniji oblik disperzije koji možemo pojmiti je **raspon** (engl. *Range*), koji je jednak razlici najveće (x_{max}) i najmanje (x_{min}) opažene vrijednosti. Matematički, raspon je jednak:

$$Raspon = x_{max} - x_{min} \tag{2-2}$$

Iako je ovako definiranu raspršenost jednostavno izračunati, upravo zbog njene jednostavnosti pri donošenju zaključaka treba biti oprezan jer ne uzima u obzir raspršenost opažanja u odnosu na aritmetičku sredinu. Vrijednost koja najbolje objašnjava raspršenost oko aritmetičke sredine je **uzoračka standardna devijacija**, u oznaci s. Na prvi pogled, matematički zapis ove mjere može izgledati zastrašujuće, ali uz malo objašnjenja, trebala bi postati razumljivija. Uzoračka standardna devijacija definira se formulom:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2-3}$$

Ova formula uključuje izračun razlike između svakog opažanja i aritmetičke sredine uzorka koja se potom kvadrira. Intuitivno je moguće objasniti ovaj postupak. Kvadriranje se vrši iz dva razloga; prvi razlog je kako bi se uklonile negativne vrijednosti u izračunu budući da je kvadrat bilo kojeg broja pozitivan broj, dok je drugi razlog naglašavanje velikih razlika između opaženih vrijednosti i aritmetičke sredine uzorka. Te kvadrirane razlike se zbrajaju što čini brojnik jednadžbe. Dobiveni brojnik se potom dijeli s $n-1$, gdje je n veličina uzorka. Razlog tome je izračun prosječne vrijednosti, slično dijeljenju s n pri izračunavanju aritmetičke sredine uzorka. Vrijednost 1 se oduzima od n kako bi se uzelo u obzir da se radi o podacima iz uzorka umjesto o podacima iz populacije. Oduzimanje 1 od n prilagođava vrijednost standardne devijacije uzorka na više, "napuhuje" procjenu standardne devijacije u odnosu na standardnu devijaciju populacije. Srećom, ovu formulu nije potrebno često koristiti u praksi, budući da Excel i drugi programski paketi jednostavno daju traženi izračun.

Formula za izračunavanje uzoračke varijance je kvadratna vrijednost uzoračke standardne devijacije, a definira se formulom:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2-4}$$

Može se primijetiti kako je uzoračka standardna devijacija kvadratni korijen uzoračke varijance te je uobičajena u primjeni jer se vrijednost uzoračke standardne devijacije izražava u istim jedinicama kao i opažene vrijednosti.

Formule u Excelu za izračun raspona podataka, uzoračke standardne devijacije i varijance su prikazane u tablici 2.3.

Statistički pokazatelj	Formula u Excelu
Raspon	=*MAX*(*raspon podataka*) – *MIN*(*raspon podataka*)
Uzoračka standardna devijacija	=*STDEV*(*raspon podataka*)
Uzoračka varijanca	=*VAR*(*raspon podataka*)

Tablica 2.3. Formule u Excelu za disperziju[4]

Sada kada imamo alate (formule) za opisivanje centralne tendencije i disperzije, ukratko će se objasniti pojam distribucije skupa podataka.

2.2.3 Distribucija

Kvantitativna opažanja čine skup podataka koji se mogu poredati po veličini tako da svako opažanje ima svoju poziciju u skupu podataka. Kada su podaci sortirani, možemo usporediti položaj ili poziciju opažanja u skupu podataka u odnosu na ostatak podataka. Excel, JMP i drugi programski paketi daju položaj određenog opažanja u skupu podataka bez potrebe za prethodnim sortiranjem ili bilo kojim drugim oblikom manipulacije podacima. Sada je moguće izračunati pozicijske vrijednosti unutar određenog skupa podataka koji dijele uređen niz podataka na n jednakih dijelova. Dobivene pozicijske vrijednosti

[4] U novijim verzijama Excela (2010 i novije) koristi se *stdev.s(raspon podataka)* i a *=var.s(raspon podataka)*.

nazivamo **kvantilima**.

Kvantil nekog opažanja u skupu podataka je njegov relativni položaj u odnosu na druga opažanja u tom skupu, pri čemu su u praksi najčešće korišteni **kvartili** i **percentili**. Percentil skupa podataka dijeli podatke na 100 jednakih dijelova. Kvartil dijeli skup podataka na četiri jednaka dijela. Prvi kvartil se svodi na 25. percentil, drugi kvartil na 50. percentil (što je ujedno i medijan), a treći kvartil na 75. percentil. Četvrti kvartil, koji se nikada ne koristi, je u suštini maksimalna opažena vrijednost. Koncept kvantila bi vam trebao biti poznat. Primjerice, često se nakon pristupanja ispitima znanja i sposobnosti, kao ishod ispita dobiva brojčani rezultat koji često nije od velikog značaja ako se nema s čim usporediti. Ali, ponekad su predstavljeni i podaci o percentilima koji daju informaciju o uspjehu na ispitu u odnosu na sve pristupnike. Primjerice, ako je vaš percentil bio 75, to znači da je vaš uspjeh bolji od uspjeha 75% svih pristupnika testu.

U Excelu se vrijednosti kvartila i kvantila lako izračunaju koristeći funkcije *percentile* i *quartile*.

Statistički pokazatelj	Formula u Excelu
Percentil	*=PERCENTILE(raspon podataka, vrijednost)*
Kvartil	*=QUARTILE(raspon podataka, kvart)*

Tablica 2.4. Formule u Excelu za percentile i kvartile[5]

Primjerice, ako imamo skup podataka pod nazivom X s nekoliko opažanja, funkcija "$=PERCENTILE(X, 0.63)$" će dati vrijednost u skupu podataka na poziciji 63. percentila. Ako koristimo funkciju "$=QUARTILE(X, 3)$", ona će dati vrijednost u skupu podataka na poziciji 3. kvartila (ili 75. percentila). Primjetite da za funkciju percentila "*vrijednost*" mora biti između 0 i 1. Nulti percentil je minimalna vrijednost, dok je 100. percentil najveća vrijednost u skupu podataka. Kod vrijednosti kvartila, argument funkcije kvartila mora biti 0, 1, 2, 3 ili

[5] U novijim verzijama Excela (2010 i novije) koriste se funkcije *=percentile.exc(raspon podataka, vrijednost)* i *=percentile.inc(raspon podataka, vrijednost) te funkcije =quartile.exc(raspon podataka, kvart)* i *=quartile.inc(raspon podataka, kvart)*.

4. Nulti kvartil je minimalna vrijednost, dok je četvrti kvartil najveća vrijednost u skupu podataka. Treba napomenuti da nulti i četvrti kvartil nemaju praktičnu vrijednost budući su to minimalna i maksimalna vrijednost.

Treba napomenuti da funkcije prikazane u tablici 2.4 pretpostavljaju da postoji mogućnost da se dosegne najveća vrijednost - kao primjerice rezultat ispita. Ako tražimo najmanji mogući rezultat, jednostavno zamijenimo *“=PERCENTILE(raspon podataka, vrijednost)”* s *“=1-PERCENTILE(raspon podataka, vrijednost)”*. Za kvartile, jednostavno zamijenimo vrijednosti za prvi i treći kvartil.

Ograničili smo pokrivenost "distribucije" na percentile i kvartile. Postoje i druge mjere distribucije, poput **koeficijenta asimetrije** i **mjere zaobljenosti** koji su izvan opsega ove knjige. U sljedećem poglavlju se i dalje raspravlja o distribuciji skupa podataka, ali na konceptualniji način.

2.3 Grafičko prikazivanje podataka

Slika vrijedi tisuću riječi. Što se tiče ove poslovice, statistika nije iznimka. Današnja računala i programski paketi su uvelike poboljšali našu sposobnost prikazivanja podataka. Prije trideset godina, naša sposobnost prikazivanja podataka bila je ograničena na ručni kalkulator, olovku i papir. To su loša sjećanja. ☹

Iako postoji bezbroj alata za prikazivanje podataka u grafičkom obliku, naša potraga za grafičkim prikazima podataka ograničit će se na **histogram** i kutijasti dijagram ili **boxplot**.

2.3.1 Histogram

Bez sumnje, histogram je najuobičajeniji način grafičkog prikazivanja **univarijatnih** (podaci jedne varijable) podataka. Mudro je podatke iz nekog skupa podataka organizirati na temelju sličnosti kao što je to učinjeno u potpoglavlju 2.1 gdje su traperice koje nose ljudi koji posjećuju rock koncerte kategorizirane prema spolu i prema boji traperica. Sada kad se radi o brojčanim podacima, sličnost se može puno lakše kategorizirati. Na sličnost numeričkih obilježja se može gledati kroz blizinu vrijednosti obilježja. U histogramu su dvije osi:

horizontalna os („x-os”) na kojoj su nanesene brojčane vrijednosti u uzlaznom redoslijedu i vertikalna os ("y-os”) gdje su nanesene **frekvencije** (učestalost) svakog ishoda. Histogram se ponekad naziva grafikon frekvencija.

Budući da uređivanje ishoda sadrži veliku dozu subjektivnosti, potrebno je dobro promisliti kako njime upravljati. Primjerice, na jednom ispitu, minimalni broj bodova može biti 70, a maksimalni 100. Moguće bodove na ispitu je lako nabrojati: 70, 71, 72, ..., 100. To je, međutim, samo jedan način strukturiranja ishoda. Glavni nedostatak ovog pristupa je prevelika detaljiziranost pogotovo ako malo studenata pristupi ispitu. Može se promatrati manje ishoda, primjerice samo parne brojeve: 70, 72, 74, ..., 100. Prema ovom izmijenjenom scenariju, neparni ishodi će biti smješteni u nižu ili višu kategoriju, što je svejedno, dok god se svi neparni rezultati tretiraju jednako.

Smisao gore navedenog je da ukaže na potrebu promišljanja prilikom organiziranja podataka, posebno grupiranja. Neka pravila predlažu da je odgovarajući broj kategorija funkcija broja opažanja u skupu podataka. Broj opažanja ćemo označiti s n. Tako primjerice, jedno pravilo sugerira da je $\sqrt{n}$ odgovarajući broj kategorija. Naravno, moguće je pravila podataka prilagoditi individualno. Za velike skupove podataka (približno 500 opažanja) koristi se tako oko 25 kategorija. Za veće skupove podataka (≥1000 opažanja) koristi se oko 30-ak kategorija. Srećom, većina statističkih programskih paketa kao što su JMP i R automatski biraju broj kategorija, ali broj kategorija može odabrati i sam korisnik. U Excelu broj kategorija se mora unaprijed odrediti.

Izradit ćemo histogram na temelju nekih podataka. Pretpostavimo da se odvija "Las Vegas noć” na poslovnoj školi. Osoblje, studenti i nastavnici su igrali nekoliko casino igara za male količine novca, a novac koji su izgubili su donirali u dobrotvorne svrhe. Ukupno je sudjelovalo 120 natjecatelja, a njihovi dobitci su prikazani u tablici 2.5. Negativni dobitci su, naravno, gubitci i imaju negativan predznak.

Ako se pobliže pogleda tablica 2.5., najveći gubitak iznosi 10 dolara, a najveći dobitak 2 dolara što čini raspon od -10 do 2. Čini se razumnim da kategorije počinju s -10 dolara, a završavaju s 2 dolara uz porast od 1 dolara po kategoriji što daje ukupno 13 kategorija. Usput, 13 kategorija je prilično blizu pravilu o broju kategorija jednakom $\sqrt{n}$.

-3.75	-6.00	-8.50	-2.75	-0.50	-5.50	-6.00	-4.75	-6.50	-5.25
-7.50	-6.75	-4.75	-3.25	-5.75	-1.00	-1.00	-5.50	-5.75	-3.75
-1.75	0.75	-0.75	-0.50	-2.00	1.75	-2.75	1.75	-5.75	-2.50
-2.25	-1.25	-3.25	-5.75	-5.25	-3.00	-6.25	-5.25	-4.75	-7.75
-3.25	-2.25	-2.00	-4.50	-3.00	-8.25	-4.00	0.00	0.75	-1.50
-6.75	-2.25	-4.25	-6.25	-7.25	-2.50	1.25	-1.75	-3.00	-4.75
-9.50	-4.50	-4.25	-1.50	-4.25	-8.50	-3.00	-2.25	-3.00	-5.50
-5.00	-3.75	-4.25	-2.25	-3.75	-1.25	-6.50	-5.50	-6.50	-5.25
-0.75	-3.00	-4.50	-1.75	-3.00	-4.50	-2.50	-2.25	-4.50	-6.00
-3.50	-6.00	-2.25	-6.00	-3.50	-6.00	-0.75	-2.75	-1.50	-1.50
-3.25	-8.50	-2.75	-4.75	-6.25	-1.00	0.25	-5.00	-3.50	-10.00
-7.25	-2.75	-1.75	-2.00	-2.25	-3.25	2.00	-3.50	-0.50	-8.25

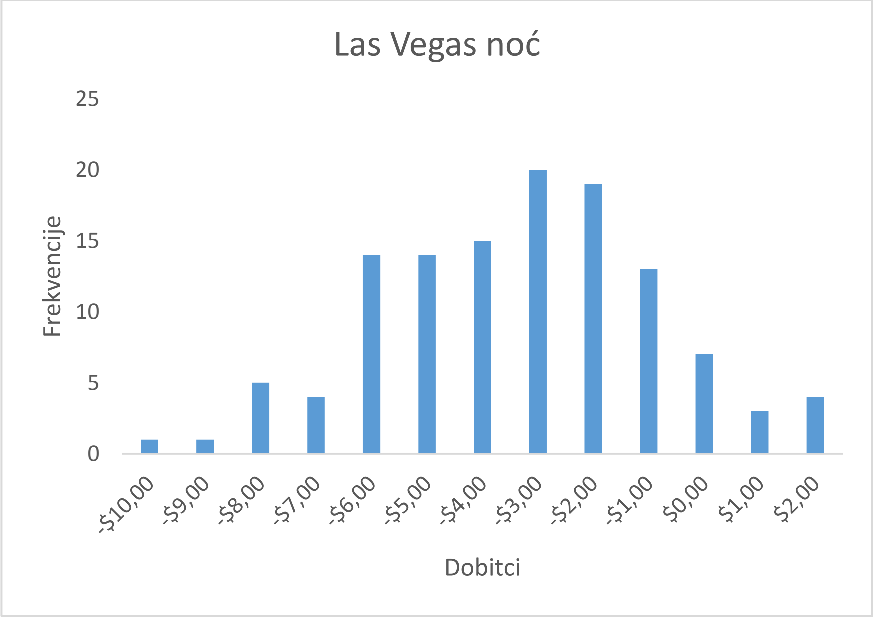
Tablica 2.5. Dobitci/gubitci u Las Vegas noći

Histogram se u Excelu može izraditi putem naredbi Data | Data analysis | Histogram. Nakon toga se odabiru dvije vrste podataka: najprije se određuju ulazni podaci (*Input Data*), koje čini skup od 120 stvarnih opažanja, a zatim raspon razreda (*Bin Range*), što je zapravo po veličini uređen popis kategorija koji u ovom slučaju glasi: -10 dolara, -9 dolara, ...,1 dolar, 2 dolara. Excel tada daje frekvencije svake kategorije kako je prikazano tablicom 2.6.

Kategorije	Frekvencije	Relativne frekvencije
-10 dolara	1	0.83%
-9 dolara	1	0.83%
-8 dolara	5	4.17%
-7 dolara	4	3.33%
-6 dolara	14	11.67%
-5 dolara	14	11.67%
-4 dolara	15	12.50%
-3 dolara	20	16.67%
-2 dolara	19	15.83%
-1 dolara	13	10.83%
0 dolara	7	5.83%
1 dolara	3	2.50%
2 dolara	4	3.33%

Tablica 2.6. Frekvencije dobitaka/gubitaka za Las Vegas noć

Histogram frekvencija iz tablice 2.6 prikazan je grafikonom 2.2.



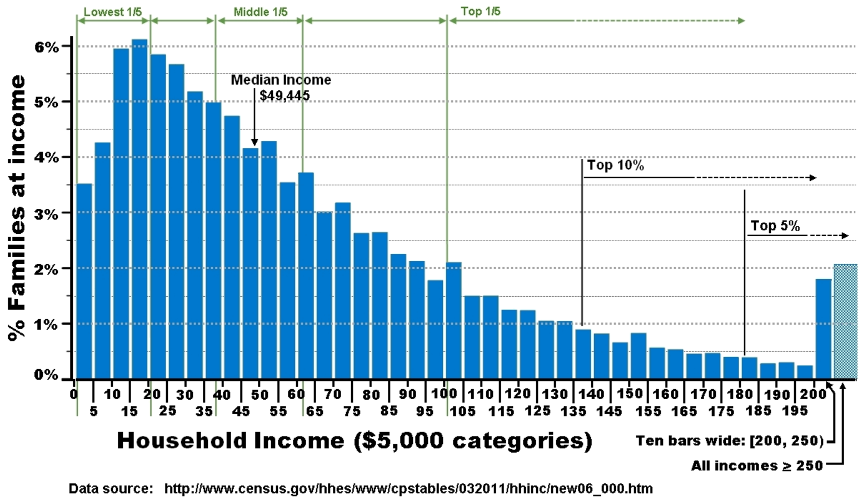
Grafikon 2.2. Histogram Las Vegas noći

Iz grafikona 2.2 se može primijetiti da histogram zapravo nije ništa drugo nego stupčasti grafikon kod kojeg su na horizontalnoj osi uređene kategorije, a na vertikalnoj osi frekvencije.

Također, treba primijetiti krajnji desni stupac tablice 2.6 pod nazivom "**relativne frekvencije**". Frekvenciju svake kategorije može se pretvoriti u relativnu frekvenciju dijeljenjem frekvencije s ukupnim brojem opažanja što je 120 u ovom slučaju. Primjerice, kategorija -4 dolara ima frekvenciju 15. Relativna frekvencija ove kategorije je 12.50% (15/120). Nema posebno velike koristi dodatno prikazati histogram relativnih frekvencija jer bi visina stupaca bila upravo proporcionalna visinama stupaca u histogramu frekvencija na grafikonu 2.2. Očita razlika između frekvencija i relativnih frekvencija je ta da je zbroj frekvencija jednak broju opažanja, dok je zbroj relativnih frekvencija jednak 1 ili 100%.

Što se tiče samog histograma, vidi se da su gubitci od 3 dolara najčešći. Ne ulazeći u deskriptivnu statistiku, opravdano se može zaključiti da je neki oblik centralne tendencije približno 3 dolara gubitka. O tome će biti više riječi u sljedećem poglavlju.

Još jedan vrlo lijep primjer histograma prikazan je grafikonom 2.3. On prikazuje distribuciju dohotka u 2011. godini gdje horizontalna os prikazuje godišnje dohotke u kategorijama po 5000 dolara, a vertikalna os prikazuje relativne frekvencije tih dohodaka.



Grafikon 2.3. Histogram distribucije dohodaka

Histogram ima nekoliko svojstava koja vrijedi istaknuti. Najprije, možemo primijetiti da se većina dohodaka nalazi na lijevoj strani što predstavlja niži dohodak, dok se s porastom prihoda frekvencije smanjuju sve dok se ne dođe do kategorije 200000 dolara i više. Mod, mjera centralne tendencije, se nalazi u šiljku distribucije kod 20 000 dolara, a dominantan rep distribucije se nalazi s desne strane. Pojam **desne asimetrije** opisuje ovu vrstu distribucije – dominantan ili teški rep je na desno. Distribucija koja je **lijevo asimetrična** je ona kojoj je dominantan rep distribucije na lijevo. Ako distribucija nije nakošena ni na jednu stranu, kaže se da je **simetrična**.

2.3.2 Box plot ili kutijasti dijagram

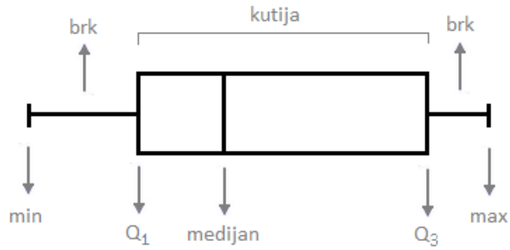
Kao što je ranije navedeno, histogram je najvažniji alat koji prikazuje opisne statističke podatke u grafičkom obliku. **Box plot** ili kutijasti dijagram daje dodatni uvid u distribuciju. Box plot je apstraktniji od histograma i zahtijeva dodatno objašnjenje. On prikazuje prvi, drugi (medijan) i treći kvartil kao vertikalne linije koje čine kutiju. Prvi kvartil je lijevi rub kutije, treći kvartil je desni rub kutije, a medijan je smješten negdje između ta dva ruba. Nadalje, ovim grafikonom su prikazani pragovi ili ograde kojima se utvrđuju **netipične**

vrijednosti. Postoji donji i gornji prag. Svaki od njih je 1.5 puta veći od interkvartila (interkvartil se računa kao razlika između trećeg i prvog kvartila, $Q_3 - Q_1$). Matematički, pragovi su definirani kako slijedi:

$$Donji\ prag = Q_1 - 1.5 \cdot (Q_3 - Q_1) \tag{2-5}$$

$$Gornji\ prag = Q_3 + 1.5 \cdot (Q_3 - Q_1) \tag{2-6}$$

Sve vrijednosti koje su ispod donjeg ili iznad gornjeg praga smatraju se netipičnim vrijednostima - ekstremnim opažanjima. Nakon toga se iscrtavaju brkovi kao na grafikonu 2.4. Donji brk se crta od prvog kvartila do odgovarajuće najmanje vrijednosti u skupu podataka iznad donjeg praga. Slično tome, gornji brk se crta od trećeg kvartila do najveće vrijednosti u skupu podataka manje od gornjeg praga. Vrijednosti izvan tih dvaju pragova smatraju se netipičnim vrijednostima i istaknute su kao takve. Moguće je da skup podataka nema netipičnih vrijednosti te stoga nisu niti prikazane box-plot dijagramom. Grafikonom 2.4 prikazan je opći box plot u opisanom kontekstu.



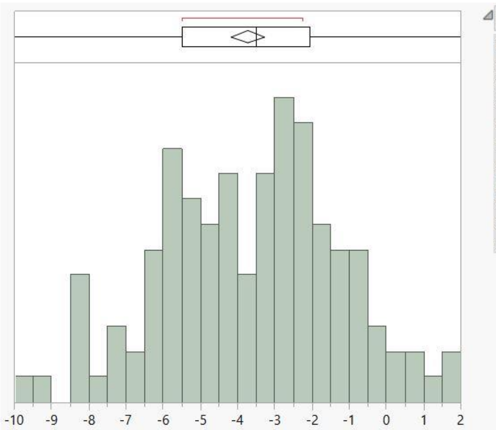
Grafikon 2.4. Općeniti primjer box plota

Važno je napomenuti da različite knjige i programski paketi daju nešto drukčije varijante box plota, osobito s obzirom na to kako se određuju brkovi i netipične vrijednosti. Ono što je jednako, međutim, za sve box plotove je da su prvi i treći kvartil uz medijan uvijek određeni na isti način.

2.4 Sistematizacija

Ovo potpoglavlje se bavi kreiranjem odgovarajućeg statističkog izvješća na temelju danog skupa podataka i korištenjem deskriptivnih, statističkih i grafičkih alata. Za ilustraciju će se koristiti podaci o Las Vegas noći, a JMP će se koristiti za analizu.

Nakon što se pokrene program JMP u njega se mogu uvesti podaci iz Excela. Tada je vrlo jednostavno u JMP-u napraviti statističku analizu. U JMP-u se odabere "Analyze", a zatim "Distribution". Odatle se odabere varijabla za "Y, stupci" opciju i klikne "Go". JMP će u tom trenutku dati ogromnu količinu rezultata, a većina toga nije od posebne potrebe. U prikazu rezultata se može odrediti koji dio rezultata ne želimo da bude prikazan i usredotočimo se na ono što želimo vidjeti na način da isključimo različite opcije u rezultatima - to (odabir koje rezultate želimo vidjeti) je nešto što zahtijeva eksperimentiranje. Na grafikonu ispod je prikazan zaslon u JMP-u koji prikazuje analizu podataka Las Vegas noći - svi ovi rezultati su standardni u JMP-u.



Quantiles

100.0%	maximum	2
99.5%		2
97.5%		1.7375
90.0%		-0.75
75.0%	quartile	-2.0625
50.0%	median	-3.5
25.0%	quartile	-5.5
10.0%		-6.75
2.5%		-8.5
0.5%		-10
0.0%	minimum	-10

Summary Statistics

Mean	-3.729167
Std Dev	2.4860078
Std Err Mean	0.2269404
Upper 95% Mean	-3.279802
Lower 95% Mean	-4.178531
N	120

Grafikon 2.5. Rezultati analize podataka Las Vegas noći u JMP-u

Prva stvar koja se primijeti je prosječni dobitak od -3.73 dolara. Drugim riječima, očekuje se gubitak igrača od 3.73 dolara i to je prilično slično medijanu od -3.5 dolara. Program daje standardnu devijaciju od 2.49 dolara i donji i gornji kvartil od -5.5 dolara i -2.06 dolara. Slijedi analiza histograma koji je u suštini isti kao na grafikonu 2.2. Ovdje je x-os histograma izmijenjena kako bi promjena iznosila 1 dolar.

Vrlo lijepo svojstvo JMP-a je da postavlja box plot na vrh histograma što daje odnos između dvaju grafičkih alata pružajući nam bolji pogled na skup podataka. JMP nudi dvije mogućnosti box plota koje ne prikazuju drugi programski paketi. U box plotu je naznačen lik u obliku dijamanta. Dio dijamanta u kojem on doseže maksimalnu visinu je aritmetička sredina skupa podataka. Širina dijamanta prikazuje 95% pouzdani interval, o čemu će biti riječi u petom poglavlju. Crvena zagrada iznad box plota se naziva najkraćom polovicom, a prikazuje raspon središnjih 50% opažanja u skupu podataka.

U ranijoj nastavničkoj karijeri autor se nikad nije bavio box plotom. Ograničio je razgovor o histogramu na grafički pristup deskriptivnoj statistici. Međutim, u posljednjih nekoliko godina autor je prigrlio box plot. Iako je definitivno manje konkretan od histograma, daje ogromnu količinu informacija na prilično jednostavan način. Kao takvog, sada ga smatra poboljšanjem u odnosu na histogram.

2.5 Zaključci

Niti jedan od pojmova u ovom poglavlju nije posebno težak - tvrdimo to iz matematičke i konceptualne perspektive. Unatoč relativnoj jednostavnosti ove teme, niti jedan dio ne treba shvatiti kao nevažan. Deskriptivna statistika je možda najvažnija tema pokrivena ovom knjigom. Kada se govori o skupu podataka, UVIJEK je potrebno sažeti podatke u brojčanoj/statističkoj formi i koristiti grafičku podršku, budući da mnogo ljudi u poslovnom okruženju, a osobito oni bez numeričkog predznanja, veću važnost daju slikama nego brojevima.

2.6 Zadaci

Za zadatke od 1. do 6. koristite skup podataka “RezultatiIspita”. Za zadatke od 7. do 12. koristite skup podataka “PromjerŽice”.

Skup podataka “RezultatiIspita” daje rezultate dva ispita kojima je pristupila grupa studenata. Studenti su najprije pristupili ispitu 1, a zatim ispitu 2.

1. Korištenjem Excela odredite:
    a. vrijednosti aritmetičkih sredina za ispit 1 i za ispit 2.
    b. vrijednosti medijana za ispit 1 i za ispit 2.
    c. vrijednosti standardnih devijacija za ispit 1 i za ispit 2.

d. Najmanju i najveću vrijednost za oba ispita.

2. Koristeći program Microsoft Excel kreirajte kombinirani histogram za svaki ispit koristeći razrede s rasponom od 2 ispitna boda.
3. Koristeći program Microsoft Excel kreirajte kombinirani histogram za svaki ispit koristeći razrede s rasponom od 1 ispitnog boda.
4. Koristeci histogram iz trećeg zadatka usporedite uspjeh na prvom i drugom ispitu.
5. Pružaju li histogrami pomoć u određivanju kako su podaci distribuirani? Komentirajte distribuciju podataka.
6. Koristeći JMP konstruirajte box plot za ispit 1 i ispit 2. Jesu li vaši rezultati u skladu s odgovorom u četvrtom zadatku?

Za skup podataka "PromjerŽice" podaci su uzeti iz dvije smjene u tvornici u kojoj se mjeri promjer 22-kalibarske žice. Cilj je da žica ima promjer 0.64 mm.

Koristeći Excel izračunajte sljedeće:

7. Aritmetičku sredinu promjera žice za smjenu 1 i za smjenu 2.
8. Medijan promjera žice za smjenu 1 i za smjenu 2.
9. Standardnu devijaciju promjera žice za smjenu 1 i za smjenu 2.
10. Koliko iznosi, promatrajući obje smjene, najmanji promjer žice, a koliko najveći?
11. Korištenjem Excela kreirajte kombinirani histogram s razredima veličine 30, smještenima između najmanjeg i najvećeg promjera žice.
12. Koristeći JMP kreirajte box plot za svaku smjenu.
13. Koristeći svoje rezultate iz 7. i 9. zadatka, komentirajte distribuciju za svaku smjenu.
14. Koja smjena bolje postiže ciljani promjer od 0.64mm?
15. Koja smjena je dosljednija u smislu promjera žice?

3 Vjerojatnost

Vjerojatnost susrećemo svaki dan u našim životima. Gledajući vijesti na TV-u saznajemo kako postoji mala vjerojatnost da će padati kiša sutra poslijepodne. Ponekad će vjerojatnost padanja kiše biti dana eksplicitno. Ako gledamo vijesti o politici, vidjet ćemo političke analitičare kako procjenjuju vjerojatnosti da će kandidati biti izabrani na različite političke funkcije.

U konkretnom smislu, vjerojatnost se bavi proučavanjem izvjesnosti događaja odnosno šansama da se događaj dogodi. U općem smislu, vjerojatnost proučava izvjesnost događaja u uvjetima nesigurnosti. Iako je ovo poglavlje usredotočeno na proučavanje vjerojatnosti da se neki događaj dogodi, ostatak knjige je svakako više usmjeren na proučavanje nesigurnosti budući da će ovaj koncept postati vrlo važan kada se počnemo baviti testiranjem hipoteza u narednim poglavljima ove knjige.

3.1 Osnove vjerojatnosti

Ovo potpoglavlje se bavi jednostavnim aspektom vjerojatnosti. Naglašeno je kako se vjerojatnost bavi izvjesnosti događaja u uvjetima nesigurnosti. Drugim riječima, promatramo zbivanja ili radnje čije ishode ne možemo sa sigurnošću predvidjeti (slučajan pokus) i ocjenjujemo izvjesnosti pojedinih ishoda ili općenitije, događaja. Intuitivno, vjerojatnost nekog događaja je broj iz intervala [0, 1], koji iskazuje određeni stupanj izvjesnosti da se taj događaj u slučajnom pokusu dogodi. Pri tomu je vjerojatnost sigurnog događaja jednaka 1, a vjerojatnost nemogućeg događaja jednaka je 0. Neka je nedvosmisleno određen neki događaj A. Vjerojatnost da će se dogoditi događaj A je omjer broja povoljnih ishoda za događaj A i broja svih mogućih ishoda u slučajnom pokusu. Jednostavan primjer je procjenjivanje vjerojatnosti dobivanja karte s brojem 6 iz standardnog snopa od 52 karte. Postoje četiri različite mogućnosti 6 u standardnom snopu karata (srce, karo, pik i tref). Dakle, postoje četiri karte s brojem 6 u snopu karata čime je vjerojatnost dobivanja karte 6 jednaka četiri prilike u 52 mogućnosti ili 4/52 što je 7.69%.

Vjerojatnost dobivanja karte 6 tref se razlikuje od gore dobivene vjerojatnosti. Snop karata ima samo jednu kartu koja odgovara ovom opisu. Zbog toga je vjerojatnost dobivanja karte 6 tref iz standardnog snopa karata 1/52 što je 1.92%. Ovakav pristup vjerojatnosti naziva se vjerojatnost a priori. Do vjerojatnosti je moguće doći različitim pristupima: subjektivno, statistički (a posteriori) ili klasičnim matematičkim pristupom (računanjem a priori ili geometrijski).

Vjerojatnost a priori je vjerojatnost u kojoj se unaprijed zna koliko je mogućih elementarnih[6] ishoda slučajnog pokusa. Ako su ishodi slučajnog pokusa jednako mogući, onda je vjerojatnost nastupa događaja A jednak omjeru broja povoljnih ishoda m i broja mogućih ishoda n odnosno

$$P(A) = \frac{m}{n}.$$

S druge strane, u pristupu vjerojatnosti a posteriori se ne polazi od unaprijed poznatog broja elementarnih događaja što je u praksi često slučaj i tad se pristupa ponavljanju pokusa i bilježenju ishoda pokusa. Ako se broj ponavljanja pokusa izvedenih u istim uvjetima povećava u beskonačnost, onda je vjerojatnost nastupa događaja A granična vrijednost relativne frekvencije povoljnog ishoda događaja A odnosno

$$P(A) = \lim_{n\to\infty} \frac{m}{n}.$$

Pravila vjerojatnosti u sljedećem potpoglavlju daju neke alate koji pomažu u razumijevanju nešto složenijih pitanja iz područja vjerojatnosti.

3.2 Pravila vjerojatnosti

Prije predstavljanja konkretnih pravila, postoji nekoliko pojmova koje treba definirati kako bi pravila vjerojatnosti bila mnogo jasnija.

Kao što je ranije navedeno, $P(A)$ je zapis koji se koristi za opisivanje vjerojatnosti da će se dogoditi događaj A. Vezano na to je

[6] događaje koji se ne mogu pojednostaviti – rastaviti na jednostavnije događaje

zapis $P(\bar{A})$, koji se naziva vjerojatnost komplementa događaja A ili "**vjerojatnost A-komplement"**.

Važan odnos koji se mora definirati je odnos **međusobne isključivosti događaja** i **nezavisnosti događaja**. Međusobno isključivi događaji su događaji koji ne mogu nastupiti u isto vrijeme. Primjer međusobno isključivih događaja možemo proučiti na primjeru uzimanja uzorka iz populacije. Naš odabir iz populacije može biti muškarac ili žena – nikad oboje. Kao takav, spol pojedinca se može shvatiti kao međusobno isključiv događaj. Drugi primjer je prijava učenika na fakultet. Postoje tri moguća ishoda: prihvaćanje, odbijanje ili lista čekanja. Kad učenik primi odluku fakulteta, ishod će biti samo jedan od tri moguća ishoda, a nikako neka kombinacija mogućih ishoda.

Nezavisni događaji *mogu* se dogoditi u isto vrijeme, ali nastupaju neovisno jedni od drugih. Pretpostavimo da večeras New York Yankeeji igraju utakmicu protiv Boston Red Soxa. U međuvremenu, Pittsburgh Piratesi igraju protiv Cincinnati Redsa. Moguće je da i Yankeeji i Piratesi pobijede jer ne igraju međusobno odnosno imaju različite protivnike te je (u poštenim uvjetima) za očekivati da ishod jedne utakmice ne ovisi o ishodu druge. Stoga, utakmice Yankeeji/Red Soxi i Piratesi/Redsi su neovisne jedna o drugoj.

Da bismo razumjeli pravila vjerojatnosti prikazana u nastavku, od velike je važnosti shvatiti razliku između međusobno isključivih događaja i nezavisnih događaja. Slijede pravila vjerojatnosti, koji su jednostavnosti radi, označena kao prvo, drugo, pa sve do šestog pravila vjerojatnosti.

3.3.1 Prvo pravilo vjerojatnosti

Vjerojatnost nekog događaja je broj između nule i jedan. Vjerojatnost ne može nikad biti manja od nule niti može biti veća od jedan. Matematički se to može zapisati kao:

$$0 \leq P(A) \leq 1 \tag{3-1}$$

3.3.2 Drugo pravilo vjerojatnosti

Međusobno isključivi događaji ne mogu nastupiti istodobno. Primjerice, ako postoje samo četiri moguća događaja (A, B, C i D) koji su isključivi, točno jedan od njih mora nastupiti. Matematički se to može zapisati kao:

$$P(A) + P(B) + P(C) + P(D) = 1 \tag{3-2}$$

3.3.3 Treće pravilo vjerojatnosti

Ako događaj A ne nastupi, tada mora nastupiti njegov komplement $\bar{A}$. Matematički to znači sljedeće:

$$P(A) + P(\bar{A}) = 1 \tag{3-3}$$

$$P(\bar{A}) = 1 - P(A) \tag{3-4}$$

3.3.4 Četvrto pravilo vjerojatnosti

Za međusobno isključive događaje vjerojatnost nastupanja događaja A ili događaja B jednaka je zbroju vjerojatnosti. Matematički se to može zapisati kao:

$$P(A\ ili\ B) = P(A) + P(B) \tag{3-5}$$

Matematički način navođenja $P(A\ ili\ B)$ je $P(A \cup B)$, gdje $A \cup B$ označava **uniju** skupova A i B. Primjerice, 10% je vjerojatnost da će se za večeru jesti pizza i 15% je vjerojatnost da će se za večeru jesti tjestenina. Stoga, postoji vjerojatnost od 25% (10% + 15%) da će se za večeru jesti pizza ili tjestenina.

