Bojan Basrak

STATISTIKA

s primjerima u R-u

PMF Sveučilište u Zagrebu

Sadržaj

1 PODACI

Naše znanje o svijetu se razvija kroz proces postavljanja i opovrgavanja znanstvenih hipoteza. Suprotstavljene teorije ocjenjujemo na osnovu njihove sukladnosti s dostupnim podacima. Nažalost, podaci koje posjedujemo su uglavnom uvijek nepotpuni, i moramo računati s mogućnošću da ćemo u budućnosti susresti podatke koji će nas natjerati da promijenimo ili sasvim odbacimo prihvaćene teorije. Ne poznajemo li metode koje omogućuju usporedbu teorije i podataka, ne možemo razumijeti puno od suvremene znanosti.

Sličan problem imamo i u svakodnevnom životu kada god odlučujemo na osnovu nepotpunih opažanja. U takvim okolnostima donosimo gotovo sve privatne, javne ili poslovne odluke. Ignorirati podatke ni na ovom nivou se ne čini razumnim. Projektante građevina koji bi zapostavljali podatke o potresima i klimatskim uvjetima na danom području npr., morali bismo zvati i više nego nerazumnim. Liječnici su posebno često prinuđeni donositi važne odluke na osnovu nepotpunih opažanja o pacijentu i nepotpunih znanja o provedenim medicinskim istraživanjima. Članovi komisija za odobravanje novih lijekova donose odluke koje su još dalekosežnije. Danas su nerijetko i pacijenti u situaciji da na osnovu sličnih podataka mogu utjecati na izbor metode liječenja. Kako količina dostupnih podataka u svim ovakvim situacijama nezaustavljivo raste, i naše metode zaključivanja iz njih postaju važnije. Te metode proučava statistika.

Statistika se bavi podacima, posebno njihovim prikupljanjem, prikazom, i konačno zaključivanjem na osnovu podataka. Statistika je ozbiljna znanost i ima naravno i stručne termine za ove tri aktivnosti, pa ih nazivamo: **dizajn pokusa** (engl. experimental design), **deskriptivna statistika** (engl. descriptive statistics) i **statističko zaključivanje** (engl. statistical inference).

Prvom od ovih područja, iako je vrlo važno, posvećen je tek jedan dio ovog poglavlja. U njemu upozoravamo na česte greške u prikupljanju podataka. Deskriptivne statističke metode su tema cijelog idućeg poglavlja. Konačno, nakon što uvedemo termine teorije vjerojatnosti, sva ostala poglavlja su posvećena statističkom zaključivanju. Ovakva podjela je uobičajena, no može navesti na krivu pomisao da je statističko zaključivanje važnije od ostala dva područja ili da npr. prikupljanje podataka možemo shvatiti donekle olako. To nije tako. Sva tri statistička zadatka su podjednako važna i u primjenama i u znanosti. Čak ih i ne provodimo uvijek nužno ovim uobičajenim redom. Nakon prikaza inicijalnih podataka mogli bismo npr. dobiti novu ideju o tome kakvi nam podaci zapravo trebaju, nakon čega prikupljanje možemo ponoviti ili proširiti. Slično, nakon statističke analize u trećem koraku, možemo naslutiti kako jasnije prikazati podatke na grafu ili sl.

1.1 Uzorak i populacija

Podaci su ishodište svih statističkih metoda i ideja. Podacima zovemo skup mjerenja na grupi jedinki iz neke populacije. Pri tome **populacija** predstavlja sve one jedinke/jedinice o kojima nešto želimo znati. Primjeri su: skup svih studenata u Hrvatskoj

ili svih primjeraka neke biljne vrste na određenom području, ili čak skup svih pokusa koje bismo mogli izvesti pod istim uvjetima kao i određeni broj pokusa koje ćemo zaista izvesti. Primjetite ova posljednja populacija je tek hipotetska. Grupu jedinki na kojoj vršimo opažanja nazivamo **uzorak**. Katkad i skup svih prikupljenih mjerenja tijekom istraživanja zovemo uzorak. Cilj statističkog zaključivanja je odgovoriti na pitanja o cijeloj populaciji na osnovu uzorka s kojim raspolažemo.

Mjerenja odn. opažanja razlikujemo po tome jesu li kvantitativna ili kvalitativna. Prije nego prikupimo konkretne podatke uobičajeno je rezultate mjerenja zvati i varijablama. Za varijable čije vrijednosti izražavamo brojem kažemo da su **kvantitativne** ili **numeričke**. To su npr. prosjek ocjena studenata, sadržaj nekog vitamina ili minerala u plodovima biljaka i sl. Kod numeričkih varijabli dodatno razlikujemo diskretne i neprekidne varijable. Ova razlika postaje vrlo bitna kasnije kad predlažemo statistički model za svoje podatke. Za sad je dovoljno znati da diskretnim numeričkim podacima smatramo one kod kojih je jasno određena skala na kojoj vršimo mjerenja kao što je npr. broj položenih kolegija na studiju ili članova obitelji danog studenta. Ovdje nikakvu značajnu novu informaciju ne dobijamo profinjujući skalu, tj. mjerenjem ovih brojeva sa sve većim brojem decimalnih mjesta. Neprekidnim varijablama smatrat ćemo one kod kojih su preciznija mjerenja moguća i dodaju novu informaciju o uzorku. Takve su npr. visine studenata ili biljaka koje možemo mjeriti u centimetrima, milimetrima ili potencijalno još manjim jedinicama.

Kvalitativne ili **kategorijalne** varijable su one koje ne možemo prirodno predstaviti na nekoj numeričkoj skali. To su npr. spol studenata, genotip ili boja cvijeta za uzorkovane biljke, i tome sl. Takve se varijable nekada nazivaju i faktorima. Naglasimo da ako kategorije (npr. spolove) označimo brojevima (npr. 1 i 2) podaci i dalje ostaju kategorijalni iako su izraženi brojem. Uočimo još, na svakoj jedinki možemo mjeriti jednu odn. više varijabli pa razlikujemo i jednodimenzionalne odn. višedimenzionalne podatke, a pri tom možemo miješati kategorijalne i numeričke podatke.

Istraživači često pretpostavljaju neki teorijski model o ponašanju varijabli u cijeloj populaciji uključujući pri tom slučajnu komponentu. Podatke tada smatramo realizacijom takvog **slučajnog** ili **statističkog modela**, pri tome i ne precizirajući točnu populaciju iz koje dolazi. Slučajna komponenta modela tada predstavlja izvor varijabilnosti u podacima (npr. za nivo ekspresije nekog gena u stanicama). U tom slučaju cilj je statističkog zaključivanja odrediti što bolji statistički model za podatke vodeći pri tome računa o neizvjesnostima.

Bez obzira da li podatke promatramo kao dio mjerenja koje bismo potencijalno mogli obaviti na cijeloj populaciji ili kao realizaciju nekog slučajnog modela, svi statistički zaključci nužno će biti bar dijelom neizvjesni. Uzrok neizvjesnosti je činjenica da uzorak reprezentira tek dio populacije, odn. tek jednu od mnogo mogućih realizacija slučajnog modela. U ovom drugom slučaju moramo voditi računa i o tome da su korišteni modeli najčešće tek aproksimativni. **Neizvjesnost** je dakle osnovna karakteristika statističkih analiza, pa statistički zaključci nužno moraju izraziti i tu neizvjesnost.

Izuzetak od ovog pravila su opisne statističke metode, kod kojih ne pravimo velike pretpostavke i ne donosimo rigorozne zaključke.

1.2 Prikupljanje podataka

Pretpostavimo da želimo ispitati utjecaj koji obrok bogat šećerima može imati na rezultate studenata na testu iz matematike. Populacija koja nas interesira mogu biti svi studenti prve godine nekog sveučilišta. Uzorak može biti npr. $n = 200$ studenata koje smo izabrali na neki način, i zamolili da dođu na test ujutro natašte. Nakon toga studenti bivaju podjeljeni u dvije grupe – jedna dobiva obrok bogat šećerima, posebno glukozom, a druga npr. obrok bogat bjelančevinama. O studentima možemo imati više varijabli, npr.– spol, dob, regiju iz koje dolaze, uspjeh na maturalnom ispitu iz matematike, fakultet na kojem studiraju, itd. Te naravno, postignuti broj bodova na testu i vrstu obroka koji im je poslužen.

Možemo postaviti razna pitanja na ovoj grupi studenata. Npr. – da li razdioba spolova odgovora razdiobi spolova u ostatku generacije? – da li se razlikuju uspjesi na maturalnom ispitu po spolu, ili regiji iz koje dolaze? – da li je razdioba dobi ista na svim fakultetima? – da li su uspjeh na maturi i na testu povezani i koliko? itd.

Nas zanima pitanje – ima li vrsta obroka za prosječne studente utjecaj na rezultat na matematičkom testu? Usporedba različitih tretmana kao ovdje jedan je od najčešćih statističkih problema u mnogim znanostima. No za odgovor na naše pitanje važno je razmisliti i o tome kako su prikupljeni podaci.

Iako se korisnici statističkih metoda često pribojavaju grešaka u statističkoj analizi, mnoge greške zapravo nastaju već u procesu dizajniranja pokusa odn. prikupljanja podataka. Iako mi nećemo govoriti o ovom problemu vrlo detaljno, možemo upozoriti na česte tipove grešaka.

Nereprezentativni uzorak

Pretpostavimo da smo studente izabrali samo među studentima jednog fakulteta. Naš odgovor bi mogao biti obojan ovakvom odlukom. Npr. studenti nekog tehničkog fakulteta bi mogli biti svi izvrsni na sličnim testovima jer upravo ponavljaju gradivo koje pokriva test, pa ne bismo uočili bitnu razliku među grupama iako ona možda postoji u ostatku populacije.

Kontrolna skupina nije odgovarajuća

Ukoliko je test težinom i zahtjevima sličan testu iz matematike na maturi, mogli bismo i svim studentima podijeliti obrok bogat šećerom. Tada bismo njihove rezultate usporedili s rezultatima na maturalnom ispitu. No mi ne možemo znati koliki efekt ima to što je npr. prošlo pola godine od tog ispita, ili što studenti novi test pišu pod puno manjim pritiskom nego maturalni ispit. Postojanje kontrolne grupe (u medicini tipično placebo grupe) koja je po svemu usporediva s grupom pod tretmanom je dobar običaj, koji je značajno unaprijedio medicinska istraživanja. Iz istraživanja u kojem nema kontrolne skupine ili je ona neodgovarajuća, puno je teže donositi ozbiljne zaključke.

Krivi izbor kontrolne skupine često uvodi pristranost u podatke. Ako bismo studente podijelili u grupe po onom redu kojim dolaze na ispitivanje, to bi moglo značiti da

veći broj "ranoranioca" dodijelimo u prvu grupu. Što dakako može rezultate učiniti pristranima, baš kao i kad bismo studente podijelili po spolovima. Kako bi se izbjegla pristranost jedinke je u pokusu razumno **randomizirati** - odn. dodijeliti u grupe na slučajan način, ovakve pokuse na engleskom zovemo **randomized control trials**.

U opisanom pokusu studenti neizbježno znaju kakav su obrok pojeli, no ponekad se to može od njih i sakriti npr. dodjelom napitaka koji su zaslađeni pravim i umjetnim šećerima. U medicinskim istraživanjima to se čini dodjelom **placebo tretmana**. Time postižemo kontrolu tzv. placebo efekta. Kad je to moguće, razumno je sakriti vrstu tretmana i od onih koji provode istraživanje. Kod ispravljanja testa npr. moguće je da istraživači blaže odn. strože ocjenjuju neku od grupa, ovisno o tome kakav efekt žele pokazati. Zato je razumno testove šifrirati i slučajno ih podijeliti po ispravljačima. U medicini to znači tipično da ni medicinsko osoblje koje provodi istraživanje do konačne obrade rezultata ne zna tko je od pacijenata u placebo skupini a tko nije. Ovakve pokuse nazivamo i **double blind**.

Istraživači nemaju jasno postavljene hipoteze

Istaživači bi mogli obaviti pokus i mjerenja korektno te prikupiti veliki broj podataka, ali postaviti svoje hipoteze tek nakon inicijalne statističke analize. Tako bi npr. mogli otkriti da je obrok bogat šećerom imao pozitivan efekt na rezultate testa, no samo za studentice i to one koje dolaze iz kontinentalnih dijelova zemlje. Ovakav pristup istraživanju može biti zanimljiv, no ponovljeni pokus će rezultate rijetko replicirati. Ovakav način analize je karakterističan u području **data mininga**. Mogao bi zapravo tek poslužiti za postavljanje hipoteza koje bi trebalo potvrditi na novim nezavisno prikupljenim podacima.

Ovo su tek neke osnovne greške koje nastaju na početku – pri pokretanju istraživanja. Greške su moguće i u statističkoj analizi, npr.: korištenje neodgovarajućih statističkih metoda, kriva interpretacija vjerojatnosti (posebno uvjetnih), miješanje koreliranosti i kauzalnosti, itd. Pogrešna analiza se u načelu može lakše popraviti. S druge strane, prikupljanje podataka je često skuplji i bitno sporiji dio istraživanja, pa je oprez u tom postupku vrlo važan.

Zadaci

Zadatak 1. Kod istraživanja o utjecaju šećera na rezultate matematičkih testova opisanog na početku odjeljka 1.2 prikupljeno je više varijabli za svakog od studenata: težina, visina, dob, spol, broj bodova na maturalnom ispitu iz matematike, regija u kojoj je završena srednja škola, broj bodova na testu, trajanje studiranja u godinama. Razmislite koje su od ovih varijabli kategorijalne, a koje numeričke.

Zadatak 2. Pronađite u literaturi, na webu, ili na svom studiju bar jedno istraživanje koje se oslanja na statistiku i razmislite o mogućim greškama u dizajnu pokusa.

2 OPISNA STATISTIKA

Kod opisnih statističkih metoda ne brinemo o neizvjesnosti koja prati donošenje statističkih zaključaka. Ovdje ne pravimo ni bitne pretpostavke o slučajnim modelima koji generiraju naše podatke. Cilj je stoga manje ambiciozan, jasno i indikativno iskazati brojevima, grafovima i drugim sredstvima varijabilnost mjerenja u našem uzorku. Sva ova sredstva zovemo **opisnim statistikama**. Kako je varijabilnost jedinki pojednostavljeno govoreći i osnovni predmet interesa istraživača u gotovo svim prirodnim i društvenim istraživanjima, ovaj početni korak u statistici ne treba shvatiti olako.

2.1 Grafički prikaz kategorijalnih podataka

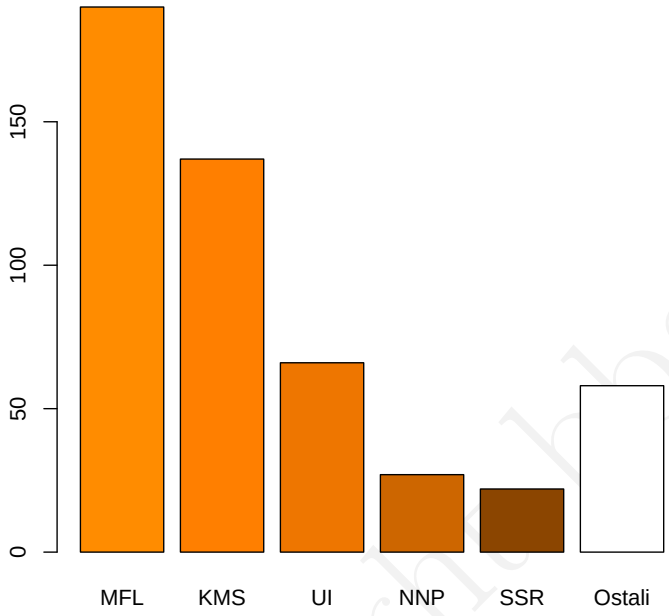
Na kategorijalne podatke poput onih o raspodjeli krvnih grupa ili regionalne pripadnosti osoba često nailazimo i u popularnim medijima. Njih je relativno jednostavno prikazati - naime ako prebrojimo jedinke u svakoj od kategorija dobijemo tzv. **apsolutne frekvencije** pojedinih kategorija. To je jednostavno broj jedinki iz uzorka koje pripadaju danoj kategoriji. Ako taj broj podijelimo s ukupnim brojem jedinki u uzorku dobivamo tzv. **relativne frekvencije** kategorija, tj.

$$\text{relativna frekvencija} = \frac{\text{apsolutna frekvencija}}{\text{veličina uzorka}}$$

Primjer 2.1.1 Istraživanje o političkom opredjeljenju provedeno je na uzorku od 500 ispitanika. Rezultati ispitivanja su dani u sljedećoj tablici.

Stranka	apsolutna frekvencija	relativna frekvencija
Moderna feministička liga (MFL)	190	0.380
Konzervativna muška stranka (KMS)	137	0.274
Umirovljenička inicijativa (UI)	66	0.132
Napredna narodnjačka partija (NNP)	27	0.054
Samostalna stranka rada (SSR)	22	0.044
Ostale stranke i neopredjeljeni	58	0.116
Ukupno	500	1

Uobičajeno je (u medijima npr.) relativne frekvencije izricati i u postocima, tako da vodeću stranku MFL prema podacima npr. podržava 38% ispitanika u uzorku. Primjetimo da je maksimalna relativna frekvencija 1 (ako sve jedinke pripadaju u istu kategoriju), a ne 100% kako bi možda netko očekivao. Mi ćemo relativne frekvencije računati po gornjoj formuli iz razloga koji će biti jasniji uvođenjem pojma vjerojatnosti. U znanosti naime vjerojatnosti uobičajeno izražavamo brojevima između 0 i 1, a ne postocima.



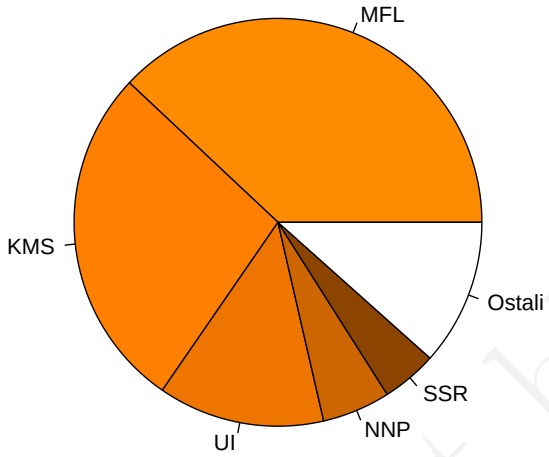
Slika 2.1: Stupčasti dijagram za podatke o popularnosti stranaka iz primjera 2.1.1.

Gornja tablica ilustrira još jednu čestu pojavu u istraživanjima, stranke s podrškom manjom od SSR nisu niti navedene u tablici. Oni ispitanici koji podržavaju takve stranke pridruženi su neopredjeljenim ispitanicima i čine kategoriju – ostali. Na taj način izgubljen je dio informacija koje su prikupljene istraživanjem. U prikazu podataka mi možemo odlučiti da nam kategorije malih frekvencija nisu osobito zanimljive, već nam je važnije sažetije prikazati podatke.

⌟

Kategorijalne podatke tipično grafički prikazujemo na dva načina, jedan je korištenjem tzv. **stupčastog dijagrama** (engl. bar-plot) koji može izražavati apsolutne ili relativne frekvencije. Primjer takvog dijagrama s apsolutnim frekvencijama za podatke iz primjera 2.1.1 je na Slici 2.1.

Alternativno iste podatke možemo prikazati i na tzv. **kružnom** ili **tortnom dijagramu** (engl. piechart). Na njemu je teže uočiti razlike u relativnoj frekvenciji kategorija. Stoga, iako tortne dijagrame jako često vidimo u popularnom tisku, istraživači uglavnom prednost daju stupčastim dijagramima. Slika 2.2 donosi tortni dijagram za podatke iz primjera 2.1.1.



Slika 2.2: Tortni dijagram za podatke o popularnosti stranaka iz primjera 2.1.1.

2.2 Grafički prikaz numeričkih podataka

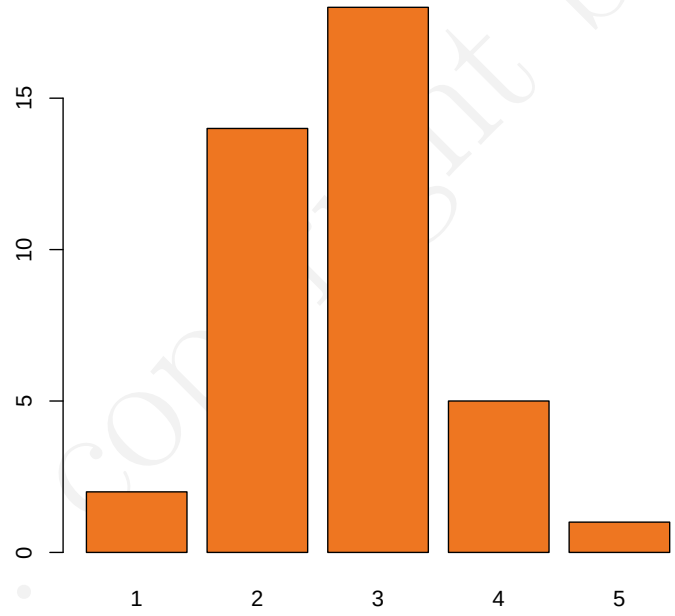
Diskretni numerički podaci

Ocjene koje su slučajno izabrani studenti u uzorku postigli na ispitu, iz statistike npr., su dakako numerička opažanja. No kako je mogućih ocjena relativno malo i sve su cjelobrojne, štoviše poprimaju vrijednosti samo u skupu $\{1, 2, 3, 4, 5\}$ smatramo ih diskretnim numeričkim podacima. Njih grafički također možemo prikazati stupčastim dijagramom kao i kategorijalne podatke. Bitna je razlika da su ovdje kategorije uređene, pa je stupce jedino razumno urediti od 1 do 5 (ili obrnuto). Ovakve podatke nekad nazivamo i ordinalnim. Za kategorijalne podatke poput onih iz primjera 2.1.1 poredak stupaca i nije toliko važan. ⊖

Primjer 2.2.1 Pretpostavimo da smo na uzorku od 40 mačaka različitih vlasnika opažali broj mačića koje je mačka okotila prvi prvom okotu. Dobili smo sljedeća opažanja

⌟

Broj mačića	apsolutna frekvencija	relativna frekvencija
1	2	0.050
2	14	0.350
3	18	0.450
4	5	0.125
5	1	0.025
Ukupno	40	1



Slika 2.3: Stupčasti dijagram za broj mačića pri prvom okotu – diskretni numerički podaci

Neprekidni numerički podaci

Podatke o visini studenata, težini ploda neke biljke ili trajanju putovanja možemo (u teoriji bar) mjeriti na sve finijoj skali. Tako težinu npr. možemo mjeriti u gramima, miligramima, mikrogramima, itd. Ovakve podatke smatramo **neprekidnim numeričkim podacima** i prikazujemo na više različitih načina, a najosnovniji je **histogram**.