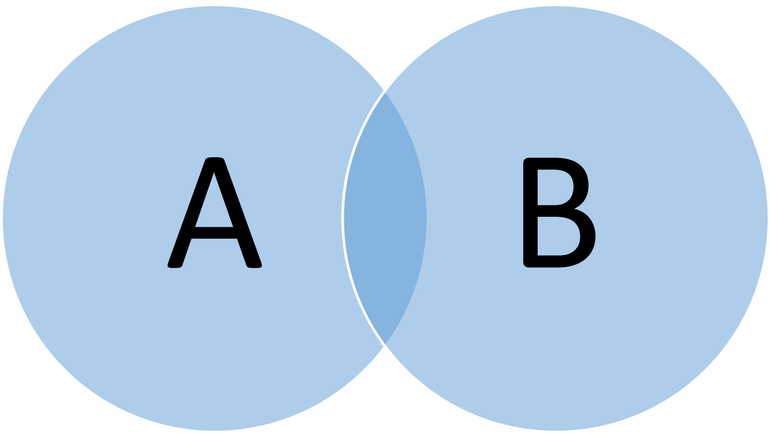
3.3.5 Peto pravilo vjerojatnosti

Za nezavisne događaje se može izračunati vjerojatnost nastupanja oba događaja A i B korištenjem sljedećeg izraza:

$$P(A\ i\ B) = P(A) \cdot P(B) \tag{3-6}$$

Također, $P(A\ i\ B)$ se može zapisati kao $P(A \cap B)$, gdje se $A \cap B$ naziva "presjekom" događaja A i B čime se označava događaj koji nastupa ako nastupe oba događaja A i B.

Grafički alat posebno prikladan za prikazivanje događaja je **Vennov dijagram** koji nam može pomoći i u analizi odnosa događaja. Primjerice, pretpostavimo da postoji šansa od 60% da će Yankeeji pobijediti Red Soxe i šansa od 55% da će Piratesi pobijediti Redse. Vjerojatnost pobjede Yankeeja i Piratesa je, dakle, 0.60·0.55=0.33 odnosno 33%. Ovim izračunom se u biti računa dio krugova koji se preklapaju na grafikonu 3.1.



Grafikon 3.1. Primjer Vennovog dijagrama

3.3.6 Šesto pravilo vjerojatnosti

Za nezavisne događaje vjerojatnost događaja A ili događaja B jednaka je:

$$P(A\ ili\ B) = P(A) + P(B) - P(A) \cdot P(B) \tag{3-7}$$

Ponovno, oznaka $P(A\ ili\ B)$ može glasiti $P(A \cup B)$. Vratimo li se na naš primjer, dobivamo da je vjerojatnost pobjede Yankeeja ili Piratesa jednaka $0.60 + 0.55 - 0.60 \cdot 0.55$ što je 0.82 odnosno 82%.

3.3 Tablice kontingence

Postoje slučajevi u kojima imamo priliku proučavati događaje u kojima se promatraju dva faktora, za razliku od slučaja s jednim faktorom boje traperica s ishodima da li ih nosi muška ili ženska osoba.

Ako događaj ima jedan faktor s a mogućih međusobno isključivih ishoda i drugi faktor s b mogućih međusobno isključivih ishoda, tada treba formirati **tablicu kontingence** s a redaka i b stupaca i $a \cdot b$ ćelija u tablici. Svaka ćelija u tablici prikazuje broj (izražen apsolutno ili relativno) ishoda za odgovarajuću jedinstvenu kombinaciju dvaju faktora. Tablica kontingence prikazuje informacije o vjerojatnosti vezano za svaki faktor i sve kombinacije faktora.

Razmotrimo primjer glasovanja u Zastupničkom domu SAD-a o zastupničkoj odluci - glasovanje o tome hoće li ili neće predloženi zakon biti usvojen. U ovom slučaju imamo dva faktora. Prvi je stranačka pripadnost koji ima dva ishoda: republikanac i demokrat. Drugi faktor predstavlja kako je zastupnik glasovao i također ima dva ishoda: za i protiv (u primjeru uzimamo da nije moguće biti suzdržan). Budući da svaki faktor ima dva ishoda, moguće su četiri kombinacije. Vezano za naš primjer proučavamo zastupničku odluku koja je predložena Zastupničkom domu 23. srpnja 2015. godine. Naziv zakona je "Zakon o sigurnom i točnom označavanju hrane". Tablica kontingence glasovanja je sljedeća:

	Za	**Protiv**	**Ukupno**
Republikanac	230	12	242
Demokrat	45	138	183
Ukupno	275	150	425

Tablica 3.1. Glasovanje u Zastupničkom domu

Prvi cilj je utvrditi je li predloženi zakon usvojen ili ne. Prema ukupnom rezultatu, zakon je usvojen s 275 glasova za i 150 protiv. Kako bi se to odredilo, gledamo redak Ukupno za obilježje "Za" i uspoređujemo ga sa retkom Ukupno za obilježje "Protiv". Također, primijetimo da je ukupno glasovalo 242 republikanca i 183 demokrata. Ovo se može primijetiti gledajući zbrojeve redaka. Konačno, dolazimo do vrijednosti od 425 ukupnih glasova - ova vrijednost se može odrediti zbrajanjem vrijednosti u zbirnom retku ili zbirnom stupcu. Ista vrijednost se može odrediti i zbrajanjem četiriju vrijednosti iz tablice (230 + 12 + 45 + 138). Ukupno je 425 opažanja što možemo označiti s n.

Ako bismo podijelili sve brojeve u tablici kontingence s n, dobili bismo relativne frekvencije kojima aproksimiramo vjerojatnosti. Zapravo, time standardiziramo dobivene apsolutne frekvencije. Na taj način se u osnovi gradi tablica relativnih frekvencija. Tablica 3.2 prikazuje rezultate tog postupka.

	Za	Protiv	Ukupno
Republikanac	54.12%	2.82%	56.94%
Demokrat	10.59%	32.47%	43.06%
Ukupno	64.71%	35.29%	100%

Tablica 3.2. Relativne frekvencije za glasovanje u Zastupničkom domu

Iz tablice 3.2. je vidljivo da se 56.94% svih glasova odnosi na republikance, a 43.06% na demokrate. Također, vidljivo je da je 64.71% zastupnika glasovalo "Za", dok je 35.29% glasovalo "Protiv". Zbroj glasova političkih stranaka iznosi 100%, kao i zbroj glasova "Za" i "Protiv". Sve ove nabrojane vrijednosti se nazivaju **graničnim ili marginalnim vjerojatnostima** jer se nalaze na "margini" odnosno na graničnom dijelu tablice kontingence. U biti, one nam govore kako je svaki od faktora distribuiran. Granične vjerojatnosti se mogu zapisati i matematički. Primjerice, o demokratima možemo reći sljedeće: *P*(demokrat)=43.06%.

Također, tablica 3.2. prikazuje procjene **zajedničkih vjerojatnosti** odnosno vjerojatnosti kombinacija svih mogućih ishoda. Primjerice, 2.82% svih glasova pripada republikancima koji su glasovali protiv navedenog zakona. Primijetite da je moguće odrediti zajedničku vjerojatnost za oba faktora što se matematički može zapisati kao: *P*(republikanac i protiv)=0.0282. Oznaka *P*(republikanac ∩ protiv) također ima smisla.

Posljednja tema koju ćemo raspraviti vezano za tablice kontingence je koncept **uvjetne vjerojatnosti**. Uvjetna vjerojatnost događaja je vjerojatnost nastupanja nekog događaja uz uvjet prethodnog nastupanja nekog drugog događaja. Matematički se to može zapisati kao:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3-8}$$

U ovom zapisu oznaka "|" znači "uz uvjet" što podrazumijeva vjerojatnost događaja A uz uvjet da je nastupio događaj B. Najjednostavnije rečeno, cilj je odrediti vjerojatnost događaja A znajući da se događaj B već dogodio.

Primjerice, možemo izračunati vjerojatnost da je zastupnik glasovao "protiv" uz prethodno poznatu informaciju da je demokrat. U kontekstu ovog primjera vrijedi:

$$P(protiv \mid demokrat) = \frac{P(protiv \cap demokrat)}{P(demokrat)} \tag{3-9}$$

Izračun za ovo bio bi 0.3247/0.4306=0.7541. Drugim riječima, postoji 75.41% šanse da će zastupnik glasovati "protiv" uz uvjet da je demokrat.

Ako netko glasa "Za", kolika je vjerojatnost da je republikanac? To pitanje možemo riješiti na sljedeći način:

$$P(republikanac \mid za) = \frac{P(za \cap republikanac)}{P(za)} \tag{3-10}$$

Dobivamo 0.5412/0.6471 što je jednako 0.8364. Drugim riječima, ukoliko je poznato da je zastupnik glasovao "Za", vjerojatnost da je zastupnik republikanac iznosi 83.64%.

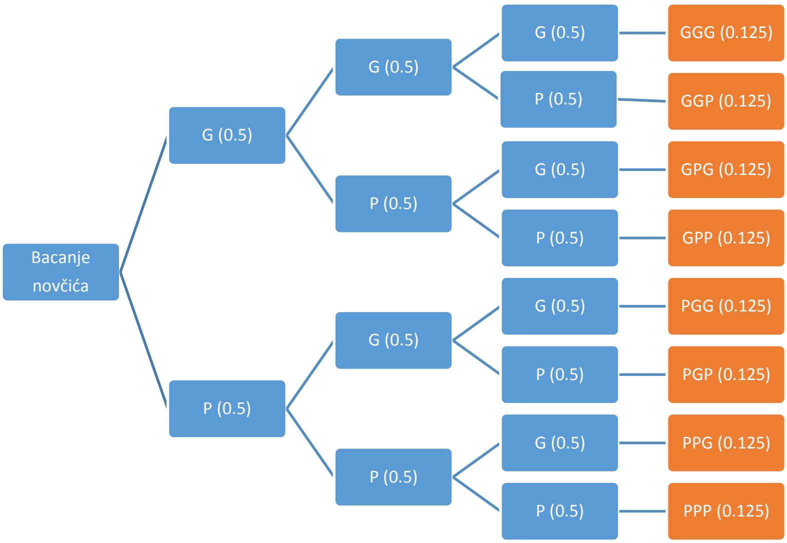
Tablice kontingence su dobra vježba za bolje razumijevanje vjerojatnosti. Često se mogu otkriti korisne informacije kad se proučavaju ishodi kod kojih su dvije međusobno isključive varijable. Ovo poglavlje ćemo zaključiti fokusirajući se na broj mogućih ishoda putem kratkog uvoda u prebrojavanje.

3.4 Stabla vjerojatnosti i ponavljanje pokusa

Često postoje slučajevi u kojima se neki vjerojatnosni entitet ponavlja nekoliko puta. Bacanje novčića više puta, igranje partije šaha i slično. U tim slučajevima može se koristiti alat poznat kao **stablo vjerojatnosti** za analizu ishoda. Kada se slučajni događaji ponavljaju, oni u suštini postaju nezavisni događaji. Na primjeru novčića to ima smisla ako se uzme da ishod bacanja drugog novčića nije povezan s ishodom bacanja prvog novčića.

U konstrukciji stabla vjerojatnosti svaki čvor stabla vjerojatnosti predstavlja pokus, a svaka grana stabla vjerojatnosti predstavlja ishod. Kada postoji nekoliko ishoda, grafički prikaz po strukturi sliči stablu. Kao primjer, promotrimo bacanje "poštenog" novčića tri puta. Ovdje termin "pošten" novčić znači da je vjerojatnost dobivanja glave i pisma jednaka i iznosi 50%.

Grafikon 3.2 prikazuje stablo vjerojatnosti za pokus bacanja novčića tri puta.



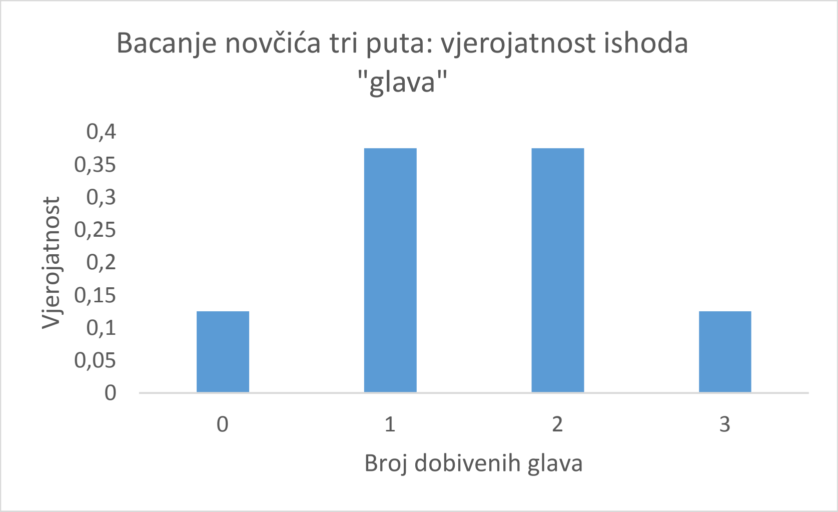
Grafikon 3.2. Stablo vjerojatnosti za bacanje novčića tri puta

Na temelju grafikona 3.2. primijetite da postoje tri grananja - jedno za svaki pokus što je rezultiralo s osam grana. Označeni čvorovi na desnoj strani prikazuju svaki od osam ishoda u smislu glave (G) ili pisma (P) i pripadajuću vjerojatnost svakog ishoda. Broj grana u stablu vjerojatnosti može se odrediti pomoću sljedećeg izraza

$$broj\ grana = broj\ ishoda^{broj\ pokusa} \tag{3-11}$$

U našem jednostavnom primjeru imamo dva moguća ishoda i tri pokusa što je rezultiralo s ukupno osam grana.

Možemo uzeti čvorove i pridružiti im broj uspjeha (kao uspjeh možemo promatrati pojavu "glave" u našem primjeru) i objasniti ih histogramom tako da možemo vizualizirati rezultate ponavljanja pokusa. Histogram je prikazan na grafikonu 3.3.



Grafikon 3.3. Histogram bacanja tri novčića

3.5 Osnove prebrojavanja

Kao što je definirano na početku ovog poglavlja, vjerojatnost pojavljivanja određenog događaja računamo kao broj mogućih uspjeha (broj ishoda koji odgovaraju određenom događaju) podijeljen s brojem mogućih ishoda. Često je utvrđivanje broja mogućih ishoda komplicirano i vrlo često je ovaj broj iznenađujuće velik. Uz malo promišljanja možemo se disciplinirati u određivanju broja ishoda tako da shvatimo koje alate treba koristiti kako bismo dobili odgovarajući izračun. Problemu ćemo pristupiti određujući tri osnovna pravila: pravilo umnoška, primjena kombinacija i primjena permutacija. Ova tema pripada grani vjerojatnosti koja se često naziva **kombinatorika** koja je prerijetko zastupljena u knjigama iz statistike.

3.5.1 Pravilo umnoška

Često promatramo niz pokusa, a svaki od njih ima određeni broj ishoda. Nadalje, ponekad su od interesa pojave koje možemo razložiti na više faktora, a svaki od faktora može poprimiti više različitih vrijednosti. Da bi se odredio ukupan broj mogućih ishoda niza pokusa, primjena pravila umnoška često je od velike koristi. U svom najjednostavnijem obliku, a iz matematičke perspektive, pravilo umnoška glasi:

$$broj\ mogućih\ ishoda = \prod_{i=1}^{m} n_i, \qquad (3\text{-}12)$$

gdje je n_i broj ishoda za i-ti pokus (ili broj različitih vrijednosti i-tog faktora), a veliko slovo "pi", oznaka (Π), označava umnožak, slično kako veliko slovo "sigma", simbol (Σ), označava zbroj.

Za primjer, razmotrimo karakteristike pizza. Postoje četiri faktora: kora, sir, preljev i veličina. Tablica 3.3. daje detaljni prikaz vrijednosti koje faktori mogu poprimiti. Za ovaj primjer se pretpostavlja da je za svaku pizzu dozvoljena točno jedna vrijednost svakog faktora.

Faktor	Broj mogućih vrijednosti	Vrijednosti faktora
Kora	4	tanka, debela, Deep Dish, Pan
Sir	4	Mozzarella, Provalone, Romano, Gouda
Preljev	30	Inćuni, paprika, …, ananas
Veličina	4	mala, srednja, velika, obiteljska

Tablica 3.3. Karakteristike pizza

Broj jedinstvenih kombinacija karakteristika pizze jednostavno je određen umnoškom broja mogućih vrijednosti svih faktora: $4 \cdot 4 \cdot 30 \cdot 4 = 1920$. Ovaj primjer pokazuje da postoji 1920 mogućih ishoda. Naravno, neki od tih ishoda imaju veću vjerojatnost nego neki drugi, ali ta činjenica je izvan dosega naše sadašnje rasprave.

3.5.2 Kombinacije

U prethodnom odlomku riječ "kombinacija" se koristila donekle neformalno. U ovom poglavlju koristit će se nešto formalnije. Pretpostavimo da imamo n različitih elemenata u skupu i želimo odabrati podskup veličine r iz tog skupa pri čemu nam poredak elemenata nije važan i u kojima ne dozvoljavamo ponavljanje elemenata. Broj takvih različitih odabira nazivamo broj jedinstvenih kombinacija i često se označava s $\binom{n}{r}$ koje čitamo "n *povrh* r". Drugi način na koji se to može zapisati je $C(n,r)$ ili ${}_nC_r$. U nastavku knjige koristit će se zapis $C(n,r)$. S ovako navedenim definicijama broj kombinacija se može odrediti na sljedeći način:

$$broj\ kombinacija = C(n,r) = \frac{n!}{r!\,(n-r)!} \tag{3-13}$$

Ovdje $n!$ označava “n faktorijela” što se definira formulom:

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 \tag{3-14}$$

Koncept faktorijela će biti detaljnije opisan u sljedećem odjeljku.

Kao primjer kombinacija pogledajmo kolekciju od deset knjiga. Uskoro idemo na dugo putovanje, a imamo mjesta samo za tri knjige. Koliko ima kombinacija knjiga koje možemo ponijeti? U ovom problemu imamo skup od deset knjiga ($n = 10$) i tražimo podskup veličine tri ($r = 3$). Korištenjem izraza (3-13) za izračun broja kombinacija, možemo utvrditi da postoji 120 mogućih kombinacija knjiga koje možemo ponijeti na putovanje.

3.5.3 Varijacije

Naglašeno je kako kombinacije ne ovise o poretku elemenata. Za gore naveden primjer s knjigama smo izračunali koliko mogućih tročlanih podskupova koji nastaju iz skupa od ukupno deset elemenata. Tih 120 kombinacija nije osjetljivo na poredak elemenata. Primjerice, skup knjiga $\{A,B,C\}$ i isti je kao i skup knjiga $\{C,A,B\}$. Varijacije su osjetljive na poredak. Varijacija skupa je svaki poredak od

r elemenata uzet iz skupa od n elemenata. Zapis za varijacije je sličan onom za kombinacije – koriste se oznake ${}_nV_r$ i $V(n,r)$. S matematičkog gledišta, broj jedinstvenih varijacija, u kojima ne dozvoljavamo ponavljanje elemenata, za podskup veličine r iz skupa od n elemenata je kako slijedi:

$$broj\ varijacija = V(n,r) = \frac{n!}{(n-r)!} \tag{3-15}$$

Kao primjer varijacija, pogledajmo moguće šifre lokota s tri koluta pri čemu je na svakom kolutu moguće četrdeset različitih elemenata (znamenaka ili slova). Imamo skup veličine četrdeset i podskup veličine tri. Koristeći izraz (3-15) dobivamo 59280 različitih šifri. Formalno, matematički izraz "kombinacija za lokot" je neprikladan, dok je "varijacija za lokot" prikladnija zbog bitnog poretkatriju znamenki.

Poseban slučaj varijacija su permutacije. Permutacija skupa je svaki poredak od n elemenata uzet iz skupa od n elemenata. Stoga za permutacije vrijedi da je $r = n$. Broj permutacija za skup od n elemenata gdje ne dopuštamo ponavljanje elemenata jednak je $n!$.

3.5.4 Prebrojavanje pomoću Excela

Microsoft Excel pomaže u računima kombinatorike. Dok su gore prikazane formule korisne u izradi odgovarajućeg izračuna, Excel pojednostavnjuje postupak s funkcijama prikazanim u tablici 3.4.

Funkcija u Excelu	Namjena
=product(raspon podataka)	Umnožak svih vrijednosti u obuhvaćenom rasponu podataka
=factorial (n)	$n!$ ili “n faktorijela”
=combin(n,r)	$C(n,r)$ ili ${}_nC_r$
=permut(n,r)	$P(n,r)$ ili ${}_nP_r$

Tablica 3.4. Kombinatorne funkcije u Excelu.

3.6 Zaključci

Kombinatorika i prebrojavanje su teme koje se obično ne nalaze u uvodnim knjigama iz statistike. Ipak, razumijevanje osnovnih principa kombinatorike neophodno je za razumijevanje vjerojatnosti.

Primjerice, u velikom broju primjera nije moguće odrediti broj mogućih ishoda događaja (nazivnik kod izračuna vjerojatnosti događaja). Često je taj broj iznenađujuće velik. Znati ove informacije je vrijedno, a može biti i korisno u bilo kojoj organizaciji za bolje razumijevanje okoline u kojoj djelujemo.

3.7 Zadaci

1. Baca se par kockica. Napišite tablicu koja prikazuje sve moguće ishode i vjerojatnosti svakog ishoda.
2. Kolika je, na temelju gornje tablice, vjerojatnost dobivanja zbroja brojeva na kockicama jednakim pet ili devet?
3. Jesu li ishodi iz zadatka 1 međusobno isključivi? Zašto da ili zašto ne?
4. Jesu li ishodi iz zadatka 1 neovisni? Zašto da ili zašto ne?
5. Kolika je vjerojatnost da se iz standardnog snopa od 52 karte izvuče karta as pik?
6. Kolika je vjerojatnost da se iz standardnog snopa od 52 karte izvuče karta s brojem deset?
7. Pomoću svojih odgovora u zadacima 5 i 6 odredite koliko iznosi vjerojatnost izvlačenja karte 10 boje pik. Ima li ovaj odgovor smisla?
8. Kolika je vjerojatnost da se iz standardnog snopa od 52 karte izvuče karta pet ili šest?
9. Večeras Houston Astrosi igraju s New York Yankeejima. Analitičari daju Yankeejima 59% šanse za pobjedu. Također, večeras Chicago Cubsi igraju s Pittsburgh Piratesima. Analitičari prognoziraju pobjedu Pittsburgha s vjerojatnošću od 55%. Neriješeni ishodi nisu dozvoljeni, dakle tim će ili pobijediti ili izgubiti. Koristeći ove informacije, kolika je vjerojatnost da će Yankeeji i Cubsi pobijediti večeras?
10. Koristeći podatke iz zadatka 9, kolika je vjerojatnost da Houston ili Cubsi pobijede?
11. Koristeći podatke iz zadatka 9, kolika je vjerojatnost da ni Yankeeji ni Piratesi ne pobijede?

12. Nedavno sam prikupio neke podatke o sklonostima ljudi prema trgovinama Wal-Marta i Targeta. 103 muškarca je preferiralo Target, dok je 67 muškaraca preferiralo Wal-Mart. 158 žena je preferiralo Target, a 27 žena je preferiralo Wal-Mart. Koristeći ove informacije, koji je postotak svih ispitanika preferirao Target?
13. Koristeći podatke iz zadatka 12, koliki su postotak ispitanika činili muškarci?
14. Koristeći podatke iz zadatka 12, koliki su postotak ispitanika činile žene koje preferiraju Wal-Mart?
15. Koristeći podatke iz zadatka 12, od svih žena, koliki je postotak preferirao Target?
16. Koristeći podatke iz zadatka 12, koliki je postotak muškaraca među onima koji preferiraju Wal-Mart?
17. Možete li donijeti neke općenite zaključke na temelju podataka iz zadatka 12?
18. Ako Yankeeji i Red Soxi igraju 4 utakmice u nizu uz vjerojatnost od 54% da će Yankeeji pobijediti u svakoj utakmici, nacrtajte histogram koji prikazuje vjerojatnost pobjede Yankeeja do 4 utakmice.
    a. Kolika je vjerojatnost da će Yankeeji izgubiti sve utakmice?
    b. Kolika je vjerojatnost da će Yankeeji pobijediti u barem jednoj utakmici?
    c. Kolika je vjerojatnost da će Yankeeji pomesti Red Soxe (pobijediti u sve četiri utakmice)?
19. S obzirom na podatke o pizzama u tablici 3.3, koliko vrsta preljeva na jednoj pizzi je moguće ako su dopuštena najviše dva preljeva na jednoj pizzi? Pretpostavimo da dvostruki preljevi na svakoj pojedinačnoj pizzi (kao primjerice dvostruke kobasice) nisu dopušteni.
20. S obzirom na podatke o pizzama u tablici 3.3, koliko vrsta preljeva je moguće ako su dopuštena najviše dva preljeva?

Pretpostavimo da su dvostruki preljevi na svakoj pojedinačnoj pizzi (kao primjerice dvostruke kobasice) dopušteni.

21. Imam 15 fotografija. Tražili su me da odaberem četiri za obiteljski foto album. Koliko je kombinacija fotografija moguće?
22. Moram posjetiti 10 gradova samo jednom. Grad iz kojeg počinjem turneju mora biti isti onaj u kojem ću turneju završiti. Koliko je mogućih različitih tura?
23. Ispitu je pristupilo 100 studenata. Dobivaju se nagrade za prvo, drugo i treće mjesto. Koliko je mogućih ishoda dobitnika?
24. Svakih 10 godina izrađuje se obiteljski zbornik McMullena. Za sljedeći broj su me tražili da pošaljem četiri fotografije, tri recepta i dvije kratke priče. Imam deset fotografija, devet recepata i pet kratkih priča između kojih moram izabrati. Koliko je mogućih kombinacija koje mogu poslati?

4. Slučajne varijable

Vjerojatnost i statistiku se može opisati na mnogo načina. Suočavanje s nesigurnošću je jedan od aspekata znanosti. Postoje pojave u prirodi, ekonomiji i životu za koje se nikad neće moći ishod predvidjeti sa sigurnošću. Konačna dnevna cijena dionica tvrtke, mjesečna prodaja određenih kemikalija i pravna pitanja mogu se također smatrati izvorom nesigurnosti. Ovakve objekte nazivamo **slučajnim varijablama**. Slučajne varijable su objekti koje poprimaju vrijednosti koje ne možemo predvidjeti sa sigurnošću. Možda imamo općenitu ideju o njihovom ponašanju, ali ne možemo sa sigurnošću reći koji će biti ishod.

Uzmimo u obzir cijenu dionica neke tvrtke. Neka je današnja cijena dionice tvrtke 10 dolara po dionici. Čini se razumnim pretpostaviti da će u isto vrijeme sutra, uz pretpostavku izostajanja velikih šokova tržišta, cijena dionice tvrtke biti slična onoj danas - možda malo niža od 10 dolara po dionici, a možda nešto viša od 10 dolara po dionici. Cijena dionice na kraju jednog dana trgovanja se može smatrati slučajnom varijablom. Naš cilj je bolje razumijevanje pojma slučajne varijable s obzirom na očekivanje i disperziju tako da možemo donijeti stratešku odluku temeljenu na što više informacija o svim pitanjima koja se odnose na konkretnu slučajnu varijablu.

Sljedeće poglavlje daje različite vrste slučajnih varijabli i opisuje njihova svojstva.

4.1 Diskretne slučajne varijable

Diskretna slučajna varijabla ima konačan (ili prebrojiv) broj poznatih mogućih. Primjer je bacanje para kockica. Najmanji mogući ishod zbroja na kockama je dva (dobivene dvije jedinice), a najveći mogući ishod je dvanaest (dobivene dvije šestice). Ukupno postoji jedanaest mogućih ishoda. Svaki mogući ishod ima određenu vjerojatnost pojavljivanja i te vjerojatnosti je moguće izračunati.

4.1.1 Diskretna distribucija

Konkretna vrijednost diskretne distribucije X je vrijednost diskretne slučajne varijable koja se označava s x, a vjerojatnost pojavljivanja događaja x je $P(x)$. Te vjerojatnosti su poznate. U slučaju konačnog broja mogućih vrijednosti slučajne varijable označimo broj mogućih ishoda s n. Očekivana vrijednost slučajne varijable ($\mu = E(x)$) i njena standardna devijacija ($\sigma = std(x)$) određuju se formulama:

$$\mu = E(x) = \sum_{i=1}^{n} x_i P(x_i), \tag{4-1}$$

$$\sigma = std(x) = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 P(x_i)}, \tag{4-2}$$

gdje je x_i vrijednost i-tog ishoda. Također, treba napomenuti da je kumulativna vjerojatnost[7] (ili funkcija distribucije) i-tog mogućeg ishoda $\sum P(x_i)$ jednaka:

$$\sum P(x_i) = \sum_{j=1}^{i} P(x_j). \tag{4-3}$$

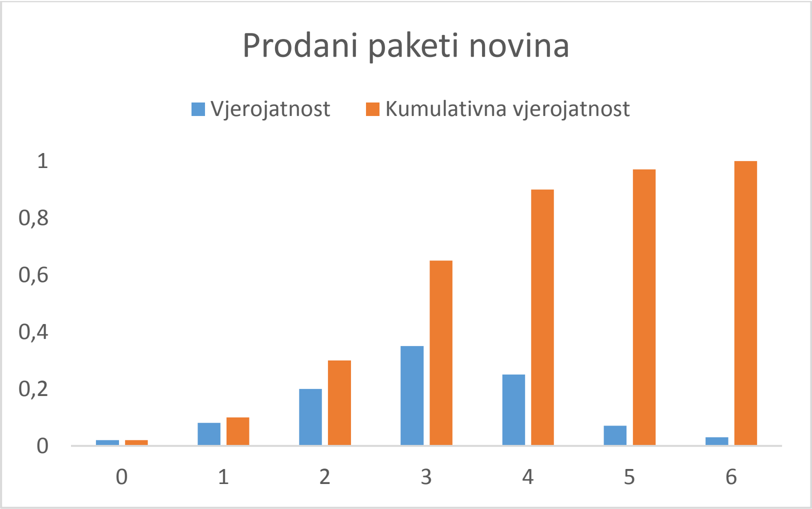
Razmotrimo primjer srednjih škola u kojima učenici prodaju novine u centru grada prije nastave. Novine se prodaju u paketu. Svakog dana moguće je ne prodati niti jedan paket ($x = 0$) ili prodati do čak šest paketa novina ($x = 6$). Tablica 4.1 prikazuje moguće ishode zajedno s pripadajućim vjerojatnostima i kumulativnim vjerojatnostima:

[7] Kumulativna vjerojatnost $\sum P(x_i)$ označava vjerojatnost da slučajna varijabla poprimi vrijednost manju ili jednaku od x_i.

Broj prodanih paketa	$P(x)$	$\sum P(x_i)$
0	0.02	0.02
1	0.08	0.10
2	0.20	0.30
3	0.35	0.65
4	0.25	0.90
5	0.07	0.97
6	0.03	1.00

Tablica 4.1. Vjerojatnosti broja prodanih paketa novina

Koristeći gornje jednadžbe učenik može očekivati da će prodati 3.06 novina sa standardnom devijacijom od 1.22 novine. Grafikon 4.1 prikazuje distribuciju vjerojatnosti i pripadne kumulativne vjerojatnosti za sedam mogućih ishoda. Kumulativna vjerojatnost za određeni broj ishoda je vjerojatnost da je nastupio najviše taj određeni broj ishoda. Primjerice, kumulativna vjerojatnost od 0.90 povezana je s vjerojatnosti prodaje najviše četiri paketa novina ili da je vjerojatnost da će biti prodana 4 paketa novina ili manje jednaka 0.90.



Grafikon 4.1. Vjerojatnost i kumulativna vjerojatnost broja prodanih paketa novina

4.1.2 Binomna distribucija

Binomna distribucija je posebna vrsta diskretne distribucije koja opisuje slučajne pokuse koji imaju samo dva moguća ishoda: uspjeh i neuspjeh. Na primjer, u bacanju novčića rezultat će biti ili glava ili pismo. Kretanje cijene dionica je takvo da će na kraju dana cijena dionice ili dobiti na vrijednosti ili neće dobiti na vrijednosti. Ponavljamo takav pokus više puta pri čemu su ishodi od pokusa do pokusa nezavisni, a vjerojatnost uspjeha se ne mijenja. Korištenjem svojstava binomne distribucije možemo određivati vjerojatnost pojave određenog broja uspjeha (k) u nekom određenom broju ponavljanja pokusa (n) za danu vjerojatnost uspjeha (p). Nakon što odredimo vrijednost p, možemo odrediti njegov komplement za određivanje vjerojatnosti neuspjeha q, koji je jednak $1 - p$.

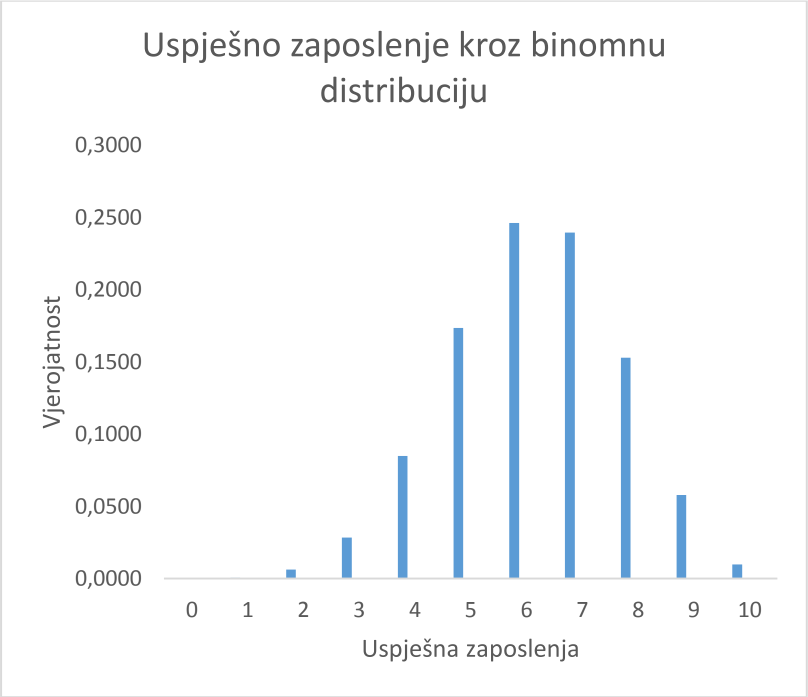
Formula kojom se može odrediti vjerojatnost pojave k uspjeha pri ponavljanju n nezavisnih pokusa s vjerojatnošću uspjeha p dana je s

$$P(k) = \frac{n!}{(n-k)!k!} p^k q^{n-k}. \tag{4-4}$$

Na prvi pogled gore navedena formula izgleda zastrašujuće.

Kombinatorni objekt $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ koji prethodi vjerojatnosti daje broj kombinacija k uspjeha za danu vjerojatnost uspjeha $p.$ Excel ima definiranu funkciju za izračun vjerojatnosti za k uspjeha u n pokusa za binomnu distribuciju. Ta funkcija glasi: "$= BINOMDIST(k, n, p, FALSE)$", gdje "FALSE" označava vjerojatnost, a "TRUE" kumulativnu vjerojatnost. U novijim verzijama Excela koristi se funkcija "$= BINOM.DIST(k, n, p, FALSE)$".

Uzmimo jednostavan primjer upravljanja ljudskim potencijalima. Menadžera ljudskih potencijala ocjenjuje viši menadžment. Dio procjene uspješnosti se temelji na broju stalno zaposlenih osoba pod njegovim vodstvom. Zaposlenje se smatra uspješnim ako traje tri ili više godina. Kao takvo, u godišnjem ocjenjivanju menadžera ljudskih potencijala analiziraju se poslovi tri godine unazad, pa ako je zaposlenik od prije tri godine i dalje zaposlen, posao se smatra uspješnim. U suprotnom, posao se smatra neuspješnim. Uzmimo da je procijenjeno kako je vjerojatnost da je jedno zaposlenje od prije tri godine i dalje ugovoreno jednako 65%. Razmotrimo deset zaposlenja u posljednje tri godine ($n = 10$). Korištenjem funkcije binomne distribucije, vjerojatnost uspješnih zapošljavanja u posljednje tri godine je opisano sljedećim grafikonom:



Grafikon 4.2. Primjer binomne distribucije

Na grafikonu 4.2 možemo vidjeti detalje vjerojatnosti uspješnog zapošljavanja u posljednje tri godine. Na temelju navedene tablice mogu se odrediti kumulativna i funkcija inverzne kumulativne distribucije.

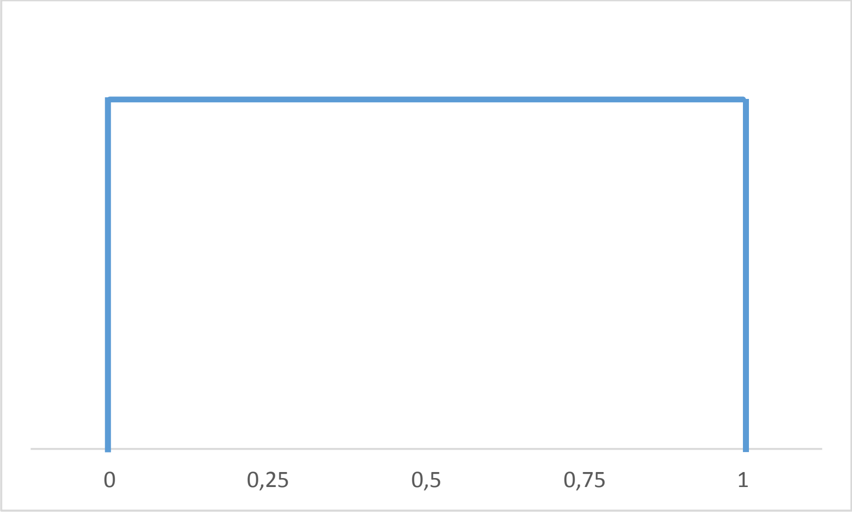
4.2 Neprekidne slučajne varijable

Prethodno opisane distribucije vjerojatnosti imaju konačan broj ishoda. Neprekidne distribucije vjerojatnosti imaju beskonačan (i neprebrojiv) broj mogućih ishoda, stoga nema smisla promatrati vjerojatnosti u točki. Ipak, ima smisla promatrati koliko je gusto koncentrirana vjerojatnost u nekoj točki. Pojam gustoće opisan je funkcijom gustoće slučajne varijable. Nadalje, nema smisla vjerojatnosti promatrati u točkama, već ima smisla promatrati vjerojatnosti intervala koji se računaju pomoću površina ispod krivulje gustoće. Iako postoje mnoge distribucije koje bismo mogli proučavati, prikazat ćemo uniformnu i normalnu distribuciju.

4.2.1 Uniformna distribucija

Uniformna distribucija ima zadanu najmanju moguću vrijednost i najveću moguću vrijednost, a sve vrijednosti između i uključujući ove dvije ekstremne vrijednosti su jednako moguće.

Generator slučajnih brojeva na kalkulatoru je primjer uniformne distribucije. Kada tražite slučajni broj, dobit ćete neki broj između 0 i 1. Broj nikad neće biti manji od 0 niti veći od 1. Korištenje funkcije "*=RAND()*" u Excelu će dati isti rezultat. Ako se generira beskonačan broj slučajnih brojeva korištenjem kalkulatora ili Excela, histogram će izgledati ovako:



Grafikon 4.3. Uniformna distribucija jednostavnog generatora slučajnih brojeva

Primijetite da grafikon 4.3 prikazuje da se sve vrijednosti između 0 i 1 pojavljuju s istom vjerojatnošću.

Jedna od posebnih korisnosti je mogućnost simuliranja koristeći rezultate uniformne distribucije kao ulazne podatke za simulaciju složenijih distribucija.

4.2.2 Normalna distribucija

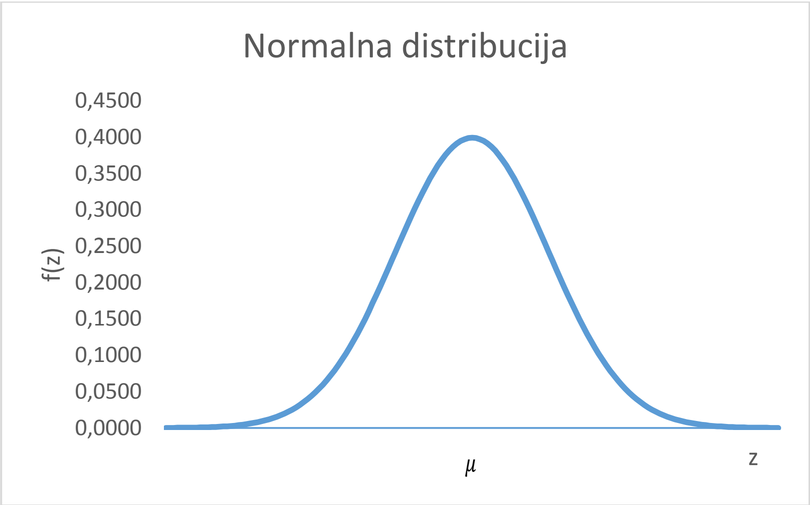
Od svih distribucija, normalna distribucija je najvažnija. Često se pojavljuje u prirodi te smo često i sami svjedoci pojavnosti normalne distribucije: distribucija nečije visine ili težine, distribucija rezultata ispita, distribucija broja sati koje je netko proveo spavajući sinoć i slično. Normalna distribucija je česta i u prirodi jer distribucija mnogih pojava slijedi normalnu distribuciju. Kod skiciranja funkcije gustoće normalne distribucije nezavisna varijabla z označava odstupanje od sredine u broju standardnih devijacija. Zavisna varijabla je $f(z)$, funkcija gustoće koja je ovisna o varijabli z. Ovdje koristimo unaprijed određenu veličinu za sredinu (μ) i standardnu devijaciju (σ) pa stoga razlikujemo različite normalne distribucije $N(\mu, \sigma^2)$[8] i kažemo da je normalna distribucija dvoparametarska. U praktičnijem smislu, z se može shvatiti kao ishod, a $f(z)$ kao gustoća vjerojatnosti pojavljivanja tog ishoda. Matematički se vezu između z i funkcije gustoće normalne slučajne varijable $N(0,1)$ može zapisati kao:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \tag{4-5}$$

Očito, ovo je složena formula. Srećom, Excel nam može pomoći generirati graf funkcije normalne distribucije korištenjem funkcije *"=NORMDIST*$(\mu, \sigma, z, FALSE)$*"* u starijim verzijama odnosno *"=NORM.DIST*$(\mu, \sigma, z, FALSE)$*"* u novijim verzijama Excela, gdje je μ aritmetička sredina populacije, σ standardna devijacija populacije, z je z - vrijednost ili broj standardnih devijacija udaljenih od srednje vrijednosti, "FALSE" očekuje funkciju gustoće odnosno funkciju vjerojatnosti, dok opcija "TRUE" vraća kumulativnu funkciju gustoće. Na grafikonu 4.4. prikazana je funkcija gustoće normalne slučajne varijable. Može se primijetiti kako je vjerojatnost gušće koncentrirana u točkama po sredini prikazanog dijela x - osi odnosno u točkama oko srednje vrijednosti budući je vrijednost funkcije gustoće tu najveća.

[8] Normalna slučajna varijabla s očekivanjem 1 i standardnom devijacijom (i varijancom) 1, u oznaci $N(0,1)$, naziva se jedinična ili standardizirana normalna slučajna varijabla.

Isto tako je i površina ispod krivulje (vjerojatnost) za intervale oko srednje vrijednosti najveća.



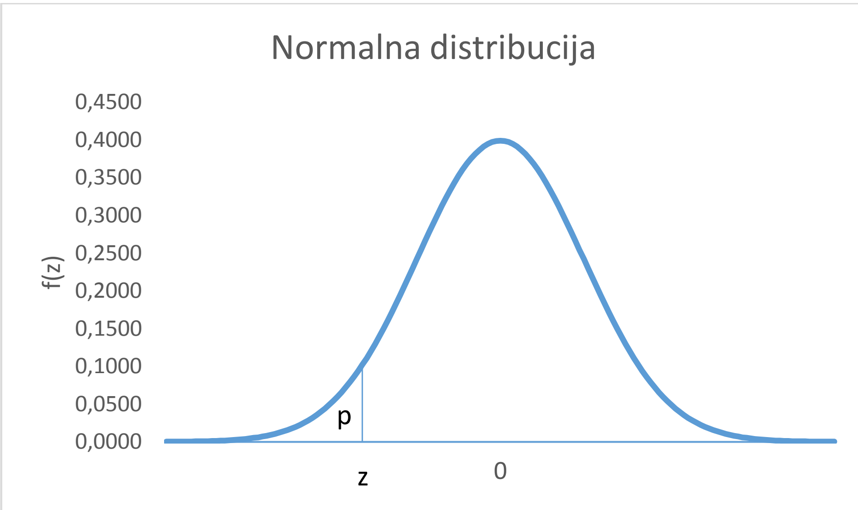
Grafikon 4.4. Normalna distribucija

Često je od interesa za neku vrijednost x odrediti njeno odstupanje od srednje vrijednosti, ali uzimajući u obzir standardnu devijaciju (raspršenost) distribucije. Z - vrijednost ili odstupanje neke vrijednosti x od srednje vrijednosti populacije μ u broju standardnih devijacija dana je formulom:

$$z = \frac{x - \mu}{\sigma} \tag{4-6}$$

Dobivena z - vrijednost daje udaljenost između vrijednosti x i μ izraženu u jedinicama standardne devijacije. Nadalje, primjenom ove formule moguće je standardizirati normalnu distribuciju $N(\mu, \sigma^2)$ tako da poprima srednju vrijednost 0 i standardnu devijaciju 1.

z - vrijednost se zatim koristi za određivanje površine ispod krivulje lijevo od z ili točnije površine ispod krivulje između $-\infty$ i z koja je integral funkcije (4-5). Odrediti površine ispod krivulje između $-\infty$ i z zapravo znači odrediti vjerojatnost da slučajna varijabla poprimi vrijednost manju ili jednaku z (kumulativna vjerojatnost). Funkcija (4-5) ne može se integrirati uobičajenim načinima te se mora provesti odgovarajući numerički postupak. Srećom, Excel ima funkciju i za to. Funkcija *"=NORMSDIST(z)"* u starijim verzijama, a funkcija *"=NORM.S.DIST(z)"* u novijim verzijama Excela računa površinu ispod standardne normalne krivulje lijevo od z - vrijednosti, drugim riječima, računa vjerojatnost da slučajna varijabla $N(0,1)$ poprimi vrijednost manju ili jednaku od z. Grafikon 4.5 ovo ilustrira pri čemu p predstavlja područje lijevo od z - vrijednosti.



Grafikon 4.5. Z - vrijednosti normalne distribucije

Primjerice, ako imamo rezultate ispita s aritmetičkom sredinom populacije 83 i standardnom devijacijom 5, rezultat ispita od 87 će dati z-vrijednost 0.80 i područje ispod krivulje *"=NORMSDIST(0.80)"* je 0.7881. Drugim riječima, studenti s uspjehom od 87 bodova su bolji od 78.81% studenata u populaciji što ujedno povlači da su lošijeg uspjeha od (100% - 78.81%) 21.19% studenata na ispitu.[9]

Isto tako, mogli bismo također saznati rezultat ispita vezan za neki konkretan percentil. Postoje dostupni alati koji nam mogu pomoći u

[9] Primjenom funkcije $normdist$ u Excelu moguće je izravno očitavati vjerojatnosti povezane uz varijablu $N(\mu, \sigma^2)$ bez potrebe za prethodnim standardiziranjem.

ovom izračunu. Prije svega, funkcija *“=NORMSINV(p)”* (u starijim verzijama Excela) odnosno funkcija *“=NORM.S.INV(p)”* daje z - vrijednost koja određuje područje ispod krivulje s površinom p pri čemu je to područje lijevo od z - vrijednosti, odnosno vraća vrijednost z za koju vrijedi da je vjerojatnost da slučajna varijabla $N(0,1)$ poprimi vrijednost manju ili jednaku z jednaka p. Nakon što smo izračunali z - vrijednost, možemo koristiti izraz (4-7) kako bismo izračunali vrijednost x povezanu s danom z - vrijednosti.[10]

$$x = \mu + z\sigma \tag{4-7}$$

Naravno, ova jednadžba izvedena je iz jednadžbe (4-6).

Nastavljajući s istim primjerom kao i prije, pretpostavimo da želimo pronaći rezultat ispita određenog s 75. percentilom. To bi značilo da je površina ispod krivulje p jednaka 0.75 i funkcija *“=NORMSINV(p)”* daje z - vrijednost jednak 0.6745. Korištenjem ove *z*-vrijednosti, vrijednosti μ jednake 83 i σ jednake 5 izraz (4-5) daje $x = 86.37$.

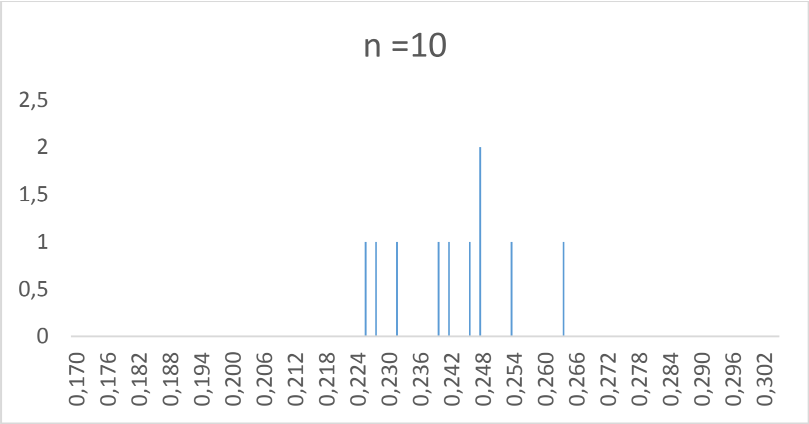
4.3 Centralni granični teorem

Veliki dio ostatka ove knjige odnosi se na prikupljanje uzorka iz veće populacije, pri čemu se prikupljanjem informacija iz uzorka nastoji donijeti zaključke o populaciji kao cjelini. Statističko zaključivanje o populacijskim karakteristikama zasniva se na **centralnom graničnom teoremu**, jednom od najvažnijih rezultata teorije vjerojatnosti. Centralni granični teorem u osnovi tvrdi da će distribucija aritmetičkih sredina dovoljno velikih slučajnih uzoraka iz jedne populacije s konačnom varijancom biti normalna.

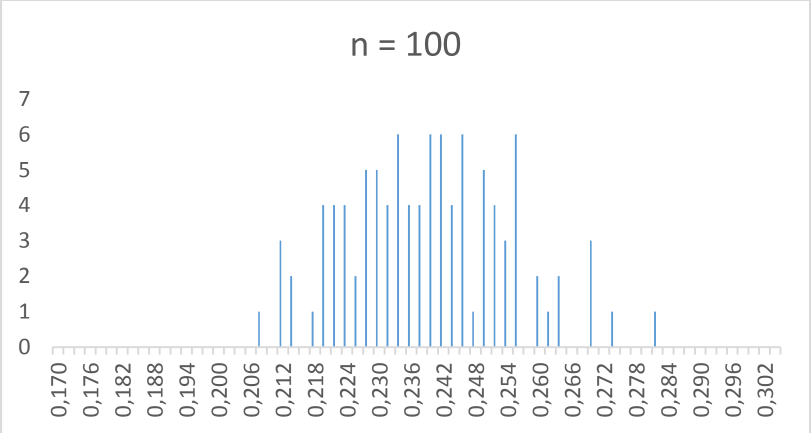
Pretpostavimo da iz neke hipotetske populacije uzimamo uzorke i za svaki uzorak računamo aritmetičku sredinu uzorka. Bilježimo

[10] Primjenom funkcije $norminv$ u Excelu moguće je izravno očitavati vrijednosti x povezane uz varijablu $N(\mu, \sigma^2)$ bez potrebe za standardiziranjem.

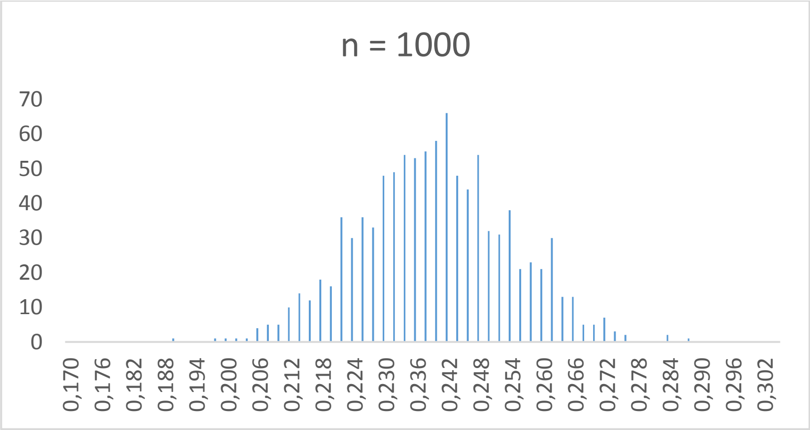
vrijednosti aritmetičkih sredina i prikažemo distribuciju aritmetičkih sredina grafički. Grafikon 4.6 dočarava tvrdnju centralnog graničnog teorema. Grafikon (a) prikazuje histogram distribucije aritmetičkih sredina kao rezultat uzimanja uzorka iz hipotetske populacije s veličinom uzorka $n = 10$. Grafikon (b) prikazuje histogram kao rezultat uzimanja uzorka iz iste populacije s veličinom uzorka $n = 100$. Grafikoni (c) i (d) pokazuju isto, ali s veličinama uzorka od 1000 i 10000. Kao što se može vidjeti, povećanjem veličine uzorka povećava se sličnost s normalnom distribucijom.



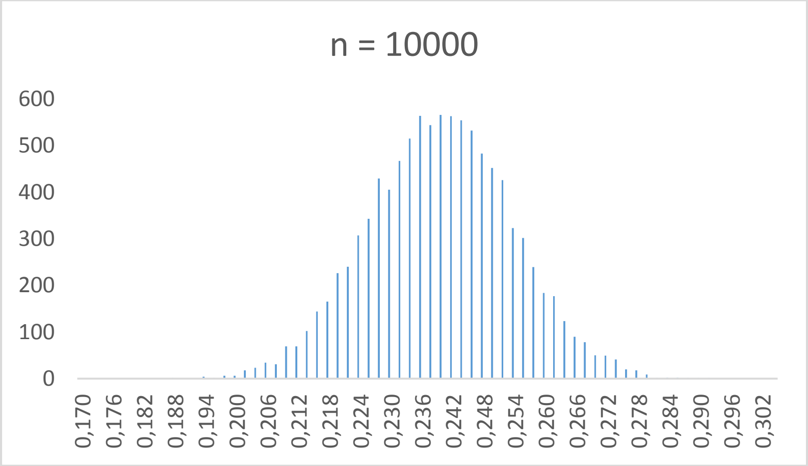
Grafikon 4.6a. Veličina uzorka $n = 10$



Grafikon 4.6b. Veličina uzorka $n = 100$



Grafikon 4.6c. Veličina uzorka $n = 1000$



Grafikon 4.6d. Veličina uzorka $n = 10000$

Kao što se može vidjeti iz grafikona 4.6, veći uzorci pomažu u razumijevanju populacije. Primijetimo, na temelju uzorka donosimo zaključke o populaciji, tj. o populacijskim veličinama kao primjerice o aritmetičkoj sredini populacije. Pri tome radimo grešku budući da su u izračun uključeni elementi uzorka, a ne cijele populacije. Dakle, uzimanjem većih uzoraka činimo manju pogrešku pri zaključivanju o populacijskim veličinama. Učinjenu pogrešku možemo kvantificirati kroz standardnu pogrešku procjene aritmetičke sredine koja je definirana formulom:

$$standardna\ pogreška(se) = \frac{s}{\sqrt{n}}, \tag{4-8}$$

gdje je s uzoračka standardna devijacija, a n veličina uzorka. Kao što se može vidjeti, standardna pogreška se smanjuje povećanjem veličine uzorka.

Dakle, na temelju uzorka procjenjujemo parametre populacije (kao što je primjerice aritmetička sredina populacije). Uzimanje različitih uzoraka rezultira različitim vrijednostima procjena stoga svaka procjena ovisi o uzetom uzorku. Drugim riječima, rezultat procjene parametra je rezultat slučajnog uzorka, pa na procjenu parametra gledamo kao na realizaciju slučajne varijable. Distribuciju te slučajne varijable nazivamo **sampling distribucijom**. Ukoliko je populacija normalno distribuirana tada je i sampling distribucija aritmetičkih sredina normalno distribuirana. Nadalje, čak i u slučajevima kada populacija nije normalno distribuirana, a uzorak je dovoljno velik (obično pod dovoljno velik uzimamo da je veći od 30), sampling distribucija će biti (približno) normalna – ovo je posljedica centralnog graničnog teorema.

Dakle, sampling distribucija je distribucija procjenitelja o kojima će biti više riječi u idućem poglavlju. Za sada rezimirajmo:

(1) Ukoliko je populacija normalno distribuirana $N(\mu, \sigma^2)$, tada je i sampling distribucija aritmetičkih sredina normalno distribuirana.

(2) Ukoliko populacija nije normalno distribuirana, a imamo velik uzorak, tada je sampling distribucija aritmetičkih sredina približno normalno distribuirana.

Ostaje vidjeti o kojoj se normalnoj distribuciji radi – odrediti parametre μ i σ^2. Ukoliko je varijanca σ^2 poznata, tada je sampling distribucija aritmetičkih sredina normalna $N(\mu, \frac{\sigma^2}{n})$ slučajna varijabla, dok je standardizirana sampling distribucija

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

jedinična (standardizirana) normalna slučajna varijabla $N(0,1)$. No, u praksi nam je varijanca populacije rijetko poznata pa se češće susrećemo sa slučajem kada i samu varijancu populacije treba procijeniti i to, ponovo, na temelju uzorka. Tada se sampling distribucija aritmetičkih sredina ravna po Studentovoj distribuciji s $(n-1)$ stupnjeva slobode koju označavamo s $t(n-1)$ i stoga je često kraće nazivamo t-distribucijom, gdje n označava veličinu uzorka. Napomenimo da se Studentova distribucija za velike n ponaša približno kao normalna $N(0,1)$ distribucija.

T-vrijednost možemo interpretirati kao udaljenost između aritmetičkih sredina uzoraka ($\bar{x}$) i stvarne sredine populacije (μ) izraženu u jedinicama standardne pogreške. Matematički se to može zapisati kao:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{4-9}$$

Ova jednadžba nema primjenu u ovom trenutku u knjizi, ali u poglavlju o testiranju hipoteza će imati. U ovom trenutku, razumno je samo shvatiti da je t-vrijednost dobivena kao rezultat odabira uzorka, a njena vrijednost označava za koliko standardnih pogreški aritmetička sredina uzorka odstupa od aritmetičke sredine populacije.

4.4 Zaključci

Unatoč činjenici da su mnoge različite vrste distribucija vjerojatnosti obrađene u ovom poglavlju, samo smo zagrebali površinu u odnosu na mnoštvo distribucija vjerojatnosti koje u prirodi postoje. Srećom, normalna distribucija opisuje velik broj pojava pa će se upravo normalna distribucija koristiti u ostatku ove knjige.

Osim toga, diskretne distribucije se također često pojavljuju i sad imamo mogućnost da ih bolje razumijemo.

4.5 Zadaci

1. Baca se jedan par kockica. Koliko iznosi očekivana vrijednost ishoda?
2. Baca se jedan par kockica. Koliko iznosi standardna devijacija ishoda?

Koristite informacije iz zadatka 3 u nastavku kako biste riješili zadatke 3 - 7.

3. Kupio sam 16 tropskih riba vrste “cichlisomanegrofasciata”. Ne mogu odrediti spol ribe, a rečeno mi je da vjerojatnost da jedna riba bude ženka iznosi 53%. Konstruirajte odgovarajuću distribuciju vjerojatnosti za ishode od 0 do 16 ženki.
4. Kolika je vjerojatnost da bude šest ili manje ženki?
5. Kolika je vjerojatnost da bude više od devet ženki?
6. Koliki je očekivani broj ženki?
7. Kolika je standardna devijacija broja ženki?

Koristite sljedeće podatke za rješavanje zadataka 8. - 12. Populacija muškaraca u Americi ima prosječnu visinu od 70 inča sa standardnom devijacijom od 1.5 inča.

8. Kolika je vjerojatnost da je muškarac visok 72 inča ili više?
9. Kolika je vjerojatnost da je muškarac visok 68 inča ili manje?
10. Kolika je vjerojatnost da je muškarac visok između 69 i 71 inča?
11. Koja visina je povezana s 85. percentilom?
12. Koja visina je povezana s 25. percentilom?

5. Procjena

Utvrdili smo mnogo puta da statistika uključuje korištenje podataka iz uzorka kako bi se moglo nešto reći o populaciji. Ova tvrdnja je posebno istinita u ovom poglavlju. Ovdje želimo odrediti obilježje populacije na temelju podataka iz uzorka. Potrebno je još jednom istaknuti da zaključke o populaciji donosimo na temelju nepotpunih podataka. Ukratko, *procjenjuje* se parametar populacije jednim brojem, a ukoliko želimo procjenu iskazati s nekom pouzdanosti tada nam **pouzdani interval** pomaže iskazati koliko smo uvjereni u našu procjenu parametra populacije.

Da bi se to postiglo, potrebno je poznavati osnovna svojstva Studentove *t*-distribucije. Studentova distribucija ima oblik sličan normalnoj, nešto je šira i položenija. S porastom broja stupnjeva slobode oblik Studentove distribucije se približava normalnoj, a za broj stupnjeva slobode veći od 30 one postaju približno jednake.

5.1 Procjena aritmetičke sredine

Procjena aritmetičke sredine populacije (μ) na temelju podataka iz uzorka prvi je postupak s kojim ćemo se upoznati. Kada se prikupe podaci iz uzorka, poznata je veličina uzorka (n) i moguće je izračunati aritmetičku sredinu uzorka ($\bar{x}$) i uzoračku standardnu devijaciju (s). Te podatke možemo koristiti, zajedno s odgovarajućom t-vrijednosti, za određivanje pouzdanog intervala, računajući donju granicu (DG) i gornju granicu (GG) procjene μ. Da bi se granice odredile, potrebno je zadati razinu pouzdanosti uz koju je stvarna aritmetička sredina populacije unutar granica. Razina pouzdanosti je $1-\alpha$, gdje je vrijednost α ili dana ili pretpostavljena. Vrijednost α se obično naziva **razina značajnosti** te je ulazni podatak za određivanje veličine granica intervala. Matematički taj interval se može zapisati:

$$P(DG \leq \mu \leq GG) = 1-\alpha \tag{5-1}$$

Izraz (5-1) govori da se stvarna aritmetička sredina populacije nalazi između donje i gornje granice s vjerojatnošću $1-\alpha$. Izraz (5-1) se može preformulirati tako da prikazuje vrijednosti koje se odnose na uzorak:

$$P(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}) = 1 - \alpha. \tag{5-2}$$

Ovdje vrijednost $t_{\alpha/2}$ predstavlja $1 - \alpha/2$ kvantil Studentove distribucije te je u Excelu računamo putem funkcije:

$$t_{\alpha/2} = T.INV.2T(\alpha, n - 1) \tag{5-3}$$

Ova konkretna vrijednost se također primjenjuje u postupku testiranja hipoteza aritmetičkih sredina što je objašnjeno u sljedećem poglavlju.

Zbog prikladnosti, donja i gornja granica se često prikazuju u skraćenom zapisu:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{5-4}$$

Simbol "$\pm$" upućuje da se oduzima ili zbraja određen broj standardnih pogrešaka od aritmetičke sredine uzorka $\bar{x}$.

Pretpostavimo da nas zanima procjena 95% pouzdanog intervala godišnjeg povećanja plaća zaposlenika neke tvrtke. Nemamo podatke o povećanju svih plaća pa moramo nasumce odabrati nekoliko radnika i pitati ih o njihovom godišnjem povećanju plaća. Naš slučajni uzorak rezultira s 25 ispitanika ($n = 25$) na temelju čega je dobivena aritmetička sredina uzorka od 1800 dolara godišnjeg povećanja uz uzoračku standardnu devijaciju od 475 dolara. Ako se zahtijeva 95% pouzdani interval, tada je signifikantnost $\alpha = 0.05$. Koristeći funkciju (5-3) u Excelu, dobiva se $t_{\alpha/2} = 2.06$. Uvrštavanjem tih vrijednosti u izraz (5-4), dobivaju se granice za prosjek populacije:

$$1800 \pm (2.06)\frac{475}{\sqrt{25}} \tag{5-5}$$

Tako je donja granica prosječnog povećanja plaće 1603.93 dolara, a gornja granica 1996.07 dolara. Tako formalni zaključak glasi:

$$P(1603.93\, dolara \leq \mu \leq 1996.07\, dolara) = 0.95 \qquad (5\text{-}6)$$

Riječima se navedeni izraz navodi da se sa *95% sigurnosti može tvrditi da je stvarno prosječno povećanje plaće cijele populacije između 1603.93 dolara i 1996.07 dolara.*

5.2 Procjena proporcije

Određivanjem pouzdanog intervala moguće je procijeniti **proporciju** odnosno udio populacije s nekom istaknutom osobinom. Proporcije se često koriste u razumijevanju preferencija potrošača, kao i u politologiji kad se pokušava saznati kako kandidati kotiraju kod svojih birača.

Za procjenu proporcije najprije je potrebno uvesti nekoliko oznaka koje su navedene u tablici 5.1.

Pojam	Objašnjenje
$\hat{p}$	procijenjena proporcija populacije
P	(stvarna) proporcija populacije
n	veličina uzorka
$z_{\alpha/2}$	koeficijent pouzdanosti

Tablica 5.1. Vrijednosti za pouzdani interval proporcije

Standardna pogreška (engl. *Standard Error, SE*) procjene proporcije dana je formulom:

$$se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (5\text{-}7)$$

Pouzdani interval proporcije je vrlo sličan pouzdanom intervalu aritmetičke sredine. Granice se računaju tako da se procjeni doda, odnosno oduzme umnožak koeficijenta pouzdanosti i standardne pogreške. Formalni zapis sličan onom iz (5-4):

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (5\text{-}8)$$

Sličnosti između izraza (5-8) i (5-4) su prilično očite, ali postoji jedna razlika koju je potrebno naglasiti. Pouzdani interval procjene aritmetičke sredine koristi t-distribuciju, a pouzdani interval procjene proporcije, međutim, koristi z-distribuciju. Potrebno je naglasiti da se pouzdani interval proporcije u prikazanom obliku koristi isključivo za velike uzorke!

Z-vrijednost koja daje koeficijent pouzdanosti pri procjeni proporcije u Excelu se može izračunati primjenom funkcije:

$$z_{\alpha/2} = NORMSINV(1 - \alpha/2) \tag{5-9}$$

Pogledajmo primjer koji objašnjava kako konstruirati pouzdani interval procjene proporcije. Pretpostavimo da smo proveli testiranje okusa na 1572 ispitanika. Od svih ljudi u uzorku, njih 832 je preferiralo proizvod A više nego proizvod B. To znači da je procijenjena proporcija ljudi koji su preferirali proizvod A više nego B jednaka $\hat{p}$ = 832/1572 = 0.5292. Ako želimo procijeniti 99% pouzdani interval, tada je $\alpha = 0.01$. Koristeći izraz (5-9) za dobivanje z -vrijednosti dobivamo da je $z_{\alpha/2}$ jednak 2.5758. Uvrštavanjem tih vrijednosti u izraz (5-8) dobivamo pouzdani interval:

$$P(0.4968 \ \leq p \leq 0.5617) = 0.99 \tag{5-10}$$

Drugi način na koji se može izreći gornja tvrdnja je da s pouzdanošću od 99% tvrdimo da je stvarna proporcija potrošača koji preferiraju proizvod A više nego proizvod B između 0.4968 i 0.5617.

5.3 Procjena razlike između aritmetičkih sredina dviju populacija

Posljednji tip pouzdanog intervala kojeg ćemo spomenuti je pouzdani interval razlike između aritmetičkih sredina dviju populacija. U tom slučaju se uzimaju uzorci iz dvije populacije i određuje pouzdani interval razlike u njihovim očekivanim vrijednostima. Za ovu vježbu uvedeni su neki novi, ali poznati pojmovi, navedeni u tablici 5.2. Također, pretpostavimo da varijance populacija nisu poznate te da su uzorci mali, nezavisni i potječu iz normalno distribuiranih populacija.

Vrijednost	Uzorak 1	Uzorak 2
Veličina uzorka	n_1	n_2
Aritmetička sredina populacije	μ_1	μ_2
Aritmetička sredina uzorka	$\bar{x}_1$	$\bar{x}_2$
Standardna devijacija uzorka	s_1	s_2

Tablica 5.2. Vrijednosti korištene za pouzdani interval razlike aritmetičkih sredina

Standardna pogreška za ovaj pouzdani interval je:

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5-11}$$

Stupnjevi slobode (engl. *degrees of freedom, df*) su:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2} \tag{5-12}$$

Formula za granice pouzdanog intervala procjene razlike aritmetičkih sredina dviju populacija je:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5-13}$$

Ovdje je $t_{\alpha/2}$ vrijednost jednaka onoj definiranoj izrazom (5-3).

Kao primjer, konstruirat će se 90% pouzdani interval ($\alpha = 0.10$) za razlike u prosječnim rezultatima ispita za dvije različite grupe studenata. Dani su sljedeći podaci:

Vrijednost	Uzorak 1	Uzorak 2
Veličina uzorka	$n_1 = 47$	$n_2 = 52$
Aritmetička sredina uzorka	$\bar{x}_1 = 81$	$\bar{x}_2 = 83.5$
Standardna devijacija uzorka	$s_1 = 7.2$	$s_2 = 6.2$

Tablica 5.3. Primjer podataka za razlike u prosječnim rezultatima ispita

Korištenjem danih vrijednosti i formula (5-11), (5-12) i (5-13), pouzdani interval je:

$$P(-4.76 \leq \mu_1 - \mu_2 \leq -0.24) = 0.90 \quad (5\text{-}14)$$

Znači da je, uz pouzdanost od 90%, razlika u srednjoj vrijednosti populacija dvije grupe studenata između -4.76 i -0.24.

Nadalje, u slučaju kada su varijance populacija poznate i uzorci veliki, u procjeni intervala koristi se normalna distribucija, a ne t-distribucija. Također, u slučaju kada su varijance populacija poznate i populacije normalne, u procjeni intervala koristi se normalna distribucija, a ne t-distribucija.

5.4 Zaključak

Intervali pouzdanosti su jednostavan način procjene parametara populacije. Što višu pouzdanost tražimo, to će interval biti širi. Istovremeno, izuzimanjem većeg uzorka, širina intervala se smanjuje.

Problemi generiranja doći će do izražaja kada se bude provodilo dvosmjerno testiranje hipoteza u sljedećem poglavlju. Zbog ovog poglavlja se sljedeće poglavlje, koje bi inače bilo pozamašnog sadržaja, značajno smanjilo.

5.5 Zadaci

1. Skup podataka "NFLLinijaš" sadrži podatke o težini nasumično odabranih napadačkih linijaša Američke nogometne lige. Korištenjem $\alpha = 0.01$ odredite odgovarajući pouzdani interval procjene aritmetičke sredine populacije.
2. Korištenjem $\alpha = 0.05$ odredite pouzdani interval broja otkucaja srca nasumično odabranih sveučilišnih sportaša. Skup podataka je dan

u datoteci “OtkucajiSrca”.

3. U uzorak je na slučajan način odabrano 2234 stanovnika Sjeverne Karoline kojima je postavljeno pitanje smatraju li Sveučilište Wake Forest najboljim fakultetom u državi. 1465 ispitanika je izjavilo kako misle da je Wake Forest najbolji fakultet u državi. Odredite 95% pouzdani interval i procijenite stvarnu proporciju onih koji smatraju da je Wake Forest najbolji fakultet u državi.
4. 3689 stanovnika Oregona su nasumično ispitani o tome podržavaju li legalizaciju marihuane u medicinske svrhe. 2983 ljudi je odgovorilo potvrdno. Koristeći ove podatke odredite 93% pouzdani intervali procijenite stvarnu proporciju stanovnika Oregona koji podržavaju legalizaciju marihuane u medicinske svrhe.
5. Profesor Prijevara i profesor Brzić podučavaju odvojene dijelove Načela financija. Dali su isti ispit studentima. Skup podataka “ProfesoriFinancija” daje rezultate ispita za obje skupine. Pomoću danih podataka odredite 90% pouzdani interval za prosječnu razliku u ostvarenom uspjehu.
6. Jesu li rezultati kod jednog profesora bolji u odnosu na rezultate studenata drugog profesora? Zašto jesu ili zašto nisu?
7. Prehrambena tvrtka upravo je stavila na tržište dva nova proizvoda za dijabetičare: proizvod X i proizvod Y. U tvrtci su zabrinuti zbog količine ugljikohidrata u proizvodu budući da dijabetičari moraju biti oprezni s unosom ugljikohidrata. Datoteka “HranaZaDijabetičare” sadrži podatke o ugljikohidratima (u gramima) za nasumično odabrana pakiranja dvaju različitih proizvoda. Odredite 96% pouzdani interval za razliku u količini ugljikohidrata između ta dva proizvoda.
8. Bez obzira na vrstu određenog pouzdanog intervala, što široki pouzdani interval sugerira?
9. Bez obzira na vrstu određenog pouzdanog intervala, što uski pouzdani interval sugerira?
10. Koliko iznosi standardna pogreška procjene u zadatku 3?
11. Koliko iznosi standardna pogreška procjene u zadatku 4?