Histogram je vrlo blizak stupčastom dijagramu. Razlika je da su podaci prikazani histogramom uvijek numerički, a mi ih tijekom procedure grupiramo u razrede. Procedura za izradu histograma može se podijeliti u nekoliko koraka.

▷ U uzorku $x_1, x_2, \ldots, x_n$ odredimo najmanju i najveću vrijednost (nazovimo ih $x_{\min}$ i $x_{\max}$), te željeni **broj razreda**, recimo k.
▷ Odredimo **razrede** $I_1, \ldots, I_k$, odn. disjunktne intervale oblika $I_j = [a_{j-1}, a_j)$ za neke brojeve $a_0, \ldots, a_i$, koje zovemo **granice razreda**. Pri tome vodimo računa da svaki podatak ulazi u točno jedan razred.
▷ Za svaki razred izračunamo njegovu apsolutnu odn. relativnu frekvenciju (u oznakama f_j odn $r_j = f_j/n$) jednostavno brojeći podatke koji u njega ulaze. Postupak možemo sažeti u obliku tablice sljedećeg oblika.

razred	granice	f_a	r_a
I_1	a_0, a_1	f_1	r_1
I_2	a_1, a_2	f_2	r_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	a_{k-1}, a_k	f_k	r_k

Tablica 2.1: Grupiranje numeričkih podataka u razrede kod izrade histograma

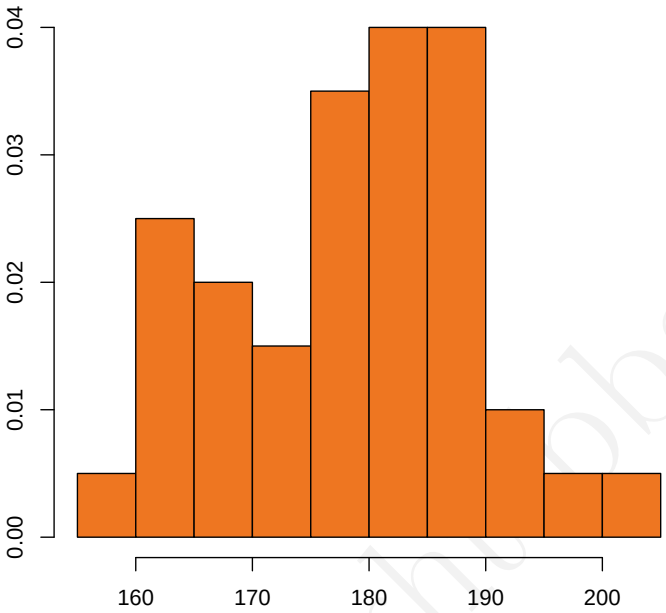
▷ Konačno nacrtamo graf na kojem iznad pojedinog razreda povučemo liniju na visini koja odgovara relativnoj frekvenciji tog razreda podijeljenoj s duljinom razreda. Dakle, iznad razreda I_j, liniju povlačimo na visini

$$\frac{r_j}{a_j - a_{j-1}}.$$

Vrlo često koristimo razrede jednake širine. Tada je dovoljno odrediti a_0 i a_k, a ostale a_j izračunamo tako da svi razredi budu širine $(a_k - a_0)/k$. Za takve razrede crtamo nekad i histogram apsolutnih frekvencija jednostavno povlačeći liniju na visini apsolutne frekvencije pojedinog razreda.

Konačni izgled dijagrama kod raznih statističkih paketa u praksi ovisi o granicama razreda, ali i o tretmanu podataka koji leže na granici dva razreda,. Statistički programi obično sami određuju broj i granice razreda (ostavljajući mogućnost korisniku da to izmijeni). Primjetite ipak da uvijek mora vrijediti $a_0 \leq x_{\min} \leq x_{\max} < a_k$. U R-u se npr. po defaultu koriste razredi jednake širine čiji je broj određen posebnim algoritmom, a procedura proizvodi histogram apsolutnih frekvencija ako korisnik drugačije ne odredi, vidi odjeljak 2.6. Primjetimo na kraju da preveliki i premali broj razreda mogu dati vrlo neinformativne histograme. ⊖

Prije raširene upotrebe računala za prikaz numeričkih podatka često je korišten tzv. dijagram debla i lista (engl. *stem and leaf*) (Tukey, 1977). On je jednostavnija (tekstualna i horizontalna) verzija histograma, u kojoj su razredi podijeljeni prema vodećim



Slika 2.4: Histogram za podatke s visinama studenata. Podaci se mogu iščitati u tablici 2.2 odn. u odjeljku 2.6. Predstavljaju visine muških studenata prve godine u cm, izmjerene pri inicijalnom liječničkom pregledu.

znamenkama, a stupci su izraženi nabrajanjem iduće značajne znamenke. Npr. podaci o visinama studenata sa slike 2.4 se ovim dijagramom mogu prikazati kao u tablici 2.2.

Grafički prikaz dvodimenzionalnih numeričkih podataka

Katkad istraživanjem prikupimo i više numeričkih podataka o svakoj jedinki - poput npr. prosjeka ocjena na studiju i iznosa prvog dohodka za uzorak bivših studenata. Tada bismo svaku od varijabli mogli posebno analizirati, no zanimljivo je nakon istraživanja nešto reći i o međuovisnosti ovih varijabli. Simbolički takvi podaci predstavljaju niz uređenih parova realnih brojeva

$$(x_i, y_i),\ i = 1, \ldots, n.$$

Njihova međuovisnost neće biti vidljiva ako varijable prezentiramo odvojeno. Kada se radi o dvije varijable podatke zovemo **dvodimenzionalnima** i uobičajeno ih prvo prikazujemo u tzv. **dijagramu raspršenosti** (engl. scatter-plot).

Zanimljiv skup podataka koji sadrži 2 numeričke varijable za svaku jedinku, nalazi se u R-u. Uzorak je nastao mjerenjima opsega i volumena 31 srušenog debla tzv. crne

Vodeće znamenke	najmanje značajna znamenka
15	9
16	1223
16	5899
17	012
17	567889
18	001223334
18	577788889
19	1
19	59
20	2

Tablica 2.2: *Stem and leaf* dijagram za podatke o visinama studenata – v. sliku 2.4

trešnje (*Prunus serotina*). Primjetite da volumen naravno raste kako se povećava opseg, no ta veza ne izgleda posve linearna. Slika sugerira, a to je čest slučaj, da jednu ili obje varijable u uzorku ima smisla i transformirati (npr. logaritmom u ovom slučaju), ako time veza između varijabli postaje jednostavnija ili je podacima lakše naći odgovarajući statistički model.

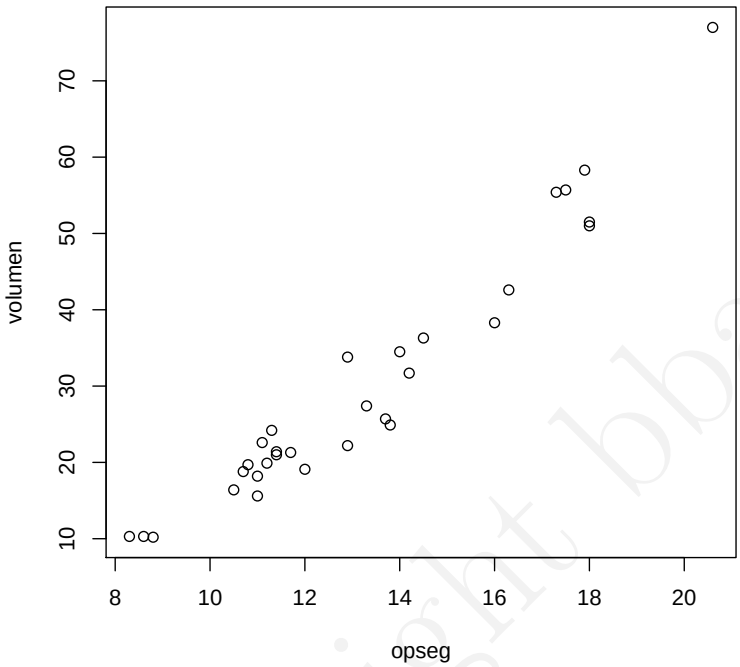
Ako promatramo dvije varijable od kojih je jedna numerička, a druga kategorijalna uzorak je smisleno prikazati usporednim *boxplot* dijagramima koje uvodimo kasnije u poglavlju, no primjer takvog dijagrama se može vidjeti na slici 2.7.

Ukoliko varijabli ima više od dvije, podatke nije uvijek lako prezentirati na papiru ili ekranu računala. To je pogotovo problem ako varijabli imamo zaista puno, što je sve češće slučaj u mnogim primjenjenim znanostima. Vrlo ilustrativan primjer takvih podataka su npr. podaci o ekspresiji gena dobiveni tzv. *DNA microarray* tehnikom, kada o svakoj jedinki o uzorku prikupljamo i više od 20 tisuća varijabli. Sličan problem bismo imali i ako bismo o bivšim studentima osim visine i dohodka znali i još desetke drugih socioekonomskih varijabli. Vizualizacija ovakvih podataka upotrebom računala je moguća, ali odabir zaista dobre metode ovisi o stručnosti i iskustvu istraživača. Ovom problemu posvećeno je puno pažnje u modernoj znanstvenoj literaturi. Inspirativni primjeri korištenja grafičkih deskriptivnih metoda mogu se pronaći u knjigama McCandlessa [4] i Wickhama [8].

Primjetimo na kraju da se gotovo sve procedure za grafički prikaz podataka dijelom zasnivaju i na subjektivnim odlukama istraživača.

2.3 Sredina uzorka

Nakon što smo prikupili numeričke podatke npr. o visini dohodka bivših studenata često bismo ovaj evt. dugi niz brojeva htjeli opisati korištenjem samo par brojčanih vrijednosti. Jedna takva vrijednost bi mogla indicirati gdje se nalazi *sredina* tog numeričkog uzorka.



Slika 2.5: Dijagram raspršenosti za opsege (u inčima) i volumene (u kubičnim stopama) stabala crne trešnje

Iako pojam sredine nismo jasno definirali, mnogi će se složiti da je **prosjek** ili aritmetička sredina jedan od najboljih izbora.

Aritmetička sredina uzorka $x_1, x_2, \ldots, x_n$ je

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

Ako bismo iste numeričke podatke prikupili i u ostatku populacije, aritmetička sredina nam, grubo govoreći, pokazuje oko koje vrijednosti možemo očekivati da se grupiraju ti novoprikupljeni podaci, stoga je ponekad nazivamo i očekivanjem uzorka.

Primjer 2.3.1 Za uzorak 6, 3, 3, 6, 3, 5, 6, 1, 4, 6, 3, 5, 5, 2, 2, 2, 2, 3, 2, 3, koji je Vilim student statistike , pretpostavimo, prikupio bacanjem kocke, aritmetičku sredinu možemo naći na sljedeći način

$$\bar{x} = \frac{1}{20}(1 \cdot 1 + 2 \cdot 5 + 3 \cdot 6 + 4 \cdot 1 + 5 \cdot 3 + 6 \cdot 4) = 3.6.$$

Razmislite koliki biste prosjek mogli očekivati ako kocku bacite veliki broj puta, npr. tisuću ili milijun? ⌟

Primjetite da je aritmetička sredina vrlo osjetljiva na ekstremne podatke u uzorku. Ako se neki poznati internacionalni nogometaš odluči uz karijeru studirati i upiše predmet Statistika, prosječni standard studenata, izražen npr. kao aritmetička sredina iznosa koje studenti godišnje mogu potrošiti, bi izuzetno skočio. Dakako za ostale studente se pri tom ništa značajno nije promijenilo. Ponekad je poželjna mjera za sredinu podataka koja je manje osjetljiva na ekstremne vrijednosti u uzorku odn. mjera koju možemo smatrati "robusnijom". Ako uredimo podatke $x_1, x_2, \ldots, x_n$ iz uzroka po veličini dobit ćemo uzlazni niz

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Ove brojeve uobičajeno zovemo **uređajnim statistikama**. Najveća i najmanja vrijednost u uzorku posebno zadovoljavaju

$$x_{\max} = x_{(n)} \quad \text{i} \quad x_{\min} = x_{(1)}.$$

Primjetite da se pojedini brojevi u ovom nizu mogu pojaviti i više puta. Za podatke iz primjera 2.3.1 uređajne statistike se nalaze u primjeru 2.3.2.

Medijan se opisno definira kao onaj broj za koji je 50% uzorka i manje ili jednako od njega, a 50% veće ili jednako od njega. Kako takvih brojeva može biti više, mi precizno definiramo **medijan** na sljedeći način:

▷ ako je n neparan, medijan je točno središnja uređajna statistika, tj.

$$m = x_{(\frac{n+1}{2})},$$

▷ ako je n paran, medijan je aritmetička sredina dvije središnje uređajne statistike, tj.

$$m = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Preferiramo li medijan ili aritmetičku sredinu kao mjeru centra našeg uzorka ponekad ovisi i o stavovima istraživača, posebno ako razmišljamo o ekonomiji. Naime ako usporedimo aritmetičku sredinu osobnih dohodaka na uzorku ispitanika u nekoj zemlji (npr. Hrvatskoj) s medijanom istog uzorka vidjet ćemo tipično da je aritmetička sredina značajno veća od medijana. Ta je razlika posljedica manjeg broja ekstremno velikih dohodaka. Slično, ako uspoređujemo današnje dohodke s onima od prije 20 ili 30 godina (svodeći ih pri tome na današnje dolare, npr. u SAD) vidjet ćemo da medijan dohodaka raste vrlo sporo (štoviše postoje indicije da u nekim relativno dugim periodima on čak i pada). Aritmetička sredina s druge strane pokazuje jasan rast (uz tek kraće periode pada), i stoga evt. ukazuje na ekonomski napredak. Zagovaratelji medijana će naravno upozoriti da je razlog tome enormni rast prihoda iznimno malog dijela populacije, te da se ne može govoriti o napretku sve dok medijan pada ili stagnira. Druga strana može argumentirati da se novostvoreni dolar treba prikazati kao napredak društva u cjelini bez obzira u čijem džepu se trenutno nalazi. Na taj način tehnička diskusija o određivanju centra razdiobe postaje prvorazredno pitanje političke ekonomije i na neki način odražava bitne razlike između suprotstavljenih političkih stavova.

Primjer 2.3.2 Ako rezultate bacanja kocke iz primjera 2.3.1 uredimo dobijamo niz:

$$1, 2, 2, 2, 2, 2, 3, 3, 3, \mathbf{3}, \mathbf{3}, 3, 4, 5, 5, 5, 6, 6, 6, 6.$$

Kako je $n = 20$ paran broj medijan je sredina između $x_{(10)} = 3$ i $x_{(11)} = 3$, dakle 3. ⌟

Kao **mod** uzorka $x_1, x_2, \ldots, x_n$ definiramo vijednost koja se pojavljuje s najvećom frekvencijom. Za naš uzorak iz primjera 2.3.1 i mod bi dakle bio 3. Primjetite da u uzorku diskretnih, no katkad i neprekidnih varijabli, mod može biti čak i najmanja ili najveća vrijednost u uzorku, dakle sasvim na njegovom rubu, a ne u centru. Naravno modova može biti i više. Iz ovih razloga mod se i ne koristi jako često kao mjera sredine uzorka.

2.4 Raspršenost uzorka

Lako je zamisliti dva jednodimenzionalna numerička uzorka iste duljine, istog medijana i aritmetičke sredine od kojih jedan pokazuje znatno veću varijabilnost od drugoga. Važnost varijabilnosti kod usporedbi uzoraka prikupljenih u različitim populacijama ilustrira Slika 2.6. Na njoj vidimo dvije usporedbe histograma dvaju uzoraka različitih aritmetičkih sredina. U oba slučaja jedan uzorak ima aritmetičku sredinu 51, a drugi 56. Dok se možemo sporiti o tome da li su uzorci bitno različiti u prvom slučaju, u drugom to slika prilično jasno sugerira.

Zbog ovakvih razloga je upravo varijabilnost (disperzija ili raspršenost) uzoraka predmet interesa i matematičkog modeliranja u statistici, pa i nju nekako moramo mjeriti.

Najjednostavnija mjera raspršenosti je vjerojatno raspon. **Raspon uzorka** je razlika najveće i najmanje vrijednosti, dakle

$$d = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Uočite: uzorci $1, 1, 1, 2, 10$ i $1, 4, 7, 8, 10$ imaju isti raspon, premda očito nisu jednako disperzirani. Zato ćemo tražiti bolju mjeru za disperziju.

Kvartili i kvantili uzorka

Bilo koji realan broj s izmedju 1 i n, možemo jednoznačno prikazati u obliku

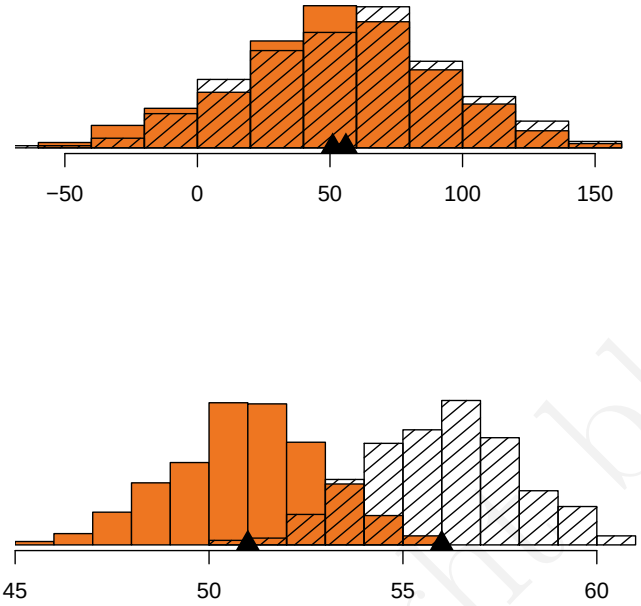
$$s = k + r$$

tako da je $k = 1, 2, \ldots, n$ njegov cijeli, a $0 \leq r < 1$ njegov razlomljeni dio. Koristeći gornji prikaz, oznaku $x_{(s)}$ možemo koristiti i za brojeve $1 \leq s \leq n$ koji nisu nužno cijeli tako da definiramo

$$x_{(s)} = (1 - r)x_{(k)} + rx_{(k+1)} = x_{(k)} + r(x_{(k+1)} - x_{(k)}). \tag{2.4.1}$$

Uz ove oznake, medijan i za parne i za neparne n možemo definirati jednim izrazom

$$m = x_{(\frac{1}{2}(n+1))}.$$



Slika 2.6: Usporedba histograma za dva para uzoraka sa istim aritmetičkim sredinama (označenim trokutićima). Uočite i bitno drugačije skale na x-osi.

Gornji i **donji kvartil** definiramo kao veličine

$$Q_3 = x_{(\frac{3}{4}(n+1))}, \quad Q_1 = x_{(\frac{1}{4}(n+1))}.$$

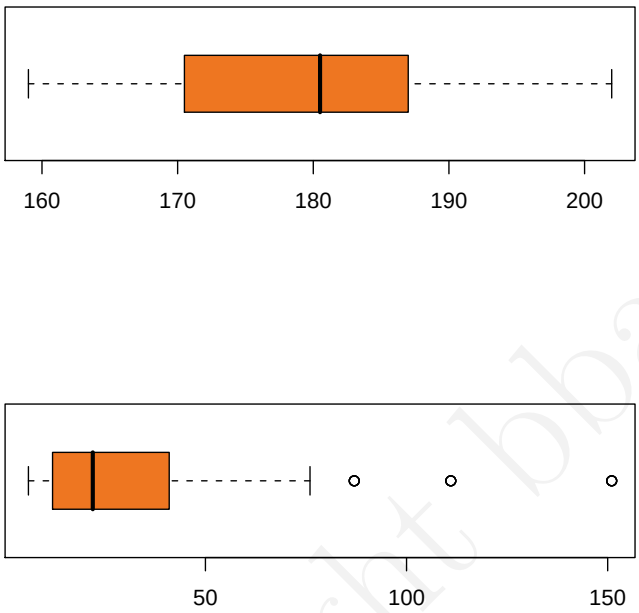
Udaljenost izmedju ova dva broja zovemo **interkvartilni raspon**

$$d_Q = Q_3 - Q_1.$$

Primjetite da interkvartilni raspon ovisi o mjernim jedinicama u kojima su izraženi naši podaci, da bismo to izbjegli za nenegativne podatke možemo koristiti koeficijent kvartilne varijacije

$$v_Q = \frac{d_Q}{Q_3 + Q_1}.$$

Postoji elegantan grafički prikaz numeričkih uzoraka koji koristi upravo uvedene veličine. Nazivamo ga **box and whiskers dijagram** (doslovno dijagram "kutije i brkova" ili pravokutni dijagram). On nastaje tako što na brojevnom pravcu označimo medijan m i oko njega izgradimo pravokutnik ("kutiju") od Q_1 do Q_3. Nakon toga s obje strane medijana pronađemo podatak koji se nalazi što dalje od njega, ali ne dalje od jednog i pol interkvartilnog raspona d_Q. Povučemo linije od tih podataka do kutije. Sve



Slika 2.7: Pravokutni dijagrami za podatke o visinama studenata (gore) i o broju epileptičkih napada (dolje)

podatke koji se nalaze izvan tog intervala oko medijana posebno označimo na dijagramu i smatramo "ekstremnim". Ovi podaci se na engleskom nazivaju *outliers*.

Ilustracija ovog dijagrama je na slici 2.7. Ovdje vidimo pravokutne dijagrame za dva uzorka. Jedan je već poznati uzorak visina studenata, a drugi se tiče broja epileptičkih napada unutar 8 tjedana za 59 pacijenata (skup podataka se nalazi i paketu MASS u R-u). Lako je uočiti veliki razliku između dva uzorka. U prvom su primjeru kutija i pogotovo brkovi relativno simetrični u odn. na medijan, u drugom su brkovi produljeni na desnu stranu, gdje se nalazi i nekoliko ekstremno velikih podataka. Za razdiobu drugog uzorka ćemo reći ponekad i da je desno nagnuta odn. da ima teži desni rep.

Brojeve

$$x_{\min}, Q_1, m, Q_3, x_{\max}, \quad \text{kao i } \bar{x}$$

možemo dobiti ako zatražimo tzv. pregledne (ili sumarne) statistike u statističkim programima. U odjeljku 2.6 je ilustriran izračun ovih statistika u R-u. Važno je istaći da se brojevi Q_1, Q_3 i m na istom uzorku mogu razlikovati ako su izračunati u raznim statističkim programima. Ipak, te razlike tipično nisu velike, Izvor ovih razlika je različita definicija tzv. kvantila slučajnog uzorka. Jedna jednostavnija definicija kvartila npr. nastaje tako da ponovo stavimo

$$Q_3 = x_{(\frac{3}{4}(n+1))}, \quad Q_1 = x_{(\frac{1}{4}(n+1))},,$$

a indekse koji nisu cijeli zaokružimo na najbliži cijeli broj između 1 i n.