6. Testiranje hipoteza

Niti u znanosti, niti u trgovini, nije dopušteno iznositi neutemeljene tvrdnje. Primjerice, ako radimo za elektroničku tvrtku, ne možemo svojim kolegama i menadžerima reći sljedeće: “Naši Q55 elektronski prekidači ne mogu podnijeti napon od 25 volti”. Što nije u redu s ovom tvrdnjom? Izrekli smo neku tvrdnju bez formalnog znanstvenog ispitivanja. U cilju iznošenja izjava iznad razine žutog tiska, mora biti znanstveno dokazana njena autentičnost. Kako bi se osigurao znanstveni dokaz, potrebno je provesti formalno testiranje tvrdnje ili hipoteze.

Ovo poglavlje opisuje alate za provođenje testiranja hipoteza. Testirat ćemo hipoteze koje uključuju testove o parametrima: aritmetičkoj sredini, proporciji i razlikama između aritmetičkih sredina. Drugim riječima, proširit ćemo područje iz prethodnog poglavlja s namjerom potvrđivanja ili opovrgavanja neke tvrdnje.

6.1 Općenito o testiranju

Pretpostavka je **tvrdnja**, **zaključak** ili **nagađanje**. Ovi termini se koriste kao sinonimi. Šesto poglavlje ove knjige je “**Testiranje hipoteza**”, dok se u drugim knjigama može pronaći naslov “**Statističko zaključivanje”**. Bez obzira na korištenu terminologiju, testira se tvrdnja o populaciji radi provjere može li se tvrdnja poduprijeti znanstvenim dokazom. Utvrđivanje istinitosti ili neistinitosti tvrdnje zahtijeva formalno testiranje hipoteza.

Postoje tri različite vrste testova o parametru. Prva vrsta testa je testiranje tvrdnje o jednakosti parametra populacije nekoj specifičnoj vrijednosti (odnosno provjeravamo je li parametar *jednak* ili *nije jednak* nekoj unaprijed određenoj vrijednosti). Druga vrsta testa se bavi pitanjem je li neki parametar populacije *jednak* ili je *manji od* neke vrijednosti. Treća vrsta testa se bavi pitanjem je li neki parametar populacije jednak ili je *veći od* neke vrijednosti. Ove tri vrste testa su formalno opisane u narednim potpoglavljima.

6.1.1 Nulta i alternativna hipoteza

Svaki test hipoteza uključuje dvije različite hipoteze: **nultu hipotezu** ($\boldsymbol{H_0}$) i **alternativnu hipotezu** ($\boldsymbol{H_A}$). Obje hipoteze se iskazuju istovremeno i proturječne su.

Primjer nulte hipoteze: prosječna težina kutije Cheeriosa je 14 unci (što je približno 397 grama). Slijedi formalni matematički zapis:

$$H_0 : \mu = 14 \tag{6-1}$$

Aritmetička sredina populacije uključena je u H_0. To je zato što se u H_0 daje tvrdnja o aritmetičkoj sredini populacije. Također, primijetite da je u H_0 izražena jednakost. Konačno, treba napomenuti da prilikom donošenja zaključaka o hipotezi, zaključak se uvijek donosi u odnosu na H_0: hipotezu H_0 ćemo *odbaciti* ili ju *nećemo odbaciti*.

Alternativna hipoteza H_A je suprotna hipotezi H_0. H_A se obično odnosi na tvrdnju oko koje je istraživač najviše zainteresiran. Postoje tri vrste testova o parametru koji se razlikuju s obzirom na sadržaj alternativne hipoteze. Tako postoje tri vrste alternativnih hipoteza o parametru: prva koja tvrdi da “nije jednak”, druga koja tvrdi da je “manje od” i treća koja tvrdi “veće od”. Odbacivanje neistinite tvrdnje presudno ovisi o hipotezi H_A. Naime, nulta hipoteza je istinita sve dok nema dovoljno prikupljenih dokaza za dokazati suprotno – da je neistinita i tek tad ju se smije odbaciti. Ako vam kazneni postupak pada na pamet na temelju ovog opisa, definitivno ste na pravom putu: pretpostavljamo da je H_0 istinita (nevina) osim ako se dokaže suprotno (krivnja je dokazana izvan svake razumne sumnje). Ovo nas dovodi do mogućih pogrešaka. Kao što je slučaj s kaznenim postupkom, možemo učiniti pogreške u našem odlučivanju vezano za H_0. Pogreška koju možemo učiniti je odbaciti istinitu nultu hipotezu (pogreška prve vrste). Druga vrsta pogreške koju možemo učiniti je ne odbaciti neistinitu hipotezu. Uvedimo oznake: pogreška tipa I se pojavljuje s vjerojatnošću α, dok je vjerojatnost pogreške tipa II jednaka β. Tablica 6.1 opisuje ove moguće pogreške.

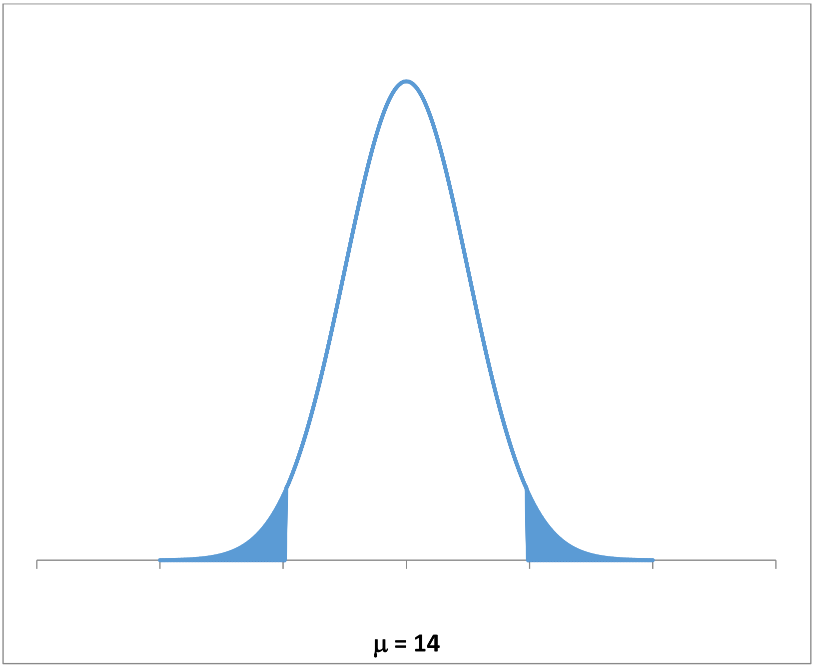
	Odbacuje se H_0	Ne odbacuje se H_0
H_0 **istinita**	Pogreška tipa I	Ispravna odluka
H_0 **lažna**	Ispravna odluka	Pogreška tipa II

Tablica 6.1. Rezultati odluke o H_0

Prvi tip alternativne hipoteze H_A koju ćemo proučiti je ona koja tvrdi da parametar “nije jednak" nekoj unaprijed određenoj vrijednosti. Kad se prikazuje H_0 u našem primjeru s Cheeriosima, imamo:

$$H_0: \mu = 14; \; H_A: \mu \neq 14 \tag{6-2}$$

Izraz “nije jednak” podrazumijeva nejednakost - ne navodi se posebno koja vrsta nejednakosti se uzima u obzir (ni “<” niti “>”). Postavlja se pitanje tko bi mogao biti zainteresiran za ovakvo testiranje. U kontekstu primjera, to su ljudi za kontrolu kvalitete koji proizvode Cheeriose. Oni jednostavno žele dosljedno provjeriti da posudice sadrže 14 unci žitarica. Grafikon 6.1 prikazuje scenarij odbijanja H_0 grafički s obzirom na to da aritmetička sredina populacije posudica iznosi 14 unci. Osjenčana područja nazivaju se “područjima odbacivanja”. Budući da H_A tvrdi "nije jednak", imamo područje odbacivanja "<" i područje odbacivanja ">". Budući da je područje odbacivanja s obje strane distribucije, ovakav test se naziva "dvosmjerni test".

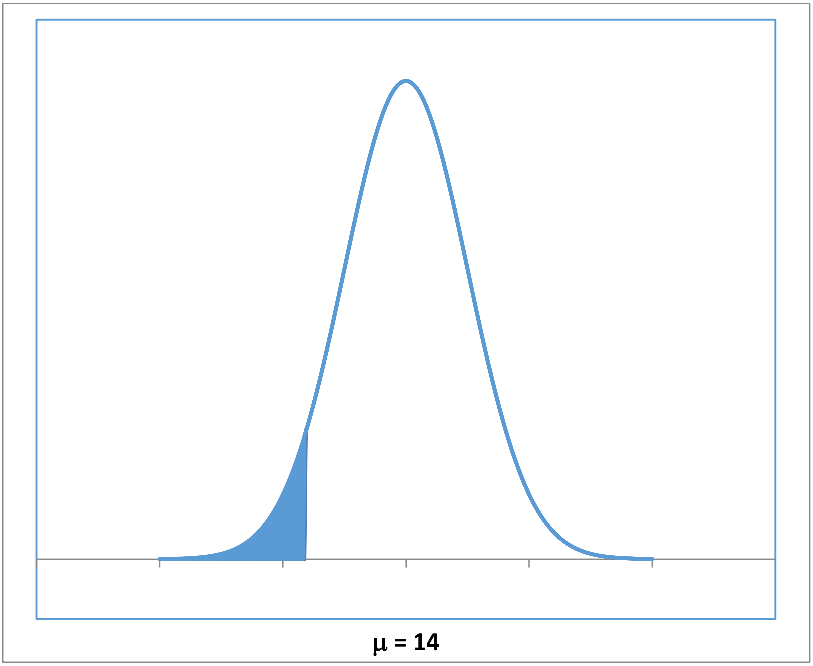


Grafikon 6.1. $H_0: \mu = 14;\ H_A: \mu \neq 14$

Drugi tip hipoteze H_A je "manje od". Za primjer o Cheeriosu imamo:

$$H_0: \mu = 14;\ H_A: \mu < 14 \tag{6-3}$$

Vezano za ovaj primjer, zainteresirana strana mogli bi biti odvjetnici potrošača koji žele pokazati da Cheerios "smanjuje" količine Cheeriosa u svojim kutijama, u slučaju da je hipoteza H_A istinita. Grafikon 6.2 grafički prikazuje ovaj scenarij uz područje odbacivanja na lijevoj strani. Budući da je područje odbacivanja na samo jednoj strani distribucije, ovaj test nazivamo "jednosmjerni test".

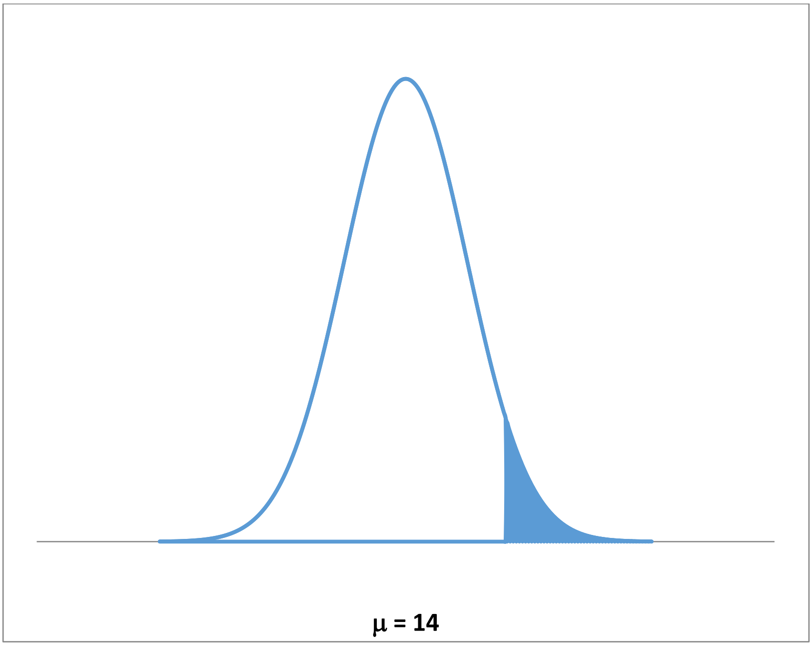


Grafikon 6.2. $H_0: \mu = 14;\ H_A: \mu < 14$

Posljednji tip hipoteze H_A koji ćemo proučavati je “veći od”. Nastavljajući s primjerom Cheeriosa imamo sljedeće:

$$H_0: \mu = 14;\ H_A: \mu > 14 \tag{6-4}$$

Za ovaj scenarij viši rukovoditelji u General Millsu (tvrtka koja proizvodi Cheerios) bi mogli biti zainteresirani jer ako bi H_A bila istinita, žitarice bi se dijelile što je neoprostivo djelo u očima viših rukovoditelja i dioničara. Grafikon 6.3 prikazuje taj scenarij na kojem možete primijetiti da se područje odbacivanja nalazi na desnoj strani. Budući da je područje odbacivanja na samo jednoj strani distribucije, ovaj test nazivamo "jednosmjerni test".



Grafikon 6.3. $H_0: \mu = 14;\ H_A: \mu > 14$

Uz opće pojmove hipoteza H_0 i H_A koje su opisane, sada možemo raspravljati o koracima testiranja hipoteza.

6.1.2 Koraci pri testiranju

Studenti često smatraju da je testiranje hipoteza najteže područje statistike - to je svakako bio slučaj i kod autora. Iskustvo i dobar profesor, postupak testiranja mogu učiniti jednostavnim.

6.1.2.1 Postavljanje H_0 i H_A

Ispravno postavljanje hipoteza H_0 i H_A je najteži dio postupka. Važno je napomenuti da istraživač ima vrlo veliku odgovornost u jasno postavljenim pitanjima. Pogledajte sljedeća hipotetska pitanja:

a) Odjel za kontrolu kvalitete u Cheeriosu istražuje stavlja li se 14 unci žitarica u pakiranje. Ako se stavi premalo Cheeriosa, kupci ne dobivaju odgovarajuću vrijednost za svoj novac. Ako se stavi previše žitarica u pakiranje, General Mills gubi novac. Stoga je stavljanje 14 unci žitarica u pakiranje iznimno važno. Koristeći skup podataka "Cheerios" utvrdite stavlja li General Mills odgovarajuću količinu žitarica u pakiranja.

b) Dostavna tvrtka Dostavljači tvrdi da je vrijeme isporuke manje od 6 sati. Koristeći skup podataka “Dostavljači” utvrdite je li njihova tvrdnja istinita.

c) CircuitKings proizvodi sklopove za daljinske upravljače za TV. Kao vodiče koriste legure srebra. Svaka pločica mora sadržavati 0.25mg legure srebra. Ako se koristi više od tog, CircuitKings gubi novac jer je legura srebra jako skupa. CircuitKings je zabrinut da se stavlja previše legure u pločice. Koristeći skup podataka CircuitKings, utvrdite stavlja li se previše legura srebra u pločice.

Svaka od ovih tvrdnji je korektno i jasno napisana. Za tvrdnju a) hipoteza H_A treba biti tipa “$\neq$” jer je stavljanje premalo žitarica u pakiranje jednako loše kao i stavljanje previše žitarica u pakiranje. Za tvrdnju b) hipoteza H_A treba biti tipa “<” jer tvrdnja tako to implicira. Hipoteza H_A izričito je navedena što se povremeno događa. Tvrdnja c) je tipa “>” jer se raspravljalo o opasnosti stavljanja previše legura srebra. Ukratko, sljedeće hipoteze su spremne za testiranje:

a) $H_0: \mu = 14;\ H_A: \mu \neq 14$
b) $H_0: \mu = 6;\ H_A: \mu < 6$
c) $H_0: \mu = 0.25;\ H_A: \mu > 0.25$

Postoje dva važna pravila kojih se ovdje treba pridržavati. Prvo, H_0 uvijek uključuje jednakost - pisanje H_0 je jednostavno. Drugo, H_A uvijek uključuje nejednakost. Još jedna važna stvar koju treba zapamtiti je da je H_A zanimljivija od H_0. Konačno, ako imate problem koji nije jasno definiran, gdje je nejasno formulirana "nejednakost", hipoteza H_A tipa “$\neq$” je vjerojatno najbolji izbor.

U ovom poglavlju ćemo koristiti tvrdnju a) kao primjer:

$$H_0: \mu = 14;\ H_A: \mu \neq 14$$

Dodatno, u ovom primjeru ćemo koristiti $\alpha = 0.05$ i skup podataka “Cheerios” koji sadrži 25 opaženih masa nasumično odabranih Cheerios pakiranja za koje se tvrdi da su teške 14 unci.

6.1.2.2 Određivanje područja odbacivanja

Kod provođenja testiranja hipoteza vrijednost α je zadana ili ju je potrebno proizvoljno odabrati. U domaćim zadaćama i ispitima je α zadana. U pisanim zadacima tipa analize slučaja, razumno je da student odabere vrijednost α. Vrijednost α određuje veličinu područja odbacivanja. Ako je α mala, onda je područje odbacivanja malo. Ako je α velika, onda je područje odbacivanja veliko. Na grafikonima 6.1. – 6.3. vrijednosti α su prikazane osjenčanim područjima koja obuhvaćaju područje odbacivanja. U Excelu se koristi funkcija "$T.INV$" kako bi se dobila kritična vrijednost za danu α. Unošenjem α i stupnjeva slobode (df) u funkciju "$T.INV$" kao rezultat se dobiva kritična vrijednost koja definira područje odbacivanja.

Vrsta H_A	Funkcija u Excelu	Funkcija u Excelu (za starije verzije Excela*)
$\neq$	$+/-T.INV.2T(\alpha, n-1)$	$+/-TINV(\alpha, n-1)$
<	$T.INV(\alpha, n-1)$	$TINV(2\alpha, n-1)$
>	$T.INV(1-\alpha, n-1)$	$TINV(1-2\alpha, n-1)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 6.2. Funkcije u Excelu za područja odbacivanja

Ova kritična vrijednost će se usporediti s t-vrijednosti (što će biti kasnije objašnjeno) kako bi se utvrdilo treba li odbaciti hipotezu H_0.

U primjeru "Cheerios", neka je $\alpha = 0.05$ i *T.INV.2T(0.05,25-1)* što rezultira kritičnom vrijednosti od +/-2.06. Dobiva se i pozitivna i negativna vrijednost jer se traži kritična vrijednost na oba repa distribucije.

Postoji jedan završni komentar koji treba reći o vrijednosti α. Studenti se često pitaju koju vrijednost α treba koristiti. To je dobro pitanje. Najčešće se koristi $\alpha = 0.05$, ali to varira ovisno o pojedinim djelatnostima. Medicinska industrija često koristi vrijednost od 0.01 ili manju. Manje vrijednosti α daju konzervativnije rezultate testova, dok veće vrijednosti α daju liberalnije rezultate testova. To znači da konzervativni testovi otežavaju odbacivanje hipoteze H_0, a liberalniji testovi olakšavaju odbacivanje hipoteze H_0. Kod liberalnijih testova, međutim, vrijednost α prelazi 0.10 i to se smatra neprikladnim. Nikad se ne koristi α veća od 0.10.

6.1.2.3 Određivanje testne veličine

Kritična vrijednost se uspoređuje s testnom veličinom. Testna veličina je rezultat formule uvedene u petom poglavlju:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{6-5}$$

Iz skupa podataka “Cheerios” imamo aritmetičku sredinu uzorka $\bar{x} = 14.03$ s uzoračkom standardnom devijacijom $s = 0.1087$. Skup podataka sadrži $n = 25$ opažanja. Nadalje, pretpostavljena aritmetička sredina populacije je $\mu = 14$. Uvrštavanjem tih vrijednosti u izraz (6-5) dobivamo sljedeću testnu veličinu:

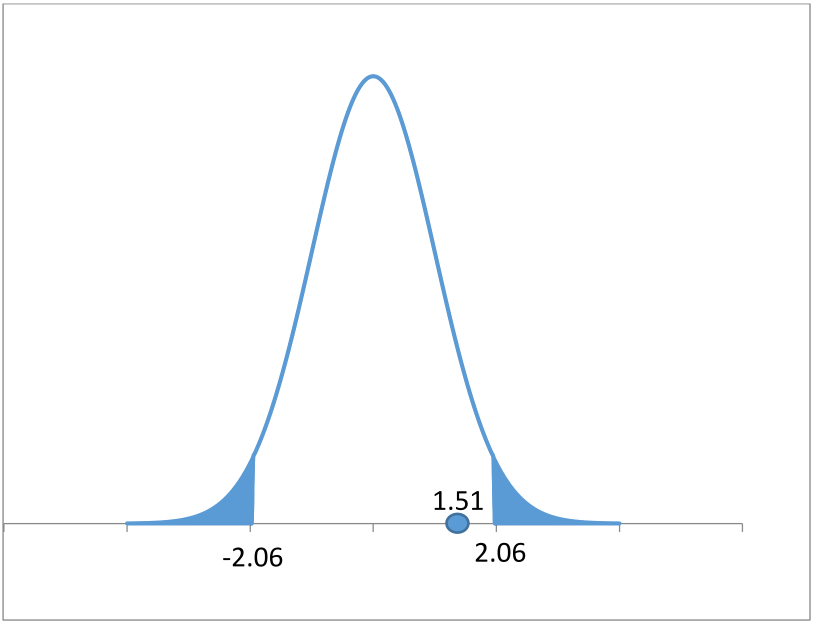
$$t = \frac{14.03 - 14}{0.1087/\sqrt{25}} = 1.51 \tag{6-6}$$

Vrijednost 1.51 govori da je aritmetička sredina uzorka 1.51 standardnu pogrešku veća od pretpostavljene aritmetičke sredine populacije. Ona je veća od pretpostavljene aritmetičke sredine populacije jer ima pozitivan predznak - drugim riječima, $\bar{x}$ premašuje μ. Pitanje je premašuje li vrijednost $\bar{x}$ pretpostavljenu vrijednost μ dovoljno da možemo tvrditi da postoji stvarna razlika i da vrijedi H_A odnosno $\neq$.

Ova testna veličina se uspoređuje s kritičnom vrijednosti kako bi se mogla donijeti odluka o H_0. Budući ona iznosi +/-20.6, nulta hipoteza se ne odbacuje.

6.1.2.4 Odluke koje se odnose na H_0

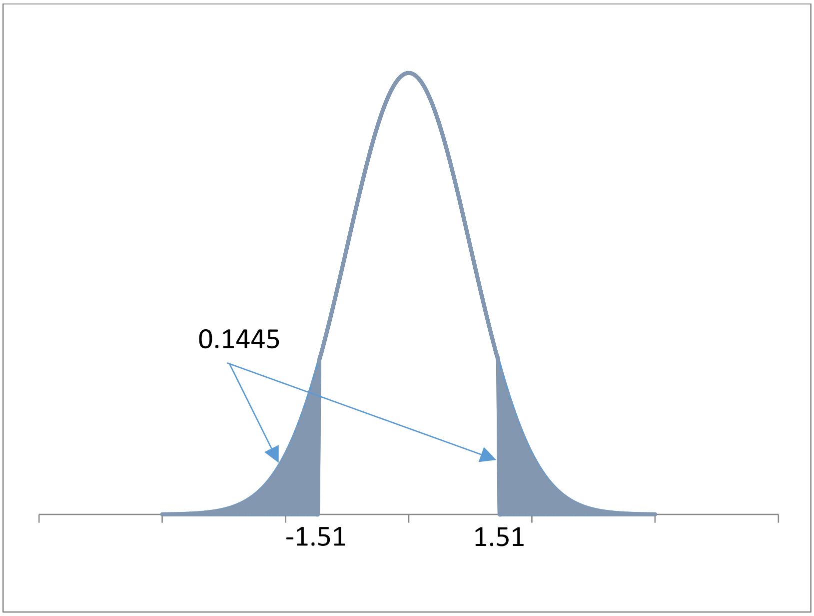
Kako je ranije navedeno, uspoređuje se kritična vrijednost s testnom veličinom. Grafikon 6.4 prikazuje grafički rezultat testa koji prikazuje da testna veličina ne pada u zasjenjeno područje odbacivanja i zbog toga se H_0 ne odbacuje. Dakle, nema dovoljno dokaza da H_0 nije istinita. General Mills stavlja 14 unci žitarica u pakiranje. Kad bi testna veličina pala u područje odbacivanja, odbacili bismo H_0 i tvrdili da General Mills ne stavlja 14 unci žitarica u pakiranje.



Grafikon 6.4. Rezultati primjera testiranja hipoteza

6.1.2.5 Izračunavanje p - vrijednosti

P - vrijednost nekog testa je posljednje što je potrebno izračunati. p - vrijednost je površina područja odbacivanja određena nekom testnom veličinom. Grafikon 6.5 prikazuje p - vrijednost.



Grafikon 6.5. p - vrijednost u primjeru testiranja hipoteza

Budući da se radi o dvosmjernom testu, zasjenjeno područje koje se odnosi na testnu veličinu također mora biti na obje strane distribucije. Da bi se izračunala p - vrijednost, koristi se funkcija *TDIST* prikazana u tablici 6.3.

Vrsta testa	Funkcija u Excelu	Funkcija u Excelu (za starije verzije Excela*)
jedosmjerni	$T.DIST(-\|t\|; n-1) = T.DIST.RT(\|t\|; n-1)$	$TDIST(\|t\|; n-1; 1)$
dvosmjerni	$T.DIST(-\|t\|; n-1) * 2 = T.DIST.2T(\|t\|; n-1)$	$TDIST(\|t\|; n-1; 2)$

***u novijim verzijama dostupno u Compatibility skupu formula**

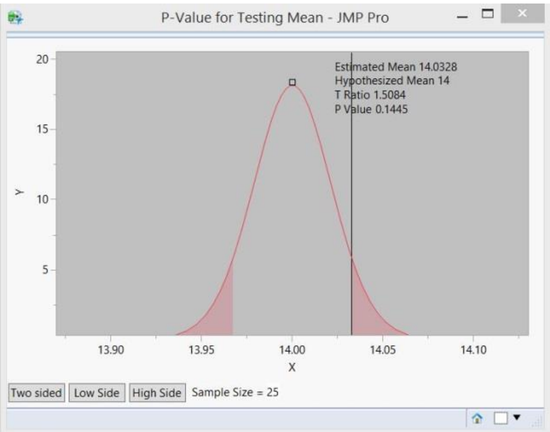
Tablica 6.3. Funkcije za p - vrijednosti

Može se primjetiti da je p - vrijednost veća od vrijednosti α. Ovo je slučaj kada se ne odbacuje H_0. Ako je testna veličina u području odbacivanja, tada je p - vrijednost manja od vrijednosti α. Ako je poznato da se H_0 neće odbaciti p - vrijednost se računa zato što daje više informacija. Daje prijelomnu vrijednost koja bi se mogla koristiti kao α - točku ravnoteže između odbacivanja i neodbacivanja H_0.

Naime, ako je p manji od α, odbacujemo H_0. Ako je p veći od α, ne odbacujemo H_0. Pjesnički rečeno:

Ako je p mali, H_0 ne pali. Ako je p velik, H_0 je k'o čelik.

Usporedbe p - vrijednosti i α nisu namijenjene zbunjivanju studenata. Primjetio sam tijekom godina da gornja izreka daje studentima pravu perspektivu. P - vrijednost je važnija od α jer se p - vrijednost može usporediti s bilo kojim pragom α. Naime, statistički programski paketi ni ne pitaju za vrijednosti α - oni samo ispisuju p - vrijednost. Grafikon 6.6 prikazuje ispis u JMP-u vezano za navedeni primjer: dana je samo p - vrijednost, a vrijednost α se ne prikazuje.



Grafikon 6.6. Izračun p - vrijednosti u JMP-u

Iako s tehničkog stajališta nije u potpunosti istinito, p - vrijednost se može shvatiti kao vjerojatnost da je H_0 istinita.

6.2 Testiranje hipoteze o sredini

Primjer istraživanja tijekom prethodnog odjeljka bio je primjer testiranja hipoteze o aritmetičkoj sredini. Postoje i testovi koji se odnose na udjele u populaciji i razlike između aritmetičkih sredina populacije. Ovo potpoglavlje će se baviti još jednim testom hipoteza koji uključuje aritmetičke sredine.

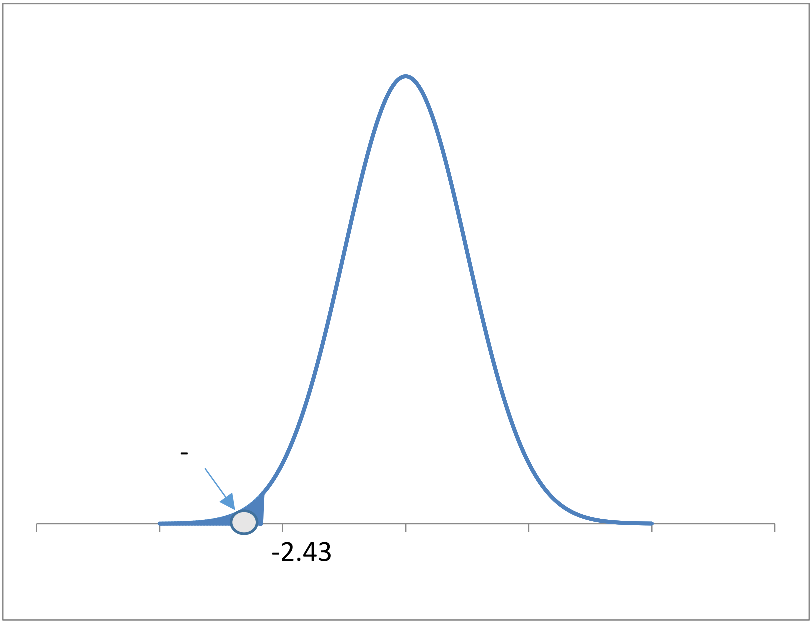
Razmatra se primjer Dostavljača koji je opisan u prethodnom potpoglavlju u kojem Dostavljači tvrde da je prosječno vrijeme isporuke manje od šest sati. Hipoteze su:

$$H_0: \mu = 6;\ H_A: \mu < 6$$

Skup podataka "Dostavljači" sadrži $n = 40$ opažanja, a zadano je $\alpha = 0.01$. Aritmetička sredina uzorka je $\bar{x} = 5.89$, a uzoračka standardna devijacija $s = 0.2836$. Pretpostavljena aritmetička sredina populacije je $\mu = 6$. Excel formula $T.INV(0.01{,}40 - 1)$ daje kritičnu vrijednost jednaku -2.43. Koristeći izraz (6-6) za izračunavanje testne veličine dobiva se:

$$t = \frac{5.89 - 6}{0.2836/\sqrt{40}} = -2.45 \tag{6-7}$$

Budući da je testna veličina još ekstremnija od kritične vrijednosti, odbacuje se H_0 i tvrdi da je vrijeme isporuke manje od 6 sati. Grafikon 6.7 prikazuje odnos između kritične vrijednosti i testne veličine.



Grafikon 6.7. Rezultati jednosmjernog testa

Testna veličina pada u područje odbacivanja. Pomoću funkcije *TDIST* izračunava se p - vrijednost i dobiva se:

$$p - vrijednost = TDIST(|-2.45|, 40 - 1, 1) = 0.0094 \quad (6\text{-}8)$$

Dobivena p - vrijednost manja je od $\alpha = 0.01$ što podržava odluku o odbacivanju H_0.

Treba naglasiti da izraz (6-8) uzima apsolutnu vrijednost testne veličine kao argument budući da funkcija "$TDIST(t, n-1{,}1)$" nije definirana za negativne testne veličine. Isto je navedeno u tablici 6.3.

6.3 Testiranje hipoteze o proporciji

Druga vrsta testiranja hipoteza odnosi se na udio u populaciji, bez obzira daje li nam uzorak vrijednost jednaku nekoj pretpostavljenoj vrijednosti. Analogno testiranju aritmetičke sredine, postoje tri različite vrste H_A: nije jednako, manje od i veće od. Za razliku od testiranja hipoteze o aritmetičkoj sredini, kod testiranja hipoteze o proporciji koristimo z - distribuciju umjesto t – distribucije jer se za testiranje proporcija uvijek pretpostavlja veliki uzorak. Kritične vrijednosti za danu vrijednost α su prikazane u tablici 6.4.

Vrsta H_A	**Funkcija u Excelu**	**Funkcija u Excelu** (za starije verzije Excela*)
≠	$+/-NORM.S.INV(1-\alpha/2)$	$+/-NORMSINV(1-\alpha/2)$
<	$NORM.S.INV(\alpha)$	$NORMSINV(\alpha)$
>	$NORM.S.INV(1-\alpha)$	$NORMSINV(1-\alpha)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 6.4. Funkcije u Excelu za područje odbacivanja kod testiranja hipoteza o proporciji

Dvosmjerni test za kritičnu vrijednost poprima obje +/- vrijednosti. Iz poglavlja o pouzdanim intervalima za proporcije, $\hat{p}$ je proporcija u uzorku, dok p_0 predstavlja pretpostavljenu proporciju populacije. Tako je testna veličina je prikazana izrazom (6-9).

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad (6\text{-}9)$$

Uz dane osnovne formule pokušajmo riješiti sljedeći problem. Hank i Keenan se natječu za gradonačelnika Louisvillea u Kentuckyju. Pobjednik izbora dobiva većinu glasova - više od 50% glasova. 352 čovjeka odabrana su u uzorak i 182 su glasovala za Keenana. Može li se na razini $\alpha = 0.03$ zaključiti da je Keenan pobjednik?

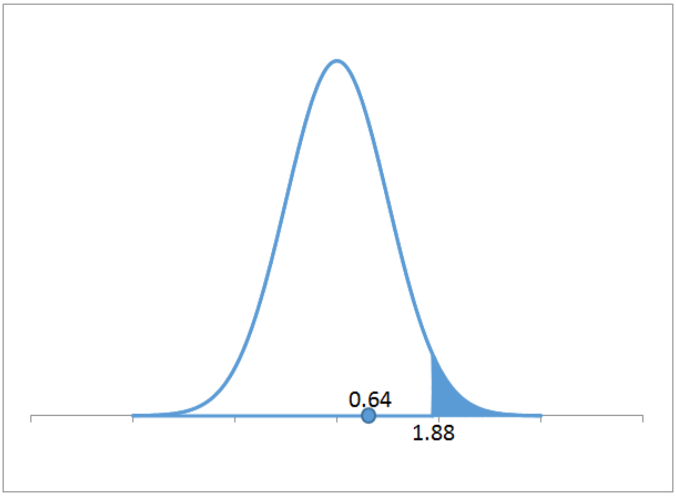
Ovaj problem se svodi na pitanje ima li Keenan više od 50% glasova, pa slijede hipoteze:

$$H_0: p = 0.50;\ H_A: p > 0.50 \tag{6-10}$$

Budući da se radi o jednosmjernom testu, kritična vrijednost dobiva se iz “$NORMSINV(1-0.03)$” što je jednako 1.88. Vrijednost $\hat{p}$ je jednaka 182/352 = 0.5170. To daje testnu veličinu jednaku

$$z = \frac{0.5170 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{352}}} = 0.64 \tag{6-11}$$

Testna veličina ne pada u područje odbacivanja, tako da ne odbacujemo H_0 jer nemamo dovoljno dokaza tvrditi da je Keenan pobjednik. Grafikon 6.8 navedeno prikazuje grafički.



Grafikon 6.8. Rezultati testiranja za izbore za gradonačelnika

Što se tiče p - vrijednosti, za ovu vrstu testa koristimo funkciju "*NORMSDIST*". Ovisno o vrsti H_A koristimo sljedeće funkcije:

Vrsta H_A	**Funkcija u Excelu**	**Funkcija u Excelu** **(za starije verzije Excela*)**
$\neq$	$2 \cdot NORM.S.DIST(-\|z\|;1) = 2 \cdot (1 - NORM.S.DIST(\|z\|;1)$	$2 \cdot NORMSDIST(-\|z\|) = 2 \cdot (1 - NORMSDIST(\|z\|)$
<	$NORM.S.DIST(-\|z\|;1) = 1 - NORM.S.DIST(\|z\|;1)$	$NORMSDIST(-\|z\|) = 1 - NORMSDIST(\|z\|)$
>	$1 - NORM.S.DIST(\|z\|;1) = NORM.S.DIST(-\|z\|;1)$	$1 - NORMSDIST(\|z\|) = NORMSDIST(-\|z\|)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 6.5. Funkcije u Excelu za *p*-vrijednosti kod testiranja proporcija

Za promatrani problem $p = 1 - NORMSDIST(0.64) = 0.2611$.

Uz gore opisan problem pojavljuje se pitanje koje je relevantno za svakodnevni život. U vijestima iz politike vezano za izbore, puno se priča o glasovanju. U detaljima glasovanja, ponekad se spominje pojam "procjene uz pogrešku (engl. *margin of error*) od primjerice +/- 3.5%". Ta "procjena uz pogrešku" koja se spominje je upravo standardna pogreška: $\sqrt{p_0(1-p_0)/n}$. Uzorci manjih veličina imaju veće "pogreške procjene", što proizlazi iz centralnog graničnog teorema.

Razmotrimo primjer proizvodnje hrenovki za koju je Američko ministarstvo poljoprivrede postavilo gornju granicu količine ostataka mesa koja se može staviti u hrenovke. "Ostaci od mesa" su jezici, repovi, njuške itd. Hrenovke moraju imati manje od 5% ostataka od mesa. Lokalni proizvođač hrenovki se odlučio za testiranje je li njegova proizvodnja u skladu s pravilima. Dokumentirao je posljednjih 200 serija proizvodnje i dobio da je u prosjeku 1.5% ostataka od mesa u njima. Na razini od $\alpha = 0.05$ odredite je li u skladu s pravilima.

Očito, hipoteze su kako slijedi: $H_0: p = 0.05$; $H_A: p < 0.05$. Korištenje funkcije $NORMSINV(0.05)$ daje kritičnu vrijednost od -1.65 koja definira područje odbacivanja. Korištenjem izraza (6-9) dobivamo testnu veličinu kako slijedi:

$$z = \frac{0.015 - 0.05}{\sqrt{\frac{0.05(1-0.05)}{200}}} = -9.9 \qquad (6\text{-}12)$$

Dobivena testna veličina je vrlo mala te je u ekstremnom dijelu područja odbacivanja. Korištenje $NORMSDIST(-9.9)$ za određivanje p - vrijednosti daje vrlo malu vrijednost. U mom slučaju je dobiven broj "2.08E-23". To znači $(2.08)(10)^{-23}$ što je vrlo mala vrijednost, tako da se jednostavno navodi da je $p < 0.0001$ radi šireg razumijevanja.

6.4 Testiranje hipoteze o razlici aritmetičkih sredina

Postoje slučajevi kada je potrebno utvrditi jesu li aritmetičke sredine dviju različitih populacija jednake. Nadalje, postoje slučajevi kada je potrebno utvrditi postoji li *određena* razlika u aritmetičkim sredinama dviju različitih populacija. To se može učiniti testiranjem hipoteza. Prije prikazivanja odgovarajućeg izračuna, treba sistematizirati relevantne termine.

Vrijednost	Uzorak 1	Uzorak 2
Veličina uzorka	n_1	n_2
Aritmetička sredina populacije	μ_1	μ_2
Aritmetička sredina uzorka	$\bar{x}_1$	$\bar{x}_2$
Uzoračka standardna devijacija	s_1	s_2

Tablica 6.6. Vrijednosti korištene za razlike između aritmetičkih sredina

Kada se testira jednakost aritmetičkih sredina dviju različitih populacija, mogu se postaviti hipoteze:

$$H_0: (\mu_1 - \mu_2) = 0; H_A: (\mu_1 - \mu_2) \neq 0 \qquad (6\text{-}13)$$

Za određenu razliku u aritmetičkim sredinama možemo koristiti bilo koju unaprijed određenu vrijednost d kao razliku u sredinama pri čemu d predstavlja vrijednost te određene razlike:

$$H_0: (\mu_1 - \mu_2) = d; H_A: (\mu_1 - \mu_2) \neq d \tag{6-14}$$

Oba prikazana skupa hipoteza su primjeri dvosmjernih testova. Kod jednosmjernih testova hipoteza H_A je tipa < ili >, no jednosmjerni testovi nisu toliko uobičajeni kao dvosmjerni. Vrijednost α je zadana ili pretpostavljena. Tablica 6.7 prikazuje funkcije $T.INV$ koje se koriste za određivanje kritične vrijednosti područja odbacivanja:

Vrsta H_A	**Funkcija u Excelu**	**Funkcija u Excelu (za starije verzije Excela*)**
≠	$+/-T.INV.2T(\alpha, n-1)$	$+/-TINV(\alpha, n-1)$
<	$T.INV(\alpha, n-1)$	$TINV(2\alpha, n-1)$
>	$T.INV(1-\alpha, n-1)$	$TINV(1-2\alpha, n-1)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 6.7. Funkcije u Excelu za područja odbacivanja

Vrijednosti za broj stupnjeva slobode df su kako slijedi:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_2}\right)^2} \tag{6-15}$$

Opća veličina za testiranje razlika između aritmetičkih sredina glasi:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{6-16}$$

Ako se testira jednakost aritmetičkih sredina, izraz (6-16) se pojednostavnjuje budući da jednakost aritmetičkih sredina populacija povlači $\mu_1 = \mu_2$ i tada je:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{6-17}$$

Ako se testira je li određena razlika u aritmetičkim sredinama jednaka nekom d, gdje je $d = \mu_1 - \mu_2$, tada se izraz (6-16) pojednostavnjuje na sljedeći:

$$t=\frac{(\bar{x}_1-\bar{x}_2)-d}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad (6\text{-}18)$$

P - vrijednosti za ove vrste testova su slične onima prikazanim u tablici 6.3. Ovdje su izrazi modificirani kako bi se prikazalo stupnjeve slobode specifične za ovu vrstu problema:

Vrsta testa	Funkcija u Excelu	Funkcija u Excelu(za starije verzije Excela*)
Jednosmjerni	$T.DIST(-\|t\|;n-1)$ $=T.DIST.RT(\|t\|;n-1)$	$TDIST(\|t\|;n-1;1)$
Dvosmjerni	$T.DIST(-\|t\|;n-1)*2$ $=T.DIST.2T(\|t\|;n-1)$	$TDIST(\|t\|;n-1;2)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 6.8. Funkcije za izračunavanje p-vrijednosti

Riješimo primjer korištenjem skupa podataka “RezultatiIspita”. Ovaj skup podataka uspoređuje rezultate dva različita ispita održana na jednom državnom sveučilištu s vrlo velikim grupama studenata. Ispit 1 je proveden prije ispita 2. Želimo ispitati je li uspjeh na ta dva ispita jednak na razini od $\alpha=0.05$. Slijede hipoteze:

$$H_0:(\mu_1-\mu_2)=0; H_A:(\mu_1-\mu_2)\neq 0 \qquad (6\text{-}19)$$

Deskriptivna statistika dobivena iz skupa podataka daje:

Vrijednost	Ispit 1	Ispit 2
Veličina uzorka	$n_1=500$	$n_2=500$
Aritmetička sredina uzorka	$\bar{x}_1=80.00$	$\bar{x}_2=77.00$
Uzoračka standardna devijacija	$s_1=4.00$	$s_2=6.99$

Tablica 6.9. Zbirni podaci za razlike u uspjesima na ispitu

Skup podataka daje $df=796.23$. Dobivena kritična vrijednost korištenjem Excela iznosi $T.INV.2T(0.05{,}796.23)=+/-1.96$, a testna veličina je $t=8.33$. Budući da je testna veličina ekstremnija od kritične vrijednosti, odbacujemo H_0 i tvrdimo da uspjesi između prvog i drugog ispita nisu jednaki. p - vrijednost dobivena na temelju tablice

6.8 iznosi $TDIST(|8.33|,796.23)$, što je manje od 0.0001.

Još jedan primjer razlike između aritmetičkih sredina predstavlja nadmetanje Alberta i Guillerma, trkača svjetske klase specijaliziranih za trčanje na 400 metara. Alberto tvrdi da je on u prosjeku za jednu sekundu brži na natjecanjima nego što je njegov konkurent Guillermo. Na temelju tih informacija, možemo postaviti sljedeće hipoteze:

$$H_0\text{: } (\mu_G - \mu_A) = 1;\ H_A\text{: } (\mu_G - \mu_A) \neq 1 \tag{6-20}$$[11]

Neka je zadana $\alpha = 0.03$. Skup podataka "Vrijeme400m" daje sljedeće pokazatelje:

Vrijednost	Guillermo	Alberto
Veličina uzorka	$n_G = 13$	$n_A = 15$
Aritmetička sredina uzorka	$\bar{x}_G = 45.8s$	$\bar{x}_A = 44.6s$
Uzoračka standardna devijacija	$s_G = 0.23s$	$s_A = 0.18s$

Tablica 6.10. Zbirni podaci za vrijeme na 400 metara

Stupnjevi slobode dobiveni primjenom izraza (6-15) iznose 22.65. Kritična vrijednost $T.INV.2T(1-0.03{,}22.65)$ iznosi +/-2.32. Testna veličina dobivena primjenom izraza (6-18) iznosi

$$t = \frac{(45.8 - 44.6) - 1}{\sqrt{\frac{0.23^2}{13} + \frac{0.18^2}{15}}} = 2.53 \tag{6-21}$$

Budući da je dobivena testna veličina ekstremnija od kritične vrijednosti, odbacuje se H_0 i tvrdi da je Alberto više od jedne sekunde brži od Guillerma. Kao što je navedeno u strukturi stvarne testne veličine, Guillermo je više od jedne sekunde sporiji od Alberta budući je $p = TDIST(2.32, 22.65, 2) = 0.0189$.

---

[11] Od Guillermovog pretpostavljenog slabijeg vremena se oduzima Albertovo pretpostavljeno bolje vrijeme kako bi se dobila pozitivna vrijednost testne veličine. U biti, tvrdimo da je Guillermo više od jedne sekunde sporiji od Alberta. Također primjetite da je vrijednost $d = 1$, što je pretpostavljena razlika prosječnih vremena.

6.5 Pouzdani intervali i dvosmjerni testovi

Postoje sličnosti između dvosmjernih testova i pouzdanih intervala. To nije slučajnost nego postoji jasna veza. Definicija pouzdanog intervala, za koji se s vjerojatnošću $(1-\alpha)\%$ može tvrditi da se u njemu nalazi stvarna vrijednost populacije, koristi procijenjenu vrijednost očekivanja (kao što je $\bar{x}$). Ovaj postupak je vrlo sličan dvosmjernom testu. U stvari, ako se odredi pouzdani interval koristeći neku vrijednost α i taj interval sadrži pretpostavljeni parametar populacije (kao što je μ), onda se ne može odbaciti H_0 za zadanu vrijednost α. Isto tako, ako interval ne sadrži parametar populacije, odbacuje se H_0 za zadanu vrijednost α.

Koristeći formulu iz potpoglavlja 5.1 i skup podataka "Cheerios", određen je 95% pouzdani interval za koji vrijedi:

$$P(13.99 \le \mu \le 14.08) = 0.95 \tag{6-22}$$

Pretpostavljena aritmetička sredina populacije $\mu = 14$ uključena je u dobiveni interval i H_0 ne treba odbaciti. Kada pretpostavljena aritmetička sredina populacije ne bi bila u intervalu (6-22), H_0 se odbacuje.

Zbog ovog jedinstvenog odnosa između pouzdanog intervala i dvosmjernog testa, pouzdani intervali dodatno dobivaju na vrijednosti.

6.6 Zaključak

Testiranje hipoteza je najteži dio učenja statistike, ali je ujedno i najvažniji dio statistike. Iznošenje svake tvrdnje zahtijeva formalnu provjeru koja bi tu tvrdnju podržala. To tako mora biti u znanstvenoj zajednici koja je važna zajednica u društvu. Zbog toga je vrlo važno razumjeti testiranje hipoteza.

Tema nije nerazumno teška. Najvažniji dio je razumijevanje i vizualizacija H_0 i H_A. Kada se to jednom postigne, ostalo samo dolazi na svoje mjesto. Naravno, potrebno je iskustvo kako bi se odgovarajuće formule i funkcije u Excelu ispravno koristile pri rješavanju problema.

6.7 Zadaci

Za probleme koji uključuju testiranje hipoteza provode se sljedeći koraci:

a) Navedite odgovarajuće hipotezu.
b) Odredite kritičnu vrijednost.
c) Izračunajte testnu veličinu.
d) Navedite svoju odluku o odbacivanju ili neodbacivanju H_0.
e) Navedite p - vrijednost.

Slijede zadaci:

1. Koristite kaznenopravni sustav u SAD-u i navedite primjer pogreške tipa I.
2. Koristite kaznenopravni sustav u SAD-u i navedite primjer pogreške tipa II.
3. Mike Trout iz Anaheim Angelsa tvrdi da njegovo bacanje u prosjeku prelazi 0.320. Koje su odgovarajuće hipoteze u vezi s ovom tvrdnjom?
4. Motorola je izumila koncept šest sigma - tvrdi da se proizvodi manje od 3.4 defekata na milijun proizvedenih jedinica. Koje su odgovarajuće hipoteze u vezi s ovom tvrdnjom?
5. LeBron James tvrdi da kada Cleveland igra protiv Chicaga košarkašku utakmicu, Cleveland pobjeđuje u prosjeku sa šest koševa razlike. Koje su odgovarajuće hipoteze u vezi s ovom tvrdnjom?
6. Koristite skup podataka "PromjerŽice" i testiranjem hipoteza na razini $\alpha = 0.05$ utvrdite postoji li razlika u promjeru žice.
7. Dostavna tvrtka u reklami tvrdi da je za lokalnu dostavu prosječno vrijeme isporuke manje od 6 sati. Slučajni uzorak čine vremena za 12 isporuka na adrese u gradu. Podaci su prikazani u nastavku. Je li to dovoljno dokaza koji bi poduprli reklamu tvrtke na razini 5% značajnosti? Podaci o vremenu isporuke su: 3.03, 6.33, 6.50, 5.22, 3.56, 6.76, 7.98, 4.82, 7.96, 4.54, 5.09, 6.46.
8. "Mikeovi bicikli" iz Columbusa u Georgiji prodaje jako puno cestovnih bicikala. Jedna od stvari koja zabrinjava Mikea

Reynoldsa (vlasnika trgovine) je da mu se kupci vraćaju s problemima s gumama. Ako u gumama ima previše zraka, one mogu eksplodirati što može biti opasno. Dodatno, unutarnje gume treba zamijeniti što je teško. Ako se premalo zraka nalazi u gumama, loše su performanse bicikla, a češće su zamjene guma. Mike se pita stavljaju li njegovi mehaničari ispravnu količinu zraka u gume. Ispravna količina zraka u gumama je 115 psi (otprilike 7.9 bara koji se očitavaju na mjeraču pri benzinskim crpkama). Mike je nasumce u uzorak izabrao 100 napumpanih guma i dobio da je srednji tlak u gumama 113.58 psi uz standardnu devijaciju od 8.1 psi. Što možemo zaključiti o tlaku u gumama kod Mikea na razini $\alpha = 0.08$?

9. Poduzeće ThreeGuys Chemical Corporation iz Cincinnatija ima novi inhibitor korozije. Proizvod (nazvan Moisture Basher) se prodaje kao dodatak za strojne dijelove koji se koriste u okruženjima visoke vlažnosti. U poduzeću tvrde da kada se koristi Moisture Basher, strojni dijelovi mogu trajati 90 dana prije nego što ih je potrebno zamijeniti zbog hrđanja. Udruga potrošača sa sjedištem u Bostonu je odlučila testirati ovu tvrdnju poduzeća. Oni vjeruju da je antikorozivno djelovanje manje od 90 dana. Tretirali su 20 strojnih dijelova s Moisture Basherom i mjerili trajanje strojnih dijelova tretiranih s tim proizvodom. Dobili su da je prosječno trajanje 81 dan uz uzoračku standardnu devijaciju od 2.5 dana. Što govore ove informacije dobivene iz uzorka o tvrdnji poduzeća ThreeGuys?

10. Osobe koje zatraže povrat poreza na dohodak prije 31. ožujka prosječno dobiju 1056 dolara. Zamislite populaciju poreznih obveznika koji su poslali svoje zahtjeve u posljednjih 5 dana poreznog razdoblja (10.-15. travnja). Prikupljen je uzorak od 400 obveznika, a utvrđeno je da je njihov prosječni povrat 910 dolara sa uzoračkom standardnom devijacijom od 1600 dolara. Dobivaju li obveznici koji kasno prijave porez povrat manji od 1056 dolara? Koristite $\alpha = 0.05$.

11. Prije deset godina tvrtka A.C. Nielson je tvrdila da u prosjeku američka kućanstva gledaju televiziju 6.70 sati dnevno. Neovisna grupa istraživača tržišta smatra da se sada televizija gleda više. Da bi testirali ovu tvrdnju, anketirali su 200 kućanstava te su dobili da se u prosjeku sada televizija gleda 7.25 sati dnevno uz uzoračku standardnu devijaciju od 2.5 sata. Je li se vrijeme koje se provede gledajući televiziju tijekom posljednjih deset godina povećalo? Koristite $\alpha = 0.02$.
12. Povijesno gledano, večernji telefonski pozivi na duge udaljenosti iz određenog grada traju prosječno 15.2 minute po pozivu. U slučajnom uzorku od 35 poziva prosječno trajanje razgovora je bilo 14.3 minute po pozivu s uzoračkom standardnom devijacijom od 5 minuta. Koristite ovaj uzorak da biste testirali bilo kakve promjene u prosječnom trajanju međugradskih telefonskih poziva. Koristite $\alpha = 0.02$.
13. Proizvođač najlonskog konopa tvrdi da njihovo uže od 10 mm može izdržati opterećenje od više od 250 funti (oko 113 kilograma) prije nego što pukne. Da bi potvrdili ovu tvrdnju, nekoliko užadi su podvrgnuli opterećenju te su bilježili sile pri kojoj je užad pucala. Te sile su izražene u funtama, a to su: 280, 259, 255, 238, 245, 265, 250, 259, 241 i 260. Koristite $\alpha = 0.06$. Jesu li tvrdnje proizvođača vjerodostojne?
14. Veliki proizvođač elektronskih komponenti tvrdi da se njihov konektor od 0.5 V dosljedno proizvodi po stopi od manje od 10 neispravnih proizvoda na milijun proizvedenih komada. Uzorkovanje je pokazalo da je u posljednjih nekoliko serija bilo oštećenja na milijun proizvedenih komada kako slijedi: 3, 8, 7, 13, 8, 4, 11, 10, 9, 6, 10 i 8. Je li točna tvrdnja proizvođača da je manje od 10 neispravnih proizvoda na milijun proizvedenih komada? Koristite $\alpha = 0.04$.
15. Koristite skup podataka “HranaZaDijabetičare” i postavite odgovarajuće hipoteze.
16. Koristite skup podataka “ProfesoriFinancija” i postavite odgovarajuće hipoteze.

17. Koristite skup podataka “NFLLinijaši” i postavite odgovarajuće hipoteze.
18. Koristite skup podataka “OtkucajiSrca” i postavite odgovarajuće hipoteze.
19. Koristite podatke iz zadatka 5-4. Podržava li većina stanovnika Oregona korištenje marihuane u medicinske svrhe?
20. Koristite podatke iz zadatka 5-3. Smatra li većina stanovnika Sjeverne Karoline Sveučilište Wake Forest najboljim fakultetom u državi?
21. Luca Brazzi želi da ekstra velike pizze u njegovoj pizzeriji imaju 15 komada kobasica po svakoj kriški. Premalo komada vodi prema nesretnim kupcima, a previše smanjuje profitabilnost poslovanja. Luca je slučajno uzorkovao nekoliko extra velikih pizza i bilježio je broj komada kobasica na svakoj kriški pizze. Ovi podaci su uključeni u skup podataka “KomadiKobasice”. Stavlja li pizzeria odgovarajući broj komada kobasica na pizze? Koristite $\alpha = 0.03$.
22. Proizvođač lijekova protiv alergije tvrdi da su proizveli pilulu koja nudi više od osam sati zaštite od alergija. Budući da se tvrdnja odnosi na farmaceutsku industriju, ona zahtijeva strogi nadzor i mora se koristiti vrijednost $\alpha = 0.01$. Nekoliko teških alergičara je testirano na korištenje ovog lijeka, a sati nepojavljivanja alergije su dani u skupu podataka “NepojavljivanjeAlergije”. Vaš zadatak je da potvrdite ili opovrgnete tvrdnje farmaceutske tvrtke.
23. Pizzeria Luce Brazzija nudi pizze i sendviče svojim klijentima s Brooklyna. Luca obećava kupcima da je vrijeme dostave manje od trideset minuta. Da bi bio siguran da je rok dostave stvarno manji od trideset minuta, Luca je nasumce prikupio vremena dostave tijekom posljednjih nekoliko mjeseci rada. Dobiveni podaci su dani u skupu podataka “VrijemeDostavePizze”. Je li vrijeme dostave manje od 30 minuta? Koristite $\alpha = 0.06$.
24. Dekan prestižnog regionalnog fakulteta zabrinut je misleći da studenti prve godine studija provode premalo vremena učeći.

Smatra se da studenti prve godine uče manje od šest sati dnevno. Da bi se ispitao taj problem, prikupljen je mali uzorak studenata te su zamoljeni da napišu broj sati koje provedu učeći dnevno. Broj sati je dostupan putem skupa podataka "VrijemeZaUčenje". Profesor statistike je predložio da dekan koristi vrijednost $\alpha = 0.10$. Vaš zadatak je da testirate vjerodostojnost dekanove tvrdnje.

25. Američka agencija za nacionalne parkove tvrdi da Old Faithful Geyser u Nacionalnom parku Yellowstone prosječno eruptira svakih 78 minuta. Da bi potvrdio ovu tvrdnju, Američki geološki zavod je mjerio vrijeme između erupcija u nasumičnim intervalima. Mjerili su u minutama vrijeme koje je proteklo od početka jedne erupcije do početka sljedeće ("VrijemeIzmeđuErupcija"). Izmjerene vrijednosti su dostupne u priloženom skupu podataka. Koristite vrijednost $\alpha = 0.05$. Ispitajte tvrdnju agencije.
26. Prije približno mjesec dana ljutiti kupac se žalio Luci Brazziju da je vrijeme čekanja za stol bilo predugo. Kupac je rekao Luci da je vrijeme čekanja bilo više od 20 minuta tijekom večernje gužve (između 18 i 20 sati). Luca je htio istražiti tvrdnju ljutitog kupca pa je prikupio slučajne podatke o vremenu čekanja za vrijeme večernje gužve. Dobiveni podaci o vremenima čekanja su uključeni u skup podataka "VrijemeČekanjaURestoranu". Je li vrijeme čekanja duže od 20 minuta? Koristite $\alpha = 0.005$.
27. Tvrtka iz Omahe u Nebraski proizvodi kobasice. Tvrtka u mješavinama za kobasice koristi kombinaciju vrhunskog mesa i ostataka od mesa. Ostaci od mesa nisu toliko privlačni kao što im ime možda sugerira – to su sastojci od najmanje poželjnih dijelova - repa, tripica itd. Američko ministarstvo poljoprivrede (USDA) nalaže da mješavine kobasica moraju imati ostataka od mesa manje od 15% svoje mase. Prikupljeni su podaci o posljednjih nekoliko mješavina kobasica. Skup podataka je dan pod imenom "OstaciMesa". Korištenjem $\alpha = 0.03$ utvrdite je li

meso u skladu s pravilima koje nalaže Američko ministarstvo poljoprivrede.

28. Dva predsjednička kandidata Hillary i Bernie se natječu za demokratsku predsjedničku nominaciju u Wisconsin Primaryju. Nasumično je odabrano nekoliko birača (pri njihovom izlasku s biračkog mjesta) i upitano za koga su glasali. Rezultati s ovog biračkog mjesta su prikupljeni i dostupni su u skupu podataka "WisconsinPrimary". Vrijednost 0 znači da nisu glasovali za određenog kandidata, dok vrijednost 1 označava da su glasovali za određenog kandidata. Treba napomenuti da su svi ispitanici u uzorku glasovali ili za Hillary ili za Bernieja. Možete li, korištenjem $\alpha = 0.02$, projicirati pobjednika u Wisconsin Primaryju? Ako je tako, za koga procjenjujete da će biti pobjednik?
29. Farmaceutska tvrtka proizvodi antihistaminski lijek protiv osnovnih simptoma alergije koji se izdaje bez recepta. Aktivni sastojak lijeka je klorovodična kiselina (HCl). HCl bi trebao činiti 2.5% lijeka. Ako je premalo HCl-a u lijeku, učinak lijeka će biti smanjen. Ako je previše HCl-a u lijeku, nuspojave lijeka su vjerojatnije. S obzirom na ove činjenice, važno je da lijek sadrži točno 2.5% HCl-a. Proučavano je nekoliko nedavnih serija proizvodnje lijeka i bilježene su razine HCl-a. Ti podaci su dostupni u skupu podataka "LijekProtivAlergije". Treba naglasiti da su sve serije proizvodnje korištene u skupu podataka iste veličine. Korištenjem $\alpha = 0.05$ ispitajte može li farmaceutska tvrtka tvrditi da lijek sadrži ispravnu količinu HCl-a.
30. Skup podataka "MjesečneZaključneCijene" sadrži zaključne cijene za Microsoft i IBM na mjesečnoj razini. Provedite odgovarajući test kako biste utvrdili je li mjesečna postotna promjena različita između tih dviju dionica. Koristite $\alpha = 0.05$.
31. SuperPrep pomaže srednjoškolcima da se pripreme za matematički dio testa na ispitima znanja i sposobnosti. Oni tvrde da njihovi učenici u prosjeku imaju više od 100 bodova

više u matematičkom dijelu testa na ispitima znanja i sposobnosti u odnosu na učenike koji pohađaju BasicPrep. Korištenjem $\alpha = 0.05$ i skupa podataka “IspitiPrep” utvrdite je li njihova tvrdnja točna.

7. Snaga testa

U šestom poglavlju naglašena je činjenica postojanja rizika od donošenja pogrešnog zaključka o populaciji temeljem podataka iz uzorka. To je logično i neizbježno budući je skup informacija ograničen. Definiranjem razine značajnosti α moguće je ograničiti vjerojatnost činjenja pogreške.

U ovom poglavlju dodatno se istražuje α i njena veza s β, koja predstavlja vjerojatnost pogreške tipa II u testiranju hipoteza.

7.1 Pogreška tipa I i α

Već je navedeno da postoje dvije vrste pogrešaka kod testiranja hipoteza. Tablica 7.1 ponavlja tablicu 6.1:

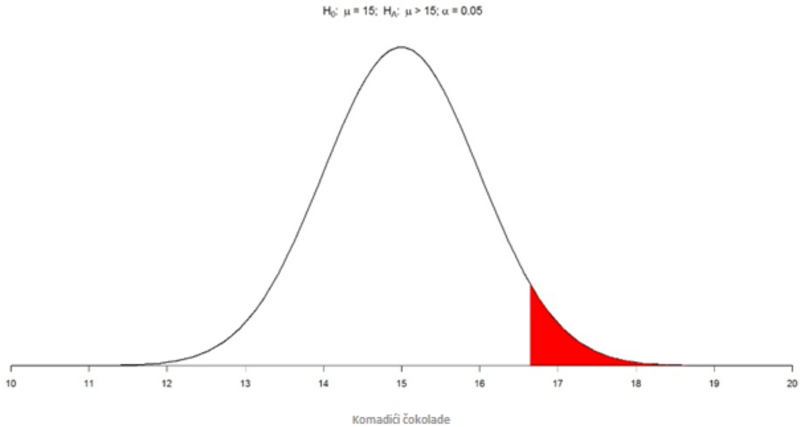
	Odbacivanje H_0	**Neodbacivanje H_0**
H_0 istinita	Pogreška tipa I	Ispravna odluka
H_0 lažna	Ispravna odluka	Pogreška tipa II

Tablica 7.1. Rezultati donošenja odluka oko H_0

Odbacivanje, u biti istinite H_0, je pogreška tipa I i njena vjerojatnost je α. Neka tvornica proizvodi kekse s komadićima čokolade tako da u svaki keks stavlja više od 15 komadića čokolade. Moguće je konstruirati odgovarajuće hipoteze:

$$H_0: \mu = 15; H_A: \mu > 15 \tag{7-1}$$

i provesti test uz korištenje razine signifikantnosti α. Problem je ilustriran grafikonom 7.1 tako da ga možemo jasno vizualizirati.



Grafikon 7.1. Vjerojatnost pogreške tipa I

Distribucija vjerojatnosti je centrirana oko vrijednosti $\mu = 15$. Budući da je $\alpha = 0.05$, 5% površine ispod krivulje je u desnom repu distribucije, što predstavlja područje odbacivanja. Za velike uzorke i konkretnu vrijednost α, područje odbacivanja je 1.645 standardnih pogrešaka iznad aritmetičke sredine. Uz pretpostavku da je standardna pogreška 1 područje odbacivanja nulte hipoteze je 16.645 komadića čokolada i više. Navedeno je detaljno prikazano grafikonom 7.1.

Sada pretpostavimo da je H_0 zapravo istinita i da sigurno postoji 15 komadića čokolade u svakom keksu. S obzirom na to da je 5% površine ispod krivulje u području odbacivanja, postoji šansa od 5% da pogrešno odbacimo istinitu H_0. To znači da imamo 5% šanse učiniti pogrešku tipa I.

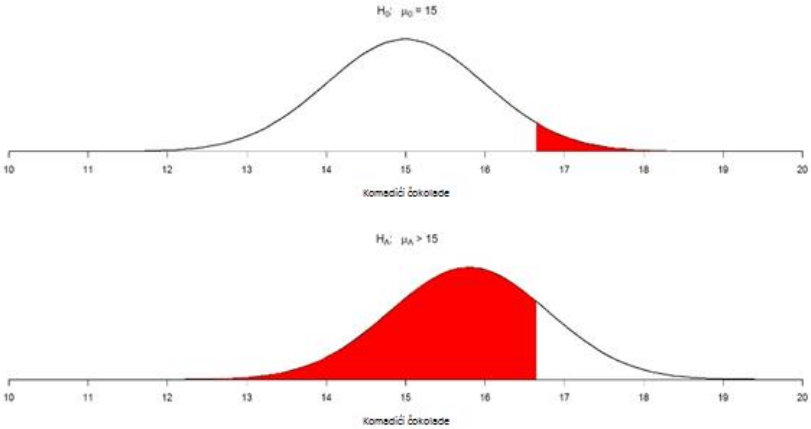
Razumijevanje pogreške tipa II i njene vjerojatnosti β je složenije od gore navedenog.

7.2 Pogreška tipa II i β

Pogreška tipa II učinjena je kada se ne odbaci H_0 iako nije istinita. Da bi se ovo razumjelo, potrebno je situaciju vizualizirati pomoću alternativne hipoteze H_A. Kada je H_0 istinita, H_A je lažna, a kada je H_0 lažna, H_A je istinita.

7.2.1 Pogreška tipa II za testove tipa “>”

U primjeru s keksima i komadićima čokolade, neka je H_0 lažna, a time H_A istinita ili ispravna. Koristeći istu vrijednost $\alpha = 0.05$ slijedi scenarij u kojem su grafikonom 7.2 prikazane pretpostavljene distribucije populacija i područja odbacivanja, odnosno neodbacivanja postavljenih hipoteza H_0 i H_A u (7-1). Sredina donje distribucije je neka vrijednost μ_A za koju se pretpostavlja da je u skladu s H_A.



Grafikon 7.2. Vjerojatnost pogreške tipa I i β za hipotezu H_A tipa “>”

Ovdje je H_0 jednako formulirana kao i ranije. Kada se raspravlja o pogreškama tipa I i tipa II, koriste se μ_0 i μ_A za pretpostavljene aritmetičke sredine populacija za H_0 i H_A, redom. Za gornji dio grafikona 7.2 vrijednost α pokazuje područje odbacivanja H_0 - zasjenjeno područje na desnoj strani distribucije. Kritična vrijednost je 16.645 komadića čokolade kao što je objašnjeno gore.

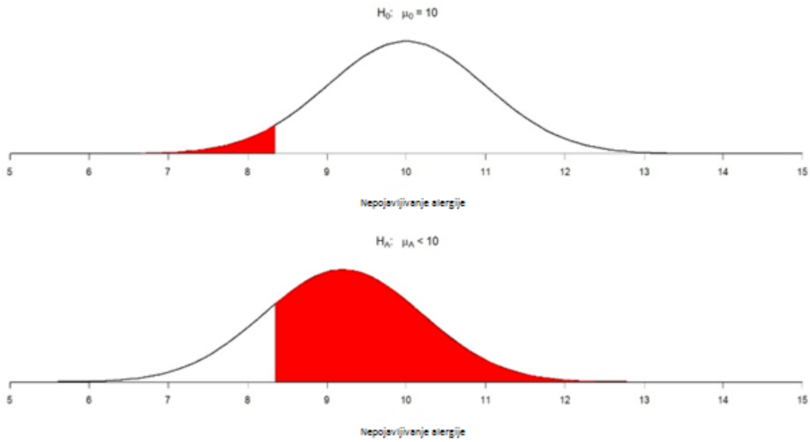
U donjem dijelu grafikona 7.2 površina ispod krivulje s lijeve strane kritične vrijednosti od 16.645 predstavlja vrijednost β. To znači da je β površina ispod distribucije populacije koja odgovara H_A i predstavlja vjerojatnost neodbacivanja H_0, upravo to je definicija β kao vjerojatnosti neodbacivanja lažne H_0.

7.2.2 Pogreška tipa II za testove tipa “<”

Drugi primjer testira pruža li lijek protiv alergija manje od 10 sati olakšanja. Odgovarajuće hipoteze za ovaj test su:

$$H_0 \colon \mu = 10; \; H_A \colon \mu < 10 \qquad (7\text{-}2)$$

Slijedi grafički prikaz testa:



Grafikon 7.3. Pogreška tipa I i β za hipotezu H_A tipa “<”

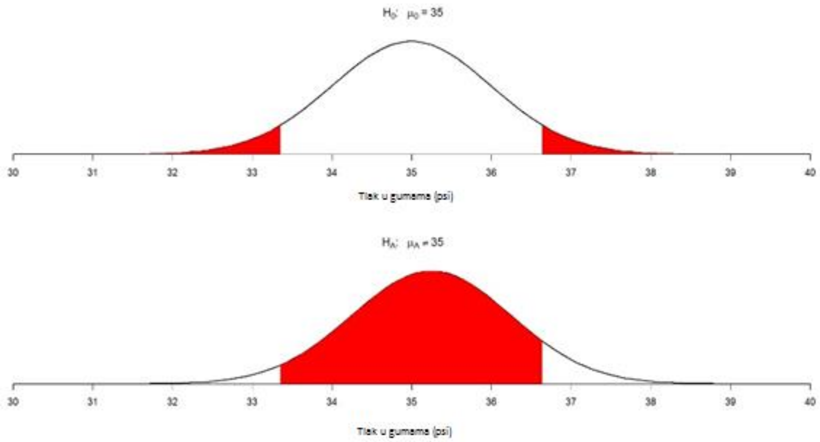
Neka je ponovno $\alpha = 0.05$ i neka je uzorak dovoljno velik i neka je standardna pogreška jednaka 1. Gornji dio grafikona 7.3 pokazuje područje odbacivanja s lijeve strane od H_0 određen gornjom granicom aritmetičke sredine uzorka od 8.355=(10-1.645). Donji dio grafikona 7.3 pokazuje vjerojatnost prihvaćanja neistinite hipoteze β ako hipoteza H_A ima takvu distribuciju u populaciji koja bi odbacila hipotezu H_A, budući područje prihvaćanja H_A odgovara površini desne strane ispod krivulje počevši od 8.355.

7.2.3 Pogreška tipa II za testove tipa “≠”

Kod testa tipa "≠" područje odbacivanja nalazi se s obje strane sredine pretpostavljene u H_0. To znači da se pri određivanju β ispituje H_A i vizualizira β kao površina između dvije granične vrijednosti povezane s H_0. Grafikon 7.4 prikazuje primjer gdje se ispitalo je li tlak u automobilskim gumama jednak 35 funti po kvadratnom inču (2.4 bara na benzinskoj crpki). Tražene hipoteze su kako slijedi:

$$H_0\colon \mu = 35;\ H_A\colon \mu \neq 35 \tag{7-3}$$

Grafički prikaz tih hipoteza i odgovarajućih vrijednosti α i β je sljedeći:



Grafikon 7.4. Pogreška tipa I i β za hipotezu H_A tipa "$\neq$"

Vjerojatnost β odgovara površini područja između dviju kritičnih vrijednosti za H_0 određeno s α.