Za proizvoljan $\alpha \in [1/(n+1), 1-1/(n+1)]$ definiramo α-**kvantil** uzorka $x_1, x_2, \ldots, x_n$, kao

$$q_\alpha = x_{(\alpha(n+1))}\,.$$

Mi pretpostavljamo dakle da je $s = \alpha(n+1)$ realan broj izmedju 1 i n. Kao i prije, da bismo našli $x_{(s)}$, rastavimo s na cijeli i razlomljeni dio i iskoristiti formulu (2.4.1). Uočite da vrijedi npr.

$$Q_1 = q_{0.25}\,,\ Q_3 = q_{0.75}.$$

Specijalno, ako je $\alpha = k/100$ za neki $k = 1, \ldots, 99$, kažemo da je q_α k%-tni **centil** (engl. percentile) uzorka. Npr. medijan je 50%-tni, a donji kvartil je 25%-tni centil uzorka.

Primjer 2.4.1 Uzorak iz primjera 2.3.1 je duljine $n = 20$, tako da je $(n+1)/4 = 5.25$, a $3(n+1)/4 = 15.75$. Stoga su donji odn. gornji kvantil ovih podataka

$$Q_1 = 0.75x_{(5)} + 0.25x_{(6)} = 0.75 \cdot 2 + 0.25 \cdot 2 = 2\,,$$

odnosno

$$Q_1 = 0.25x_{(15)} + 0.75x_{(16)} = 0.25 \cdot 5 + 0.75 \cdot 5 = 5\,.$$

Dok 5%-tni centil, kako $(n+1) \cdot 0.05 = 1.05$, iznosi

$$q_{0.05} = 0.95x_{(18)} + 0.05x_{(2)} = 0.95 \cdot 1 + 0.05 \cdot 2 = 1.05\,.$$

⌟

Varijanca i standardna devijacija uzorka

Sljedeća, vrlo često korištena mjera disperzije je varijanca. **Varijanca uzorka** $x_1, x_2, \ldots, x_n$ definirana je izrazom

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

Ona je uvijek nenegativna, baš kao i prethodne dvije mjere disperzije uočimo. Poprima vrijednost 0, ako i samo ako je i raspon 0, tj. ako su sve vrijednosti x_i jednake. Varijanca je vjerojatno najrašireniji mjera disperzije, pogledajte primjer 2.4.3 za korištenje varijance u ekonomiji.

Primjetite da varijanca mjeri prosječno kvadratno odstupanje podataka od aritmetičke sredine. Ona je dakle izražena u drugačijim mjernim jedinicama od samih podataka (u kvadratu originalnih jedinica). To možemo ispraviti koristeći njen pozitivan drugi korijen tj. standardnu devijaciju. **Standardna devijacija uzorka** definirana je na sljedeći način

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

Standardna devijacija uzorka ovisi o mjernoj jedinici u kojoj izražavamo naše podatke. Stoga se raspršenost za nenegativne podatke mjeri i tzv. koeficijentom varijacije uzorka

$$v = \frac{s}{\bar{x}}.$$

Primjer 2.4.2 Za podatke iz primjera 2.3.1, aritmetička sredina je 3.6, pa varijanca iznosi (približno)

$$\begin{aligned} s^2 &= \frac{1}{19}[1 \cdot (1-3.6)^2 + 5 \cdot (2-3.6)^2 + 6 \cdot (3-3.6)^2 \\ &+ \quad 1 \cdot (4-3.6)^2 + 3 \cdot (5-3.6)^2 + 4 \cdot (6-3.6)^2] \\ &= \quad 2.674 \end{aligned}$$

Tako da standardna devijacija iznosi

$$s = \sqrt{2.674} = 1.635.$$

⌟

Primjer 2.4.3 U financijama se varijanca i standardna devijacija uzorka koriste kao mjere za usporedbu rizika različitih ulaganja. Pretpostavite da želite uložiti novac u dionice a) tvornice keksa, b) farmaceutske kompanije ili c) turističkog poduzeća. Pretpostavite da ste kroz posljednjih 10 godina pratili godišnji relativni povrat na ove dionice i dobili sljedeće aritmetičke sredine

$$\bar{x}_a = 6\%, \bar{x}_b = 6\%, \text{ te } \bar{x}_c = 7\%\,.$$

Ako ne posjedujete neke dodatne informacije koje legalno (ili ne) možete koristiti pri izboru dionice, mogli bismo reći da dionice a) i b) predstavljaju ekvivalentni izbor, dok se dionica c) čini najboljom. Međutim iz podataka ste mogli izraziti i standardne devijacije. Ako one iznose redom

$$s_a = 4.5\%, s_b = 2.2\%, \text{ te } s_c = 5.0\%\,,$$

mogli bismo reći da je dionica b) zbog manje varijabilnosti povrata i manje rizična od ulaganja u a). Kako donose isti prosječni prinos, mogli bismo preferirati b) u odn. na a). S druge strane dionica c) ima najveći prosječni prinos ali i najveću varijabilnost. Konačan izbor u praksi nije jednoznačan. On ovisi i o individualnim sklonostima ulagača prema riziku. Ekonomisti posebno promatraju i tzv. Sharpov omjer $\bar{x}/s$ (što je zapravo vrijednost obrunto proporcionalna koeficijentu varijacije). On npr. za tri dionice u ovom primjeru iznosi redom

$$1.33, 2.72, \text{ te } 1.40\,,$$

i sugerira da je evt. upravo dionica b) najbolje ulaganje s obzirom na odnos prosječnog povrata i rizika.

⌟

Katkad nemamo dovoljno precizne numeričke podatke, pa umjesto egzaktnih mjerenja znamo tek da podaci pripadaju u pojedine razrede tj. intervale. Pitanje je kako odrediti aritmetičku sredinu, varijancu odnosno standardnu devijaciju za grupirane podatke? Odgovor daje sljedeći primjer.

Primjer 2.4.4 Pretpostavimo da smo mjerenjem duljine 14 primjeraka tzv. sijamskog krokodila (*Crocodylus siamensis*) dobili podatke u tablici 2.3 o duljini njihova trupa u metrima.

interval	f_a	r_a
1.0-1.2	1	1/14
1.3-1.4	3	3/14
1.5-1.6	6	6/14
1.7-1.8	1	1/14
1.9-2.0	3	3/14

Tablica 2.3: *Crocodylus siamensis* duljina tijela.

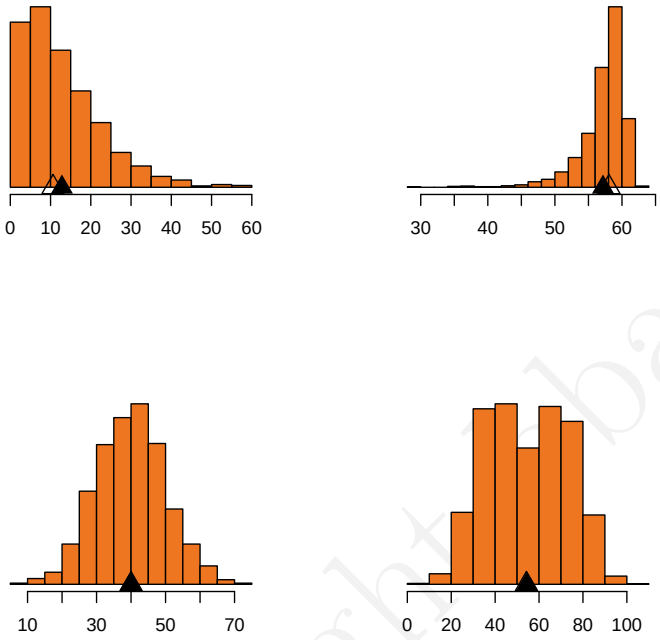
Ako nas zanima aritmetička sredina, medijan ili bilo koja druga numerička karakteristika najjednostavnije nam je zamijeniti ove neprecizne podatke s podacima koji svakom intervalu pridružuju njegovu sredinu i to s odgovarajućom frekvencijom. U ovom slučaju to znači da bismo kreirali "pomoćni uzorak":

$$1.1, 1.35, 1.35, 1.35, 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 1.75, 1.95, 1.95, 1.95, 1.95\,,$$

koji analiziramo na uobičajeni način. Tako je medijan polaznog uzorka jednostavno 1.55, a aritmetička sredina je suma ovih brojeva podijeljena s 14, tj. 1.6. Analogno bismo odredili i varijancu odn. standardnu devijaciju ili kvartile za ovakve uzorke. ⌟

2.5 Oblik razdiobe uzorka

Histogrami neprekidnih numeričkih podataka mogu imati različite oblike, a prema tim oblicima mi opisujemo i neka svojstva raspodjele ili razdiobe uzorka. Kao što ćemo kasnije vidjeti razdiobu numeričkih obilježja možemo teoretski modelirati grafom funkcije koji oblikom odgovara histogramu. Iz histograma ili grafa funkcije koji ga aproksimira možemo nasluti da li su, grubo govoreći, podaci manji od prosjeka $\bar{x}$ značajnije udaljeniji od njega nego podaci s desne strane prosjeka, tj. oni veći od $\bar{x}$. U tom slučaju kažemo da je lijevi rep razdiobe teži od desnog repa, ako vrijedi obrnuto kažemo da je teži desni rep. Za podatke udaljene od $\bar{x}$ posebno govorimo da pripadaju repu razdiobe. Nije rijedak slučaj da vidimo podatke koji su približno podjednako raspodijeljeni s obje strane prosjeka, tada govorimo i da su repovi podjednaki.



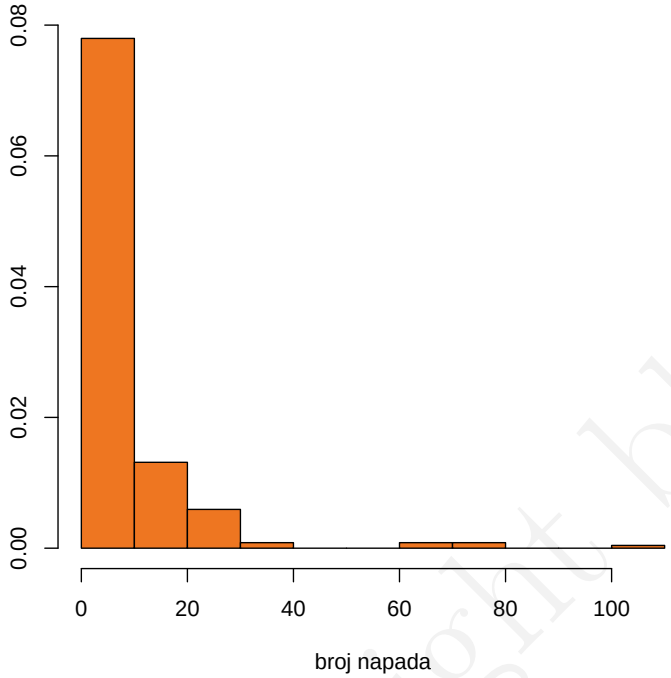
Slika 2.8: Usporedba histograma desno nagnute (gore lijevo), lijevo nagnute (gore desno), približno simetrične (dolje lijevo) i bimodalne razdiobe (dolje desno). Bijeli trokutić označava medijan, a crni aritmetičku sredinu podataka. Kod donja dva histograma oni se ne mogu razlučiti.

Ovisno o obliku histograma odn. grafa funkcije koji ga opisuje (to je tzv. *funkcija gustoće*) razdioba može biti **unimodalna** ili **višemodalna** već prema tome ima li ova funkcija odn. histogram jedan ili više lokalnih maksimuma. Unimodalne razdiobe nadalje razlikujemo i po tome da li su: **simetrične** oko svog prosjeka (očekivanja), **desno nagnute**, odn. **lijevo nagnute**. Pri tome kažemo da je razdioba desno odn. lijevo **nagnuta**, (engl. *skewed*), ako je njen desni odn. lijevi rep teži od onog drugog, v. sliku 2.8.

Pitanje je možemo li nagnutost, tj. lijevo odn. desno nagnute razdiobe definirati rigoroznije tj. ne oslanjajući se samo na promatranje histograma čiji oblik ipak ovisi i o odabranim razredima.

U tom smislu, definiramo prvo za prirodni broj k tzv. k-ti centralni moment uzorka kao

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



Slika 2.9: Histograma za podatke o broju epileptičkih napada tijekom perioda od 8 tjedana za 59 pacijenata.

Očito je $\mu_1 = 0$ i $\mu_2 = s^2$ (provjerite). Broj

$$\alpha_3 = \frac{\mu_3}{s^3}$$

zovemo **koeficijent asimetrije uzorka** (engl. skewness). Ako je $\alpha_3 < 0$ razdioba je lijevo ili negativno nagnuta, a za $\alpha_3 > 0$ ona je desno ili pozitivno nagnuta (ili asimetrična kako se još kaže). Primjetite da je za uzorke približno simetrične oko $\bar{x}$ koeficijent asimetrije približno 0. Za podatke sa slike 2.9 o broju epilepičkih napada već smo komentirali da su desno nagnuti. Za njih koeficijent asimetrije iznosi 2.186. Za podatke o visinama studenata on je bliže 0, što smo mogli naslutiti i iz histograma sa slike 2.4, približno iznosi -0.119.

Kako smo napomenuli, statistikama se nekad nazivaju sve numeričke vrijednosti izvedene iz uzorka. Na kraju poglavlja naglasimo da su sve do sada uvedene veličine

$$\bar{x},\ s,\ s^2,\ m,\ Q_1,\ Q_3,\ d_Q,\ q_\alpha,\ \mu_k,\ \alpha_3, \text{itd.}$$

zapravo primjeri različitih statistika.

2.6 Opisne statistike u R-u

Prikaz kategorijalnih podataka

Nakon što u R učitamo podatke, stupčasti dijagram nam je dostupan u jednoj naredbi.

```
> stranke = c(190,137,66,27,22,58)
> barplot(stranke)
```

Relativne frekvencije dobijemo dijeljenjem s 500 (duljina uzorka)

```
> rel.frekvencije = stranke/500
```

Želimo li pak da ispod stupaca stoje kratice stranaka, to možemo učiniti naredbom

```
> barplot(stranke,names=c("MFL","KMS","UI","NNP","SSR","Ostali")
```

Rezultat ove naredbe je na slici 2.1. Tortni dijagram je s druge strane rezultat bilo koje od sljedeće dvije naredbe

```
> pie(stranke)
> pie(stranke,labels=c("MFL","KMS","UI","NNP","SSR","Ostali"))
```

Histogram

Pretpostavite da imamo uzorak koji u centimetrima daje visine za 40 muških studenata. Podatke u R možemo učitati i korištenjem naredbe `scan`.

```
> visine = scan()
159 188 175 176 177 168 162 188
183 187 187 162 184 161 180 169
195 171 170 199 181 169 189 191
172 182 183 178 180 165 185 202
183 187 188 182 163 179 178 188
```

Primjetite da R broji ukupni broj učitanih podataka i javlja

```
Read 40 items
```

nakon što je učitavanje okončao korisnik nakon ukucavanjem praznog retka. Histogram poput onoga sa slike 2.4 rezultat je sljedeće naredbe

```
> hist(visine,prob=T)
```

Probajte i jednostavnu naredbu `hist(visine)` pa pronađite razliku između dva dijagrama. *Stem and leaf* dijagram je rezultat naredbe `stem(visine)`. Pravokutni dijagram poput onoga ("kutija s brkovima") rezultat naredbe

```
> boxplot(visine,prob=T)
```

Objasnimo na kraju i kako bismo nacrtali dijagram raspršenosti poput onoga na slici 2.6. Ovoga puta učitajmo vlastite podatke npr. o visinama prvog osobnog dohodka bivših studenata u odn. na prosjek ocjena tijekom studja

```
> dohodak=c(6700,4500,4800,5300,6100,3900)
> ocjene =c(3.2,4.1,4.2,3.9,4.8,3.7)
> plot(ocjene,dohodak)
```

Naredbe `help(plot)` i `help(par)` objašnjavaju kako svoj dijagram možemo uljepšati, mijenjajući boje točaka, natpise na osima ili dodajući naslov npr.

Sredina i raspršenost uzorka

Osnovne mjere centra – aritmetičku sredinu i medijan u R-u računamo naredbama

```
> mean(visine)
> median(visine)
```

Naredba

```
> table(visine)
```

će nam dati apsolutne frekvencije svih podataka u uzorku, uobičajeno je koristimo za kategorijalne podatke, no u ovom slučaju nam omogućava da se lako uvjerimo kako je mod ovog uzorka 188.

Raspon, varijanca i standardna devijacija uzorka se lako nađu naredbama

```
> max(visine)-min(visine)
> var(visine)
> sd(visine)
```

Osnovne statistike možemo dobiti i korištenjem jedne naredbe: `summary(visine)`.

Koeficijent asimetrije se ne može naći korištenjem osnovnih naredbi R-a, no ako proširimo R paketom "moments" možemo ga izračunati jednom naredbom

```
> library("moments")
> skewness(visine)
```

alternativno koeficijent asimetrije dobijemo složenijom naredbom

```
> (sum((visine-mean(visine))^3)/(length(visine)-1 )) / sd(visine)^3
```

Numerički se ovi koeficijenti malo razlikuju– to je posljedica nešto drugačije definicije koeficijenta asimetrije u paketu "moments". Kako smo već uvidjeli – pomalo iritantna osobina statističkih paketa (kao i knjiga) je da nemaju potpuno iste definicije za istoimene statističke pojmove. Srećom te razlike su praktično zanemarljive, pogotovo na većim uzorcima.

Zadaci

Zadatak 1. Zadani su podaci 1, 4, 3, 3, 1, 0, 2, 2, 5, 3. Odredite medijan i donji kvartil za te podatke.

Zadatak 2. Zadana je tablica podataka:

interval	frekvencija
$[2, 3)$	3
$[3, 4)$	4
$[4, 5)$	3
$[5, 6)$	2

Odredite aritmetičku sredinu i standardnu devijaciju podataka iz tablice.

Zadatak 3. Unesite podatke o broju mačića u okotu iz primjera 2.2.1 u R i nacrtajte pripadni stupčasti dijagram.

Zadatak 4. Unesite podatke o rezultatima bacanja kocke iz primjera 2.3.1 u R, a zatim odredite aritmetičku sredinu, medijan, kvartile i varijancu ovog uzorka naredbama R-a.

Zadatak 5. Pretpostavite da su mjerenjem brzine kojim zamorci savladavaju dani labirint u laboratoriju dobiveni sljedeći podaci u minutama:

$$2.08, 2.20, 1.99, 1.41, 2.38, 1.71, 1.61, 1.34, 2.58, 2.38.$$

Unesite podatke u R i nacrtajte im histogram i histogram apsolutnih frekvencija. Izvedite i naredbe `hist(x,breaks=1)` odn. `hist(x,breaks=3)` gdje `x` označava vaše podatke. Što možete reći o informativnosti ovih histograma u odn. na prva dva?

Za ove podatke nacrtajte i pravokutni dijagram, te im odredite koeficijent asimetrije.

3 VJEROJATNOST

3.1 Pojam vjerojatnosti

Neizvjesnost je značajka gotovo svih procesa, događaja i pokusa koje opažamo. U nastavku npr. želimo razgovarati o igrama na sreću, o rezultatima fizikalnih mjerenja, medicinskih tretmana ili poslovnih ulaganja. Jezik vjerojatnosti omogućuje nam da koherentno govorimo o takvim pojavama. Koristeći pojmove teorije vjerojatnosti naime, ponekad možemo napraviti razuman matematički model za slučajne pojave.

U susretu sa stvarnim neizvjesnim događajima, ne trebamo zaboraviti da matematički modeli, kao što je npr. model mutacije DNK u evolucionarnoj biologiji, samo pokušavaju modelirati tu neizvjesnost. Većina modela korištenih u znanosti uključuje tzv. vjerojatnosne ili stohastičke komponente. Oni su u pravilu samo približni. S druge strane, isto možemo reći i za razne zakone klasične fizike. Iako približni, oni su dovoljno precizni za razne znanstvene i praktične svrhe, posebno za predviđanje rezultata pokusa. Korisnost modela u predikciji jedan je od glavnih pokazatelja njihove vrijednosti.

Suočeni s podacima u nekom uzorku tj. rezultatima nekog pokusa, mi u statistici također moramo kvantificirati neizvjesnost. Naime, mi pretpostavljamo da na osnovi podataka o jedinkama u uzorku, nešto možemo reći o ukupnoj populaciji, a pri tome jasno ne možemo uvijek izbjeći greške. Ovaj problem ilustrira i sljedeći jednostavan primjer.

Primjer 3.1.1 Pretpostavimo da smo 100 puta bacili novčić. Neka je 80 puta palo pismo. Prije nego pozovemo policiju ili središnju banku da ih upozorimo kako je novčić neispravan, trebamo uočiti da je takvo što moguće čak i ako je novčić savršeno simetričan. Tada su dakako ishodi pismo i glava jednako vjerojatni. Samo, u tom slučaju naših 80 (ili više) pisama nisu jako vjerojatni. Korisno je odrediti vjerojatnost da se ovako ekstremno odstupanje od 50tak pisama koje očekujemo, uopće dogodi kod simetričnog novčića. Kad postavimo matematički model ovog pokusa pokazat ćemo da je ova vjerojatnost zapravo 0.0000000011 . . . Dakle izuzetno mala. ⊖

Važno je primjetiti da smo računali vjerojatnost ovakvog odstupanja ili većeg, a ne vjerojatnost da padne točno 80 pisama. Razlog leži u tome što je svaki pojedini numerički ishod pokusa malo vjerojatan (u ovom slučaju svi imaju vjerojatnost manju od 0.08), pa bi svaki rezultat mogli smatrati iznenađujućim. Nas je iznenadilo upravo odstupanje od očekivanih 50 pisama. ⊖

⌟

U ovom i idućem poglavlju ćemo objasniti kako se mogu izračunati vjerojatnosti i očekivanja poput onih koje spominjemo u gornjem primjeru. Ali i prije formalnog izračuna svi imamo neku intuiciju o ovim pojmovima koja može biti vrlo korisna. Pokazuje se nažalost da nam je intuicija često manjkava i da neke odgovore možemo naći tek koristeći rigorozne matematičke argumente. Jedna poznata dilema koja testira našu intuiciju je opisana sljedećim primjerom.

Primjer 3.1.2 Pretpostavite da će ujutro odlukom predsjednika biti oslobodjena dvojica od trojice zatvorenika: A, B i C. Sva tri zatvorenika imaju jednaku šansu biti oslobođena po onome što se zna o dosadašnim predsjednikovim abolicijama. Intuitivno je jasno dakle, da je za svu trojicu vjerojatnost oslobađanja 2/3.