Zasigurno, razumijevanje β je složenije nego razumijevanje α. Tajna u razumijevanju β je u vizualiziranju i razumijevanju da se vjerojatnost β računa kao površina ispod krivulje koja pokazuje razdiobu vjerojatnosti u populaciji obuhvaćenoj hipotezom H_A koja bi bila u području odbacivanja H_A i prihvaćanja H_0. To su generalizirani detalji primjera s ciljem fokusiranja na vizualizaciju vjerojatnosti β preko neke od razdioba obuhvaćene hipotezom H_A.

7.3 Izračunavanje β

Nakon vizualiziranja β, slijedi njen izračun. Prvo treba odrediti vrijednost aritmetičke sredine povezane s H_A koja će biti označena s μ_A. Vrijednost μ_A se može pretpostaviti ili se za μ_A uzima aritmetička sredina uzorka: $\mu_A = \bar{x}$. Korištenje $\bar{x}$ u tu svrhu ima smisla ako nije identična srednjoj vrijednosti μ_0 povezanoj s H_0. Osim toga, korištenje $\bar{x}$ je rezultat truda statističara oko uzorkovanja što predstavlja najbolju moguću procjenu aritmetičke sredine populacije.

Prije samog izračuna vrijednosti β, potrebno je prethodno izračunati nekoliko vrijednosti. Prvo treba izračunati kritične vrijednosti za odgovarajući test hipoteza. Kritične vrijednosti izražene u mjernim jedinicama varijable su jednake $c_1 = \mu_0 - z_{\alpha/2}\sigma_{\bar{x}}$ i $c_2 = \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}$, gdje je $\sigma_{\bar{x}}$ standardna pogreška procjene aritmetičke sredine.

Tada su izrazi za računanje β za dvosmjeran test o pretpostavljenoj aritmetičkoj sredini populacije dani sljedećim relacijama

$$\beta = P\left(\frac{c_1-\mu_A}{\sigma_{\bar{x}}} < Z < 0\right) - P\left(\frac{c_2-\mu_A}{\sigma_{\bar{x}}} < Z < 0\right) \text{ za } \mu_A > c_2$$

$$\beta = P\left(0 < Z < \frac{c_2-\mu_A}{\sigma_{\bar{x}}}\right) - P\left(0 < Z < \frac{c_1-\mu_A}{\sigma_{\bar{x}}}\right) \text{ za } \mu_A < c_1$$

$$\beta = P\left(\frac{c_1-\mu_A}{\sigma_{\bar{x}}} < Z < 0\right) + P\left(0 < Z < \frac{c_2-\mu_A}{\sigma_{\bar{x}}}\right) \text{ za } c_1 < \mu_A < c_2;$$

za jednosmjeran test na gornju granicu

$$\beta = 0{,}5 - P\left(\frac{c_2-\mu_A}{\sigma_{\bar{x}}} < Z < 0\right) \text{ za } \mu_A > c_2$$

$$\beta = 0{,}5 + P\left(0 < Z < \frac{c_2-\mu_A}{\sigma_{\bar{x}}}\right) \text{ za } \mu_A < c_2;$$

a za jednosmjeran test na donju granicu

$$\beta = 0{,}5 - P\left(0 < Z < \frac{c_1-\mu_A}{\sigma_{\bar{x}}}\right) \text{ za } \mu_A < c_1$$

$$\beta = 0{,}5 + P\left(\frac{c_1-\mu_A}{\sigma_{\bar{x}}} < Z < 0\right) \text{ za } \mu_A > c_1,$$

pri čemu je $Z \sim N(0{,}1)$ standardizirana normalna slučajna varijabla.

Kritična vrijednost (inverzna vrijednost normalne distribucije) je funkcija od α te se može izračunati pomoću Excel formula:

Vrsta H_A	Funkcija u Excelu	Funkcija u Excelu (za starije verzije Excela*)
$\neq$	$+/-NORM.S.INV(1-\alpha/2)$	$+/-\ NORMSINV(1-\alpha/2)$
<	$NORM.S.INV(\alpha)$	$NORMSINV(\alpha)$
>	$NORM.S.INV(1-\alpha)$	$NORMSINV(1-\alpha)$

***u novijim verzijama dostupno u Compatibility skupu formula**

Tablica 7.2. Funkcije u Excelu za kritične vrijednosti

Primjeri u potpoglavlju 7.2 su vrlo općeniti. Namjera nije bila matematička, nego konceptualna, kako bi se vjerojatnost pogreške tipa II mogla vizualizirati. Ovi konceptualni problemi su numerički riješeni pomoću posebnih skupova podataka koji se odnose na gore navedene primjere. Excelove datoteke podataka i deskriptivna statistika zajedno s izračunatim vrijednostima β su prikazani u nastavku. U svim slučajevima, vrijednost μ_A je postavljena da bude jednaka aritmetičkoj sredini uzorka $\bar{x}$.

Vrsta H_A	Excel datoteka	μ_0	μ_A	s	N	α	β
$\neq$	TlakUGumama	35	35.27	2.44	67	0.05	0.7699
<	NepojavljivanjeAlergije	10	9.18	2.19	62	0.05	0.1020
>	KomadićiČokolade	15	15.81	2.03	100	0.10	0.0116

Tablica 7.3. Vrijednosti β za gornje primjere

7.4 Snaga testa i β

Pouzdanost testa jednaka je $(1-\alpha)$. Snaga testa je $(1-\beta)$. U literaturi se mnogi autori neizravno pozivaju na β iskazujući snagu testa tvrdnjom: što je vrijednost β manja, to je veća snaga testa.

Postoje dva glavna faktora koji utječu na β. Prvo, što je veća razlika između vrijednosti μ_0 i μ_A, to će biti manja vrijednost β. Slično tome, što je uzorak veći, to će biti manja vrijednost β. Ponovno, što je manja vrijednost β, to je snaga testa veća.

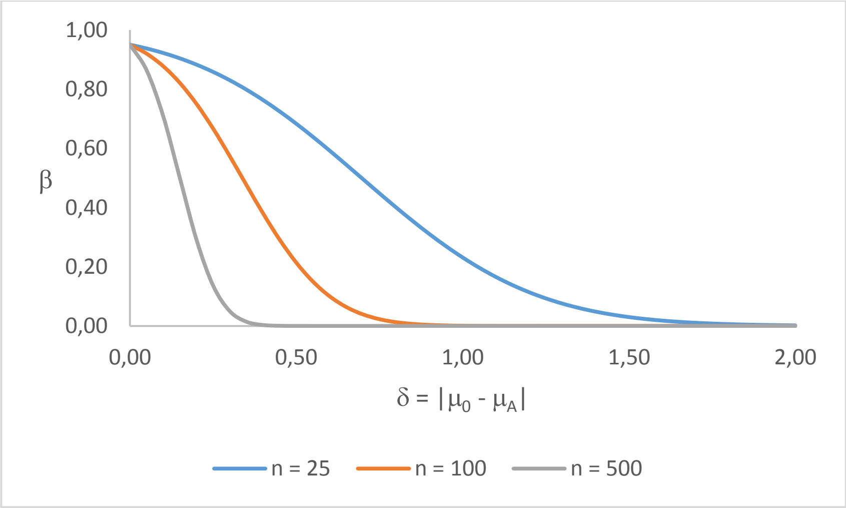
Razlika između vrijednosti μ_0 i μ_A računa se na sljedeći način:

$$\delta = |\mu_0 - \mu_A| \tag{7-4}$$

Zanimljive su promjene vjerojatnosti pogreške tipa II ovisno o promjenama aritmetičkih sredina uzoraka, odnosno promjenama μ_A. Grafikon koji na x - osi prikazuje alternativne vrijednosti aritmetičkih sredina, a na y - osi pripadne vjerojatnosti pogreške prve vrste zove se *krivulja operativne karakteristike* ili kraće OC krivulja (engl. *Operating characteristic curve – OC curve*). U ovoj knjizi je prikazana modificirana OC krivulja u kojoj se umjesto alternativnih vrijednosti aritmetičkih sredina promatraju razlike između pretpostavljene sredine u nultoj hipotezi i alternativnih sredina.

Za ilustraciju OC krivulje odabran je primjer "KomadićiČokolade" gdje se dopušta da vrijednost μ_A varira između 15 i 17 komadića čokolade što rezultira s vrijednosti δ u rasponu od 0 do 2, budući da je vrijednost μ_0 fiksna i iznosi 15 komadića čokolade. Modificirana OC krivulja je konstruirana za tri različite veličine uzorka: n =25, 100 i 500. Krivulja je prikazana grafikonom 7.5.

Jasno se može uočiti da veće veličine uzorka i veće razlike između μ_0 i μ_A smanjuju vrijednost β, a time povećavaju snagu testa. Veće veličine uzoraka i veće razlike između μ_0 i μ_A rezultirat će strmijom krivuljom što čini test snažnijim.



Grafikon 7.5. OC krivulja za primjer "KomadićiČokolade"

7.5 Odabir veličine uzorka

Kod provođenja testiranja hipoteza unaprijed se određuje vrijednost α - vjerojatnost pogreške tipa I koju nazivamo razinom značajnosti, a koja posredno određuje razinu pouzdanosti koju se želi primijeniti u testu. Može se unaprijed odrediti i vjerojatnost pogreške tipa II koja određuje snagu testa. Ako su unaprijed zadani α, β, i δ, tada je moguće izračunati veličinu uzorka n koja je potrebna za dobivanje unaprijed zadanih vjerojatnosti pogrešaka. Za jednosmjerni test veličina uzorka se određuje prema:

$$n = \left[\left(z_\alpha + z_\beta\right)^2 s^2\right] / \delta^2. \tag{7-5}$$

Za dvosmjerni test veličina uzorka se određuje prema:

$$n = \left[\left(z_{\alpha/2} + z_\beta\right)^2 s^2\right] / \delta^2. \tag{7-6}$$

Za ove dvije jednadžbe korištene su z - vrijednosti pri izračunu kritičnih vrijednosti povezanih s α i β. Već je naglašeno kako se pri velikim uzorcima umjesto t - vrijednosti često koriste z - vrijednosti. Te vrijednosti z analogne su kritičnim vrijednostima dobivenim pomoću t.

Vrijednosti z dobivaju se analogno kao kod kod testiranja hipoteza o proporcijama. Tablicom 7.4 prikazane su funkcije u Excelu koje se koriste za određivanje tih vrijednosti.

Vrsta H_A	z_α ili $z_{\alpha/2}$	z_β
≠	$NORMSINV(1-\alpha/2)$	$NORMSINV(1-\beta)$
<	$NORMSINV(1-\alpha)$	$NORMSINV(1-\beta)$
>	$NORMSINV(1-\alpha)$	$NORMSINV(1-\beta)$

*u novijim verzijama Excela koristi se NORMSINV(__;1)

Tablica 7.4. Funkcije u Excelu za z-vrijednosti

Još jednom razmotrimo skup podataka "TlakUGumama" opisan u tablici 7.3. Za ovaj primjer ćemo unaprijed odrediti $\alpha = 0.05$, $\beta = 0.03$, $\delta = 0.25$ i iz uzorka već znamo da je $s = 2.44$. Budući da se radi o dvosmjernom testu, koristit ćemo izraz (7-6) i funkcije iz tablice 7.4. Dobivamo da je $z_{\alpha/2} = 1.96$ i $z_\beta = 1.88$. Korištenjem izraza (7-6) dobivamo:

$$= [(1.96 + 1.88)^2 2.44^2]/0.25^2 = 1404.93 \tag{7-7}$$

Da bi pogreška tipa I iznosila 0.05 i pogreška tipa II iznosila 0.03 u ovom slučaju, potrebna veličina uzorka je 1405 jedinica.

7.6 Zaključci

Vrlo malo predavanja i/ili knjiga iz uvodne statistike daju mnogo prostora temi o pogrešci tipa II. Kao takva, ova tema se često tretira fakultativno. Ipak, ona daje više informacija o testiranju hipoteza u cjelini, a također ilustrira odnos između mogućih pogrešaka (tipa I i tipa II) i veličine uzorka.

7.7 Zadaci

Za sljedeće zadatke pretpostavite da je vrijednost razlike između pretpostavljene i alternativne sredine δ jednaka $|\mu_0 - \bar{x}|$.

1. Kolika je vrijednost β u testiranju hipoteze u zadatku 6-7?

2. Koju veličinu uzorka treba koristiti u zadatku 6-7 ako želite da β

bude 0.025?

3. Kolika je vrijednost β u testiranju hipoteze u zadatku 6-8?

4. Koju veličinu uzorka treba koristiti u zadatku 6-8 ako želite da β bude 0.05?

5. Koja je vrijednost β kod testiranja hipoteza u zadatku 6-9?

6. Koju veličinu uzorka treba koristiti u zadatku 6-9 ako želite da β bude 0.01?

7. Kolika je vrijednost β kod testiranja hipoteza u zadatku 6-11?

8. Koju veličinu uzorka treba koristiti u zadatku 6-11 ako želite da β bude 0.04?

9. Koliko iznosi β kod testiranja hipoteza u zadatku 6-12?

10. Koju veličinu uzorka treba koristiti u zadatku 6-12 ako želite da β bude 0.02?

11. Kolika je vrijednost β kod testiranja hipoteza u zadatku 6-13?

12. Koju veličinu uzorka treba koristiti u zadatku 6-13 ako želite da β bude 0.06?

13. Kolika je vrijednost β kod testiranja hipoteze u zadatku 6-14?

14. Koju veličinu uzorka treba koristiti u zadatku 6-14 ako želite da β bude 0.05?

8. Jednosmjerna analiza varijance

Često je važno utvrditi jesu li aritmetičke sredine iz različitih populacija slične. Navedeno je opisano u potpoglavlju 6.4. no nažalost, alati prikazani u potpoglavlju 6.4 nas ograničavaju u istraživanju sličnosti u aritmetičkim sredinama na slučaj kad su uključene samo dvije populacije. Ovo poglavlje daje alate za otkrivanje sličnosti u aritmetičkim sredinama dviju ili više populacija.

Pri određivanju razlikuju li se populacije u pogledu srednjih vrijednosti, potrebno je definirati nekoliko novih termina. **Faktor** je kategorička varijabla, također poznata kao **tretman**. Može se shvatiti kao vrsta podataka. U hipotetskom primjeru utvrđivanja je li broj neispravnih proizvoda na milijun proizvedenih komada isti u svakoj smjeni, smjena se može promatrati kao faktor koji ima tri jedinstvene **kategorije**: prvu, drugu i treću smjenu. Veza izmjerenih vrijednosti s faktorom u matematičkom obliku glasi:

$$f(smjena) = \frac{Broj\ defektnih\ jedinica}{Milijun\ proizvedenih\ jedinica}$$

Moguće je testirati je li broj defeknih proizvoda na milijun proizvedenih jedinica povezan sa smjenama. Ako su defekti povezani sa smjenama, zaključuje se da smjene imaju utjecaj na broj defekata. Inače, stoji zaključak da smjena nema utjecaja na broj defektnih proizvoda.

8.1 Varijabilnost i F - distribucija

Da bi se utvrdilo jesu li aritmetičke sredine jednake za više populacija (neka postoji a populacija), potrebno je utvrditi odgovarajuće hipoteze. One su kako slijedi:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_a$$
$$H_A: \text{ postoje sredine koje nisu jednake} \tag{8-1}$$

Navedeni alat naziva se Analiza varijance (skraćeno "ANOVA") jer se u suštini analize zapravo uspoređuju varijabilnosti. F - distribucija to čini pomoću omjera dvije mjere varijabilnosti, odnosno disperzije. Prije prezentiranja ove distribucije potrebno je definirati nekoliko pojmova.

Pojam	Definicija
A	Broj grupa ili populacija
n_i	Veličina uzorka grupe i $(i = 1,2, \dots , a)$
$\bar{x}_i$	Aritmetička sredina uzorka grupe i
s_i	Uzoračka standardna devijacija grupe i
x_{it}	i-to opažanje grupe t
$\bar{\bar{x}}$	Sveukupna uzoračka sredina

Tablica 8.1. Pojmovi potrebni za ANOVA

Vrijednosti u tablici 8.1 se koriste za određivanje F statistike koja je zapravo omjer dviju mjera varijabilnosti. U svom najopćenitijem obliku F statistika izgleda ovako:

$$F = \frac{Varijabilnost\ između\ grupa}{Varijabilnost\ unutar\ grupe} \tag{8-2}$$

Vrijednost F statistike u odnos stavlja standardnu devijaciju aritemtičkih sredina grupa i sumu standardnih devijacija svih grupa. Što je navedeni omjer veći, to nas više upućuje na zaključak o postojanju statistički značajne razlike u aritmetičkim sredinama grupa i stoga na odbacivanje H_0. Inače, se ne može tvrditi da postoji značajna razlika između grupa (ne odbacujemo H_0). Formula F - statistike glasi:

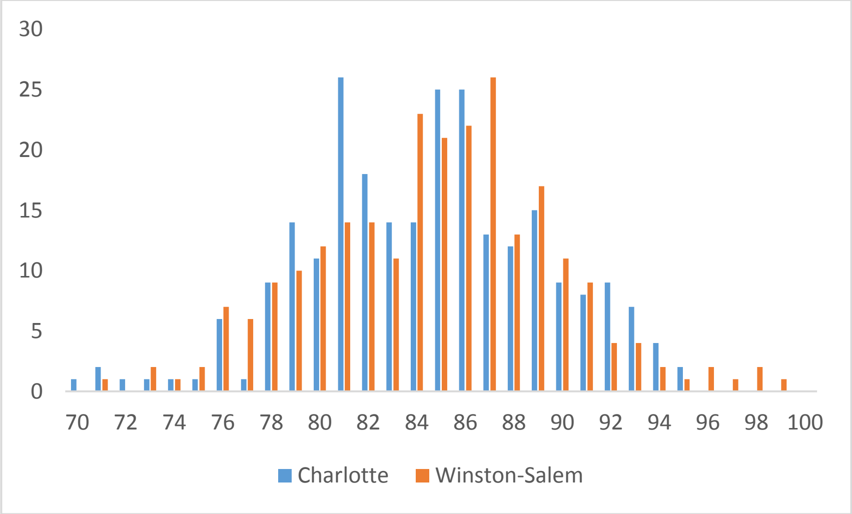
$$F = \frac{\sum_{i=1}^{a} n_i(\bar{x}_i - \bar{\bar{x}})^2/(a-1)}{\sum_{i=1}^{a}\sum_{t=1}^{n_i}(x_{it} - \bar{x}_i)^2/\sum_{i=1}^{a}(n_i - 1)} \tag{8-3}$$

Formula (8-3) je komplicirana i neprikladna za primjenu kod studenata ekonomije. Microsoft Excel i drugi programski paketi izračunavaju ove vrijednosti, a više o tome slijedi u nastavku.

Prikladno je primijetiti da, kao i za druga testiranja hipoteze, manja p-vrijednost znači da treba odbaciti H_0. p - vrijednost je vjerojatnost odbacivanja hipoteze koja odgovara površini područja odbacivanja H_0.

Kao primjer F - testa koji utvrđuje značajne razlike u aritmetičkim sredinama između grupa, promotrit će se jednostavan ANOVA primjer s dvije grupe. Problem se bavi statistikom uspjeha studenata na ispitu iz Statistike. U ponedjeljak navečer ispitivani su izvanredni studenti u Charlotteu. U utorak navečer isti su ispit pisali izvanredni studenti u Winston-Salemu. Nakon ocjenjivanja ispita, zanimljivo bi bilo testirati postoji li razlika u uspjehu između te dvije grupe studenata. ANOVA je alat pomoću kojeg je moguće riješiti to pitanje (rezultat je ekvivalentan korištenju prije opisanog t - testa).

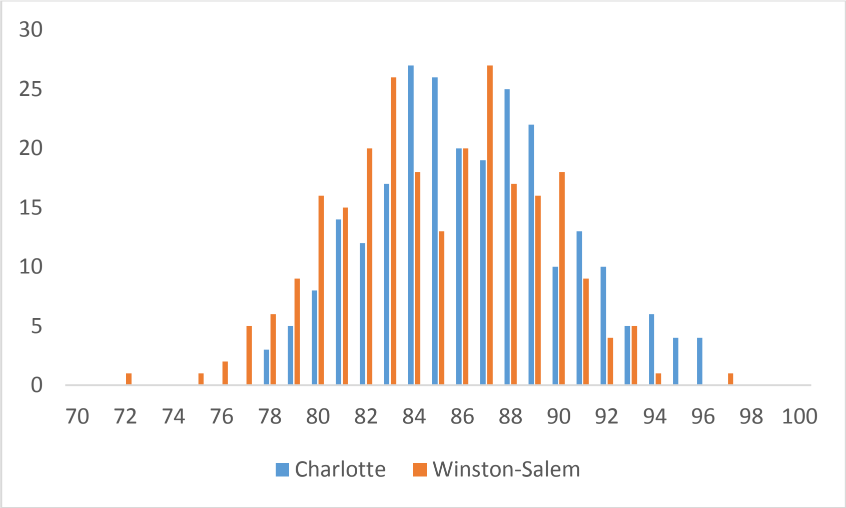
Dobiveni prosječan uspjeh studenata na ispitu iz Charlottea je oko 86 bodova, a studenata iz Winston-Salema oko 85. Standardna devijacija za obje grupe je oko 5.5. F - vrijednost u ovom slučaju je 1.52 i daje odgovarajuću p -vrijednost jednaku 0.7971 uz varijabilnosti između grupa 39.20, a unutar grupa 25.8. Tako ne postoji značajna razlika između grupa. Grafikon 8.1 podržava ovaj zaključak temeljem kombiniranog histograma za dvije grupe studenata.



Grafikon 8.1. Nema značajne razlike između grupa

Grafikon 8.1 ukazuje na zaključak o nepostojanju statistički značajne razlike između dviju grupa i može se tvrditi da su grupe podjednako uspješne.

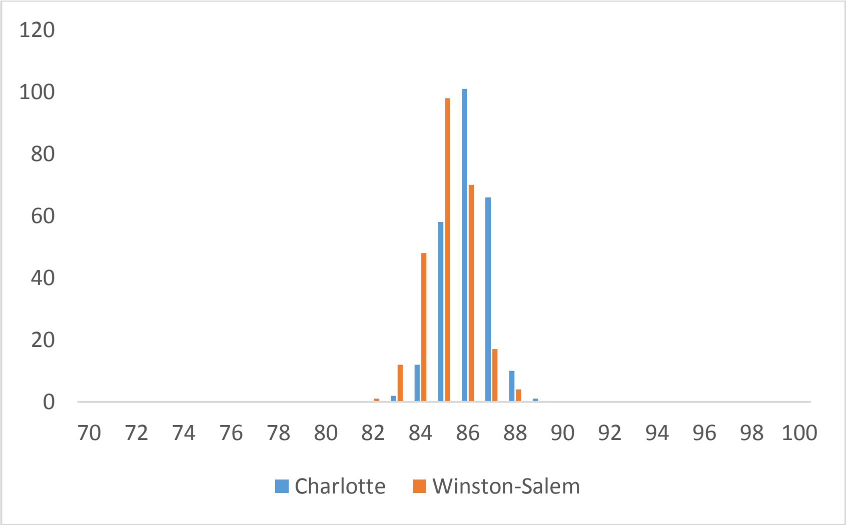
Neka su sada aritmetičke sredine dviju grupa podjednake kao što je prije navedeno, ali neka su sada standardne devijacije uzoraka oko 4 za svaku grupu. Dobiva se F - vrijednost jednaka 2.80 s varijabilnosti između grupa jednakom 45.00 i varijabilnosti unutar grupa jednakom 16.05. F - vrijednost daje p - vrijednost jednaku 0.0947. Grafikon 8.2 prikazuje kombinirani histogram ovog slučaja.



Grafikon 8.2. Značajna razlika između grupa

Ovdje možemo primijetiti da postoje određeni dokazi da studenti iz Charlottea imaju nešto bolji uspjeh od onih iz Winston-Salema budući da su plavi stupci viši desno od sredine, a narančasti na lijevoj strani. Ipak, ovaj dokaz nije dovoljno uvjerljiv.

Konačno, neka ponovno prosjek studenata iz Charlottea bude opet oko 86, a prosjek studenata iz Winston-Salema oko 85. Ovaj put neka je standardna devijacija za obje grupe studenata oko 1. Dobiva se F - vrijednost jednaka 100.66 (odgovarajuća p - vrijednost je manja od 0.0001), varijabilnost između grupa iznosi 106.72, a unutar grupa 1.06. Grafikon 8.3 prikazuje dobivene rezultate pomoću kombiniranog histograma.



Grafikon 8.3. Značajna razlika između skupina

Grafikon 8.3 prikazuje da postoji nedvosmislena razlika između grupa jer je razlika aritmetičkih sredina jednaka standardnoj devijaciji.

Poanta prikazivanja ova tri slučaja je u naglašavanju kako F - vrijednost kontrolira odluku o odbacivanju ili ne nulte hipoteze H_0. U svakom od ovih slučajeva razlika u aritmetičkim sredinama je oko 1 bod što znači da je varijabilnost između grupa u suštini ista. No, kako nastavljamo s primjerom, varijabilnost unutar grupe se smanjuje, a time se povećava F - vrijednost što povećava vjerojatnost da ćemo odbaciti H_0. Razlika aritmetičkih sredina može biti mala, ali ako postoji dosljednost unutar svake grupe, može se i dalje tvrditi da postoji značajna razlika između grupa. To je upravo ono što nam govori F - vrijednost.

8.2 Testiranje jednakosti aritmetičkih sredina više populacija

Kod usporedbe aritmetičkih sredina triju ili više populacija koristi se statistički alat poznat pod skraćenicom ANOVA (analiza varijance). Cilj je ispitati opažene razlike između sredina triju ili više uzoraka. U nultoj hipotezi pretpostavljamo da su sve aritmetičke sredine jednake, dok u alternativnoj hipotezi pretpostavljamo suprotno, a to je da postoje barem dvije koje se statistički značajno razlikuju. Distribucija u pozadini testa je Fischerova distribucija i stoga se vrijednost testne statistike označava s F. Donošenje odluke se temelji na vrijednosti testne statistike F i pripadne p - vrijednosti. Pri tome se zaključci o odbacivanju ili neodbacivanju nulte hipoteze temelje na p - vrijednosti i donose se na isti način kao i do sada. Za provođenje ANOVA testova nije komplicirano postaviti hipoteze H_0 i H_A budući je test jednak svaki put. U H_0 se tvrdi da su sve aritmetičke sredine jednake, a u H_A da nisu i to se nikada ne mijenja. Prema tome, programski paketi uvijek izračunavaju aritmetičke sredine i standardne devijacije za svaku grupu, a zatim daju F - vrijednost i odgovarajuću p - vrijednost.

Kao primjer, promatra se skup podataka "DefektiPoSmjeni". U ovom skupu podataka bilježen je broj defektnih proizvoda u tri radne smjene pri čemu su podaci grupirani po stupcima tako da svaki stupac predstavlja jednu smjenu. Svaka vrijednost u stupcu prikazuje broj defektnih proizvoda na milijun jedinica za odgovarajuću smjenu. Istražuje se je li broj defektnih proizvoda na milijun proizvedenih podjednak u svakoj smjeni. U Microsoft Excelu je to vrlo jednostavno učiniti. Skup podataka mora biti zapisan u Excelu s jedinstvenim stupcem za svaku jedinstvenu populaciju - u ovom slučaju, svaka populacija je jedna smjena. Koristimo alat Data Analysis i odabiremo "ANOVA: Single Factor". Odatle označavamo podatke koje želimo koristiti u analizi. Ponovo, svaki stupac predstavlja jedinstvenu populaciju. Potrebno je naznačiti nalaze li se u skupu podataka u Excel tablici nazivi stupaca ili ne i konačno unijeti zadana razina značajnosti ($\alpha = 0.05$). Konačno se bira željeno mjesto ispisa rezultata i ikona "U redu". Dobiva se ANOVA rezultat koji je sažet u tablici 8.2.

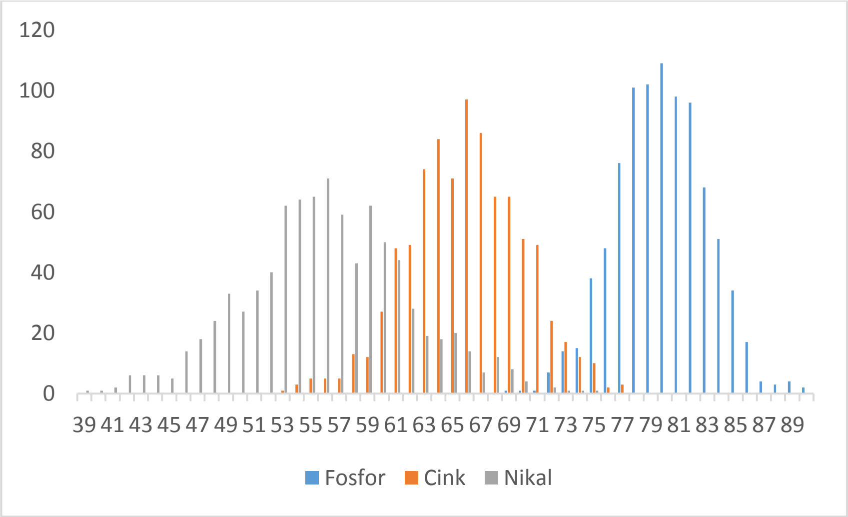
Grupe	*Broj*	*Zbroj*	*Prosjek*	*Varijanca*
Smjena 1	62	3033	48.92	219.42
Smjena 2	62	2733	44.08	174.08
Smjena 3	62	2913	46.98	258.80

Izvor varijabilnosti	SS	df	MS	F	$p-$*vrij.*
Između grupa	735.48	2	367.74	1.69	0.1871
Unutar grupa	39790	183	217.43		
Ukupno	40526	185			

Tablica 8.2. ANOVA rezultat za zadatak s brojem defekata po smjeni

Prvi dio grafikona 8.2 prikazuje deskriptivnu statistiku za svaku od tri smjene. Najviše nas zanima prosječan broj defekata po smjeni. Primijetit ćete da su defekti na milijun proizvoda za tri smjene u rasponu od oko 44 do oko 49. ANOVA tablica ispod prikazuje vrijednost testne statistike (F = 1.69) i odgovarajuću p - vrijednost. Visoka p - vrijednost (0.1871) ne odbacuje nultu hipotezu i može se tvrditi da je prosječan broj defekata na milijun proizvedenih jedinica podjednak za svaku smjenu. Zaključak je da sastavi smjena *nemaju* statistički značajnog utjecaja na defekte. Ostatak tablice ANOVA jednostavno pokazuje izračun F - vrijednosti.

Drugi je primjer vijek trajanja alata obloženih s tri različita metala: niklom, fosforom i cinkom. Želimo analizirati vijek trajanja alata za svaku skupinu metala.



Grafikon 8.4. Vijek trajanja alata prema vrsti metala

Iz kombiniranog histograma na grafikonu 8.4 se čini da fosfor daje najduži vijek trajanja alata, zatim cink, dok nikal pruža najkraći vijek trajanja alata. No bez odgovarajućeg ANOVA testa ne može se službeno tvrditi. Koristeći Excel na sličan način kao u zadnjem primjeru dobiva se ANOVA rezultat prikazan tablicom 8.3.

Grupe	*Broj*	*Zbroj*	*Prosjek*	*Varijanca*
Fosfor	890	71090	79.88	10.30
Cink	878	57864	65.90	16.38
Nikal	872	48867	56.04	33.30

Izvor varijabilnosti	SS	df	MS	F	$p-$*vrij.*
Između grupa	253026	2	126513	6352.38	0[12]
Unutar grupa	52518	2637	20		
Ukupno	305544	2639			

Tablica 8.3. ANOVA rezultat za vijek trajanja alata prema vrsti metala

Iz grafikona 8.4 i tablice 8.3 slijedi snažan dokaz da tri metalna premaza

[12] Excel ovdje prikazuje p - vrijednost kao 0, a trabalo bi biti < 0.0001. Excel daje takav ispis za p - vrijednost manju od 10^{-128}.

ne rezultiraju sličnim vijekom trajanja alata budući da je vrijednost testne statistike F = 6352.38 i da je odgovarajuća p - vrijednost manja od 0.0001. Konačan zaključak je da metalni premaz *ima* utjecaja na vijek trajanja alata.

8.3 Višefaktorska analiza varijance

U prethodnom odjeljku uspoređuju se sredine populacija s obzirom na to da je samo jedan faktor – primjerice radna smjena ili vrsta premaza. Usporedba se provodi statističkim alatom jednofaktorske ANOVA-e. Postoje situacije u kojima se analizira više faktora. Primjerice, kada se utvrđuje imaju li i spol prodavača i dan u tjednu utjecaj na prodaju automobila. Matematički:

$$f(spol\ prodavača, dan\ u\ tjednu)\ = prodani\ automobili$$

Ovaj problem ima dva faktora (spol i dan) pri čemu spol ima dvije razine (muški i ženski), a dan u tjednu ima sedam razina (svaki dan u tjednu). Problem ove vrste može se riješiti višefaktorskom ANOVA-om. Nažalost, dodavanjem faktora u ANOVA-u se problem u velikoj mjeri komplicira pa je ostavljeno da se tim problemom bave specijaliziranije knjige.

8.4 Zaključci

Primjena ANOVA-e omogućuje nam ispitati ima li određeni faktor utjecaj na vrijednost drugog faktora. Faktor se može smatrati kategoričkom varijablom i ne može se precizno mjeriti numerički. Primjerice, faktoru jedne radne smjene moguće je dodijeliti broj, ali samo u odnosu na druge smjene. Nije moguće mjeriti metalne premaze, spol prodavača ili određeni dan u tjednu, ali je moguće ustanoviti vezu kategoričke i numeričke varijable. U desetom poglavlju o jednostavnoj linearnoj regresiji, kategorička varijabla je zamijenjena numeričkom varijablom. Zbog toga postoji jaka veza između jednostavne ANOVA analize i jednostavne linearne regresije.

8.5 Zadaci

1. Analizira se odnos između temperature i doba dana zaokruženog na puni sat. Koja je varijabla kategorička, a koja je numerička? Predložite broj razina koje može imati kategorička varijabla.
2. U ovom zadatku koristite datoteku "SpolProdaja". Ima li spol, uz signifikantnost $\alpha = 0.05$, utjecaja na količinu prodaje? Prikažite aritmetičke sredine prodaje za svaki spol i objasnite svoj odgovor u skladu s tim.
3. U ovom zadatku koristite datoteku "SpolProdaja". Odredite imaju li dani u tjednu, na razini $\alpha = 0.07$, utjecaja na količinu prodaje. Prikažite aritmetičke sredine prodaje za svaki dan te u skladu s tim objasnite svoj odgovor.
4. U ovom zadatku koristite datoteku "KonobarPritužbe". Razlikuje li se (statistički značajno) postotak gostiju restorana koji se žale na uslugu s obzirom na konobara? Testirajte na razini $\alpha = 0.03$. Prikažite prosječni postotak žalbi za svakog konobara te u skladu s tim objasnite svoj odgovor.
5. U ovom zadatku koristite datoteku "IspitDan". Profesor Prijevara podučava statistiku ponedjeljkom, srijedom i petkom. Također, ispit će održati na jedan od navedenih dana. Skup podataka prikazuje nasumce odabrane rezultate ispita za studente koji su pristupili ispitu na jedan od tih dana. Razlikuju li se (statistički značajno) rezultati ispita s obzirom na dan pisanja? Test provedite na razini značajnosti $\alpha = 0.05$. Prikažite aritmetičke sredine rezultata ispita za različite dane te koristite dobiveno kako biste objasnili svoj rezultat.
6. U ovom zadatku koristite skup podataka "WhisperingPines". Arnold, Jack i Tiger su renomirani golferi, a Whispering Pines je poznati golf teren s bogatom poviješću. Na golf terenu postoje četiri Par-5 rupe. Kod rupe Par-5 se od profesionalnog golfera očekuje da baci svoju loptu na zeleno u tri hica. Ipak, najbolji golferi pokušavaju baciti loptu na zeleno u dva hica kako bi povećali vjerojatnost da će postići dobar rezultat. Ako golfer baci svoju loptu na zeleno u Par-5 u dva hica, u ovom je izazovu uspio. Skup podataka uključuje nasumično odabrane stope uspjeha za Arnolda, Jacka i Tigera za neke od njihovih partija

golfa koje su igrali na rupi Par-5 na golf terenu Whispering Pines. Jesu li stope uspjeha, uz korištenje $\alpha = 0.03$, jedinstvene za tri golfera, tj. razlikuju li se značajno?

7. Nadovezano na prethodno pitanje, je li Tigerova stopa uspjeha jedinstvena u odnosu na druga dva golfera? Ponovno koristite $\alpha = 0.03$.

8. U ovom zadatku koristite skup podataka “CijeneKućaČetiriGrada”. Uključena su četiri grada: Indianapolis, Boston, Rochester i San Diego. Za svaki grad su prikazane nasumično odabrane cijene nekretnina. Može li se, na razini $\alpha = 0.01$, reći da grad ima utjecaj na cijenu?

9. Može li se, koristeći isti skup podataka kao u prethodnom zadatku i uz $\alpha = 0.01$, tvrditi da postoji razlika u cijeni između San Diega i Bostona?

9. Testiranje χ^2 (hi-kvadrat) testom

Neka je dana distribucija opaženih podataka koji su prikupljeni iz stvarnog okruženja, a zadatak je odrediti tip teorijske distribucije populacije. Jedino pouzdano rješenje je formalni statistički test. Primjerice, ako histogram ima vrhunac u sredini podataka, simetričan je, a i niske frekvencije se nalaze na krajevima histograma, valja pretpostaviti da se radi o normalnoj distribuciji.

Pretpostavku je potrebno dokazati formalnim ispitivanjem kako bi se utvrdilo odgovara li opažena distribucija podataka nekoj pretpostavljenoj teorijskoj (očekivanoj) distribuciji podataka. U tu svrhu se provodi χ^2 (hi- kvadrat) test.

U ovom poglavlju će se opisati dvije vrste χ^2 testa. Prvi se zove test prilagodbe modela podacima ili "*goodness of fit*" test kojim se ispituje odgovara li dana distribucija podataka očekivanoj distribuciji. Druga vrsta testa je test neovisnosti koji utvrđuje jesu li varijable prikazane tablicom marginalnih vjerojatnosti neovisne jedna o drugoj.

9.1 χ^2 test (hi-kvadrat test)

Formalni test, kojim se utvrđuje odgovara li distribucija stvarno prikupljenih, opaženih ili empirijskih podataka očekivanoj, teorijskoj distribuciji podataka je χ^2 test. Odgovarajuće hipoteze su:

$$\begin{aligned} H_0&: \text{ dana distribucija odgovara očekivanoj distribuciji} \\ H_A&: \text{ dana distribucija ne odgovara očekivanoj distribuciji} \end{aligned} \tag{9-1}$$

Kao što je slučaj kod ANOVA analize, ove hipoteze su uvijek definirane na način iskazan u (9-1) pa nije uvijek potrebno službeno ih navoditi kao što je to bilo kod testova koji su provođeni u šestom poglavlju.

χ^2 distribucija je teorijska distribucija vrlo složeog formalnog zapisa čije je vrijednosti moguće dobiti Excel tablicama. U pozadini je samog testa, budući da statistika koja se u testu računa, prati χ^2 distribuciju. χ^2 distribucija ima jedan parametar koji se naziva broj stupnjeva slobode. Pri ispitivanju hipoteza navedenih u (9-1) zapravo se uspoređuje podudarnost dvaju nizova podataka. Jedan čine opaženi podaci, a drugi niz su očekivani podaci izračunati iz očekivane teorijske distribucije slučajne varijable. Očekivani podaci računaju se na temelju parametra neke od teorijskih razdioba slučajne varijable. Ukoliko nisu zadani, parametri se procjenjuju iz niza opaženih podataka. Pod pretpostavkom da svaki niz ima n mogućih vrijednosti, uspoređuju se opažene frekvencije (f_o) s očekivanim frekvencijama (f_e) i dobiva se testna veličina χ^2 testa na sljedeći način:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(f_{o(i)} - f_{e(i)}\right)^2}{f_{e(i)}}. \tag{9-2}$$

Testna veličina dobiva se tako da se kvadrirana razlika opaženih i očekivanih frekvencija za svaki podatak dijeli očekivanom frekvencijom (zbog standardizacije) i te se sve dobivene vrijednosti zbroje. Ako vrijednost zbroja prelazi neku kritičnu vrijednost, odbacujemo H_0 i tvrdimo da postoji razlika između opažene i očekivane distribucije. Tablica 9.1 prikazuje funkcije Excela koje daju kritičnu vrijednost i odgovarajuću p – vrijednost χ^2 testa na razini značajnosti α i za n mogućih ishoda.

Funkcija	Definicija
$\boldsymbol{CHIINV(\alpha, n-1)}$ [13]	kritična vrijednost χ^2 testa
$\boldsymbol{CHIDIST(\chi^2, n-1)}$[14]	p - vrijednost χ^2 testa

Tablica 9.1. Funkcije u Excelu za χ^2 test

9.2 Test prilagodbe modela podacima

Vrsta χ^2 testa kojom se ispituje odgovara li niz opaženih podataka očekivanoj, teorijskoj distribuciji naziva se test prilagodbe modela podacima ili *goodness of fit* test.

Dobar primjer za testiranje prilagodbe modela podacima je pokus bacanja pravilne kockice 100 puta i proučavanje učestalosti pojavljivanja (šest) mogućih ishoda. S obzirom na to da očekujemo da

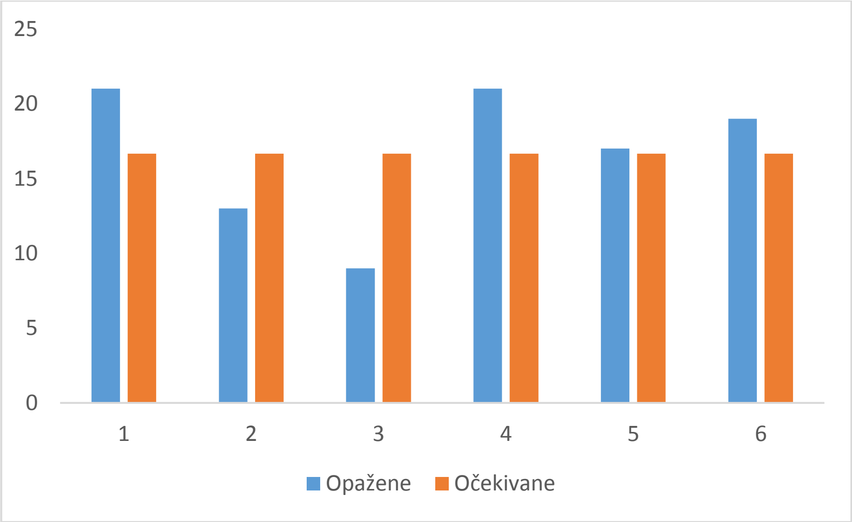
[13] U novijim verzijama Excela koriste se funkcije *„=CHISQ.INV(1-α, n-1)“* ili *„=CHISQ.INV.RT(α, n-1)“.*
[14] U novijim verzijama Excela koriste se funkcije *„=CHISQ.DIST.RT(χ^2, n-1)“* ili *„=1-CHISQ.DIST(χ^2, n-1, 1)“.*

se svaki ishod pojavi s vjerojatnosti 1/6, za svaki ishod od 100 bacanja očekivana frekvencija bit će 100/6 što je jednako 16.66. Tablica 9.2 prikazuje opažene frekvencije usporedno s očekivanim frekvencijama za 100 simuliranih bacanja. Tablica prikazuje i detalje izračuna χ^2 testne statistike prikazan izrazom (9-2).

Ishod	**1**	**2**	**3**	**4**	**5**	**6**
Opaženo	21	13	9	21	17	19
Očekivano	16.67	16.67	16.67	16.67	16.67	16.67
$(f_o - f_e)^2/f_e$	1.13	0.81	3.53	1.13	0.01	0.33

Tablica 9.2. χ^2 test za bacanje kocke

Zbrajanjem elemenata donjeg retka ove tablice dobivamo testnu χ^2 veličinu koja je jednaka 6.92. Za razinu značajnosti $\alpha = 0.05$ imamo kritičnu vrijednost od 11.07. Budući da χ^2 testna veličina od 6.92 ne prelazi kritičnu vrijednost od 11.07, dobivenu formulom Excela iz tablice 9.1, ne može se odbaciti H_0. Na razini značajnosti 5% zaključuje se da je očekivana distribucija podataka ista kao i opažena distribucija. Isto se tako može tvrditi da je bacanje kocke bilo pošteno odnosno da je kocka pravilna. Grafikon 9.1 prikazuje rezultate pomoću histograma - primijetite male razlike u visini svakog stupca.



Grafikon 9.1. Histogram bacanja kocke

Također, može se testirati slijedi li neki skup opaženih podataka neku unaprijed određenu neprekidnu, primjerice normalnu distribuciju. U četvrtom poglavlju se navodi da je normalna distribucija neprekidna s beskonačno mnogo mogućih ishoda, a to na prvi pogled nije u skladu s diskretnom prirodom χ^2 testiranja. Zato se podaci promatraju po unaprijed određenim razredima i uspoređuju se opažene frekvencije razreda s očekivanim frekvencijama. Očekivane frekvencije računaju se uz pretpostavku da je distribucija podataka normalna. Parametri μ i σ normalne distribucije procjenjuju se iz opaženih podataka. Pri tome se u Excelu mogu koristiti funkcije $NORMDIST$ ili $NORMSDIST$ koje za normalno distribuiranu varijablu $X \sim N(\mu, \sigma^2)$ računaju vjerojatnost

$$P(X < x_i) = NORMDIST(x_i, \mu, \sigma, true). \qquad (9\text{-}3)$$

Funkcija $NORMSDIST$ se koristi za standardiziranu normalnu slučajnu varijablu $N(0,1)$ gdje je unaprijed postavljeno $\mu = 0, \sigma = 1$.

Potrebno je prije računanja vjerojatnosti diskretne ishode pretvoriti u kontinuirane formiranjem razreda $[x_{i-1}, x_i]$. Vjerojatnost da slučajna varijabla poprimi vrijednost iz intervala $[x_{i-1}, x_i]$ jednaka je:

$$P([x_{i-1}, x_i]) = NORMDIST(x_i, \mu, \sigma, 1) - NORMDIST(x_{i-1}, \mu, \sigma, 1) \quad (9\text{-}4)$$

Ovo je aproksimacija vjerojatnosti pojavljivanja nekog konkretnog ishoda iz i-tog razreda. Ovu aproksimiranu vjerojatnost možemo pretvoriti u vrijednost očekivane frekvencije ($f_{e_{(i)}}$) na sljedeći način:

$$f_{e_{(i)}} = mP([x_{i-1}, x_i]), \quad (9\text{-}5)$$

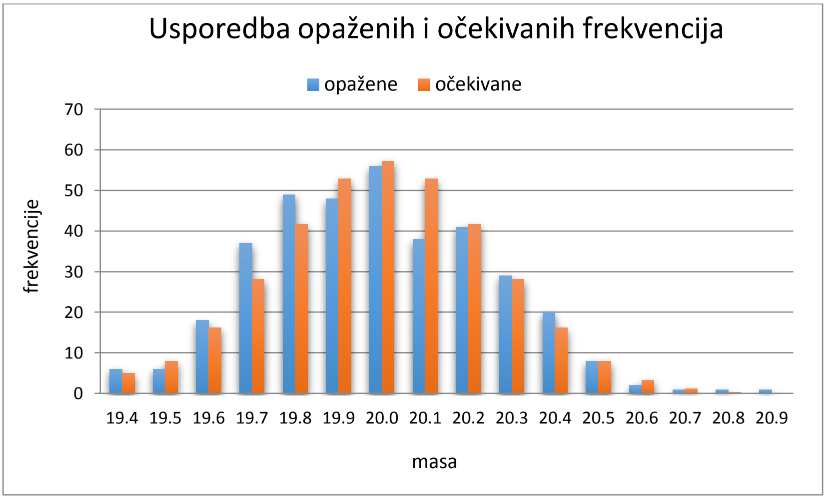
pri čemu je m ukupan broj opažanja u skupu podataka. Sada se može provesti χ^2 test jer su poznate vrijednosti f_o i f_e za sve moguće razrede podataka.

Primjena χ^2 testa za normalnu razdiobu ilustrirana je na primjeru proizvodnje lijekova u tabletama od 20 mg. Tvrtka tvrdi da postoji varijabilnost mase tableta, ali kaže da je prosječna masa jednaka 20mg uz standardnu devijaciju od 0.25 mg. Odjel za kontrolu kvalitete tvrtke nasumično je odabrao 361 tabletu i izmjerio njihove mase do 0.1 mg preciznosti. Svi mogući ishodi i opažene frekvencije (f_o) su prikazane u prva dva stupca tablice 9.3. U trećem stupcu su opaženi, empirijski dobiveni podaci grupirani u razrede. Sada iz empirijskih frekvencija f_o, treba izračunati teorijske frekvencije f_e. Izrazima (9-3) i (9-4) izračunavaju se vjerojatnosti razreda koje su prikazane u četvrtom, petom i šestom stupcu tablice 9.3. Očekivane frekvencije računaju se tako da se vrijednosti iz šestog stupca pomnože s brojem ukupnih opažanja $m = 361$ što rezultira očekivanim frekvencijama (f_e) koje su prikazane u sedmom stupcu tablice 9.3. Iz očekivanih frekvencija izračunava se testna veličina χ^2 koja je prikazana u zadnjem retku osmog stupca tablice 9.3 i iznosi 18.56121. Funkcija "$CHIDIST$" daje p - vrijednost jednaku 0.23431. Stoga se za bilo koju razinu značajnosti α za koju je vrijednost manja od 0.23431 može tvrditi da je masa tableta normalno distribuirana s očekivanom vrijednosti $\mu = 20\,mg$ i standardnom devijacijom $\sigma = 0.25\,mg$.

Reprezentant razreda	f_0	Razred	$p(x < x_i)$	$p(x < x_{i-1})$	$p([x_{i-1}, x_i])$	f_e	$\frac{(fo - f_e)^2}{f_e}$
19.4	6	-∞ do 19.45	0.013903		0.013903	5.019145	0.191682
19.5	6	19.45 - 19.55	0.03593	0.013903	0.022027	7.951701	0.479034
19.6	18	19.55 - 19.65	0.080757	0.03593	0.044826	16.18231	0.204174
19.7	37	19.65 - 19.75	0.158655	0.080757	0.077899	28.12139	2.803192
19.8	49	19.75 - 19.85	0.274253	0.158655	0.115598	41.73083	1.266231
19.9	48	19.85 - 19.95	0.42074	0.274253	0.146487	52.88187	0.450677
20.0	56	19.95 - 20.05	0.57926	0.42074	0.158519	57.22551	0.026245
20.1	38	20.05 - 20.15	0.725747	0.57926	0.146487	52.88187	4.188015
20.2	41	20.15 - 20.25	0.841345	0.725747	0.115598	41.73083	0.012799
20.3	29	20.25 - 20.35	0.919243	0.841345	0.077899	28.12139	0.027451
20.4	20	20.35 - 20.45	0.96407	0.919243	0.044826	16.18231	0.90066
20.5	8	20.45 - 20.55	0.986097	0.96407	0.022027	7.951701	0.000293
20.6	2	20.55 - 20.65	0.995339	0.986097	0.009242	3.336456	0.535333
20.7	1	20.65 - 20.75	0.99865	0.995339	0.003311	1.195376	0.031933
20.8	1	20.75 - 20.85	0.999663	0.99865	0.001013	0.365682	1.1003
20.9	1	20.85 do +∞	1	0.999663	0.000337	0.121631	6.343188
Zbroj	361				1	361	18.56121

Tablica 9.3. Izračun za test normalnosti. Ishodi su reprezentanti razreda

Grafikon 9.2 prikazuje kombinirani histogram kojim su uspoređene očekivane i opažene frekvencije. U slučajevima kada su očekivane frekvencije manje od 5, spajaju se razredi tih frekvencija sve dok ne budu već od 5. U primjeru su očekivane frekvencije zadnja 4 razreda manje od 5. Ako se grupiraju zadnja 4 razreda u jedan razred s granicama $[20.55, +\infty >$, pripadna očekivana frekvencija iznosi 5.01914, a pripadna vrijednost testne statistike iznosi $\chi^2 = 10.55$ što još više sugerira neodbacivanje hipoteze uz p - vrijednost = 0.56.



Grafikon 9.2. Usporedba opaženih i očekivanih frekvencija

Moguće je provesti test o pripadnosti nekoj drugoj teorijskoj distribuciji, primjerice ima li populacija binomnu distribuciju. Taj test je lakše provesti nego test normalnosti jer je binomna distribucija po prirodi diskretna i poprima samo konačno mnogo vrijednosti. Ako su zadane opažene frekvencije (f_o) za svaki od n različitih podataka, tada se može procijeniti vjerojatnost uspjeha p, a zatim pomoću binomne distribucije odrediti očekivane vrijednosti f_e za χ^2 test.

Razmotrimo primjer poluumirovljenog liječnika koji radi u ruralnom području. On prima pet pacijenata dnevno, ali često pacijenti ne dođu u dogovoreno vrijeme. Na kraju dana, voditeljica ordinacije broji pacijente čiji su pregledi obavljeni u dogovoreno vrijeme. Tablica 9.4 prikazuje distribuciju broja dana u kojima je odgovarajući broj pregleda obavljen u dogovoreno vrijeme, a za posljednjih 100 dana:

Na vrijeme	0	1	2	3	4	5
Frekvencije (f_o)	22	26	34	11	5	2

Tablica 9.4. Broj dana u kojima su pacijenti pregledani u dogovoreno vrijeme

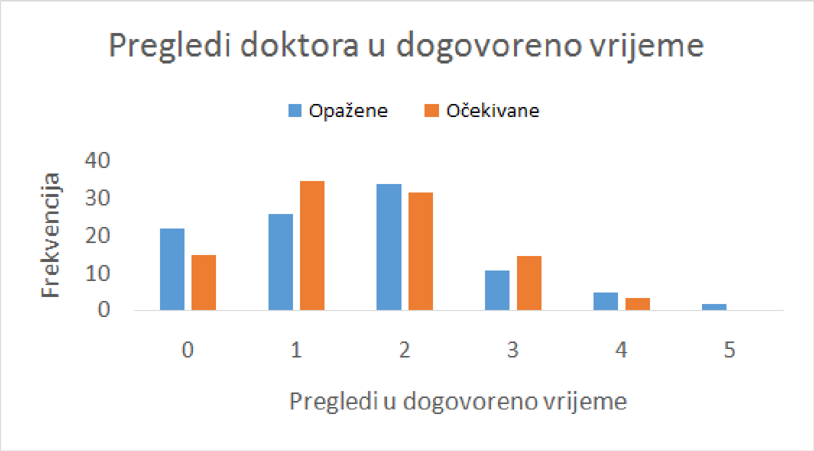
Testiranje binomne distribucije zahtijeva poznate vjerojatnosti uspjeha da pacijent bude pregledan na vrijeme, a ta vrijednost se mora procijeniti na uzorku od 100 dana. Broj pacijenata pregledanih u dogovoreno vrijeme u uzorku je 0·22+ 1·26 + 2·34 + 3·11 + 4·5 + 5·2 = 157. Budući da je bilo ukupno 500 pacijenata (5 pacijenata na dan tijekom 100 dana), vjerojatnost da pacijent bude pregledan u dogovoreno vrijeme je $\frac{157}{500} = 0.314$. Ova vrijednost se nadalje koristi u funkciji Excela "$BINOM.DIST$" koja daje vjerojatnosti broja "uspjeha" za 0 do 5 pacijenata pregledanih u dogovoreno vrijeme. Funkcija $BINOM.DIST$ opisana je u odjeljku 4.1.2. Excelom dobivene binomne vjerojatnosti pomožene su sa 100 dana čime se dobivaju očekivane frekvencije iz kojih se računaju veličine potrebne za χ^2 test:

Uspjesi	0	1	2	3	4	5
f_o	22	26	34	11	5	2
Binomne vjer.	0.15	0.35	0.32	0.15	0.03	0.00
f_e	15.19	34.77	31.83	14.57	3.33	0.31
$\frac{(f_e - f_o)^2}{f_e}$	3.05	2.21	0.15	0.87	0.83	9.41

Tablica 9.5. Rezultati χ^2 testa za preglede pacijenata u dogovoreno vrijeme

Zbroj vrijednosti u posljednjem retku tablice 9.5 daje testnu veličinu χ^2 jednaku 16.53 s odgovarajućom p - vrijednosti od 0.0055. Uz najčešće razine značajnosti α, odbacuje H_0 što znači da binomna distribucija ne opisuje populaciju iz koje je uzorkovan dani skup podataka. Ponovno, u slučajevima kada su očekivane frekvencije manje od 5, potrebno je grupirati razrede. Tako je potrebno grupirati razrede sa uspjesima 3, 4 i 5. Ukupna očekivana frekvencija grupiranih razreda iznosi 18.21, dok je ukupna opažena frekvencija 18. Testna veličina $\chi^2 = 5.41$ s odgovarajućom p - vrijednosti 0.067 što uz različite α dovodi do različitih odluka: za α manje od 0.067 ne odbacuje se hipoteza da podaci dolaze iz binomne distribucije, dok veći α od 0.067 hipotezu odbacuje.

Grafikon 9.3 prikazuje usporedbu opaženih i očekivanih frekvencija.



Grafikon 9.3. χ^2 test za preglede pacijenata u dogovoreno vrijeme

9.3 Test neovisnosti

Prilikom provođenja testova prilagodbe modela podacima uspoređujemo opaženi, empirijski skup ishoda s izračunatim, očekivanim skupom ishoda. Nadogradnja je u istraživanju je li više varijabli međusobno povezano ili zavisno. Istražiti je li više naizgled nepovezanih pojava zapravo povezano moguće je χ^2 testom koji se naziva testom neovisnosti. U tom kontekstu slijede H_0 i H_A:

$$\begin{aligned} H_0&: \text{ Pojave su neovisne jedna o drugoj} \\ H_A&: \text{ Pojave nisu neovisne jedna o drugoj} \end{aligned} \tag{9-6}$$

Provođenje testa neovisnosti bit će ilustrirano na primjeru rezultata glasovanja političkih stranaka u SAD-u prikazanih u tablici 3.1. Detalji glasovanja su izostavljeni, a u stupcima i retcima su prikazani samo marginalni redak/stupac zajedno s postocima.

	Za	Protiv	Ukupno
Republikanci			242 (56.94%)
Demokrati			183 (43.06%)
Ukupno	275 (64.71%)	150 (35.29%)	425

Tablica 9.6. Predložak glasovanja

Budući da je poznato kako je 64.71% glasova "Za" te da 56.94% birača čine republikanci, *očekuje* se da su (64.71% · 56.94% · 425) 156.59 od 425 birača republikanci koji su glasovali “Za”.

Formalno, vrijednosti očekivane frekvencije jednake su:

$$f_{e_{(ij)}} = m\left(\frac{\sum redak_i}{m}\right)\left(\frac{\sum stupac_j}{m}\right) \qquad (9\text{-}7)$$

gdje je m zbroj svih frekvencija ili veličina uzorka. Korištenjem izraza (9-7) računa se ostatak vrijednosti iz tablice 9.6:

	Za	Protiv
Republikanci	156.59	85.41
Demokrati	118.41	64.59

Tablica 9.7. Očekivani rezultati glasovanja

Vrijednosti u četiri "popunjene" ćelije tablice 9.7 temelje se samo na očekivanjima, oni su nezavisni od političkih stranaka i/ili stvarnih glasova ili očekivanih vrijednosti. Opažene vrijednosti (f_o) su izvorne vrijednosti iz tablice 3.1.

	Za	Protiv
Republikanci	230	12
Demokrati	45	138

Tablica 9.8. Opaženi rezultati glasovanja

Revidiramo izraz (9-2) kako bismo odredili testnu veličinu za χ^2 test neovisnosti i dobivamo:

$$\chi^2 = \sum_{i=1}^{retci} \sum_{j=1}^{stupci} \frac{\left(f_{o_{(ij)}} - f_{e_{(ij)}}\right)^2}{f_{e_{(ij)}}}. \qquad (9\text{-}8)$$

Iz izraza (9-8) dobivaju se sljedeće vrijednosti koje treba zbrojiti:

	Za	Protiv
Republikanci	34.42	63.10
Demokrati	45.51	83.44

Tablica 9.9. Izračun χ^2 za rezultate glasovanja

Zbroj tih vrijednosti iznosi 226.47. Broj stupnjeva slobode χ^2 testa neovisnosti određuje se prema:

$$df = (broj\ redaka - 1)(broj\ stupaca - 1) \qquad (9\text{-}9)$$

Test, dakle, ima $(2-1)(2-1) = 1$ stupanj slobode. To daje p -vrijednost <0.0001 što snažno upućuje na to da politička stranka i stav prema zakonu (za/protiv) nisu međusobno nepovezani (neovisni). Međusobna povezanost stvara političku klimu jakog strančarenja, a to nije iznenađenje.

Drugi primjer je anketa koju su ispunili studenti. Studentima je postavljeno pitanje o tome koliko su optimistični glede budućnosti. Studentima su ponuđena tri moguća odgovora: optimističan, indiferentan i pesimističan. Ispitanici su grupirani prema razredima: brucoš, student druge godine, student treće godine i apsolvent. Opaženi podaci $f_{o(ij)}$ su prikazani u tablici 9.10.

	Brucoš	student druge godine	student treće godine	apsolvent	Ukupno
Optimističan	7	3	2	8	20
Indiferentan	6	8	9	7	30
Pesimističan	15	15	14	6	50
Ukupno	28	26	25	21	100

Tablica 9.10. Opažene frekvencije u primjeru

Očekivane vrijednosti su prikazane u tablici 9.11.

	brucoš	student druge godine	student treće godine	apsolvent	Ukupno
Optimističan	5.60	5.20	5.00	4.20	20
Indiferentan	8.40	7.80	7.50	6.30	30
Pesimističan	14.00	13.00	12.50	10.50	50
Ukupno	28	26	25	21	100

Tablica 9.11. Očekivane frekvencije u primjeru

Tablica 9.12 prikazuje komponente testne veličine χ^2 za test neovisnosti.

	brucoš	student druge godine	student treće godine	apsolvent
Optimističan	0.35	0.93	1.80	3.44
Indiferentan	0.69	0.01	0.30	0.08
Pesimističan	0.07	0.31	0.18	1.93

Tablica 9.12. χ^2 detalji za primjer

χ^2 testna veličina iznosi 10.08. Ovaj test ima $(3-1)(4-1)=6$ stupnjeva slobode što daje p - vrijednost 0.1215. Prema tome, ne može se odbaciti H_0 i ostaje da su načini razmišljanja o budućnosti i godine studiranja neovisni, odnosno da nisu povezani.

9.4 Zaključak

χ^2 test je moćan, ali jednostavan statistički test. Zbog svoje se jednostavnosti može koristiti za dobivanje raznih općenitih zaključaka. Namjera χ^2 testa je usporedba dobivenih opaženih podataka s vrijednostima neke teorijske slučajne varijable i tako se mogu riješiti mnoga važna statistička pitanja.

Sljedeće poglavlje obuhvaća jednostavnu linearnu regresiju - važan alat u statistici. Pažljiva priprema podataka može regresijsko testiranje hipoteza zamijeniti jednostavnijim χ^2 testom. Motivacija je pojednostavljenje procesa i predstavlja vrijednost χ^2 testa čija snaga leži u njegovoj jednostavnosti.

Unatoč svojoj jednostavnosti i svestranosti, nedostatak χ^2 testa je njegova nestrukturiranost. Zato se mora osmisliti način koji povezuje opažene frekvencije s očekivanim.

9.5 Zadaci

1. Baca se novčić 103 puta i pismo padne 57 puta. Na razini $\alpha = 0.05$ testirajte je li novčić pošten.
2. Baca se par kockica 1000 puta i dobiveni su sljedeći rezultati. Na razini $\alpha = 0.05$ testirajte jesu li kockice poštene.

Ishod	2	3	4	5	6
Frekvencija	25	56	83	112	145

Ishod	7	8	9	10	11	12
Frekvencija	172	136	108	86	51	26

Sljedeće podatke koristite za zadatke 3 - 7. Baca se novčić 10 puta i zapisuje se broj puta koliko je dobiveno pismo. Ovaj pokus je ponovljen 1000 puta i dobivene su sljedeće frekvencije dobivanja pisma.

Pismo	0	1	2	3	4	5
Frekvencija	1	5	47	115	207	248

Pismo	5	6	7	8	9	10
Frekvencija	248	202	120	41	12	2

3. Odražavaju li podaci, na razini $\alpha = 0.05$, binomnu distribuciju?
4. Odražavaju li podaci, na razini $\alpha = 0.05$, normalnu distribuciju?
5. Možemo li, na razini $\alpha = 0.05$, tvrditi da se radi o binomnoj distribuciji s vjerojatnošću uspjeha $p = 0.5$?
6. Možemo li, na razini $\alpha = 0.05$, tvrditi da se radi o normalnoj distribuciji s očekivanom vrijednošću 5 i standardnom devijacijom 1.5811?
7. Je li novčić, na razini $\alpha = 0.05$, pošten?
8. Zamislite tvrtku koja proizvodi kosilice. Proizvodi se ista količina kosilica svaki dan. Po završetku proizvodnje kosilica, odjel kontrole kvalitete provodi konačnu provjeru. U nastavku je prikazan prosječan dnevni broj proizvoda koji su odbačeni u proteklih godinu dana.

Dan	Pon	Uto	Sri	Čet	Pet
Odbačeni proizvodi	21	25	19	22	31

Na razini značajnosti $\alpha = 0.04$ ispitajte je li broj dnevnih odbačenih proizvoda očekivan odnosno jednako distribuiran s obzirom na dane u tjednu.

9. Zamislite veliko sveučilište sa sedam fakulteta. Svake godine, sveučilište objavljuje broj prijevara na ispitima za svaki fakultet. Zabilježeni su podaci:

Fakultet	Broj prijevara	Broj upisanih
Arhitektura	6	1500
Ekonomija	51	6000
Tehnički fakultet	11	6000
Umjetnička akademija	6	1500
Društvene znanosti	10	4000
Prirodne znanosti	15	4000
Veterinarski fakultet	1	1000

Jesu li, na razini $\alpha = 0.05$, slučajevi prijevara povezani s brojem upisanih studenata po fakultetu?

10. Ljudi iz SAD-a su nasumično odabrani kako bi ih se pitalo bi li željeli da zakoni o oružju budu restriktivniji. Konkretno, pitanje je bilo: "Želite li da zakoni o oružju u SAD-u budu restriktivniji?" I muškarci i žene su odgovarali na pitanja, a rezultati su sljedeći:

	Da	Ne
Muškarci	74	51
Žene	80	32

Jesu li, korištenjem $\alpha = 0.03$, spol i odgovor na pitanje nezavisni?

11. Nekoliko Amerikanaca je nasumično odabrano i pitano odobravaju li odluku američkog Vrhovnog suda koja omogućuje istospolne brakove. Ispitanici su podijeljeni prema stranačkoj pripadnosti. Podaci su prikazani u nastavku:

	Demokrati	Neopredjeljen	Republikanac
Odobrava	96	31	41
Indiferentan	41	42	56
Ne odobrava	12	38	107

Jesu li stranačka pripadnost i odobravanje odluke da se istospolnim brakovima omogući brak, uz korištenje $\alpha = 0.06$, neovisni?

10. Jednostavna linearna regresija

Temelj statističkog zaključivanja je testiranje tvrdnje putem formalnog testa kojim se pravilno interpretiraju zaključci uvijek skeptičnoj znanstvenoj zajednici. To je važan uspjeh. U ovom se poglavlju interpretacija zaključaka nadograđuje istraživanjem potencijalnih veza između varijabli. Veze između varijabli istražuju se svakodnevno. Primjerice, veza izloženosti azbestu i mesothelioma, tipa raka pluća, zatim veza vremena koje dijete provodit pred ekranom i vjerojatnosti da to dijete bude pretilo.

Veze se iskazuju vrlo suptilno. Većinom kroz neki oblik linearne regresije koja predstavlja glavni alat ustanovljavanja vezam između varijabli ili entiteta. Ovo poglavlje obuhvaća jednostavnu linearnu regresiju kojom se istražuje linearna povezanost dviju varijabli. Mnoga informativna, ali i mnoga važna istraživanja provode se primjenom ovog moćnog alata.

10.1 Pravac regresije

Prije udubljivanja u regresiju, sa statističke točke gledišta, važno je prisjetiti se jednadžbe pravca u Kartezijevom koordinatnom sustavu. Položaj točke u Kartezijevim koordinatama određen je uređenim parom (x, y), gdje x koordinata označava položaj prema horizontalnoj, a y prema vertikalnoj osi. Opća jednadžba pravca ima dva koeficijenta: **slobodni član $\boldsymbol{a}$** i **koeficijent smjera $\boldsymbol{b}$**.

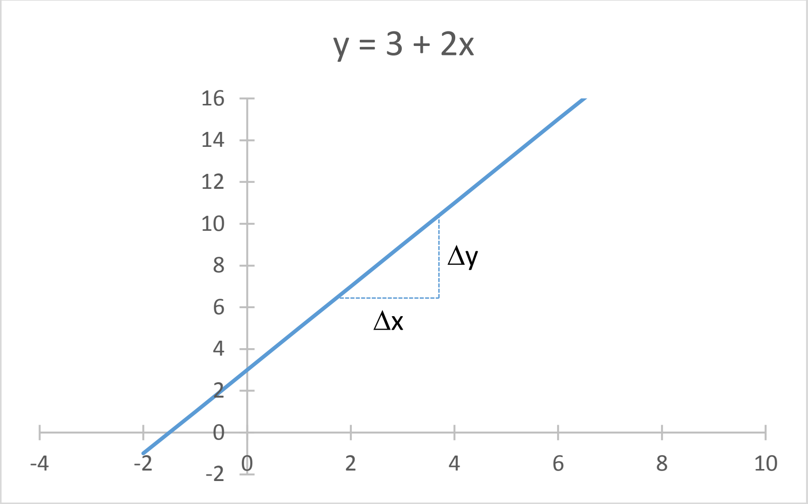
$$y = a + bx. \tag{10-1}$$

U funkcijskom smislu u jednadžbi (10-1) vrijednost x se proizvoljno zadaje i ona je **nezavisna varijabla**. Ako su poznate konstante a i b, tada vrijednost y ovisi samo o nezavisnoj varijabli x. Zato se y naziva vrijednost funkcije ili **zavisna varijabla**.

Slobodni član a je vrijednost funkcije y kada se za x odabere nula. U grafičkom prikazu, to je mjesto gdje pravac siječe vertikalnu ili y - os i zato se još naziva i odsječkom na osi y. Koeficijent smjera b je omjer promjene funkcije y nakon promjene varijable x. Matematički zapis:

$$b = \frac{\Delta y}{\Delta x}. \tag{10-2}$$

Grafikon 9.1 prikazuje graf pravca $y = 3 + 2x$.



Grafikon 10.1. Primjer linearnog modela

Slobodni član pravca na grafikonu 10.1 jednak je 3 pa pravac presijeca $y = 3 + 2 \cdot 0 = 3$. Koeficijent smjera pravca jednak je 2 pa je omjer $\Delta y : \Delta x$ jednak 2 u svakoj točki pravca, a ne samo u točki koja je istraknuta trokutom.

10.2 Metoda najmanjih kvadrata

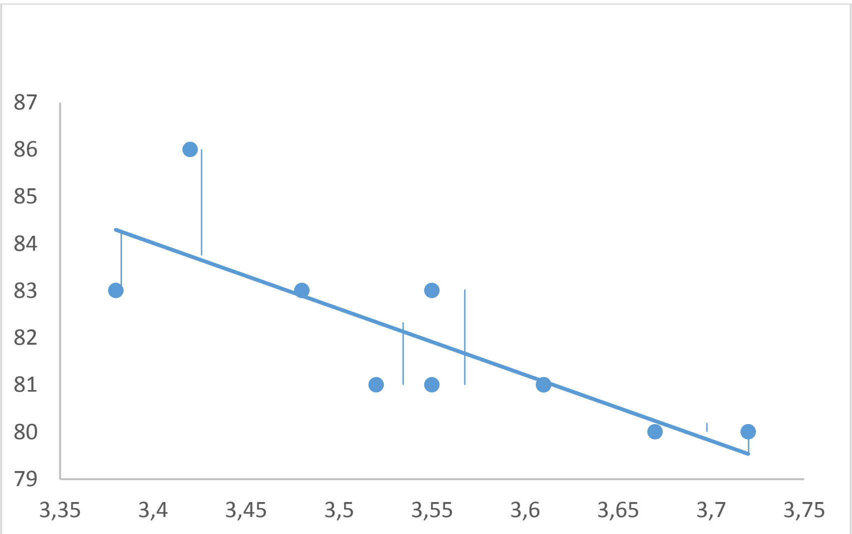
Neka je zadan skup točaka u ravnini koje se mogu poistovjetiti s naseljima u ravnici. Problem je nacrtati pravac tako da ukupan zbroj udaljenosti svih točaka do pravca bude najmanji i može se poistovjetiti sa izgradnjom ravne pruge do koje će ukupna udaljenost svih naselja biti najmanja. Problem je prikazan na grafikonu 10.2. Kada se u statistici na uzorku ispituju dvije osobine, tada se rezultat ispitivanja sastoji od dva podatka. Ako se prvi proglasi za x, a drugi za y, tada se oba podatka mogu predočiti kao jedna točka na grafikonu. Zadatak je konstruirati pravac kroz točke tako da ukupna udaljenost svih nacrtanih točaka grafikona do pravca bude minimalna.