Zatvorenik A u razgovoru sa stražarem pokuša otkriti svoju sudbinu večer prije. Stražar, iako zna, rezolutno odbija reći hoće li A biti oslobođen. Tada A zamoli stražara da mu otkrije ime bar jednog od preostale dvojice koji će otići na slobodu. Nakon dugotrajnog preklinjanja, stražar mu otkrije jedno ime. A A? Problijedi, sasvim zbunjen, pomislivši kako su se njegove šanse za aboliciju upravo spustile sa 2/3 na 1/2.

Dilemi iz ovog primjera ćemo se vratiti na kraju ovog poglavlja naoružani formalnim znanjem iz teorije vjerojatnosti. ⌟

Vjerojatnost događaja u stvarnom svijetu nije lako rigorozno definirati. Intuitivno, ako ponavljamo isti slučajan pokus veliki broj puta, recimo n, i bilježimo koliko puta se pojavio izvjestan rezultat, recimo A, te ako je broj pojavljivanja tog rezultata n_A, tada možemo očekivati da će omjer (ili relativna frekvencija)

$$\frac{n_A}{n}$$

težiti k nekom broju. Upravo takav granični broj bismo mogli zvati vjerojatnost događaja A. Iako neformalna, ovakva intuicija može motivirati naš pristup definiciji vjerojatnosti. U slučaju bacanja novčića rezultat $A = \{\text{palo je pismo}\}$ jedan je od dva moguća ishod pokusa. Ukoliko je novčić simetričan, očekivali bismo da se relativna frekvencija pojavljivanja događaja A, nakon puno bacanja, približava 1/2. No postoje i druga intuitivna shvaćanja vjerojatnosti stvarnih događaja. Za neizvjesne događaje koji se ne mogu ponavljati (poput finalne utrke nekog atletskog natjecanja) vjerojatnost očito ne možemo uvesti kao graničnu frekvenciju. Jedna raširena interpretacija, npr. u tzv. bayesovskom pristupu statistici, vjerojatnost razumijeva kao subjektivni stupanj uvjerenja.

Bez obzira na naše razumijevanje tog pojma, vjerojatnost uvijek vežemo uz pojave koji imaju neizvjestan ishod i računamo koristeći formalna pravila koja slijede. Svaku takvu neizvjesnu pojavu/proces/radnju općenito ćemo zvati **slučajni pokus**.

Primjer 3.1.3 Primjeri slučajnih pokusa

- nakon bacanja jedne igraće kocke zabilježimo rezultat,
- nakon 100 ponovljenih bacanja novčića prebrojimo pisma,
- nakon ispitivanja 1000 *slučajno* odabranih birača kroz tzv. izlazne ankete (exit polls), prebrojimo broj glasova dodjeljenih pojedinim listama,
- nakon sadnje 100 sadnica iste vrste biljaka u laboratorijskim uvjetima, bilježimo parametre njihova rasta tokom sljedeće 2 godine,
- istu vrstu bakterija izložimo u dvadeset odvojenih posuda utjecaju dva različita antibiotika, i to u dvije grupe od po 10 posuda, zatim bilježimo rezultate tijekom 48 sati.

- odabiremo na *slučajan* način bebu rođenu u Hrvatskoj tijekom protekle godine i bilježimo njene karakteristike (npr. spol i težinu).

⌟

Prije nego odredimo vjerojatnosti pojedinih rezutata slučajnog pokusa, korisno je znati sve moguće ishode takvog pokusa. Svi mogući (a nedjeljivi) ishodi slučajnog pokusa nazivaju se **elementarni događaji**. Skup svih elementarnih događaja zovemo **prostor elementarnih događaja**. Standardna matematička oznaka za ovaj skup je Ω. **Slučajan događaj** može biti bilo koji podskup od Ω. Posebno svaki slučajni događaj skup je elementarnih događaja. Slučajne događaje obilježavamo velikim slovima abecede, tipično

$$A, B, C, \text{ itd.}$$

Uočite: slučajni događaj se može sastojati i od samo jednog elementarnih događaja.

Primjer 3.1.4 i) Kod bacanja jedne kocke elementarni ishodi su 1,2,3,4,5 i 6, pa je prirodno postaviti

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

a slučajni događaj bi mogao biti npr.

$$A = \{\text{pao je paran broj}\} = \{2, 4, 6\}.$$

ii) Kod bacanja jednog novčića ishodi su pismo ili glava, (alternativno 0 ili 1), pa pišemo

$$\Omega = \{\text{P,G}\} \text{ ili } \{0, 1\}.$$

a slučajni događaj bi mogao biti npr. $A = \{\text{P}\} = \{\text{palo je pismo}\}$.

iii) Kod slučajnog odabira bebe rođene u RH prošle godine

$$\Omega = \{\omega_1, \ldots, \omega_N\},$$

gdje je N broj takvih beba, a ω_i su njihove jednoznačne oznake npr. OIB. Slučajni događaj bi mogao biti npr.

$$A = \{\text{odabrana beba je ženskog spola}\}$$

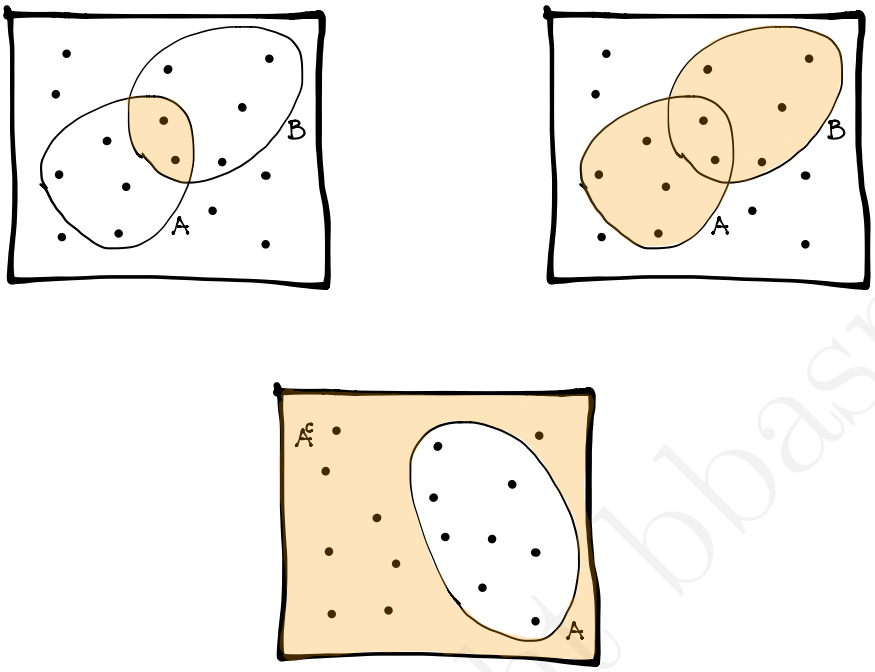
ili u drugom zapisu

$$A = \{\omega_i : \omega_i \text{ je ž. spola}\}.$$

iv) Kod bacanja para kocaka možemo staviti

$$\Omega = \{(i, j) : i, j = 1, \ldots, 6\}.$$

⌟



Slika 3.1: Vennovi dijagrami. Presjek događaja A i B. Unija događaja A i B. Događaj A i suprotni događaj A^c

Slučajni događaji su dakle samo skupovi elementarnih događaja, a skupovima je često korisno odrediti i komplement, ako je A^c komplement od A u Ω mi ga zovemo **suprotnim događajem**. Na događajima ima smisla provjeravati i različite skupovne relacije kao npr.

$$A \subseteq B \text{ kažemo da događaj } A \text{ povlači događaj } B \ ,$$

$$A = B^c \text{ kažemo događaji } A \text{ i } B \text{ su suprotni,}$$

$$A \cap B = \emptyset \text{ kažemo događaji } A \text{ i } B \text{ su disjunktni ili se isključuju.}$$

A možemo koristiti i uobičajene skupovne operacije npr.

$$A \cap B \text{ kažemo dogodili su se i događaj } A \text{ i događaj } B,$$

$$A \cup B \text{ kažemo dogodio se bar jedan on događaja } A \text{ i } B,$$

Skupovne relacije i operacije ćemo najlakše predstaviti tzv. Vennovim dijagramima.

Napomenimo da Ω može biti i neprebrojivo beskonačan skup npr. kada odabiremo slučajnu točku na nekom intervalu ili u kvadratu. No tu je matematička teorija iako analogna nešto zahtjevnija.

3.2 Vjerojatnost slučajnog događaja

Vjerojatnost slučajnog događaja A označit ćemo kao $P(A)$. Od vjerojatnosti ćemo tražiti da zadovoljava sljedeća svojstva:

- Za svaki slučajan događaj A
$$0 \leq P(A) \leq 1.$$
- Vjerojatnost da će se dogoditi jedan od svih mogućih elementarih događaja je 1, tj.
$$P(\Omega) = 1.$$
- Ako se događaji $A_1, A_2, \ldots$ medjusobno isključuju tada vrijedi
$$P(\cup_i A_i) = \sum_i P(A_i).$$

Pravila za računanje vjerojatnosti

Direktno iz matematičkog opisa vjerojatnosti slijedi

$$P(A^c) = 1 - P(A).$$
$$A \subseteq B \text{ povlači } P(A) \leq P(B)$$
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

te

$$P(\emptyset) = 0.$$

Posebno ako su A i B disjunktni

$$P(A \cup B) = P(A) + P(B).$$

Unatoč ovim formalnim pravilima, nije jasno kako doći do broja $P(A)$ za zadani slučajni događaj A. Pokušajmo ipak s nekim primjerima

Primjer 3.2.1 Francuski plemić i bonvivan C. de Mere je u 17 st. pitao više matematičara – što je vjerojatnije da će u 4 bacanja kocke pasti bar jedna šestica (pokus 1) ili u 24 bacanja para kocaka pasti bar jedna dvostruka šestica (pokus 2).

Mogli bismo pretpostaviti da je očekivani broj uspješnih pokusa u prvom slučaju je (intuitivno, za sada)

$$\frac{4}{6},$$

a u drugom

$$\frac{24}{36},$$

što je dakle isto. No vjerojatnosti nisu jednake naslućivao je de Mere. Odgovor na ovu dilemu je nastao u korespondenciji B. Pascala i P. de Fermata i smatra se jednim od važnih koraka u razvoju teorije vjerojatnosti.

⌟

Uniformni vjerojatnosni model

Primjer 3.2.2 Odredimo vjerojatnost da će nakon bacanja igraće kocke pasti paran broj? Već smo odredili sve moguće ishode tj.

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Mi tražimo $P(A)$ za $A = \{$ pao je paran broj$\} = \{2, 4, 6\}$. Ako je kocka simetrična, svi elementarni događaji su jednako vjerojatni, tj.

$$P(1) = P(2) = \cdots = P(6).$$

Kako se elementarni događaji međusobno isključuju definicija vjerojatnosti nam kaže

$$P(\Omega) = P(\{1\} \cup \ldots \cup \{6\}) = \sum_{i=1}^{6} P(i) = 6P(1).$$

No $P(\Omega) = 1$ pa dakle vrijedi

$$P(i) = \frac{1}{6}$$

za sve i. Stoga je

$$P(A) = P(\{2\} \cup \{4\} \cup \{6\}) = P(2) + P(4) + P(6) = \frac{3}{6}.$$

⌟

Ovaj nam primjer ukazuje na vrlo važno općenito pravilo.

Ako možemo pretpostaviti da su elementarni događaji jednako vjerojatni i ako ih je konačno, npr. $n \in \mathbb{N}$, tad je vjerojatnost svakog od njih

$$\frac{1}{n}.$$

Nadalje, vjerojatnost bilo kojeg slučajnog događaja A u tom slučaju je

$$P(A) = \frac{\text{broj elementarnih događaja koji su povoljni za } A}{\text{broj svih elementarnih događaja}} \tag{3.2.1}$$

Dakle jednako vjerojatni elementarni događaji daju vrlo jednostavan način za računanje vjerojatnosti.

Prema definiciji vjerojatnosti, vjerojatnost bilo kojeg događaja A možemo naći kao

$$P(A) = \sum_{i:\omega_i \in A} P(\omega_i)\,,$$

čak i ako nisu svi ω_i jednako vjerojatni. No ovu vjerojatnosti je teže izračunati kada elementarni događaji nisu jednako vjerojatni.

U praksi je potreban **oprez**. Naime, česta je greška pretpostaviti da su događaji jednako vjerojatni iako oni to nisu. Npr. ako predizbornu anketu provodimo telefonski, nije razumno pretpostaviti da svi birači imaju jednaku vjerojatnost biti članovima uzorka.

Složeni pokusi

Ponekad se pokus zapravo sastoji od mjerenja ishoda dva ili više pokusa kao u sljedećem primjeru. Pitanje je kako zapisati sve ishode odn. elementarnih događaje u ovakvim slučajevima.

Primjer 3.2.3 Pretpostavimo da se slučajni pokus sastoji u promatranju spola prvo dvoje djece koja će se roditi 1. siječnja iduće godine u Zagrebu i Splitu. Odredimo vjerojatnost da će se prve dvije bebe biti različitog spola u različitim gradovima. Pretpostavimo (aproksimativno) da su kod beba oba spola jednako vjerojatna. Prije određivanja vjerojatnosti korisno je odabrati dobar matematički model. Nameću se dvije različite ideje.

Prva ideja: postavimo

$$\Omega = \{bb, bg, gg\}$$

i

$$P(bb) = P(bg) = P(gg) = 1/3.$$

Druga ideja

$$\Omega = \{(b,b), (b,g), (g,b), (g,g)\}$$

i

$$P(b,b) = P(b,g) = P(g,b) = P(g,g) = 1/4.$$

Iako su u oba modela elementarni događaji jednako vjerojatni, samo je drugi model razuman. Mi iz iskustva znamo da je u složenim pokusima koji promatraju rezultate više odvojenih pokusa rezultate uvijek razumno urediti, odn. pisati kao: rezultat 1. pokusa, rezultat 2. pokusa,... Prvu ideju bismo stoga odbacili.

⌟

Općenito, pretpostavite da pratimo k pokusa te da se svi elementarni događaji koji odgovaraju i-tom pokusu nalaze u skupu Ω_i, $i = 1, \ldots, k$. Naš složeni pokus tada opisuje skup svih k-torki oblika

$$\Omega = \{(\omega_1, \ldots, \omega_k) : \quad \omega_i \in \Omega_i\},$$

Zadati vjerojatnost sada možemo tako da zadamo vjerojatnost svakoj pojedinoj k-torci.

Najvažniji poseban slučaj je nezavisno **ponavljanje** istog pokusa. Matematički pišemo tada $\Omega = \Omega_1^k$. A ako pokuse ponavljamo neovisno i ako na svakom vrijedi uniformni vjerojatnosni model sa n mogućih ishoda, razumno je pretpostaviti da svaka k-torka ima istu vjerojatnost

$$P(\omega_1, \ldots, \omega_k) = \frac{1}{n^k} = \frac{1}{\text{broj svih mogućih } k\text{-torki}}.$$

Dakle, tada i složeni pokus tada prati uniformni vjerojatnosni model. Tako možemo u gornjem primjeru argumentirati postavljanje vjerojatnosti elementarnih događaja na

$$\frac{1}{2^2} = \frac{1}{4},$$

dok je recimo vjerojatnost svakog pojedinog ishoda kod bacanja para kocaka

$$\frac{1}{6^2} = \frac{1}{36}\,.$$

Primjer 3.2.4 Pretpostavite da iz urne s 4 bijele i 4 crne kuglice slučajno biramo dvije i to na dva načina: i) vraćajući kuglice natrag i ii) bez vraćanja. U oba slučaja elementarne događaje možemo zapisati npr. kao

$$\Omega = \{(b,b),(b,c),(c,b),(c,c)\}\,.$$

Uočite, $P(\text{u prvom izvlačenju dobili smo } c) = P(\text{u prvom izvlačenju dobili smo } b) = 1/2$, no zbog simetričnosti isto vrijedi i za drugi izvlačenje bez obzira na vraćanje. No važno je primjetiti da nam samo prvi način izvlačenja daje uniformni vjerojatnosni model. Zaista u tom slučaju ponavljamo isti pokus i to **neovisno**, dok u drugom načinu ishod drugog izvlačenja ovisi o rezultatu prvog. Izračunajmo npr. vjerojatnost da smo dobili kuglice iste boje.
i) Prvim načinom to je zbog uniformnosti modela $P(\{(b,b),(c,c)\}) = 2/4 = 1/2$.
ii) Bez vraćanja, situacija je složenija, mogli bismo bijele kuglice numerirati brojevima od 1 do 4, a crne brojevima 5 do 8. Pokusu bismo sad mogli pridružiti

$$\Omega = \{(i,j): \quad i,j = 1,\ldots,8,\ i \neq j\}\,,$$

koji sadrži $8 \cdot 7 = 56$ elemenata, a na kojem sad zaista (zbog simetričnosti) vrijedi uniformni model. Sad lako računamo (vidi varijacije bez ponavljanja u idućem odjeljku)

$$P(\{(b,b)\}) = P(\{(c,c)\}) = \frac{4\cdot 3}{8\cdot 7} = \frac{3}{14}\,,$$

tako da je $P(\{(b,b),(c,c)\}) = P(\{(b,b)\}) + P(\{(c,c)\}) = 3/7$. ⌟

Prebrojavanje

Formula (3.2.1) sugerira kako je vrlo važno u modelu s jednako vjerojatnim elementarnim događajima znati prebrojati elemente proizvoljnog skupa $A \subseteq \Omega$. Ponovit ćemo stoga neka općenita pravila o brojanju elemenata pojedinih skupova.

Ako imamo dva skupa npr. A_1 i A_2 koji imaju n_1 odn. n_2 elemenata. Tada skup svih uređenih parova (x_1,x_2) za koje je $x_1 \in A_1,\ x_2 \in A_2$ ima

$$n_1 \cdot n_2$$

elemenata. Ovo nekad zovu i pravilom množenja.

Slično, ako imamo više konačnih skupova npr. $A_1, A_2, \ldots A_k$, i i pri tom sa n_i označimo broj elementarnih događaja koji su povoljni za $A_i,\ i = 1,\ldots,k$, tada skup uređenih k-torki $(x_1,x_2,\ldots,x_k)$ kod kojih je $x_1 \in A_1,\ x_2 \in A_2,\ldots,x_k \in A_k$ ima

$$n_1 n_2 \cdots n_k$$

elemenata.

Posebno, ako iz istog skupa A od n elemenata izabiremo jednu uređenu k-torku, $k \geq 1$, dopuštajući pri tom ponavljanja, to možemo učiniti na

$$n^k$$

načina. Ovakve k-torke se nazivaju i **varijacije s ponavljanjem.**

Primjer 3.2.5 Ako bacamo simetričnu kocku k puta vjerojatnost da ćemo svaki puta dobiti paran broj je

$$\frac{\text{broj } k\text{-torki sastavljenih od parnih brojeva}}{\text{broj svih } k\text{-torki}} = \frac{3^k}{6^k} = \frac{1}{2^k}$$

⌟

Pretpostavimo da iz skupa A od $n \in \mathbb{N}$ elemenata izabiremo uređenu k-torku $(x_1, \ldots, x_k)$ $(k \leq n)$, tako da su svi x_i međusobno različiti. Dakle izabiremo njih k od n bez ponavljanja, ali pazeći pri tom na poredak. To možemo učiniti na

$$n(n-1)\cdots(n-k+1)$$

načina. Taj broj nazivamo broj **varijacija bez ponavljanja** k-tog razreda (ili broj permutacija duljine k) skupa od n elemenata.

Ako se pitamo samo na koliko načina možemo poredati n elemenata u niz, to je specijalni slučaj varijacija bez ponavljanja za $k = n$. Stoga je taj broj koji nazivamo i broj **permutacija** jednak

$$n(n-1)\cdots 2 \cdot 1 = n!,$$

što čitamo n faktorijela. Po definiciji je $0! = 1$, $n! = n(n-1)!$ za $n \in \mathbb{N}$. Gore uvedeni broj varijacija bez ponavljanja k-tog razreda skupa od n elemenata sada možemo računati i kao $n!/(n-k)!$.

Primjer 3.2.6 Na koliko načina možete posložiti 3 od 4 slova A,C,G,T kako biste kreirali sve kodone u kojima nema istih nukleotida? Prisjetite se – kodoni su zapravo riječi od 3 slova u ovom alfabetu.

Odgovor je dakako

$$4 \cdot 3 \cdot 2 = 24.$$

Usput, razmislite koliko je kodona sveukupno? ⌟

Pretpostavimo da iz skupa od n elemenata izabiremo uzorak od njih r **ne** pazeći pri tom na poredak. To možemo učiniti na

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} = \binom{n}{r}$$

načina. Taj broj zovemo broj **kombinacija** duljine r iz skupa od n elemenata. Brojevi oblika $\binom{n}{r}$ nazivaju se i binomni koeficijenti.

Primjer 3.2.7 Pretpostavimo da je u diploidnoj populaciji 8 alela nekog gena. Preciznije, svaka jedinka u populaciji ima dva primjerka danog gena koji u populaciji dolazi u 8 različitih oblika – alela. Broj genotipova koji odgovara heterozigotima je

$$\binom{8}{2} = 28\,,$$

a homozigotima naravno 8.

Ovdje je jako važno uočiti da nisu svi genotipovi jednako vjerojatni, naime aleli tipično nemaju istu frekvenciju u populaciji. No i kad bi je imali, bilo koji od heterozigotnih genotipova bio bi dvostruko vjerojatniji od homozigotnih. Stoga ovdje uniformni vjerojatnosni model ne možemo koristiti. ⌟

Primjer 3.2.8 Na koliko načina možete izabrati 7 brojeva od 39? Odgovor je dakako

$$\binom{39}{7} = 15380937\,,$$

Ovdje je razumno pretpostaviti uniformni vjerojatnosni model, odn. da su sve kombinacije jednako vjerojatne. Stoga je vjerojatnost da ćete uplatom jedne kombinacije u igri lota 7 od 39 osvojiti glavni dobitak izuzetno mala i iznosi otprilike

$$6.5 \cdot 10^{-8}\,.$$

⌟

3.3 Uvjetna vjerojatnost i nezavisnost

Ponekad imamo djelomičnu informaciju o ishodu slučajnog pokusa. Na primjer kod bacanja simetrične igraće kocke, iako nam je nepoznat točan ishod, mi bismo mogli saznati da je rezultat broj veći od 3. Vjerojatnost elementarnog događaja 4 uz ovaj uvjet je

$$\frac{1}{3}\,.$$

Naime, sad znamo da su događaji 4,5 i 6 jedini mogući, ali su još uvijek (zbog simetričnosti) jednako vjerojatni. Tako da je vjerojatnost zapravo omjer broja povoljnih elementarnih događaja i broja događaja koji se uopće mogu dogoditi uz informaciju koju mi imamo. Slično vjerojatnost da je pao paran broj je sada 2/3, a ne više 1/2 uočimo.

Općenito ako imamo pokus s jednako vjerojatnim elementarnim ishodima (ili tzv. uniformni vjerojatnosni model), i znamo da se dogodio neki od ishoda u skupu B te se pitamo kolika je vjerojatnost da se dogodio slučajni događaj A uz ove uvjete, odgovor je

$$P(A|B) = \frac{\text{broj elementarnih događaja u } A \cap B}{\text{broj elementarnih događaja u } B} = \frac{P(A \cap B)}{P(B)}.$$

Motivirani ovim primjerom, u općenitom vjerojatnosnom modelu uz pretpostavku da je $P(B) > 0$, definiramo **uvjetnu vjerojatnost** proizvoljnog događaja A uz uvjet da se dogodio događaj B kao

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,.$$

Nije se teško uvjeriti da uvjetna vjerojatnost nasljeđuje svojstva vjerojatnosti. Posebice vrijedi

$$P(\Omega|B) = 1 \quad \text{i } P(A^c|B) = 1 - P(A|B),$$

za sve događaje A, B takve da je $P(B) > 0$.

Uvjetna vjerojatnost pokazuje kako naše prethodno znanje o ishodu pokusa mijenja vjerojatnosti svih ostalih događaja Npr. vjerojatnost da će slučajno odabrani student u RH biti student pomorskog fakulteta se mijenja ako nam je poznato da student dolazi iz Rijeke (ili Đakova) npr.

S druge strane, vjerojatnost da će ako taj student baci kocku pasti 6, ostaje 1/6 čak i ako znamo ovu činjenicu (razumno je prepostaviti). Rekli bismo da su događaji {slučajno odabrani student bacio je 6 na kocki} i {slučajno odabrani student dolazi iz Rijeke} nezavisni.

Matematički precizno, slučajni događaji A i B su **nezavisni** ako vrijedi

$$P(A \cap B) = P(A)P(B)\,.$$

Posebno je tada

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

Dakle vjerojatnost događaja A se u tom slučaju ne mijenja čak i ako znamo da se dogodio događaj B.

Primjer 3.3.1 Kod bacanja dvije kocke događaji da je na prvoj odn. drugoj kocki pala šestica su nezavisni. Zaista, kako smo vidjeli vjerojatnosni model za ovaj složeni pokus je

$$\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 = 1, \ldots, 6\}$$

uz jednake vjerojatnosti svih 36 elementarnih događaja. Šestica na prvoj odn. drugoj kocki može se simbolički zapisati kao

$$A = \{(6, \omega_2) : \omega_2 = 1, \ldots, 6\}$$

odn.

$$B = \{(\omega_1, 6) : \omega_1 = 1, \ldots, 6\}.$$

Uočite $A \cap B = \{(6, 6)\}$. Jasno je da je $P(A) = P(B) = 6/36 = 1/6$, ali i

$$P(A \cap B) = \frac{1}{36}$$

stoga su A i B zaista nezavisni događaji.

⌟

Prethodni primjer možemo poopćiti. Konkretnije, ako neovisno izvodimo 2 ili više pokusa i promatramo događaje koji se tiču različitih pokusa ti događaji su nezavisni u matematičkom smislu. Dakle, vjerojatnost njihovog presjeka (odn. vjerojatnost da će se svi ti događaji dogoditi simultano) je produkt vjerojatnosti svakog od njih.

Primjer 3.3.2 Sad možemo riješiti i de Mereov problem. Pretpostavimo da pokus bacanja dvije igraće kocke opisuje isti model kao u prethodnom primjeru $\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 = 1, \ldots, 6\}$, te da su svi elementarni događaji jednako vjerojatni. Prema tome za svaki ishod (ω_1, ω_2) imamo

$$P((\omega_1, \omega_2)) = \frac{1}{36}\,.$$

Za vjerojatnost događaja da će u 24 bacanja dvije kocke pasti barem jedan par 6-ica vrijedi

$$P(A) = 1 - P(\text{niti u jednom od 24 bacanja nisu pale dvije 6}) = 1 - P(A^c).$$

Formalno vrijedi $A, A^c \subseteq \Omega^{24}$. Jasno je

$$P(A^c) = P(\cap_{i=1}^{24}\{\text{u bacanju } i \text{ palo je nešto drugo od dvije 6}\}).$$

Ova vjerojatnost je zbog nezavisnosti između bacanja (vidi prethodni primjer i napomenu nakon njega) zapravo umnožak 24 vjerojatnosti oblika

$$P(\text{u bacanju } i \text{ palo je nešto drugo od dvije 6})\,.$$

Za sva bacanja $i = 1, \ldots, 24$ ova vjerojatnost je ista (jer ponavljamo isti pokus), a kako su svi ishodi u svakom pokusu jednako vjerojatni ona iznosi

$$\frac{35}{36}.$$

Pa je tražena vjerojatnost

$$P(A) = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914039.$$

Slično se može pokazati (provjerite) da je vjerojatnost bar jedne 6-ice u 4 bacanja jedne kocke

$$1 - \left(\frac{5}{6}\right)^{4} = 0.5177469.$$

Dakle ovakva oklada je povoljnija.

⌟

Bayesova formula

Direktno iz definicije uvjetne vjerojatnosti može se dobiti sljedeća tzv. **Bayesova formula**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}\,.$$

Ako se skup Ω može napisati kao unija disjunktnih podskupova $H_i,\ i = 1, 2, \ldots$, tada je

$$P(B) = \sum_i P(B \cap H_i) = \sum_i P(B|H_i)P(H_i)\,.$$

koristeći svojstva vjerojatnosti i činjenicu da su $B \cap H_i$ disjunktni. Zbog ove relacije se za $A = H_1$ Bayesova formula može napisati i u ovom često korisnom obliku

$$P(H_1|B) = \frac{P(B|H_1)P(H_1)}{\sum_i P(B|H_i)P(H_i)}\,.$$

Primjer 3.3.3 Postoji test koji ispituje postojanje neke rijetke bolesti na osnovu uzorka krvi. Test nije savim precizan, pa je poznato da vrijedi

$$P(\text{test je pozitivan}|\text{pacijent je bolestan}) = 0.99,$$

ali i

$$P(\text{test je negativan}|\text{pacijent je zdrav}) = 0.98$$

U dijagnostičkoj medicini se ove vjerojatnosti nazivaju osjetljivost i specifičnost testa. Pretpostavite da je slučajno odabrana osoba pozitivna na test, ako se bolest javlja u 1 od 10000 osoba u populaciji, izračunajmo vjerojatnost da je ta osoba bolesna.

Iskoristit ćemo Bayesovu formulu. Postavimo $B = \{\text{test je pozitivan}\}$, te

$$H_1 = \{\text{pacijent je bolestan}\},\ H_2 = \{\text{pacijent je zdrav}\}\,.$$

Nama je zadano $P(B|H_1) = 0.99$ ali i $P(B^c|H_2) = 0.98$. Kako i uvjetne vjerojatnosti zadovoljavaju $P(B|H_2) = 1 - P(B^c|H_2)$ imamo

$$P(B|H_2) = P(\text{test je pozitivan}|\text{pacijent je zdrav}) = 1 - 0.98 = 0.02$$

Slijedi

$$P(H_1|B) = \frac{0.99 \cdot 1/10000}{0.99 \cdot 1/10000 + 0.02 \cdot 9999/10000} = 0.0049$$

⌟

Na kraju se možemo vratiti i primjeru sa zatvorenicima s početka poglavlja.

Primjer 3.3.4 (nastavak Primjera 3.1.2) Pretpostavimo da zatvorski čuvar napravi sljedeće: ako je B onaj koji neće biti pušten, on otkriva da će pušten biti C i obrnuto. Jedino u slučaju kada A neće biti pušten, stražar odabire (pretpostavimo slučajno i jednako vjerojatno) jednog od B i C i njegovo ime govori zatvoreniku A.

Događaji {B je otkriven} i {A je zadržan} su uz ove uvjete nezavisni. Da bismo to pokazali uočite {B je otkriven} = {C je otkriven}^{c}. Pa je zbog simetričnosti naših uvjeta

$$P(\text{B je otkriven}) = P(\text{C je otkriven}) = \frac{1}{2}\,.$$

Nadalje, prema pretpostavkama je $P(\text{B je otkriven} \mid \text{A je zadržan}) = 1/2$. Pa Bayesova formula daje

$$P(\text{A je zadržan} \mid \text{B je otkriven}) = \frac{1/2 \cdot 1/3}{1/2}.$$

Dakle

$$P(\text{A je zadržan}|\text{B je otkriven}) = P(\text{A je zadržan}) = \frac{1}{3}.$$

Baš kao i prije nego što je A išta doznao od čuvara. ⌟

Zadaci

Zadatak 1. Neka je $P(A) = \frac{1}{4}$ i $P(A \cup B) = \frac{1}{3}$. Odredite $P(B)$ ako znamo da su događaji A i B nezavisni.

Zadatak 2. Odredite vjerojatnost da će u 4 bacanja simetrične igraće kocke pasti bar jedna 6-ica.

Zadatak 3. U nekom društvu je 20% ljudi koji su ili kratkovidni, ili ljevoruki, ili oboje. Ako znamo da je 10% ljudi u tom društvu ljevoruko, a 3% i ljevoruko i kratkovidno, koja je vjerojatnost da slučajno odabrana osoba nije kratkovidna?

Zadatak 4. Iz standardnog skupa od 52 karte odjednom izvlačimo 5 karata. Kolika je vjerojatnost da smo izvukli točno 3 tref karte?

Zadatak 5. U nekoj kutiji nalazi se 8 kuglica od čega su 2 plave, a u drugoj imamo 6 kuglica od čega su 3 plave. Sadržaj obje kutije pomiješamo i stavimo u treću. Ako iz treće kutije izvučemo plavu kuglicu, koja je vjerojatnost da je ta kuglica prije toga bila u prvoj kutiji?

Zadatak 6. Vilim i Marko idu u kino sa djevojkama. Ako se na slučajan način rasporede na 4 susjedna sjedala, koja je vjerojatnost da obojica sjede kraj svojih djevojaka?

Zadatak 7. Postoji test koji ispituje prisutnost prisutnost određene nedopuštene supstance na osnovu uzorka krvi sportaša. Poznato je da vrijedi

$$P(\text{test je pozitivan}|\text{supstanca je prisutna}) = 0.99$$

ali i

$$P(\text{test je pozitivan}|\text{supstanca nije prisutna}) = 0.03.$$

Pretpostavite da je slučajno odabrani sportaš pozitivan na test, te da supstancu koristi u danom trenutku tek 1 u 8000 sportaša u populaciji, izračunajte vjerojatnost da je ovaj sportaš koristio ovo nedopušteno sredstvo?

Zadatak 8. Prepostavite da za neki gen u diploidnoj populaciji postoje dva oblika – alela: A i a. Neka su vjerojatnosti da je slučajno odabrani gen u populaciji prvi odn. drugi od njih p odn $q = 1 - p$. Tri su moguća genotipa za osobe u ovoj populaciji: AA, aa i Aa. Pretpostavite model **slučajnog parenja** (engl. random mating) odn. da su pateralni G_p i materalni gen G_m svake jedinke u novoj generaciji nezavisno izabrani. Pokažite da je vjerojatnost genotipova AA, aa i Aa, redom p^2, q^2 i $2pq$.

Ako frekvencija genotipova i alela u populaciji zadovoljava ovakve pretpostavke govorimo o **Hardy–Weinbergovoj ravnoteži**.

Zadatak 9. Pokažite da ne postoji populacija kao u prethodnom zadatku koja zadovoljava uvjete slučajnog parenja i Hardy–Weinbergove ravnoteže, no takva da sva tri genotipa imaju istu frekvenciju odn. 1/3.

Zadatak 10. Promotrite u diploidnoj populaciji gen sa dva alela A i a, koji recesivno utječu na neku osobinu jedinki, dakle jedino jedinke genotipa aa posjeduju ovu osobinu. Pretpostavite da su aleli jednako zastupljeni ($p = q = 1/2$), a genotipovi u Hardy-Weinbergovoj ravnoteži. Pokažite da je vjerojatnost da će potomak dvije jedinke koje nemaju ovu osobinu imati traženu osobinu (npr. kuštravost) 1/9.

Zadatak 11. Pretpostavite da studiramo monogenetsku bolest koja ovisi o genu koji se u populaciji nalazi u alelima A i a. Pretpostavite da je 20% jedinki u populaciji homozigot genotipa AA, te da je 5% posto populacije bolesno i homozigot genotipa AA. Izračunajte vjerojatnost da je slučajno odabrana jedinka genotipa AA bolesna. Ova uvjetna vjerojatnost se na engl. naziva i *penetrance* odn. prodornost.

Zadatak 12. Pretpostavite da studiramo monogenetsku bolest koja ovisi o genu koji se u populaciji pojavljuje u dva alela: A i a, te da su poznate vjerojatnosti

$$\begin{aligned} f_0 &= P(\text{osoba je bolesna}|\ \text{osoba je genotipa } aa), \\ f_1 &= P(\text{osoba je bolesna}|\ \text{osoba je genotipa } Aa), \\ f_2 &= P(\text{osoba je bolesna}|\ \text{osoba je genotipa } AA). \end{aligned}$$

Ako alel A povećava vjerojatnost prisutnosti bolesti očekujemo $f_0 \leq f_1 \leq f_2$. Ako je $f_0 < f_1 = f_2$ utjecaj alela A je dominantan, a ako je $f_0 = f_1 < f_2$ on je recesivan. ⊖

Neka su frekvencija genotipova AA, aa i Aa u populaciji 0.01%, 98.01% i 1.98% i neka je utjecaj alela A dominantan na proučavanu bolest. Pretpostavite $f_0 = 0.1$, $f_1 = 0.4$, odredite vjerojatnost da će slučajno odabrana osoba u ovoj populaciji imati danu bolest.

4 RAZDIOBE

4.1 Slučajne varijable

U mnogim primjenama tijekom slučajnog pokusa prikupljamo podatke o nekom numeričkom obilježju jedinki u populaciji. Visina, težina ili uspjeh na ispitu za slučajno odabranog studenta su primjer takvih obilježja. Matematički model za numeričke rezultate slučajnog pokusa je **slučajna varijabla**.

Slučajna varijabla, npr. X je funkcija koja svakom elementarnom ishodu pokusa ω pridružuje broj $X(\omega)$. Dakle za $x \in \Omega$

$$\omega \to X(\omega)\,.$$

Primjer 4.1.1 Pretpostavimo da nakon bacanja novčića ako padne pismo dobijamo 1 euro, a ako padne glava gubimo 2 eura. Pokus opisuje npr. $\Omega = \{\mathrm{P},\mathrm{G}\}$. Našu zaradu opisuje slučajna varijabla definirana sa

$$X(\mathrm{P}) = 1,\ X(\mathrm{G}) = -2.$$

Primjetite da je kod bacanja simetričnog novčića razumno pretpostaviti $P(\mathrm{P}) = P(\mathrm{G}) = 1/2$ pa je dakako i

$$P(X = 1) = P(X = -2) = 1/2.$$

⌟

Jasno je da slučajne varijable primaju konačno međusobno različitih vrijednosti ako je i sam skup Ω konačan. No to se može dogoditi dakako i na beskonačnim Ω, ako X razne elementarne ishode iz Ω preslika u iste realne brojeve. Tada, kao u gornjem primjeru možemo jednostavno navesti sve međusobno različite vrijednosti koje poprima X, kao i pripadne vjerojatnosti. Ako je Ω prebrojivo beskonačan, odn. ako X prima najviše prebrojivo beskonačno međusobno različitih vrijednosti, mi katkad možemo zadati kolike su vjerojatnosti da X poprimi bilo koju od mogućih vrijednosti. U oba slučaja kažemo da je X diskretna slučajna varijabla. U gornjem primjeru slučajna varijabla X može primiti samo dvije vrijednosti 1 i -2, pa je stoga i diskretna. Prisjetite se: prebrojivo beskonačni skupovi su oni čije članove možemo ispisati u niz.

Važna razlika se pojavljuje tek kod onih slučajnih varijabli koje mogu poprimiti neprebrojivo beskonačno mnogo međusobno različitih vrijednosti, npr. sve realne brojeve između 0 i 1. Tada ne možemo jednostavno ispisati sve vrijednosti koje X može poprimiti i pripadne vjerojatnosti, što ovakve slučajne varijable čini bitno složenijim.

4.2 Diskretne slučajne varijable

Općenito međusobno različite vrijednosti koje poprima **diskretna slučajna varijabla** možemo napisati kao niz: $a_1, a_2, a_3, \ldots$. Definirajmo niz

$$p_i = P(X = a_i) = P(\omega \in \Omega : X(\omega) = a_i), \quad i = 1, 2, \ldots$$

Kažemo da ova dva niza odn. tablica

$$\begin{pmatrix} a_1 & a_2 & a_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix},$$

predstavljaju **razdiobu** ili **distribuciju slučajne varijable** X. Pišemo

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}. \tag{4.2.1}$$

Ponekad koristimo i tablicu sljedećeg oblika

x	a_1	a_2	a_3	$\dots$
$P(X=x)$	p_1	p_2	p_3	$\dots$

Primjetite da su brojevi p_i uvijek nenegativni, te da vrijedi

$$\sum_i p_i = 1.$$

No primjetimo i da svaka tablica oblika kao u (4.2.1) zadaje razdiobu neke slučajne varijable čim su svi p_i nenegativni i vrijedi ovaj posljednji uvjet.

Primjer 4.2.1 Ako bacamo dva novčića, pokus opisuje npr. $\Omega = \{PP, PG, GP, GG\}$. Ako nas zanima broj pisama koji je pao, to je slučajna varijabla zadana sa $X(PP) = 2$, $X(PG) = X(GP) = 1$, i $X(GG) = 0$. Ako je novčić nepristran, svi su elem. dogadjaji jednako vjerojatni. Tako da X ima sljedeći zakon razdiobe

$$X \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}.$$

Zakon razdiobe možemo takodjer prikazati i grafički stupčastim dijagramom ili tzv. vjerojatnosnim histogramom na jednostavan način. Napravite to za broj pisama u prethodnom primjeru, ali i za broj pisama nakon bacanja 3 novčića. ⌟

Primjer 4.2.2 Ako neki gen u diploidnoj populaciji dolazi u dva oblika – alela: A i a. Za slučajno odabranu osobu pateralni G_p i materalni gen G_m mogu imati jednu od dvije vrijednosti u skupu $\{A, a\}$. Bez obzira što ne poprimaju numeričke vrijednosti, G_p i G_m možemo smatrati diskretnim slučajnim varijablama (tako da $\{A, a\}$ zamijenimo sa skupom $\{1, 0\}$ ako baš želimo). Slično možemo napraviti i sa svim drugim kategorijalnim varijablama, pa možemo i govoriti o njihovim razdiobama. Ako su npr. vjerojatnosti da je slučajno odabrani gen u populaciji A odn. a od njih p odn. $q = 1 - p$, možemo pisati

$$G_p \sim \begin{pmatrix} A & a \\ p & q \end{pmatrix}.$$

⌟

Matematičko očekivanje

Neka je X slučajna varijabla takva da vrijedi

$$X \sim \begin{pmatrix} a_1 & a_2 & a_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}.$$

Matematičko očekivanje slučajne varijable X definiramo kao sumu

$$E\,X = E(X) = \sum_i a_i p_i\,. \tag{4.2.2}$$

Uočite, suma u gornjem izrazu može biti i beskonačna, no i tada očekivanje definiramo kao pripadnu sumu reda ako ona postoji. Kako postoje divergenti redovi čija suma nije odrediva, postoje i slučajne varijable za koje očekivanje ne definiramo.

Primjer 4.2.3 Nađimo matematičko očekivanje za broj pisama nakon bacanja dva simetrična novčića. Prema Primjeru 4.2.1

$$E(X) = 0\frac{1}{4} + 1\frac{1}{2} + 2\frac{1}{4} = 1.$$

Izračunajte očekivanje i za broj pisama nakon bacanja 3 odn. 4 simetrična novčića. ⌟

Ako su X i Y dvije slučajne varijable, a α proizvoljan realan broj, matematičko očekivanje zadovoljava

$$E(\alpha X) = \alpha E(X)$$
$$E(X+Y) = E(X) + E(Y)$$

Može se pokazati da za proizvoljnu funkciju $g : \mathbb{R} \to \mathbb{R}$ vrijedi

$$E[g(X)] = \sum_i g(a_i) p_i\,, \tag{4.2.3}$$

kad god očekivanje slučajne varijable $g(X)$ postoji.

Varijanca i standardna devijacija

Ako slučajna varijabla X ima očekivanje EX, stavimo $g(u) = (u - E(X))^2$, tada definiramo **varijancu slučajne varijable** X kao broj

$$\operatorname{var} X = E[g(X)] = E[(X - EX)^2]\,.$$

Prema (4.2.3) vrijedi

$$\operatorname{var} X = \sum_i (a_i - E(X))^2 p_i\,.$$

Može se pokazati da za varijancu vrijedi

$$\operatorname{var}(X) = E(X^2) - (E(X))^2,$$

a za sve realne brojeve a, b

$$\operatorname{var}(aX + b) = a^2 \operatorname{var} X.$$

Varijanca je intuitivno mjera raspršenosti razdiobe slučajne varijable oko njenog očekivanja. Kao i uzoračka varijanca, i varijanca razdiobe katkad se koristi kao mjera rizika, tako npr. ako dvije investicije imaju slučajan dobitak istog očekivanja, manje rizičnom se smatra ona koja ima manju varijancu tog dobitka. Uočimo još da smo varijancu definirali kao očekivanje slučajne varijable $g(X) = (X - EX)^2$. Kako ne možemo definirati očekivanje baš za sve slučajne varijable, može se dogoditi da ni varijanca nekih slučajnih varijabli ne postoji, no nas razdiobe takvih slučajnih varijabli nas za sada u načelu neće zanimati.

Standardnu devijaciju slučajne varijable X definiramo kao broj

$$\sigma_X = +\sqrt{\operatorname{var} X}.$$

Primjer 4.2.4 Nadjite varijancu odn. standardnu devijaciju za broj pisama nakon bacanja 3 novčića. Pokažite

$$\operatorname{var} X = \frac{3}{4}$$

⌟

Primjetite matematičko očekivanje, varijanca, i standardna devijacija imale su svoj ekvivalent medju deskriptivnim statistikama. Za diskretne razdiobe možemo definirati i medijan, kvartile odn. kvantile slično kao što smo radili i za slučajni uzorak. Pokušajte pogoditi kako bismo definirali medijan npr.

Zajednička razdioba dvije slučajne varijable

Zanimljivo nam je promatrati i slučajne pokuse kod koji dvije slučajne varijable X i Y istom elementarnom dogadjaju pridružuju dva realna broja.