Primjer je ispitivanje odnosa cijene i potražnje proizvoda nasumce odabranih prodavaonica u Winston-Salemu, NC. U svakoj trgovini je ispitivana cijena paketa od 12 limenki dijetalnog soka Dr. Pepper za koji je uprava dala podatke o prodaji prethodni tjedan. Dva su niza dobivenih podataka:

Cijena	3.38	3.42	3.48	3.52	3.55	3.55	3.61	3.61	3.67	3.72
Potražnja	83	86	83	81	81	83	81	81	80	80

Tablica 10.1. Cijena i potražnja dijetalnog soka Dr. Pepper

Dijagram rasipanja je prikazan grafikonom 10.2.



Grafikon 10.2. Dijagram rasipanja za podatke o potražnji dijetalnog soka Dr. Pepper

Točke pokazuju potražnju za svaku zadanu razinu cijene slučajnog uzorkovanja. Očito postoji inverzna veza između cijene i potražnje - kako se cijena povećava, potražnja se smanjuje.

Za najpribližniji pravac koji bi povezivao cijenu i potražnju, treba odrediti koeficijent smjera i slobodni član tako da **zbroj kvadrata pogrešaka** bude minimalan. Grafikon 10.2 prikazuje takav pravac. Pogreške su prikazane pomoću tankih vertikalnih linija koje spajaju točke na pravcu i stvarne vrijednosti podataka. Pogreške se kvadriraju da se izbjegnu negativne vrijednostii te da se pojača utjecaj velikih pogrešaka. Statistički programi izračunavaju vrijednosti koeficijenta smjera i slobodnog člana koji minimiziraju zbroj kvadrata pogrešaka. Sama metoda se naziva **"metodom najmanjih kvadrata"**.

U sljedećem potpoglavlju je detaljno prikazana metoda najmanjih kvadrata uz podršku Microsoft Excela, koja u primjeru soka Dr. Pepper rezultira s koeficijentom smjera -14.13 dolara i slobodnim članom od 132.05. Interpretaciju ne treba uvijek shvatiti doslovno: na primjer, temeljem procijenjenih vrijednosti lako bi se (neispravno) zaključilo kako bi uz cijenu $p = 0$ potražnja za sokom bila 132.05. Naglasak se ovdje stavlja na koeficijent smjera koji pokazuje da za svako povećanje cijene od 1 dolara možemo očekivati pad *potražnje veći od 14 paketa*. Negativni predznak sugerira smanjenje.

10.3 Značenje koeficijenta smjera i slobodnog člana

U poglavlju se na primjeru dijetalnog soka Dr. Pepper demonstrira primjena Excelovih funkcija za izračun koeficijenta smjera i slobodnog člana regresijskog pravca. Zatim se testira statistička značajnost koeficijenta smjera i slobodnog člana.

10.3.1 Regresija u Excelu

Vrijednosti koeficijenta smjera i slobodnog člana određuju se Excel alatom Data Analysis | Regression option. Unose se podaci za obilježja odnosno varijable x i y. U primjeru je cijena nezavisna varijabla x, a potražnja je varijabla y zavisna o x. Definiramo područje gdje će se ispisati rezultat, a zatim pokrenemo alat regresije koji daje opširan rezultat. Dio rezultata je prikazan u tablici 10.2.

	Koeficijenti	*Standardna pogreška*	*t-statistika*	*p-vrijednost*
Slobodni član	142.05	12.87	10.26	< 0.0001
Cijena	14.13	3.62	-3.90	0.0045

Tablica 10.2. Koeficijent smjera i slobodni član za podatke o dijetalnom soku Dr. Pepper

Tablica prikazuje vrijednosti koeficijenta smjera i slobodnog člana. Ostatak prikazanih vrijednosti slijedi u nastavku.

10.3.2 Testiranje koeficijenta smjera i slobodnog člana

Kada su poznate vrijednosti koeficijenta smjera i slobodnog člana, postavlja se pitanje jesu li oni statistički značajni. Da bi se to provjerilo, najprije je potrebno prikazati linearni model koji daje vezu između x i y za cijelu populaciju:

$$Y = \alpha + \beta X. \tag{10-3}$$

Ovdje x i y predstavljaju vrijednosti za cijelu populaciju u smislu slučajne varijable. Član α predstavlja slobodni član, a β predstavlja koeficijent smjera u smislu uniformnih slučajnih varijabli za cijelu populaciju. Slijede hipoteze za testiranje statističke značajnosti koeficijenta smjera i slobodnog člana:

$$H_0\colon \alpha = 0;\ H_A\colon \alpha \neq 0 \tag{10-4}$$

$$H_0\colon \beta = 0;\ H_A\colon \beta \neq 0 \tag{10-5}$$

Izraz (10-4) prikazuje test značajnosti slobodnog člana, a izraz (10-5) test značajnosti koeficijenta smjera. Test značajnosti slobodnog člana je važan u smislu minimiziranja zbroja kvadrata pogrešaka, ali u testiranju postojanja veze između varijabli x i y, statistička značajnost slobodnog člana nije predmet interesa. Važan je test statističke značajnosti koeficijenta smjera. U izrazu (10-5) H_0 pretpostavlja da je koeficijent smjera jednak nuli. To znači da slučajna varijabla Y nije osjetljiva na promjenu slučajne varijable X. Hipoteza H_A pretpostavlja da je slučajna varijabla Y osjetljiva na promjene varijable X.

Testovi značajnosti koeficijenta smjera i slobodnog člana hipoteze su uvijek dvosmjerni testovi, zadani s dva gore navedena para hipoteza. Zbog ove dosljednosti hipoteze nije potrebno uvijek navoditi. To omogućuje Excel ili bilo koji drugi program koji se koristi kako bi se prikazali odgovarajući statistički podaci za interpretaciju. Potrebno je usporediti hipotetske vrijednosti s procijenjenim vrijednostima slobodnog člana a i koeficijenta smjera b. Testna veličina za slobodni član:

$$t = \frac{a - \alpha}{se_a} = \frac{a - 0}{se_a} = \frac{a}{se_a} \tag{10-6}$$

Testna veličina za koeficijent smjera je:

$$t = \frac{b - \beta}{se_b} = \frac{b - 0}{se_b} = \frac{b}{se_b} \tag{10-7}$$

Pretpostavka da su α i β jednaki nula uvrštava nulu u obje gore navedene jednadžbe i daju omjere procjena slobodnog člana a i koeficijenta smjera b podijeljene s vlastitim standardnim pogreškama. Krajnji rezultat je t -statistika koja podsjeća na t - vrijednost definiranu formulom (6-18).

Rezultati ovih dvosmjernih testova također su prikazani tablicom 10.2. Kad se procjene koeficijenta smjera i slobodnog člana, izračunate na uzorku, podijele vlastitim standardnim pogreškama, dobije se odgovarajuća t - statistika i može se izračunati pripadna p - vrijednost, koja je kod slobodnog člana <0.0001. Ovdje test odbacuje H_0 i slijedi da je slobodni član statistički značajan. t-statistika procjene koeficijenta smjera je -3.90, a odgovarajuća p - vrijednost iznosi 0.0045. To govori da je veza između potražnje i cijene statistički značajna.

10.4 Procjena / Predviđanje

Neka u primjeru dijetalnog soka Dr. Pepper zadovoljava činjenica da je veza između cijene i potražnje statistički značajna. Tada se vrijednost potražnje za neku unaprijed određenu razinu cijena određuje formulom:

$$potražnja(cijena) = 132.05 - 14.13 \cdot cijena \tag{10-8}$$

Budući je model izgrađen na temelju podataka minimalne cijene od 3.38 dolara i maksimalne cijene od 3.72 dolara, ovaj model se smije koristiti samo za vrijednosti cijena u tom intervalu. Ako se uvrsti cijena izvan tog intervala, procjena postaje nagađanje. Procijenjivanje veličine x vrijedi samo u slučaju kada je $x_{min} \leq x \leq x_{max}$.

Koeficijent determinacije R^2 je udio varijabilnosti zavisne varijable y koja je posljedica varijabilnosti nezavisne varijable x i mjeri prediktivnu sposobnost modela. Vrijednost R^2 može biti najmanje nula i najviše jedan. Što je veća vrijednost R^2 to je veći udio varijabilnosti zavisne varijable y koja je objašnjena varijabilnošću nezavisne varijable x. U primjeru s dijetalnim sokom vrijednost R^2 jednak je 0.6553 i uključen je u ispisu rezultata Excelom. To znači da cijena objašnjava 65.53% varijabilnosti potražnje, dok 34.47% varijabilnosti potražnje nije objašnjeno modelom. U skladu s tim, treba ublažiti očekivanja o kvaliteti prediktivne sposobnosti jer velik dio varijabilnosti potražnje objašnjavaju neke druge varijable.

10.5 Zaključak

Linearna regresija je alat za proučavanje odnosa među varijablama i najvažnije je jesu li dvije varijable povezane. Ako su povezane, onda se odbacuje H_0 i onda i samo onda se smije nastaviti u smjeru predviđanja. Ukratko, značajnost je mnogo važnija od predviđanja.

Višestruka linearna regresija proučava više nezavisnih varijabli i tako povećava prediktivne sposobnosti pa se vrijednost R^2 u modelu povećava.

Jednostavna linearna regresija uglavnom se koristi za ispitivanje povezanosti zavisne s nezavisnom varijablom, dok višestruka linearna regresija proučava više varijabli i time povećava prediktivne sposobnosti modela.

10.6 Zadaci

1. Ranije u ovom poglavlju, a vezano za primjer s dijetalnim sokom

Dr. Pepper, spomenuto je da povećanje veličine uzorka smanjuje $p-$ vrijednost koeficijenta smjera. Zašto je tome tako?

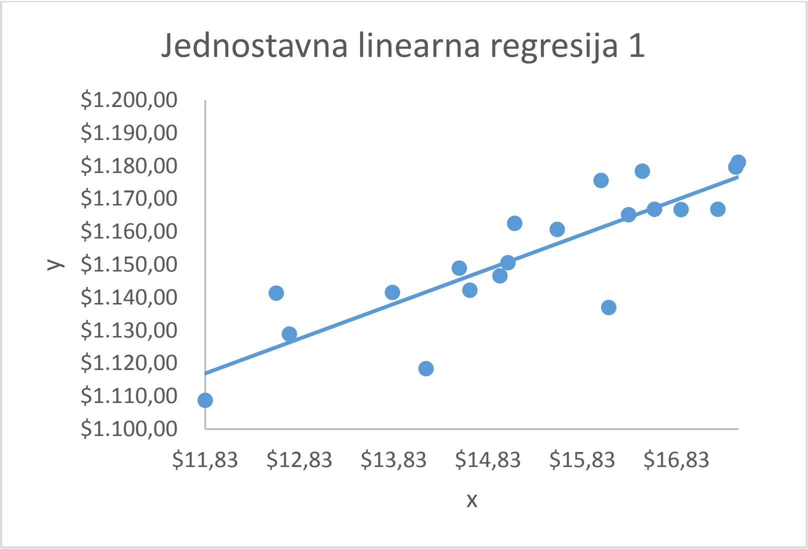
2. Skup podataka “RezultatiIspita” ima dva stupca. Stupac "Ispit 1" prikazuje rezultate ispita 1 za određene studente. Stupac "Ispit 2" prikazuje rezultate ispita 2 za te iste studente. Objasnite kako se može koristiti jednostavna linearna regresija kao alat za proučavanje uspjeha studenata.
3. Koristeći skup podataka “RezultatiIspita” ispitajte jesu li rezultati ispita 1 povezani s rezultatima ispita 2.
4. Koristeći skup podataka “RezultatiIspita” odredite R^2 s rezultatom na ispitu 1 kao nezavisnom varijablom i rezultatom na ispitu 2 kao zavisnom varijablom.
5. Koristeći skup podataka “RezultatiIspita” odredite R^2 s rezultatom na ispitu 2 kao nezavisnom varijablom i rezultatom na ispitu 1 kao zavisnom varijablom.
6. Kako uskladiti rezultate iz gornja dva zadatka?
7. Koji rezultat očekujete, koristeći skup podataka “RezultatiIspita”, da će student ostvariti na ispitu 2, ako ostvari 83 boda na ispitu 1?
8. Koristeći skup podataka “NovineOglasi”, odredite vezu između Oglasi/mjesec i Kupci/mjesec. Ima li ta veza smisla?
9. Koristeći skup podataka “NovineOglasi” procijenite broj kupaca u jednom mjesecu koji obilazi trgovine kada je poznato da je napravljeno 15 oglasa na mjesec. Jeste li zadovoljni s tom procjenom? Zašto?
10. Koristeći skup podataka “NovineOglasi” procijenite broj kupaca koji obilazi trgovine u jednom mjesecu kada je poznato da je napravljeno 23 oglasa na mjesec. Jeste li zadovoljni s tom procjenom? Zašto?
11. Kolika je prediktivna sposobnost, korištenjem skupa podataka “NovineOglasi”, ovog modela?
12. Koristeći skup podataka “BejzbolPlaće2014” odredite vezu između plaće cijele ekipe i postotaka pobjede. Ima li ta veza smisla? Iznenađuju li vas ovi rezultati? Zašto?
13. Koristeći skup podataka “BejzbolPlaće2014” procijenite postotak pobjeda ekipe s godišnjom plaćom od 150000 dolara. Jeste li zadovoljni

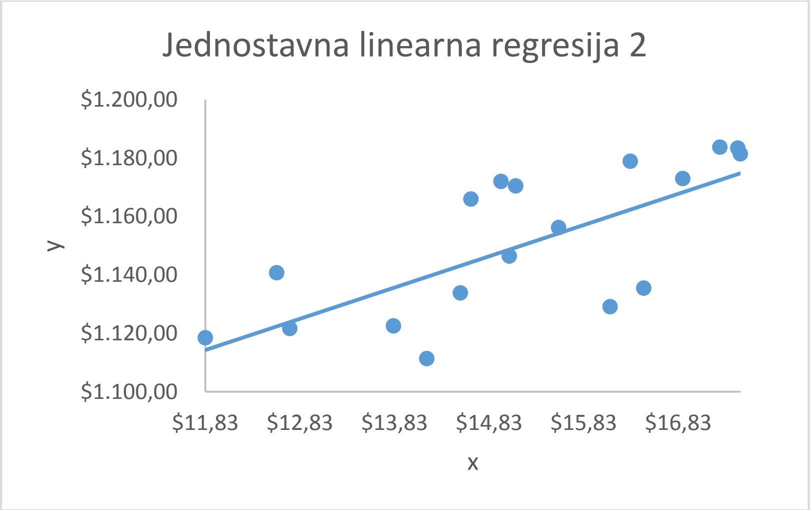
s tom procjenom? Zašto?

14. Koristeći skup podataka “BejzbolPlaće2014” procijenite postotak pobjeda ekipe s godišnjim plaćom od 300000 dolara. Jeste li zadovoljni s tom procjenom? Zašto?

15. Korištenjem skupa podataka “BejzbolPlaće2014” opišite prediktivnu sposobnost modela.

16. Pogledajte dva grafa regresije prikazana ispod. Na svakom grafu je prikazan regresijski pravac najmanjih kvadrata zajedno s podacima. Uspoređujući dva grafa, odredite koji ima veći R^2. Objasnite svoj odgovor.




Jednostavna linearna regresija 2
$1.200,00
$1.180,00
$1.160,00
$1.140,00
$1.120,00
$1.100,00
y
$11,83
$12,83
$13,83
$14,83
$15,83
$16,83
x

11. Analiza korištenjem stabla odlučivanja

Životne odluke mogu biti jednostavne, ali uglavnom su teške. U ovom poglavlju se bavimo teškim odlukama, uglavnom kvantitativnim odlukama. Jednostavna je odluka o izboru između dva posla, A i B, ako su poslovi A i B jednaki u svakom pogledu osim u smislu naknade. Odluka je trivijalna jer se prihvaća posao koji nudi višu naknadu.

Posao A nudi višu naknadu nego posao B, ali posao B izgleda obećavajuće u smislu profesionalnog razvoja. Koja je vjerojatnost za razvoj u poslu B? Koliki razvoj nudi posao B? Odjednom je ova odluka postala manje trivijalna zbog neizvjesnosti povezane s budućim profesionalnim mogućnostima razvoja. Ovo je primjer teške odluke.

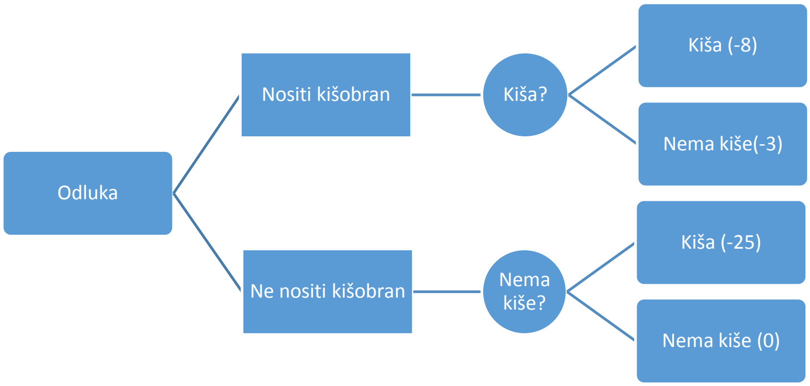
Ovo poglavlje opisuje neke alate koji se koriste u analizi odluka koje uključuju nesigurnost.

11.1 Stabla odlučivanja

Za donošenje odluka koje uključuju nesigurnost najprije se sastavi popis mogućnosti ili **alternativa**. Neka je na popisu n jedinstvenih alternativa i neka svaka alternative $a_i, i = 1, \ldots, n$ ima m mogućih, ali nesigurnih **ishoda.** Svaki ishod ima određenu vjerojatnost pojavljivanja $p_j, j = 1, \ldots, m$. Nadalje, svaki ishod ima odgovarajući **profit**, **isplatu ili korisnost** P_{ij} koji se odnosi na alternativu i i ishod j.

Sve se informacije omogućuju konstruiranjem stabla odlučivanja kao grafičkog prikaza svih alternativa, pripadajućih ishoda i isplata. U prikazu stabla čvor označen pravokutnikom predstavlja čvor odlučivanja ili alternative odlučivanja (mjesto gdje odluka treba biti donesena). Okrugli čvor predstavlja čvor okolnosti ili čvor mogućnosti. tj. čvor i stanje koje nije pod kontrolom donositelja odluke iz kojeg izviru grane za različite moguće ishode. Svaki ishod ima određenu vjerojatnost pojavljivanja i isplate izražene u novčanim jedinicama ili nekim drugim jedinicama.

Za odluku o tome treba li ili ne ponijeti kišobran na posao u jutarnjim satima postoje dvije alternative: ponijeti kišobran ili ne ponijeti kišobran. Za svaku alternativu postoji nesigurnost: kiša će padati ili kiša neće padati. Odluka se donosi pod utjecajem mogućih ishoda. Problem je predstavljen stablom odlučivanja i prikazan je grafikonom 11.1.



Grafikon 11.1. Stablo odlučivanja za primjer s kišobranom

U ovom problemu postoje četiri ishoda: (1) nosimo kišobran i kiša pada; (2) nosimo kišobran i kiša ne pada; (3) ne nosimo kišobran i kiša pada i (4) ne nosimo kišobran i kiša ne pada. Moguće isplate povezane s tim rezultatima su: -8, -3, -25 i 0. Ovdje isplate predstavljaju **korisnosti** subjektivno iskazane nenovčanim jedinicama (pojam korisnosti čest je u ekonomiji).

11.2 Strategije odlučivanja

Odluka (izbor alternative) se donosi tek kad je poznato što se želi ostvariti. Prije donošenja odluke treba osmisliti strategiju. U analizi odlučivanja postoji više strategija, a slijede tri najosnovnije.

11.2.1 Optimistična strategija (Maximax)

Optimistična pretpostavka je da će se dogoditi najbolje od najboljeg. Za svaku alternativu bira se najpovoljniji ishod, a zatim se bira alternativa s najboljim od najpovoljnijih ishoda. U našem primjeru s kišobranom, dobiva se:

$$max[max(-8,-3), max(-25,0)], \qquad (11\text{-}1)$$

Pojednostavnjenjem se dobiva:

$$max(-3, 0) = 0 \tag{11-2}$$

Dobivena **maximax** vrijednost 0 je isplata povezana s nenošenjem kišobrana jer je izabran ishod koji ima najveću korisnost. Kišobran se ne nosi zato jer se misli da neće padati kiša.

11.2.2 Pesimistična strategija (Maximin)

Pesimistična pretpostavka je da će se dogoditi najgore, a onda u skladu s tim za svaku alternativu pretpostavlja se najnepovoljniji ishod, nakon čega se bira alternativa koja daje najpovoljniji od najnepovoljnijih ishoda. U primjeru s kišobranom tako je:

$$max[min(-8, -3), min(-25, 0)] \tag{11-3}$$

Pojednostavnjenjem se dobiva:

$$max(-8, -25) = -8 \tag{11-4}$$

Dobivena **maximin** vrijednost -8 je korisnost koja se odnosi na nošenje kišobrana kad pada kiša. To znači da je najbolja pesimistična odluka ponijeti kišobran (jer se pretpostavlja da će padati kiša).

11.2.3 Strategija očekivane vrijednosti

Dvije spomenute strategije su važne strategije, a rezultati su očigledni. Optimist neće ponijeti kišobran jer misli da neće padati kiša, a pesimist će ponijeti kišobran jer misli da će padati kiša.

Druga strategija primjenjuje teoriju vjerojatnosti i temelji se na očekivanoj vrijednosti svake alternative. Očekivana vrijednost alternative a_i, u oznaci $E(a_i)$ i iznosi:

$$E(a_i) = \sum_{j=1}^{m} p_j \, P_{ij} \tag{11-5}$$

Neka je u primjeru s donošenjem odluke o nošenju kišobrana vjerojatnost kiše jednaka 0.25. Tada je vjerojatnost da neće padati kiša 0.75.

Za alternative nošenja kišobrana očekivana vrijednost je:

$$(0.25)(-8) + (0.75)(-3) = -4.25 \tag{11-6}$$

Za alternativu bez nošenja kišobrana očekivana vrijednost je:

$$(0.25)(-25) + (0.75)(0) = -6.25 \tag{11-7}$$

Uspoređujući očekivane vrijednosti dviju alternativa, uzima se ona najveće očekivane vrijednosti:

$$max(-4.25, -6.25) = -4.25 \tag{11-8}$$

Najbolja alternativa ima očekivanu vrijednost -4.25 i odnosi se na alternativu "bez kišobrana". Koristeći metodu očekivane vrijednosti odluka bi bila ne nositi kišobran. Vrijednost -4.25 naziva se **očekivanom vrijednosti**.

11.3 Očekivana vrijednost savršene informacije

Očekivana vrijednost (engl. *Expected Value*, EV) dobivena u (11-8) naziva se **očekivanom vrijednosti u uvjetima neizvjesnosti** (eng. *Expected Value Under Uncertainty*, EVUU) budući nije bilo sigurno hoće li kišiti ili ne. Kada bismo unaprijed znali da će padati kiša, ponijeli bi kišobran, a ako bismo unaprijed znali da neće padati kiša, ne bismo ponijeli kišobran. S matematičkog pristupa tom problemu, **očekivana vrijednost u uvjetima izvjesnosti** (engl. *Expected Value Under Certainty*, EVUC) bi bila:

$$EVUC = \sum_{j=1}^{m} p_j \cdot \max_i(P_{ij}) \tag{11-9}$$

Neka je u primjeru s kišobranom unaprijed poznato da će padati kiša. Tada se odabire nošenje kišobrana jer je korisnost veća od korisnosti nenošenja kišobrana kad pada kiša ($max[-8, -25] = -8$). Ako je unaprijed poznato da neće padati kiša, odluka o nenošenju kišobrana ima veću korisnost ($max[-3, 0] = 0$). Budući kiša pada u 25% vremena, a ne pada u 75% vremena, očekivana vrijednost u uvjetima izvjesnosti jednaka je:

$$EVUC = -8 \cdot 0.25 + 0 \cdot 0.75 = -2 \tag{11-10}$$

Drugim riječima, ako je unaprijed poznato što će se dogoditi, možemo očekivati korisnost jednaku -2.

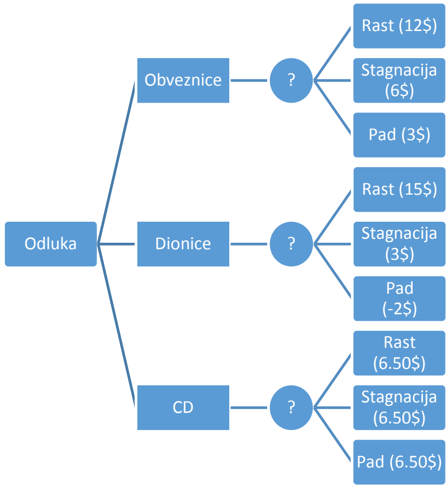
Zaključno, očekivana vrijednost u uvjetima neizvjesnosti (EVUU) jednaka je -4.25. U uvjetima izvjesnosti očekivana vrijednost (EVUC) jednaka je -2. To se naziva Razlika između tih vrijednosti naziva se očekivanom vrijednosti savršene informacije (engl. *Expected Value of Perfect Information*, EVPI):

$$EVPI = EVUC - EVUU. \tag{11-11}$$

U našem primjeru je EVPI jednaka $-2-(-4.25)=2.25$. Vrijednost je uvijek pozitivna, budući predstavlja razliku između uvijek veće očekivane vrijednosti uslijed sigurnosti i uvijek manje očekivane vrijednosti uslijed nesigurnosti (neizvjesnosti). EVPI predstavlja maksimalni iznos koji bi trebalo platiti kako bi se uklonila nesigurnost (neizvjesnost).

11.4 Primjer

Uzmimo u obzir monetarni primjer s tri investicijske alternative: obveznice, dionice i certifikate o depozitu (CD). Svaka alternativa ima tri moguća ishoda: rast, stagnacija i pad. Pretpostavimo da je vjerojatnost rasta 50%, vjerojatnost stagnacije 30% i vjerojatnost pada 20%. Grafikon 11.2 prikazuje detalje ovog problema što uključuje isplate u obliku stabla odlučivanja.



Grafikon 11.2. Stablo odlučivanja za primjer s investicijama

Tablica 11.1 prikazuje ove podatke u tabličnom obliku radi lakšeg računanja.

Alternative	Ishodi			Rezultati		
	Rast ($p = 0.5$)	Stagnacija ($p = 0.3$)	Pad ($p = 0.2$)	Max	Min	EVUU
Obveznice	12$	6$	3$	12$	3$	8.4$
Dionice	15$	3$	-2$	15$	-2$	8$
CD	6.5$	6.5$	6.5$	15$	6.5$	6.5$
Max vrijednost	15$	6.5$	6.5$	15$	6.5$	8.4$
Odluka	N/A			Dionice	CD	Obveznice

Tablica 11.1. Detalji odlučivanja u monetarnom primjeru

Za svaku alternativu, maksimalnu, minimalnu i očekivanu vrijednost određene su isplate. Optimist će izabrati dionice jer je 15 dolara najbolji mogući ishod koji će se dogoditi zbog te odluke. Pesimist će primjenom maximin metode odabrati CD jer je 6.5 dolara najbolji od najgorih ishoda koji će se dogoditi zbog te odluke. Kada se pogleda očekivana vrijednost, obveznice su najbolje rješenje jer one daju najveću očekivanu vrijednost (8.4 dolara).

Očekivana vrijednost u uvjetima izvjesnosti računa se na sljedeći način:

$$EVUC = 15 \cdot 0.5 + 6.5 \cdot 0.3 + 6.5 \cdot 0.3 = 10.75 \qquad (11\text{-}12)$$

Analogno je dobivena očekivana vrijednost u uvjetima neizvjesnosti u iznosu od 8.4 dolara. Budući je očekivana vrijednost u uvjetima izvjesnosti 10.75 dolara, očekivana vrijednost savršene informacije računa se kao njihova razlika i iznosi $10.75 - 8.4 = 2.35$ dolara. Monetarno tumačenje je da vrijednost eliminiranja neizvjesnosti iznosi 2.35 dolara. Ako zaposlimo nekog tko može ukloniti neizvjesnost, njegove usluge možemo platiti najviše 2.35 dolara.

11.5 Zaključak

Analiza odlučivanja je važna tema koja spaja teorijski i primijenjeni svijet. Stablo odlučivanja stvara strukturu odluke u smislu mogućnosti, povezanih neizvjesnosti i njihovih međusobnih odnosa i važno je za osmišljavanje odluke.

Isplate ili korisnosti podvrgnute su interpretaciji. U primjeru s nošenjem kišobrana kiša bez kišobrana je loša stvar za nekoga tko putuje na posao ili na nastavu. Kada nervoznog farmera u Južnoj Dakoti zabrinutog zbog sušenja kukuruza u kolovozu uhvati pljusak bez kišobrana, osjetit će olakšanje i bit će sretan. Isplate ili korisnosti treba interpretirati relativno.

U navedenim primjerima ishodi su diskretni, dogode se ili ne. U primjeru imamo ishode da kiša pada ili ne pada. Postoje slučajevi, posebno u gradovima kao što su Seattle, San Francisco ili London, gdje u magli nije sigurno pada li kiša ili ne i takve situacije nisu diskretne. Kvantificiranje isplata ili korisnosti ishoda je dio veće društvene znanosti poznate kao teorija odlučivanja/korisnosti.

Koncept očekivane vrijednosti savršene informacije je vrijednost eliminiranja neizvjesnosti. Kada bi se mogla eliminirati neizvjesnost, tada ne bi bilo razloga za izdavanje ove knjige. Ipak, organizacija i pojedinci bolje rade ako smanje neizvjesnost. Tvrtke koja pružaju financijske usluge reklamiraju istraživanja koja smanjuju neizvjesnost, poboljšavajući performanse portfelja. Tvrde da mogu smanjiti nesigurnost sugerirajući investitorima vrijednost reklamiranog istraživanja koje će smanjiti neizvjesnosti u procesu odlučivanja.

11.6 Zadaci

Za ove zadatke odredite što će optimist učiniti, odredite što će pesimist učiniti, odredite što će učiniti onaj tko koristi očekivanu vrijednost kao strategiju te odredite očekivanu vrijednost savršene informacije.

1. Imam nešto novca za ulaganje i imam tri investicijske mogućnosti: ChemicalBrothers, Mötley Crüe i ConcreteBlonde. ChemicalBrothers ima dobit od 10 dolara u povoljnim tržišnim uvjetima, a gubitak od 5 dolara u nepovoljnim tržišnim uvjetima. Mötley Crüe ima dobit od 7 dolara u povoljnim tržišnim uvjetima, a gubitak od 4 dolara u nepovoljnim tržišnim uvjetima. ConcreteBlonde ima dobit od 5 dolara u povoljnim tržišnim uvjetima te gubitak od 2 dolara u nepovoljnim tržišnim uvjetima. Povoljni tržišni uvjeti nastupaju s vjerojatnošću od 40%.
2. Dobili ste nešto novca za ulaganje i razmatrate tri različite opcije kao izbor ulaganja: AlphaStuds, BetaStuds i GammaStuds. U povoljnim tržišnim uvjetima AlphaStuds će vratiti 100 dolara, ali izgubiti 30 dolara u nepovoljnim tržišnim uvjetima. U povoljnim tržišnim uvjetima BetaStuds će vratiti 75 dolara, ali izgubiti 25 dolara u nepovoljnim tržišnim uvjetima. GammaStuds će vratiti 60 dolara u povoljnim tržišnim uvjetima, ali izgubiti 15 dolara u nepovoljnim tržišnim uvjetima. Povoljni tržišni uvjeti nastupaju s vjerojatnošću od 60%, a nepovoljno tržište s vjerojatnošću od 40%. Sav novac mora biti uložen u jednu od tri opcije - distribucija investicija između više investicija nije dopuštena.
3. Imam nešto novca koje želim uložiti. Investirat ću sav novac ili u RustStuds ili u JetElectro. RustStuds će vratiti dobit od 10 dolara u povoljnim tržišnim uvjetima i imati gubitak od 3 dolara u nepovoljnim tržišnim uvjetima. JetElectro će vratiti dobit od 14 dolara u povoljnim tržišnim uvjetima te gubitak od 5 dolara u nepovoljnim tržišnim uvjetima. Povoljni tržišni uvjeti nastupaju s vjerojatnošću od 60%, a nepovoljni s vjerojatnošću od 40%.
4. Imam nešto novca za ulaganje i razmišljam o tri moguća ulaganja: AcmeCorp, BestCorp i CoolioCorp. Kada je isplata povoljna, mogu očekivati 1000 dolara isplate od AcmeCorpa, 500 dolara isplate od

BestCorpa i 400 dolara isplate od CoolioCorpa. Kada je isplata nepovoljna, mogu očekivati gubitak od 300 dolara od AcmeCorpa, gubitak 200 dolara od BestCorpa i gubitak od 100 dolara od CoolioCorpa. Vjerojatnost povoljne isplate je 55%, a nepovoljne 45%.

5. Raspolažem s 1000000 dolara za ulaganje. Suzio sam svoje investicijske odluke na dvije moguće alternative: Alphatron i OmegaTron. Moram uložiti ukupno 1000000 dolara u jednu od opcija - dijeljenje ulaganja između dvaju vrijednosnih papira nije dopušteno. Uz povoljne uvjete na tržištu AlphaTron će osigurati povrat 100000 dolara, a OmegaTron povrat od 60000 dolara. Uz prosječne tržišne uvjete AlphaTron će osigurati povrat od 20000 dolara, a OmegaTron će osigurati povrat od 10000 dolara. Uz loše uvjete na tržištu AlphaTron će izgubiti 60000 dolara dok će OmegaTron izgubiti 30000 dolara. Procijenjeno je da će se povoljni uvjeti na tržištu pojaviti uz vjerojatnost od 40%, a prosječni i loši uvjeti na tržištu svako uz vjerojatnost od 30%.

Literatura

Budući da nikad prije nisam napisao knjigu, nisam u potpunosti siguran kako citirati literaturu. Moje znanje o statistici se nakupilo tijekom godina do točke u kojoj je većina onog što je ovdje napisano došlo iz iskustva. Prema tome, literatura koju navodim u nastavku su knjige koje volim, knjige koje sam koristio u prošlosti ili knjige u mojoj kolekciji koje sam u prošlosti smatrao važnima.

Albright, C., Winston, W., Zappe, C. "Data Analysis and Decision Making, 4th Edition." Cengage Learning. Cincinnati, OH. 2011.

Bowerman, B., O'Connell, R., Murphee, E. "Business Statistics in Practice." Irwin/McGraw-Hill. New York, New York. 2011.

Brightman, H. "Statistics in Plain English." Southwestern Publishing. Cincinnati, OH. 1986.

Gaither, N., Frazier, G. "Operations Management." Southwestern Publishing. Cincinnati, OH. 2002.

Sharpe, N., DeVeaux, R., Velleman, P. "Business Statistics: A First Course." Pearson Higher Education. New York, New York. 2014.

Sternstein, M. "Statistics." Barron's College Review Series. Hauppauge, New York. 1996.

Wonnacott, T.H., Wonnatott, R.J. "Introductory Statistics." John Wiley & Sons. New York, New York. 1990.

Pogovor prevoditeljica

Ovaj prijevod poznatog udžbenika *Undergraduate Business Statistics* autora Patricka R. McMullena (profesora na Poslovnoj školi američkog sveučilišta Wake Forest u Sjevernoj Karolini u Sjedinjenim Američkim Državama) je prvi naš pokušaj bavljenja ovim zahtjevnim poslom prevođenja stručne literature na hrvatski jezik. Prijevodom ovog udžbenika htjele smo ponuditi nešto drugačiji pristup temama iz područja statistike nego što je to uobičajeno. Stil pisanja prilagođen je studentima, a strogost matematičkog formalizma podređen je razumljivosti čitatelja i korisnika udžbenika. Materija koja se udžbenikom obrađuje u potpunosti ili u velikoj mjeri je usklađena s kolegijima iz statistike na stručnim studijima poslovnog usmjerenja i može poslužiti kao nadopuna dosadašnjoj literaturi.

Zahvaljujemo se autoru udžbenika Patricku R. McMullenu na odobrenju za korištenje njegovog udžbenika kao i svesrdnoj pomoći prilikom prevođenja.

Unaprijed se zahvaljujemo na svakom komentaru o propustima i nedosljednostima, a svaka konstruktivna primjedba i kritika je dobrodošla.

Kristina Devčić

Ana Perišić