Primjer 4.2.5 Za slučajni pokus bacanja igraće kocke neka je $X(\omega) = 1$ za parne ω, a 0 za neparne. A $Y(\omega) = 1$ za $\omega = 5$ ili 6, a 0 za sve ostale. Tada zajednički zakon razdiobe od X i Y možemo zadati tablicom

	0	1
0	$P(X = 0, Y = 0) = 2/6$	$P(X = 1, Y = 0) = 2/6$
1	$P(X = 0, Y = 1) = 1/6$	$P(X = 1, Y = 1) = 1/6$

⌟

Općenito zajednički zakon razdiobe za diskretne slučajne varijable X i Y možemo zadati tablicom

	a_1	a_2	a_3	$\dots$
b_1	$P(X=a_1, Y=b_1)$	$P(X=a_2, Y=b_1)$	$P(X=a_3, Y=b_1)$	$\dots$
b_2	$P(X=a_1, Y=b_2)$	$P(X=a_2, Y=b_2)$	$P(X=a_3, Y=b_2)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Primjetite da je suma vjerojatnosti u tablici uvijek 1.

Za diskretne slučajne varijable X i Y kažemo da su **nezavisne** ako za sve a_i i b_j iz gornje tablice vrijedi

$$P(X = a_i,\, Y = b_j) = P(X = a_i)P(Y = b_j)\,.$$

Za slučajne varijable iz Primjera 4.2.5 možemo vidjeti da su nezavisne, npr.

$$P(X = 0, Y = 0) = 1/3 = P(X = 0)P(Y = 0)\,.$$

Slično možemo provjeriti i preostale jednakosti. S druge strane, ako definiramo na slučajnom pokusu iz istog primjera i slučajnu varijablu Z koja je 1 samo za $\omega = 1$ a 0 inače. Tada

$$P(X = 1, Z = 1) = 0 \neq P(X = 1)P(Z = 1)\,,$$

pa X i Z nisu nezavisne.

Primjer 4.2.6 Za slučajno odabranu osobu u diploidnoj populaciji iz primjera 4.2.2 pateralni G_p i materalni gen G_m su slučajne varijable s vrijednostima u skupu $\{A, a\}$. Uočimo, pretpostavka o modelu **slučajnog parenja** (engl. random mating) u teorijskoj biologiji je zapravo pretpostavka da su G_p i G_m nezavisne slučajne varijable.

Slično, Mendelovi zakoni o nasljeđivanju mogu se interpretirati vjerojatnosno. Ako pratimo križanje dvije jedinke s genotipom $YyRr$ (prateći dva različita gena), prvi zakon (o segregaciji) govori da možemo u potomstvu očekivati genotip Yy s vjerojatnošću $1/2$, dok YY i yy možemo očekivati s vjerojatnošću $1/4$ (preciznije, vrijedi uniformni model u kojem su četiri ishoda za pateralni i materalni alel $\{(YY), (Yy), (yY), (yy)\}$ jednako vjerojatna). Drugi zakon govori da se to neovisno događa za dva različita gena, tj. da vjerojatnosti možemo množiti pa je vjerojatnost da ćemo u potomstvu vidjeti jedinku genotipa $yyrr$ npr. jednaka $1/4 \cdot 1/4 = 1/16$, itd.

⌟

Za nezavisne (diskretne numeričke) slučajne varijable vrijedi

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

i

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y).$$

Za slučajne varijable X i Y definiramo **kovarijancu** od X i Y kao

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Lako se vidi da vrijedi

$$\mathrm{cov}(X,Y) = E(XY) - EXEY.$$

Posebno je za nezavisne slučajne varijable X i Y kovarijanca $\mathrm{cov}(X,Y)$ uvijek jednaka 0.

Iz zajedničke razdiobe kovarijancu računamo po formuli

$$\mathrm{cov}(X,Y) = \sum_i \sum_j (a_i - EX)(b_j - EY)P(X = a_i, Y = b_j)\,.$$

Nadalje vrijedi

$$\mathrm{var}\,(X+Y) = \mathrm{var}\,X + \mathrm{var}\,Y + 2\mathrm{cov}(X,Y).$$

Za slučajne varijable X i Y definiramo **koeficijent korelacije** kao

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Ako $\mathrm{cov}(X,Y) = 0$, slijedi $\mathrm{corr}(X,Y) = 0$, u tom slučaju kažemo da su X i Y **nekorelirane**. Npr. nezavisne slučajne varijable su uvijek nekorelirane, no obrnuto ne vrijedi.

Broj $\mathrm{corr}(X,Y)$ je uvijek izmedju -1 i 1. U slučaju $\mathrm{corr}(X,Y) = \pm 1$ slučajne varijable su X i Y potpuno linearno zavisne, preciznije, za neke brojeve a, b vrijedi

$$Y = aX + b\,.$$

Binomna razdioba

Kako smo vidjeli zakonom razdiobe odredjujemo **vjerojatnosni model** za izvjesno numeričko obilježje koje očitavamo nakon slučajnog pokusa. Najjednostavniji primjer pokusa je onaj kod kojeg imamo samo dva moguća ishoda: "uspjeh" i "neuspjeh". Neka je vjerojatnost uspjeha $p \in [0,1]$. Zakon razdiobe slučajne varijable X koja iznosi 1 ako se dogodio "uspjeh", a 0 ako se dogodio "neuspjeh", izgleda tada ovako

$$X \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix},$$

gdje je $q = 1 - p$. Slučajnu varijablu X s ovakvom razdiobom zovemo **Bernoullijeva slučajna varijabla s parametrom** p.

Ako sada gore opisan pokus ponavljamo n puta *nezavisno*, tada nam je interesantno vidjeti kako se ponaša ukupan broj uspjeha, označimo i njega sa X. Jasno, X je svakako izmedju 0 i n, no kolika je vjerojatnost dogadjaja $\{X = k\}$?

Ispostavlja se za $k = 0, 1, \ldots, n$

$$P(X = k) = \binom{n}{k} p^k q^{n-k}\,.$$

Slučajnu varijablu X koja ima ovakvu razdiobu zovemo **binomna slučajna varijabla s parametrima** n i p. Pišemo katkad skraćeno $X \sim B(n,p)$. Provjerite da je binomna

razdioba dobro definirana, odn. da je suma ovih vjerojatnosti 1. Imate li argument za gornju formulu? Za $n = 4$, $p = 1/4$ napišite tablicu razdiobe za $X \sim B(4, 1/4)$. Primjetite, Bernoullijeva slučajna varijabla je specijalno i binomna, no tada je broj ponavljanja $n = 1$.

Primjer 4.2.7 Pretpostavimo da iz posijanog sjemena nikne biljka u 80% slučajeva. Kolika je vjerojatnost da će iz 5 sjemenki niknuti barem dvije? A manje od 2?

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = \cdots = 0.99328.$$

⌟ ⊖

Matematičko očekivanje za X binomnu slučajnu varijablu s parametrima n i p je

$$EX = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = np.$$

Varijanca joj je

$$\operatorname{var} X = npq\,.$$

Pa joj je standardna devijacija

$$\sigma_X = \sqrt{npq}\,.$$

Hipergeometrijska razdioba

Pretpostavimo da u populaciji od m jedinki od kojih je $r \leq m$ na neki način obilježeno biramo slučajni uzorak od njih $n \leq m$. Ako sa X označimo slučajnu varijablu koja prebroji obilježene jedinke u našem uzorku, kažemo da je X **hipergeometrijska slučajna varijabla s parametrima** m, r i n. Očito je X broj izmedju 0 i manjeg od brojeva r i n. Zakon razdiobe od X formulom zapisujemo kao

$$P(X = k) = \frac{\binom{r}{k}\binom{m-r}{n-k}}{\binom{m}{n}}\,,$$

$k = 0, 1, \ldots, r$.

Primjer 4.2.8 U kutiji su 15 bijelih i 10 crvenih kuglica, ako na slučajan način izaberemo 8 kuglica iz kutije, kolika je vjerojatnost da je medju njima bar 2 crvene?

Postavimo $m = 25,\ r = 10,\ n = 8$, a tražimo

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = \cdots$$

⌟

Hipergeometrijska slučajna varijabla X ima očekivanje

$$EX = \frac{rn}{m}.$$

Argument za ovu formulu možemo dati jednostavno: X je ovdje suma n Bernoullijevih slučajnih varijabli X_i koje određuju da li je i-ta jedinka u uzorku obilježena, a svaka od njih je to s vjerojatnošću r/m. Uočite da su ove X_i zavisne, tako da varijancu ne možemo baš tako jednostavno izračunati. Ipak pokazuje se

$$\operatorname{var} X = n\frac{r(m-r)(m-n)}{m^2(m-1)}.$$

Poissonova razdioba

Pretpostavimo da je u iznimno velikoj populaciji relativno malo označenih jedinki. Ako uzmemo veliki uzorak iz populacije ispostavlja se da razdioba broja obilježenih jedinki u uzorku, označimo ga sa X, često ima sljedeći oblik

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda},$$

za $k = 0, 1, 2, \ldots$ i neku konstantu $\lambda > 0$. Za ovakvu slučajnu varijablu X kažemo da je **Poissonova slučajna varijabla a parametrom** λ.

Ako je X Poissonova slučajna varijabla a parametrom λ tada je

$$EX = \operatorname{var} X = \lambda.$$

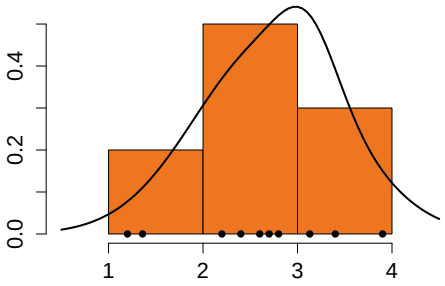
Primjer 4.2.9 Pretpostavljamo da želimo prebrojati broj izvjesnih rijetkih planktona u uzorku od 1l vode iz Jadranskog mora. Označimo taj broj sa X i pretpostavimo da znamo da je očekivani broj planktona u 1l vode 3.8. Tada bismo mogli X modelirati Poissonovom razdiobom s parametrom 3.8. ⌟

4.3 Neprekidne slučajne varijable

Diskretne slučajne varijable su nam služile kao model kod prebrojavanja raznih elemenata u uzorku. Primale su vrijednosti tipično u skupu $0, 1, 2, \ldots$.

Neka numerička obilježja, npr. visina u m, padaline u $\mathrm{l/m^2}$, duljina životnog vijeka u godinama i sl. mogu teoretski poprimiti sve vrijednosti u nekom intervalu. Kako u praksi sva mjerenja zaokružujemo na određeni broj decimalnih mjesta i njih bismo mogli modelirati diskretnim slučajnim varijablama. No pokazuje se da postoji matematički jednostavniji model, posebno kad je različitih vrijednosti koje poprimaju naša mjerenja zaista mnogo.

Sjetimo se histograma koji smo koristili kod neprekidnih numeričkih obilježja. Podijelili smo skup mogućih vrijednosti u intervale oblika $I_j = [a_j, a_{j+1})$, izračunali bismo relativnu frekvenciju podataka u j-tom razredu i podijelili ga s njegovom duljinom.



Slika 4.1: Histogram i teorijska funkcija gustoće za podatke označene točkama

Histogram je bio stepenasti dijagram kojemu je ukupna površina koju omeđuje iznosila 1. Relativna frekvencija svakog razreda bila je jednaka površini histograma iznad tog intervala. Intuitivno očekujemo da relativna frekvencija podataka u j-tom intervalu odgovara (bar približno) vjerojatnosti da će mjerenje za slučajno odabranu jedinku iz populacije pasti u taj interval.

Ova intuitivna interpretacija sugerira sljedeću **ideju**: ako poželimo svakom intervalu dodijeliti njegovu vjerojatnost, mogli bismo to učiniti preko neke općenitije funkcije f, tako da vjerojatnost upadanja obilježja (odn. slučajne varijable) u taj interval bude jednaka površini omeđenoj rubovima intervala i grafom funkcije f.

Po analogiji s histogramom, prirodno je tada da je ukupna površina ispoda grafa od f jednaka 1, te da je f nenegativna funkcija. Takvu fukciju $f : \mathbb{R} \to \mathbb{R}$ zvat ćemo **funkcija gustoće**, za njih dakle vrijedi

$$\int_{-\infty}^{\infty} f(t)\mathrm{d}t = 1\,, \;\; f(t) \geq 0.$$

Slika 4.3 sugerira kako histogram stvarnih podataka ne mora baš savršeno odgovarati zamišljenoj teorijskoj fukciji f.

Za slučajnu varijablu X kažemo da je **neprekidna** ako postoji funkcija gustoće $f = f_X$ takva da za sve $a < b$ vrijedi

$$P(a \leq X \leq b) = \int_a^b f_X(t)\mathrm{d}t\,.$$

Posebno za neprekidnu slučajnu varijablu X i bilo koji realan broj a vrijedi

$$P(X = a) = 0\,!?!$$

Nadalje

$$P(X \leq b) = \int_{-\infty}^b f_X(t)\mathrm{d}t\,.$$

Funkcija $F_X(b) = P(X \leq b)$ zove se funkcija distribucije od X. Pokazuje se, relativno lako, da je F_X uvijek neopadajuća funkcija. Jasno uvijek je

$$P(a < X \leq b) = F_X(b) - F_X(a)\,,$$

za sve $a < b$.

Očekivanje i varijanca neprekidnih razdioba

Matematičko očekivanje neprekidne slučajne varijable X definiramo kao

$$EX = \int_{-\infty}^{\infty} t f_X(t)\mathrm{d}t$$

u slučajevima kad je ovaj integral dobro definiran. Ako slučajna varijabla X prima vrijednosti samo u konačnom intervalu $[a, b]$, ovaj integral je lakše izračunati jer je tada

$$EX = \int_{a}^{b} t f_X(t)\mathrm{d}t.$$

Ako je $g : \mathbb{R} \to \mathbb{R}$ neprekidna funkcija takva da $Eg(X)$ postoji tada vrijedi

$$E[g(X)] = \int_{-\infty}^{\infty} g(t) f_X(t)\mathrm{d}t.$$

Tako da je **varijanca** neprekidne slučajne varijable X

$$\operatorname{var} X = E[(X - EX)^2] = \int_{-\infty}^{\infty} (t - EX)^2 f_X(t)\mathrm{d}t\,,$$

dok je standardna devijacija kao i prije $\sigma_X = \sqrt{\operatorname{var} X}$. Kao i u diskretnom slučaju za sve realne α i β i slučajne varijable X i Y vrijedi

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$$

te

$$\operatorname{var}(\alpha X + \beta) = \alpha^2 \operatorname{var} X.$$

Zajednička razdioba neprekidnih slučajnih varijabli

Ukoliko pratimo dvije neprekidne slučajne varijable X i Y na istom vjerojatnosnom prostoru, katkad i njihova zajednička razdioba ima gustoću. Kažemo da je $f_{X,Y}$ gustoća zajedničke razdiobe za X i Y, ako vrijedi

$$P(X \leq a\,, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, y)\mathrm{d}y\mathrm{d}x\,.$$

I u tom slučaju interesantno je izračunati i kovarijancu i koeficijent korelacije između ovih slučajnih varijabli, za što možemo koristiti formule

$$\operatorname{cov}(X,Y) = \int_{-\infty}^{a}\int_{-\infty}^{b}(x-EX)(y-EY)f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x$$

odn.

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Koeficijent korelacije uvijek pripada intervalu $[-1,1]$, a za slučajne varijable X i Y za koje $\operatorname{corr}(X,Y)=0$ kažemo da su nekorelirane.

Iako smo nezavisnost definirali samo za diskretne slučajne varijable, to možemo učiniti i za sve ostale slučajne varijable, posebno i za neprekidne. Tako su npr. dvije sl. varijable X i Y nezavisne, ako za sve a,b u $\mathbb{R}$ vrijedi

$$P(X\le a, Y\le b) = P(X\le a)P(Y\le b).$$

Može se pokazati da i ovdje nezavisne slučajne varijable imaju koeficijent korelacije 0, pa su prema tome nekorelirane.

Normalna razdioba

Neprekidna slučajna varijabla X ima **normalnu razdiobu s parametrima** $\mu\in\mathbb{R}$ i $\sigma^2>0$ ako joj je funkcija gustoće

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}.$$

Skraćena oznaka za ovu tvrdnju je $X\sim N(\mu,\sigma^2)$. Primjetimo $f_X(t)>0$ za sve $t\in\mathbb{R}$, tako da X prima vrijednosti u cijelom skupu realnih brojeva.

Matematičko očekivanje slučajne varijable $X\sim N(\mu,\sigma^2)$ je

$$EX = \int_{-\infty}^{\infty} t\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t = \mu.$$

Dok varijanca iznosi upravo

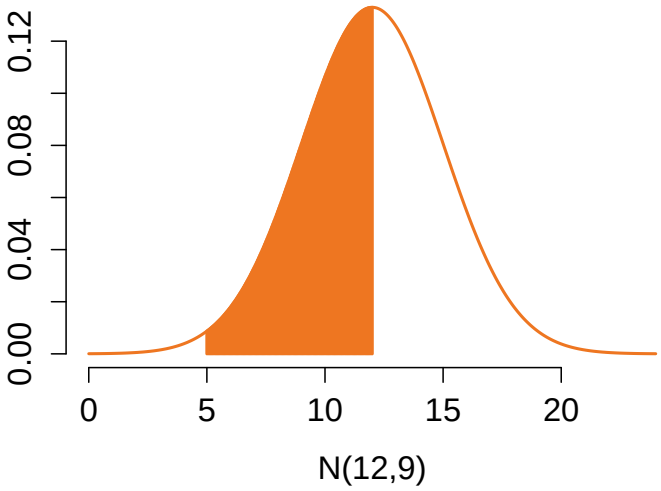
$$\operatorname{var}X = \sigma^2.$$

Ako je $X\sim N(\mu,\sigma^2)$, za zadane $a,b\in\mathbb{R}$, $a\neq 0$ vrijedi

$$Y = aX+b \sim N(a\mu+b, a^2\sigma^2).$$

Zaista, ako je $a>0$, za proizvoljne realne $c<d$, slijedi

$$\begin{aligned}
P(c\le Y\le d) &= P(c\le aX+b\le d) = P(\frac{c-b}{a}\le X\le \frac{d-b}{a})\\
&= \int_{\frac{c-b}{a}}^{\frac{c-b}{a}} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t\\
&= \text{zamijena } u = at+b\\
&= \int_c^d \frac{1}{a\sigma\sqrt{2\pi}}e^{-\frac{(u-a\mu-b)^2}{2a^2\sigma^2}}\mathrm{d}u
\end{aligned}$$



Slika 4.2: Normalna gustoća s parametrima 12 i 9. Osjenčana površina odgovara vjerojatnosti da slučajna varijabla X bude između 5 i 12.

Posebno ako je $X \sim N(\mu, \sigma^2)$, tada postavimo $a = 1/\sigma$, $b = -\mu/\sigma$ da bismo vidjeli

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Ovakva transformacija slučajne varijable X se zove **standardizacija**, a slučajna varijabla $Z \sim N(0, 1)$ ima tzv. **standardnu (ili jediničnu) normalnu razdiobu**. Njenu funkciju distribucije označavamo posebno

$$\Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t.$$

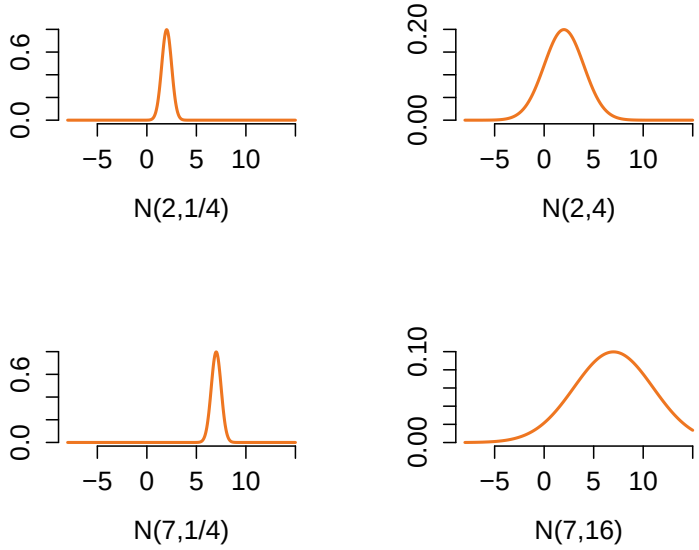
Ovu funkciju nije lako izračunati za različite x, no zato su njene vrijednosti tabelirane u svim statističkim paketima, pa i u programima šire namjene kao što su npr. spreadsheet programi.

Preko vrijednosti funkcije Φ možemo odrediti funkciju distribucije za sve normalne slučajne varijable. Naime za sve $X \sim N(\mu, \sigma^2)$ i $a < b$ nalazimo

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right),$$

a za standardiziranu slučajnu varijablu $Z = (X - \mu)/\sigma$, desna strana je jednaka

$$P(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma})$$



Slika 4.3: Normalne gustoće s raznim parametrima. Maksimum normalne gustoće uvijek je u μ, ona je simetrična oko točke μ, a raširenost funkcije gustoće određuje parametar σ^2.

Iz osnovnih pravila za računanje vjerojatnosti (ali i integrala) za $a < b$ slijedi

$$\begin{aligned} P(a \le Z \le b) &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t \\ &= P(Z \le b) - P(Z \le a) = \Phi(b) - \Phi(a)\,. \end{aligned}$$

Tako da za $X \sim N(\mu, \sigma^2)$ vrijedi

$$P(a \le X \le b) = P(Z \le \frac{b-\mu}{\sigma}) - P(Z \le \frac{a-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}).$$

Posebno uočite

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = P(-2 \le Z \le 2) = \Phi(2) - \Phi(-2) \approx 0.9545,$$

a

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = P(-3 \le Z \le 3) = \Phi(3) - \Phi(-3) \approx 0.9973,$$

te

$$P(X > \mu + 1.65\sigma) = P(Z > 1.65) = 1 - \Phi(1.65) \approx 0.05.$$

No pitanje ostaje – zašto smo kao prvu neprekidnu razdiobu uveli baš normalnu? Za razliku od binomne, hipergeometrijske ili Poissonove, do nje nismo dospjeli preko nekakvog pokusa. Ipak normalna razdioba je daleko najvažnija razdioba u statistici.

Postoje dva bitna razloga:

i) Naime mnoge pojave u prirodi su normalno distribuirane: npr. fizičke karakteristike u raznim populacijama biljaka i životinja.

ii) Postoji više matematičkih rezultata tzv. *centralnih graničnih teorema* koji pokazuju da se razne razdiobe mogu aproksimirati normalnom.

Zapravo se ii) može tumačiti i kao razlog za i).

4.4 Normalna aproksimacija

Kao prvu diskretnu razdiobu uveli smo binomnu razdiobu, tj. $X \sim B(n,p)$. Kao što znamo ova slučajna varijabla ima očekivanje np i varijancu npq, pa nam sljedeća transformacija

$$\frac{X-np}{\sqrt{npq}}$$

daje slučajnu varijablu koja ima očekivanje 0 i varijancu 1. No to i dalje ostaje diskretna razdioba. Zbog toga je možda iznenađujuće da za sve $0<p<1$ i za dovoljno velike n

$$P\left(\frac{X-np}{\sqrt{npq}} \le x\right) \approx P(Z \le x),$$

gdje je $Z \sim N(0,1)$, to je posljedica tzv. de Moivre–Laplaceovog teorema (odn. centralnog graničnog teorema općenito).

Zbog gornje aproksimacije vrijedi

$$P(a \le X \le b) \approx P\left(\frac{a-np}{\sqrt{npq}} \le Z \le \frac{b-np}{\sqrt{npq}}\right).$$

Ova aproksimacija je vrlo korisna i upotrebljava se već u slučajevima kada $np \ge 5$ i $nq \ge 5$.

Kako je binomna razdioba diskretna, a standardna normalna neprekidna u praksi se često koristi i **korekcija po neprekidnosti** koja aproksimaciju čini još boljom. Stoga umjesto

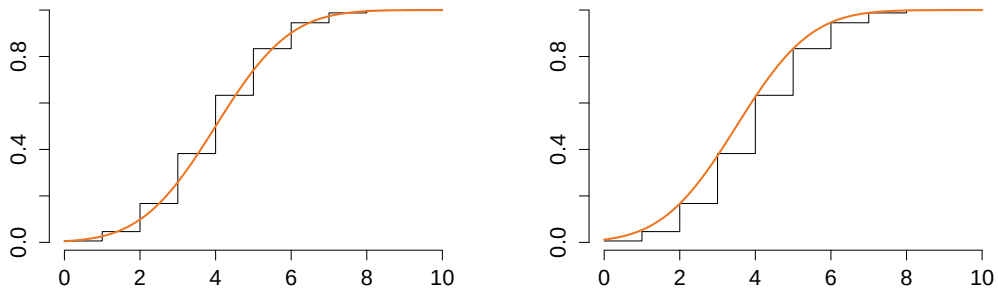
$$P(X \le b) \approx P\left(Z \le \frac{b-np}{\sqrt{npq}}\right), \tag{4.4.1}$$

koristimo

$$P(X \le b) \approx P\left(Z \le \frac{b+1/2-np}{\sqrt{npq}}\right). \tag{4.4.2}$$

Iz istog razloga koristimo i sljedeću aproksimaciju

$$P(a \le X \le b) \approx P\left(\frac{a-1/2-np}{\sqrt{npq}} \le Z \le \frac{b+1/2-np}{\sqrt{npq}}\right).$$



Slika 4.4: Aproksimacija binomne razdiobe normalnom bez i uz korištenje korekcije po neprekidnosti. Usporedba dvije strane u formulama (4.4.1) i (4.4.2) za $n = 10,\ p = 0.4$

Aproksimaciju binomne razdiobe normalnom prvi su objasnili matematičari de Moivre i Laplace. No nakon toga je dokazano da ona vrijedi i puno općenitije. Naime binomna slučajna varijabla je zapravo zbroj od n nezavisnih slučajnih varijabli koje sve imaju vrijednosti 0 ili 1, ovisno o ishodima pojedinačnih pokusa. Njena razdioba za velike n poprima isti oblik kao i normalna uz dobro odabranu varijancu i očekivanje. Ova tvrdnja vrijedi u vrlo generalnoj situaciji, a možemo je sažeti na sljedeći način.

"Centralni granični teorem"
Sva numerička obilježja koja su rezultat puno malih i nepovezanih slučajnih utjecaja imat' će približno normalnu razdiobu.

4.5 Druge neprekidne razdiobe

χ^2 razdioba

Slučajna varijabla X ima χ^2 **razdiobu s ν stupnjeva slobode** ako prima samo nenegativne vrijedosti i ima gustoću

$$f_X(t) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} t^{\nu/2-1} e^{-\frac{t}{2}},\ t > 0.$$

gdje je $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$.

Može se pokazati da vrijedi: $EX = \nu$, $\operatorname{var} X = 2\nu$. Zanimljivo, ako je $X \sim N(0,1)$, onda X^2 ima χ^2 razdiobu s 1 stupnjem slobode. Općenitije, ako su $X_1, \ldots, X_\nu$ nezavisne $N(0,1)$ distribuirane slučajne varijable, tada $X_1^2 + \cdots + X_\nu^2$ ima χ^2 razdiobu s ν stupnjeva slobode. ⊖

Studentova razdioba

Slučajna varijabla X ima **Studentovu t–razdiobu** sa ν stupnjeva slobode ako ima gustoću

$$f_X(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad t \in \mathbb{R}.$$

Ako je $\nu > 2$ za njeno očekivanje odn. varijancu vrijedi

$$EX = 0, \quad \text{odn. } \operatorname{var} X = \frac{\nu}{\nu - 2}.$$

Interesantno je da za $\nu = 1$, očekivanje od X nije definirano.

Uniformna razdioba

Slučajna varijabla X ima **uniformnu razdiobu na intervalu** $[a, b]$, $a < b$, ako prima vrijedosti u intervalu $[a, b]$ i ima gustoću

$$f_X(t) = \frac{1}{b - a}, \quad t \in [a, b].$$

Može se pokazati da vrijedi:

$$EX = \frac{a + b}{2}, \quad \operatorname{var} X = \frac{(b - a)^2}{12}.$$

Zadaci

Zadatak 1. Neka je X slučajna varijabla zadana tablicom

$$X \sim \begin{pmatrix} -2 & 0 & 2 & 4 \\ 0.1 & 0.3 & 0.4 & 0.2 \end{pmatrix}.$$

Odredite $E[2X^2]$.

Zadatak 2. Broj štakora na 100 m^2 podrumskih prostorija u nekom gradu je slučajna varijabla s očekivanjem 1.5. No broj štakora na 100 m^2 kanalizacijskog sustava je slučajna varijabla s očekivanjem 4. Koliki je očekivani broj štakora u kompleksu zgrada s 1600 m^2 podrumskih prostorija i 250 m^2 kanalizacijskog sustava?

Zadatak 3. Olivera svaki dan dolazi na posao, ali u 20% slučajeva kasni (nezavisno o prijašnjim kašnjenjima). Koja je vjerojatnost da će u 5 radnih dana zakasniti točno 2 puta?

Zadatak 4. Na ispitu je za svaki zadatak ponuđeno 5 odgovora. Student nasumice odgovara na pitanja pa je vjerojatnost da točno odgovori na neko pitanje 20%. Ako se ispit sastoji od 8 pitanja, odredite vjerojatnost da će student točno odgovoriti na barem jedno.

Zadatak 5. Zajednički zakon razdiobe slučajnih varijabli X i Y je dan sa

$Y \setminus X$	2	4	6
0	0.2	0	0.2
1	a	0.3	0.1

Odredite a i izračunajte kovarijancu $cov(X, Y)$. Jesu li X i Y nezavisne?

Zadatak 6. Neka je $X \sim N(20, 64)$. Odredite $P(15 \leq X \leq 30)$.

Zadatak 7. U tvornici čokolade težina čokolade je normalno distribuirana slučajna varijabla s očekivanjem 100 grama i standardnom devijacijom 5 grama. Odredite težinu u tako da vjerojatnost da će slučajno odabrana tabla čokolade težiti barem u grama iznosi 95%.

Zadatak 8. U nekoj populaciji je kvocijent emocionalne inteligencije normalno distribuirana slučajna varijabla s očekivanjem 100 i standardnom devijacijom 23.4. Iznad kojeg broja je kvocijent emocionalne inteligencije 2% ljudi?

Zadatak 9. U nekoj populaciji čimpanzi vjerojatnost da će mužjaci ostaviti potomke iznosi 55%. Odredite vjerojatnost da će od 100 čimpanzi njih barem 60 imati potomke.

Zadatak 10. U nekoj bolnici je vjerojatnost da se rodi žensko dijete 0.4. Koja je vjerojatnost da među 600 novorođenčadi ima između 220 i 270 djevojčica?

5 PROCJENA PARAMETARA MODELA

5.1 Procjenitelji

Nakon što smo prikupili numeričke podatke i odredili primjeren vjerojatnosni model, tipično želimo procjeniti parametre na osnovu podataka.

Primjer 5.1.1 Ako označimo sa X broj ženskih potomaka u 20 slučajeva prvorođene djece kraljevskih obitelji ili broj adenina u uzorku od 1000 nukleotida odabranih slučajno sa genoma neke vrste, mogli bismo modelirati X kao binomnu slučajnu varijablu, s parametrom $n = 20$ odn. 1000, no s nepoznatim parametrom p. Dakako, jedan prirodan procjenitelj za p je jednostavno X/n. ⌟

Primjer 5.1.2 Ako označimo s X raspon krila slučajno odabrane mušice u nekoj koloniji i prepostavimo da je X normalno distribuirana slučajna varijabla, postavlja se pitanje možemo li odrediti parametre ove razdiobe na osnovu prikupljenog uzorka. Kako su parametri normalne razdiobe upravo njeno očekivanje i varijanca, razumno je pretpostaviti da očekivanje i varijanca uzorka mogu poslužiti u procjeni. ⌟

Slučajni uzorak je niz slučajnih varijabli $X_1, \ldots, X_n$, koji zadovoljava

- slučajne varijable X_i su nezavisne,
- slučajne varijable X_i imaju istu razdiobu.

Slučajni uzorak predstavlja numeričke podatke koje namjeravamo prikupiti tokom istraži vanja.

Procjenitelji parametara su uvijek samo brojevi izvedeni iz slučajnog uzorka. Kako se proizvoljna funkcija uzorka zove i **statistika**, možemo reći da su procjenitelji tek vrste statistika. Uočite: svaka je statistika i sama slučajna varijabla.

Primjer 5.1.3 Aritmetička sredina slučajnog uzorka $X_1, X_2, \ldots, X_n$ je

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Primjetite, sada opažanja modeliramo slučajnim varijablama, pa smo u ovoj definiciji koristili velika slova X. ⌟

5.2 Procjena parametra p binomne razdiobe

Pretpostavimo da se naš slučajni uzorak sastoji od n slučajnih varijabli $X_1, \ldots, X_n$ koje sve imaju vrijednosti 0 ili 1, dakle X_i je 1 ako se u i-tom pokusu dogodio "uspjeh" inače

je $X_i = 0$. Uz oznaku

$$S = X_1 + \cdots + X_n,$$

mi znamo da S zbog nezavisnosti i jednake distribucije slučajne varijabli X_i ima binomnu razdiobu s parametrima n i p. Primjetite da je za Bernoullijeve slučajne varijable X_i, parametar p ujedno njihovo očekivanje. Pitanje je kako iz uzorka procjeniti parametar p.

Primjer 5.2.1 i) na ispitivanju uzorka od 1000 glasača utvrđeno je da izvjesnog kandidata za gradonačelnika podržava njih 513. Procjenite vjerojatnost da taj kandidat uživa podršku slučajn izabranog glasača.
ii) u slučajnom uzorku od 800 nukleotida sa genoma *C. elegans* utvrđeno je da 312 njih predstavlja baze C i G. Koliki je postotak ovih nukelotida u genomu ove vrste?
iii) na financijskim tržištima uzorak od 200 uzastopnih dana pokazuje da su dionice kompanije ZXY porasle u 143 dana. Kolika je vjerojatnost da vrijednost ove dionice poraste tokom slučajno odabranog dana?
iv) U 600 bacanja novčića od 1 eura pismo je palo u 289 slučajeva, kolika je vjerojatnost da padne pismo nakon pojedinog bacanja?

⌟

Pretpostavku da su X_i nezavisne i jednako distribuirane u ovim slučajevima (a i inače) bi u praksi trebalo na neki način i opravdati.

U svim gornjim primjerima, jedan prirodan procjenitelj za vjerojatnost uspjeha p je relativna frekvencija

$$\hat{p} = \frac{S}{n}.$$

Uočite da je $\hat{p}$ slučajna varijabla (prije nego nam je poznat uzorak). Možemo izračunati njeno očekivanje, ono jasno iznosi

$$E\hat{p} = \frac{1}{n}E(S) = \frac{1}{n}np = p.$$

Dakle očekivanje procjenitelja $\hat{p}$ je jednako vrijednosti koju želimo procjeniti, takve procjenitelje zovemo **nepristranim**.

Izračunajmo i varijancu od $\hat{p}$ ona je

$$\operatorname{var}\hat{p} = \frac{1}{n^2}\operatorname{var} S = \frac{1}{n^2}n\operatorname{var} X_1 = \frac{pq}{n},$$

gdje je $q = 1 - p$ kao i do sada. **Standardnu grešku** procjenitelja definiramo kao

$$\text{s.e.}(\hat{p}) = \sqrt{\operatorname{var}\hat{p}} = \sqrt{\frac{pq}{n}}.$$

Dakle s.e.$(\hat{p})$ je zapravo standardna devijacija slučajn varijable $\hat{p}$. Primjetite da ona teži k 0, za $n \to \infty$, jer $\operatorname{var}\hat{p} \to 0$ za $n \to \infty$. Za nepristrane procjenitelje koji imaju i ovo svojstvo kažemo da su **konzistentni**.

Kako je naš uzorak slučajan, nije jako vjerojatno da vrijedi $p = \hat{p}$, no očekujemo da vrijedi $p \approx \hat{p}$ za velike n. Dakle, procjenitelj je "blizu" prave vrijednosti no nije odmah jasno koliko blizu i što to točno znači?

Prisjetimo se de Moivre-Laplaceovog teorema koji kaže da za $0 < p < 1$ i "velike" n slučajna varijabla

$$\frac{S - np}{\sqrt{npq}} = \sqrt{\frac{n}{pq}}(\hat{p} - p)$$

ima približno N(0,1) razdiobu.

Dakle za $a < b$ vrijedi

$$P\left(a \leq \sqrt{\frac{n}{pq}}(\hat{p} - p) \leq b\right) \approx P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Ili malo drugačije zapisano

$$P\left(p + a\sqrt{\frac{pq}{n}} \leq \hat{p} \leq p + b\sqrt{\frac{pq}{n}}\right) \approx \Phi(b) - \Phi(a).$$

Koristeći činjenicu da je $\text{s.e.}(\hat{p}) = \sqrt{pq/n}$ te da vrijedi $\Phi(1.96) - \Phi(-1.96) = 0.95$ (provjerite u tablicama) možemo pisati i

$$P\left(p - 1.96\text{s.e.}(\hat{p}) \leq \hat{p} \leq p + 1.96\text{s.e.}(\hat{p})\right) \approx 0.95,$$

ili uobičajenije

$$P\left(\hat{p} - 1.96\text{s.e.}(\hat{p}) \leq p \leq \hat{p} + 1.96\text{s.e.}(\hat{p})\right) \approx 0.95.$$

Dakle, vjerojatnost da p leži u intervalu

$$(\hat{p} - 1.96\text{s.e.}(\hat{p}), \hat{p} + 1.96\text{s.e.}(\hat{p}))$$

je 95%.

Ovdje je ipak potreban oprez u interpretaciji. U praksi, nakon što prikupimo podatke i broj uspjeha poprimi neku stvarnu vrijednost $S = s$, ne možemo više govoriti o vjerojatnosti pripadanja intervalu jer interval na kraju pokusa više nije slučajan, tako da je prava interpretacija intervala pouzdanosti frekvencionistička: ako su nam pretpostavke korektne i puno puta ponovimo proceduru, naš 95% interval pouzdanosti će sadržavati stvarnu vrijednost parametra u 95% slučajeva. Postoji dakle suptilna, ali bitna razlika između pouzdanosti i vjerojatnosti.

Općenitije, definirajmo za $\alpha \in (0, 1)$, broj z_α kao $(1 - \alpha)$-kvantil razdiobe od $Z \sim N(0, 1)$, tj. kao broj koji zadovoljava

$$1 - \Phi(z_\alpha) = P(Z > z_\alpha) = \alpha.$$

Broj z_α se nekad naziva i gornji α-kvantil standardne normalne razdiobe. Gornji argumenti pokazuju da vrijedi

$$P\left(\hat{p} - z_{\alpha/2}\text{s.e.}(\hat{p}) \leq p \leq \hat{p} + z_{\alpha/2}\text{s.e.}(\hat{p})\right) \approx 1 - \alpha,$$

za "velike" n. Primjetite, već smo koristili: $z_{0.025} = 1.96$.

Zbog svega navedenog interval

$$\big(\hat{p} - z_{\alpha/2}\text{s.e.}(\hat{p}), \hat{p} + z_{\alpha/2}\text{s.e.}(\hat{p})\big)$$

zovemo $(1-\alpha)\cdot 100\%$-tni **interval pouzdanosti** za parametar p. Ipak, postoji problem u praksi, da bismo odredili gornji interval morali bismo znati standardnu pogrešku s.e., no ona sadrži nama nepoznati parametar p, zbog toga je zamijenjujemo nama dostupnom procjenom

$$\text{s.e.}(\hat{p}) \approx \sqrt{\hat{p}(1-\hat{p})/n}.$$

Tako dobijemo interval

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

koji zapravo koristimo u praksi i također zovemo $(1-\alpha)\cdot 100\%$-tni **interval pouzdanosti** za parametar p.

Ovaj interval se sužava kako raste veličina našeg uzorka n ili kada se $\hat{p}$ približava 0 ili 1. Naglasimo, ako je $\hat{p} = 0$ ili 1, ili jako blizu ovih vrijednosti, interval pouzdanosti ne konstruiramo – naime izveden je uz korištenje normalne aproksimacije koja je u takvim slučajevima neopravdana.

Primjetite da smo očekivanje slučajnih varijabli X_i procijenili preko aritmetičke sredine uzorka. Zbog toga se katkad govori da populacijsko očekivanje u ovom slučaju možemo procjeniti očekivanjem uzorka. To i nije neko iznenađenje.

Napomenimo da i intervale pouzdanosti uobičajeno zovemo procjeniteljima za traženi parametar, no oni su takozvani intervalni procjenitelji. Intervalni procjenitelji su općenito korisniji od točkovnih procjenitelja, jer nam daju i ocjenu pogreške u našoj procjeni.

Primjer 5.2.2 U uzorku od 1000 mušica pronađeno je da ih 550 nosi određeni genotip, nađimo interval pouzdanosti 95% za broj p koji predstavlja frekvenciju takvih mušica u populaciji. Uočite p predstavlja i vjerojatnost da je slučajno odabrana mušica iz populacije upravo tog genotipa.

Jasno je

$$\hat{p} = \frac{550}{1000} = 0.55\,.$$

Tako da za s.e.$(\hat{p})$ koristimo

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{0.55(1-0.55)}{1000}} = 0.0157\,.$$

Dakle traženi interval je

$$\hat{p} \pm 1.96 \cdot 0.0157 = 0.55 \pm 0.031\,.$$

Drugim riječima $0.519 < p < 0.581$ uz pouzdanost od 95%.

⌟

Primjer 5.2.3 Agencije za ispitivanje javnog mišljenja tipično koriste uzorak od 1000 ispitanika i napominju da je "moguća greška" $\pm 3\%$. Razlog se može naći u formulama za standardnu grešku našeg procjenitelja, naime

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} \leq \frac{1/2}{\sqrt{1000}},$$

jer je $\hat{p}(1-\hat{p}) \leq 1/4$ za bilo koji $\hat{p} \in [0,1]$. Stoga ako koristimo 95%–tni interval pouzdanosti $\hat{p} \pm 1.96\text{s.e.}(\hat{p})$, prava vrijednost p se s pouzdanošću 95% nalazi oko $\hat{p}$ na udaljenosti ne većoj od

$$1.96\frac{1/2}{\sqrt{1000}} \approx \frac{1}{10\sqrt{10}} \approx 0.03\,.$$

⌟

5.3 Procjena parametra μ normalne razdiobe uz poznatu varijancu

Primjer 5.3.1 Prepostavimo da nam je poznato da određeni instrument uspjeva izmjeriti udio šećera u 100g nekog proizvoda. Pri tom prilikom svakog mjerenja pravi i grešku koja je normalno distribuirana s očekivanjem 0 i standardnom devijacijom od 1.5 g. Nakon 5 mjerenja količine šećer u jednoj vrsti gumenih bombona dobiveni su sljedeći rezultati: 77.7,78.2,78.9,76.9 i 76.7 g u 100g proizvoda. Možete li procjeniti stvarni sadržaj šećera u ovom proizvodu? ⌟

U nastavku je korisno znati nekoliko činjenica o normalnim slučajnim varijablama. Uočite, ako je greška $\varepsilon \sim N(0,\sigma^2)$ distribuirana, tada uz prepostavku da je stvarna koncentracija μ, dobiveni podaci imaju prikaz $\mu + \varepsilon$, te imaju $N(\mu,\sigma^2)$ razdiobu.

Nadalje, suma nezavisnih normalnih slučajnih varijabli ponovo je normalna slučajna varijabla. Dakle, ako je $X_1 \sim N(\mu_1,\sigma_1^2)$ i $X_2 \sim N(\mu_2,\sigma_2^2)$ tada je uz prepostavku njihove nezavisnosti

$$X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)\,.$$

Slično vrijedi i za n nezavisnih normalnih slučajnih varijabli. Posebno, ako $X_1,\ldots,X_n$ čine slučajni uzorak iz normalne razdiobe s parametrima μ i σ^2 tada je njihova suma $S = X_1 + \cdots + X_n$ ponovo normalno distribuirana s očekivanjem $n\mu$ i varijancom $n\sigma^2$.

Kako je parametar μ ujedno očekivanje razdiobe od X_i i njega možemo pokušati procjeniti preko očekivanja uzorka, odn. njegove aritmetičke sredine

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Kako je $\bar{X} = S/n$ i ovo je normalna slučajna varijabla samo ima očekivanje μ i varijancu σ^2/n (pokažite). Dakle $\bar{X}$ je nepristran, ali i konzistentan procjenitelj parametra

μ, jer mu varijanca konvergira k 0.

I ovdje bismo željeli imati intervalni procjenitelj. A njega lako dobijemo ako nam je poznata varijanca σ^2. Naime, standardizacijom slučajne varijable $\bar{X}$ slijedi

$$\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \sim N(0,1).$$

Posebno za sve $a < b$ vrijedi

$$P\left(a \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq b\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a),$$

za slučajnu varijablu $Z \sim N(0,1)$. Ili

$$P\left(\mu + a\sigma/\sqrt{n} \leq \bar{X} \leq \mu + b\sigma/\sqrt{n}\right) = P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Što možemo pisati i kao

$$P\left(\bar{X} - b\sigma/\sqrt{n} \leq \mu \leq \bar{X} + a\sigma/\sqrt{n}\right) = P(-b \leq Z \leq -a) = \Phi(-a) - \Phi(-b).$$

Konačno, 95%-tni **interval pouzdanosti** za parametar μ jednostavno je

$$\left(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right).$$

Općenitije je

$$\left(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right), \tag{5.3.1}$$

$(1-\alpha)100\%$-tni **interval pouzdanosti** za parametar μ.

I ovdje se greška u procjeni smanjuje s rastom veličine uzorka, ali i s opadanjem populacijske varijance σ^2. No pretpostavka o tome da nam je varijanca poznata tipično je nerealistična u praksi, zato i nju moramo procjeniti. Naravno, razumno je pokušati **varijancom uzorka**

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Varijanca uzorka ovako definirana nepristran je i konzistentan procjenitelj za populacijsku varijancu. Primjetite da smo ponovo podatke označili velikim slovima jer su prije pokusa i oni za nas slučajne varijable.

5.4 Procjena parametra μ normalne razdiobe uz nepoznatu varijancu

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine slučajni uzorak iz normalne razdiobe s parametrima μ i σ^2, te neka su nam oba parametra nepoznata. Kako smo vidjeli tada njihova aritmetička

sredina ima $N(\mu, \sigma^2/n)$ razdiobu. Da bismo konstruirali interval pouzdanosti za μ, σ^2 možemo procijeniti sa $\hat{s}^2$. No, iako

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

ima standardnu normalnu razdiobu to ne vrijedi za slučajnu varijablu

$$\sqrt{n}\frac{\bar{X}-\mu}{\hat{s}}.$$

Za ovu slučajnu varijablu prvi je razdiobu odredio William Gosset, i objavio je 1908. u časopisu "Biometrika" pod pseudonimom Student. Gosset je inače radio za pivovaru "Guiness". Danas ovu razdiobu zovemo **Studentova t razdioba**. Ona je jedna od najvažnijih neprekidnih razdioba, tako da i za nju tabeliramo funkciju distribucije odn. njene kvantile.

Zapamtimo da slučajna varijabla

$$\sqrt{n}\frac{\bar{X}-\mu}{\hat{s}}.$$

ima Studentovu t razdiobu s $d = n-1$ **stupnjem slobode**. Broj stupnjeva slobode $n-1$ uvijek je cijeli i pozitivan, on predstavlja parametar t razdiobe. Studentova razdioba ima gustoću simetričnu oko 0, pa joj je i očekivanje 0 (izuzetak je Studentova razdioba s jednim stupnjem slobode, vidi odjeljak 4.5). Napomenimo samo da je za velike n (npr. $n > 100$), Studentova t razdioba gotovo istovjetna sa standardnom normalnom razdiobom.

U tablicama nalazimo brojeve t_α, koji nam daju $(1-\alpha)$-kvantil razdiobe slučajne varijable T koja ima Studentovu t razdiobu s d stupnjeva slobode. Dakle t_α zadovoljava

$$P(T > t_\alpha) = \alpha.$$

Pri tom, tipično različiti reci tablice odgovaraju različitim stupnjevima slobode d. Sada možemo reći da je

$$\left(\bar{X} - t_{\alpha/2}\hat{s}/\sqrt{n}, \bar{X} + t_{\alpha/2}\hat{s}/\sqrt{n}\right), \tag{5.4.1}$$

$(1-\alpha)100\%$-tni **interval pouzdanosti** za parametar μ u ovom slučaju. Ovaj interval možemo pisati i u obliku

$$\left(\bar{X} - t_{\alpha/2}\text{s.e.}(\bar{X}), \bar{X} + t_{\alpha/2}\text{s.e.}(\bar{X})\right),$$

gdje je s.e.$(\bar{X})$ standardna greška procjenitelja $\bar{X}$ tj. naša procjena za nju

$$\text{s.e.}(\bar{X}) = \hat{s}/\sqrt{n}$$

Procjena očekivanja u slučaju odstupanja od normalnosti

Važna je napomenuti da, iako su intervalni procjenitelji parametra μ izvedeni pod pretpostavkom da sami podaci dolaze iz normalne razdiobe, oni predstavljaju razuman izbor i koriste se u praksi čak i u onim slučajevima kada podaci slijede neku drugu razdiobu. Tada su ovi intervali zapravo intervalni procjenitelji očekivanja razdiobe iz koje dolaze podaci.

Ako je uzorak dovoljno velik (npr. $n \geq 30$ ili bolje $n \geq 100$), tada centralni granični teorem omogućuje da procjenimo interval pouzdanosti za μ preko normalne razdiobe dakle formulom (5.3.1). Dakako, u toj formuli poznatu standardnu devijaciju σ zamjenjujemo procjenom tj. sa $\hat{s}$.

Za manje uzorke ovaj teorem ne možemo koristiti, no ako je gustoća podataka u uzorku približno istog "zvonastog" oblika kao i normalna, u praksi možemo i dalje koristiti intervale dobivene iz formule (5.4.1), uz ogradu da su ovako dobiveni intervali zapravo aproksimativni.

5.5 Usporedba podataka s normalnom razdiobom

U praksi je ipak važno znati usporediti razdiobu prikupljenih podataka s normalnom, kako bismo znali da li su korištene statističke metode primjenjive na naše podatke. Usporediti ih možemo npr. koristeći deskriptivne statističke metode koje smo već spominjali.

Jedna je mogućnost da nacrtamo histogram podataka s funkcijom gustoće normalne razdiobe s procjenjenim vrijednostima očekivanja i varijance.

Druga metoda, koja je ujedno ilustrativnija u praksi, je metoda koja uspoređuje kvantile u slučajnom uzorku s kvantilima standardne normalne razdiobe.

Ako uredimo slučajni uzorak $X_1, X_2, \ldots, X_n$ po veličini dobit ćemo uzlazni niz

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

Za bilo koji $0 < \alpha < 1$ možemo definirati α-**kvantil** slučajnog uzorka $X_1, X_2, \ldots, X_n$, kao

$$q_\alpha = X_{(\alpha(n+1))}\,.$$

gdje pretpostavljamo da je $s = \alpha(n+1)$ realan broj između 1 i n.

Pritom kao i prije, broj s između 1 i n, prikažemo u obliku

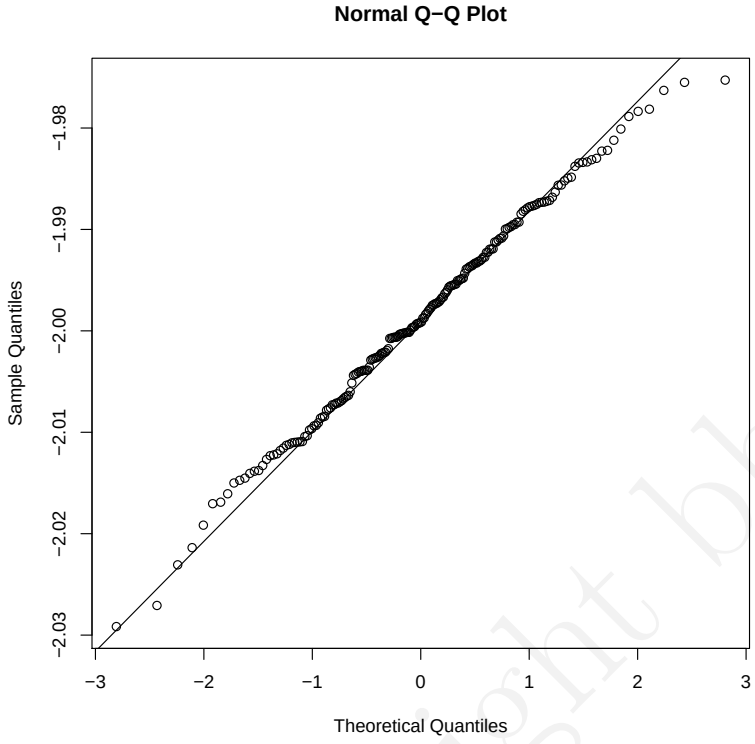
$$s = k + r$$

tako da je $k = 1, 2, \ldots, n$ njegov cijeli, a $0 \leq r < 1$ njegov razlomljeni dio, te definiramo

$$X_{(s)} = (1-r)X_{(k)} + rX_{(k+1)}\,.$$

Kvantili neprekidne slučajne varijable X ili njene razdiobe se definiraju preko funkcije distribucije F od X, tako da za $\alpha \in (0,1)$, α-kvantil od X bude broj x_α takav da vrijedi

$$P(X \leq x_\alpha) = F(x_\alpha) = \alpha.$$



Slika 5.1: Graf kvantila za slučajni uzorak simuliran iz $N(-2, 0.0001)$ razdiobe.

Ako dvije slučajne varijable imaju istu funkciju gustoće (odn. istu funkciju distribucije) tada su im i kvantili jednaki. Ova činjenica sugerira da bismo mogli usporediti kvantile slučajnog uzorka s kvantilima normalne razdiobe kako bi napravili još jednu usporedbu.

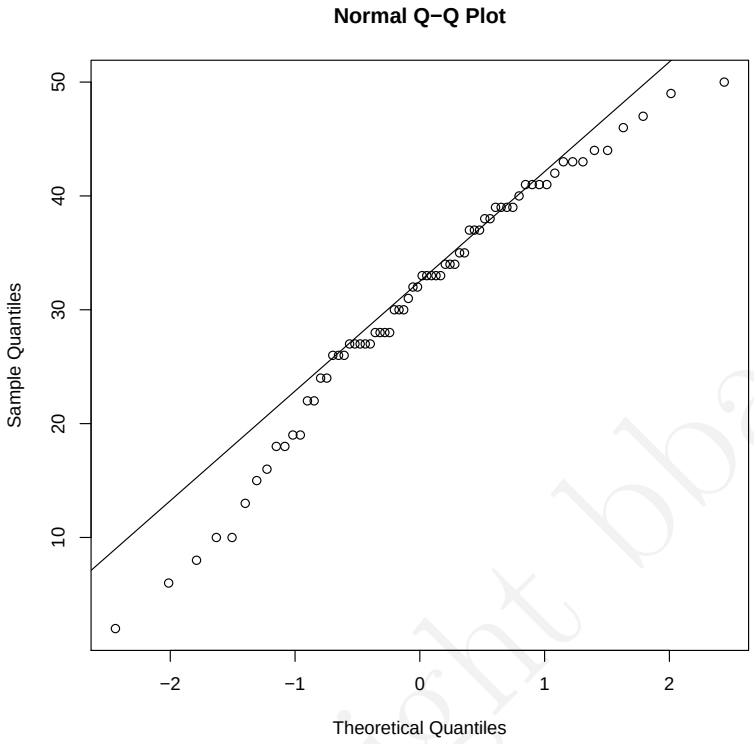
Za brojeve oblika $\alpha_i = i/(n+1)$, $i = 1, \ldots, n$, pripadni α_i-kvantil našeg uzorka upravo $X_{(i)}$. Ako je x_{α_i} α_i-kvantil normalne razdiobe iz koje dolazi uzorak očekivali bismo

$$X_{(i)} \approx x_{\alpha_i}.$$

Koliko su blizu kvantili uzorka i teorijske normalne razdiobe najbolje prikazuje *graf kvantila* ili *qq-plot* u odn. na normalnu razdiobu. Ako podaci dolaze iz standardne normalne razdiobe, točkice na grafu će se grupirati oko pravca $y = x$. A ako dolaze iz neke druge normalne razdiobe ponovo ćemo ih vidjeti kako približno leže na pravcu, istina s nekim drugim koeficijentima.

Primjer 5.5.1 Jednom kolokviju na kojem je maksimalni broj bodova bio 50, pristupilo je 68 studenata. Prilikom obrade rezultata dobili smo sljedeće deskriptivne statistike:

medijan: $m = 32.50$,aritm. sredina: $\bar{x} = 30.75$,donji kvartil: $Q_1 = 26.00$,gornji kvartil: $Q_3 = 39.00$,stand. devijacija: $s = 10.81$.



Slika 5.2: Histogram za rezultate kolokvija.

Pitanja za razmišljanje:

- da li je razdioba rezultata normalno distribuirana?
- da li je $n = 68$ dovoljno velik za normalnu aproksimaciju?
- ima li smisla ovo smatrati uzorkom ili ne (ispitali smo cijelu populaciju zapravo)?

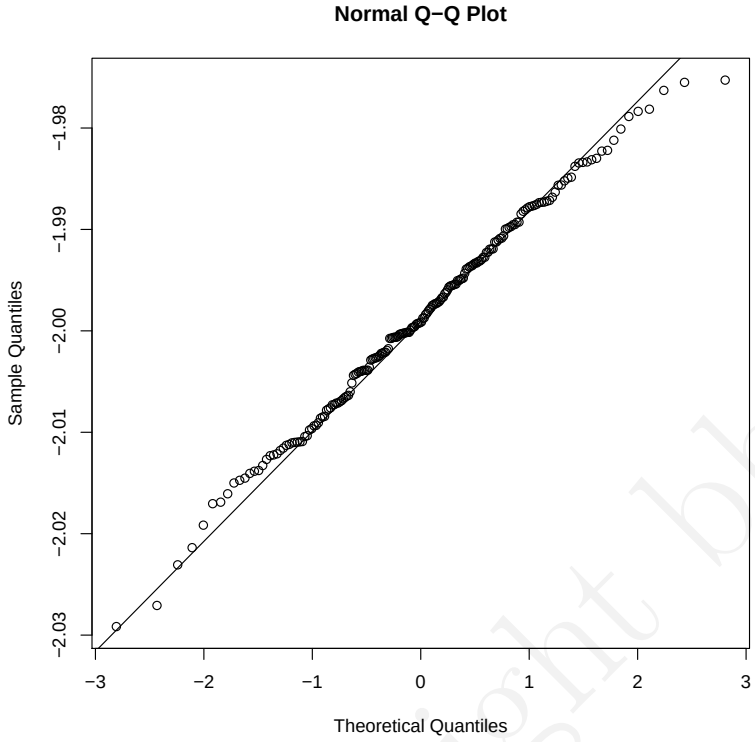
Uz prepostavku da ovo možemo smatrati uzorkom, izračunamo 95%-tni interval pouzdanosti za očekivani broj bodova kao

$$\left(\bar{x} - 1.96s/\sqrt{n}, \bar{x} + 1.96s/\sqrt{n}\right).$$

Naime, kako smo istaknuli na kraju prošlog odjeljka, centralni granični teorem opravdava ovu proceduru čak i u slučajevima kad nemamo normalno distribuirane podatke. U ovom slučaju konkretno dobijem interval:

$$(28.13185,\ 33.36815).$$

⌟



Slika 5.3: Graf kvantila za rezultate kolokvija.

5.6 Procjena parametara u R-u

Ako u uzorku od 1000 mušica, njih 550 nosi određeni genotip, interval pouzdanosti $(1-\alpha)100\%$ za proporciju p takvih mušica u populaciji možemo naći sljedećim naredbama

```
> uzorak <- 550
> alfa <- 0.05
> prop.test(uzorak, 1000, conf.level = 1-alfa, corr=F)
```

Opcija `corr=F` naglašava da ne želimo da se u konstrukciji intervala koristi korekcija zbog neprekidnosti, kako bismo dobili iste intervale kao i u tekstu. Nju bismo u praksi mogli izostaviti.

Interval pouzdanosti $(1-\alpha)100\%$ za parametar očekivanja $N(\mu, \sigma^2)$ razdiobe nalazimo naredbom `t.test`. Tako da za zadane visine studenata npr. kao

```
> visine = scan()
159 188 175 176 177 168 162 188
183 187 187 162 184 161 180 169
195 171 170 199 181 169 189 191
172 182 183 178 180 165 185 202
183 187 188 182 163 179 178 188
```

traženi interval dobijemo npr. naredbama

```
> alfa <- 0.05
> t.test(visine,conf.level=1-alfa)
```

Za vizualnu provjeru pretpostavke o normalnosti možemo koristiti histogram ili bolje graf kvantila, odn. naredbe

```
> hist(visine,prob=T)
> qqnorm(visine)
> qqline(visine)
```

Ukoliko nam je poznata varijanca σ^2, interval pouzdanosti $(1-\alpha)100\%$ za parametar očekivanja $N(\mu, \sigma^2)$ razdiobe, možemo odrediti implementirajući formulu iz teksta direktno u R.

```
> sigma <- 10
> s.greska<- sigma/sqrt(length(visine))
> alfa<-0.05
> zalfapol <- qnorm(1-alfa/2)
```

interval pouzdanosti je sada

```
> mean(visine)+c(-s.greska*zalfapol,s.greska*zalfapol)
```

Zadaci

Zadatak 1. Zadani su podaci 4.2, 4.1, 2.3, 5.1, 2.5 koji dolaze iz normalne razdiobe. Procijenite očekivanje μ i standardnu devijaciju σ.

Zadatak 2. Na temelju 5 podataka koji dolaze iz neke normalne razdiobe određena ja aritmetička sredina 130 i standardna devijacija 40. Odredite 90% -tni interval pouzdanosti za očekivanje ove razdiobe.

Zadatak 3. Razmislite za koliko biste trebali povećati veličinu uzorka da biste interval pouzdanosti za parametar p suzili na pola. Pretpostavite da biste dobili iste procjenitelje $\hat{p}$ u oba slučaja.

Zadatak 4. U industriji je važno znati otpornost metalnih dijelova na stres. Na uzorku od 298 istih metalnih dijelova korištenih u proizvodnji automobila, primjećeno je da ih je 122 pretrpilo ozbiljnu štetu (napuknuće ili koroziju) tijekom 10 godina korištenja. Nađite 95% interval pouzdanosti za frekvenciju metalnih dijelova koji će nakon 10 godina imati takav problem.

Zadatak 5. Neka je nakon prelaska na novog mobilnog operatera prosječna ušteda za 45 ispitanika iznosila $\overline{x} = 2.003$ eura. Koristeći pretpostavku o normalnoj razdiobi za nivo uštede, uz otprije poznatu standardnu devijaciju koja za uštedu u potrošnji iznosi $\sigma = 0.388$, odredite interval pouzdanosti 90% za očekivani nivo uštede.

6 TESTIRANJE STATISTIČKIH HIPOTEZA

6.1 Dvije hipoteze

Ponekad je na osnovu prikupljenih podataka poželjno donijeti binarnu odluku npr.:

i) Novi lijek je efikasniji od do sada korištenog?

ii) Podaci prikupljeni na uzorku potvrđuju do sada prihvaćenu znanstvenu teoriju?

iii) Nakon ispitivanja javnog mišljenja, moramo odlučiti da li će novi proizvod biti uspješan na tržištu?

Ako želimo odluku zasnovati na statističkoj analizi potrebno je naći vjerojatnosni model za podatke. U nekim situacijama je to relativno lako, kao u sljedećem primjeru, gdje je razumno pretpostaviti da podaci slijede binomnu razdiobu.

Primjer 6.1.1 i) Standardni lijek nakon jednomjesečne terapije uzrokuje poboljšanje u 60% pacijenata. U uzorku od 420 pacijenata novi lijek je doveo do poboljšanja u njih 343. Želimo odlučiti da li je novi lijek efikasniji.

ii) Nakon 800 bacanja novčića i opaženih 447 pisama, želimo odlučiti da li je novčić nepristran.

iii) Od 80 predloženih članova porote u nekoj regiji, samo je 4 pripadnika manjine, iako je njihov udio u populaciji regije iz koje se biraju članovi porote približno 50%. Želimo provjeriti je li to u skladu s pretpostavkom da svi građani imaju jednaku vjerojatnosti biti predloženi za člana porote ili ne. ⌟

Ako podatke koje smo prikupili možemo opisati vjerojatnosnim modelom u kojem nam je nepoznat parametar θ, tipično je moguće i hipotezu koju želimo testirati izraziti preko nepoznatog parametra. Uobičajeno je zapisati jednu od njih kao

$$H_0 : \theta \in \Theta_0,$$

a njoj suprotstavljenu hipotezu, slično možemo izraziti kao

$$H_A : \theta \in \Theta_A.$$

Ovdje Θ_0 i Θ_A predstavljaju dva disjunktna skupa u prostoru svih mogućih vrijednosti parametra θ.

U primjeru 6.1.1 prirodan model za podatke bila je binomna odn. Bernoullijeva razdioba. U takvom modelu se hipoteze mogu izraziti u obliku

$$H_0 : p \in B_0 \subseteq [0, 1],$$

a njoj suprotstavljenu hipotezu, slično možemo izraziti kao

$$H_A : p \in B_A = [0,1] \setminus B_0.$$

U postupku testiranja dvije hipoteze **nisu** ravnopravne. Prvu od njih zovemo **nul-hipoteza** i označavamo s H_0. Ona pretpostavlja tipično da se prikupljene numeričke vrijednosti daju objasniti slučajem (npr. novi lijek nije značajno bolji od standardnog, novčić je nepristran, kao i postupak izbora porote). Grubo govoreći, ako mislimo da bi statistički test mogao ukazati na istinitost neke tvrdnje, nul-hipoteza H_0 predstavlja njenu negaciju.

Testom prije svega utvrđujemo da li imamo dovoljno dokaza da bismo odbacili H_0, odn. da li su podaci sukladniji nekoj drugoj tvrdnji o parametrima. Tu drugu hipotezu zovemo **alternativna hipoteza**, u oznaci H_A. Mogli bismo reći da je H_A nama "zanimljiva hipoteza".

Primjer 6.1.2 Nastavak primjera 6.1.1. Nul odn. alternativnu hipotezu u prethodnom primjeru je prirodno postaviti na sljedeći način.

i) $H_0 : p \leq 0.6$ i $H_A : p > 0.6$.

ii) $H_0 : p = 0.5$ i $H_A : p \neq 0.5$.

iii) $H_0 : p \geq 0.5$ i $H_A : p < 0.5$.

⌟

6.2 Dvije vrste pogreške

Očito možemo napraviti dvije vrste pogreške u testiranju.

Pogreška 1. vrste

Odbacujemo H_0, a ona je istinita.

Pogreška 2. vrste

Ne odbacujemo H_0, a ona nije istinita.

Kako je uzorak slučajan, a ishod testiranja ovisi o njemu možemo se pitati koliko su vjerojatne ove pogreške uz unaprijed određenu proceduru. Stoga uvodimo dvije vjerojatnosti:

Vjerojatnost pogreške 1. vrste

$$\alpha = \alpha_\theta = P_\theta\left(H_0 \text{ odbacujemo}\right), \text{ a vrijedi } \theta \in \Theta_0.$$

Vjerojatnost pogreške 2. vrste

$$\beta = \beta_\theta = P_\theta\left(H_0 \text{ ne odbacujemo}\right), \text{ a vrijedi } \theta \in \Theta_A.$$

Primjetite da procedura koja nikad ne odbacuje H_0, ima $\alpha = 0$, slično ako uvijek odbacujemo H_0, možemo postići $\beta = 0$. Cilj je naravno naći proceduru testiranja tako da imamo obje vjerojatnosti male istovremeno. Kod testiranja tipično postavljamo gornju ogradu na vjerojatnost pogreške 1. vrste α, a zatim tražimo test koji ima malu vjerojatnost pogreške 2. vrste. Ovako postavljena ograda na vjerojatnost pogreške 1. vrste se i sama tipično označava ss α i naziva **nivo značajnosti** (signifikantnosti) testa.

Testna statistika

Nakon što formuliramo hipoteze H_0 i H_A, te nivo značajnosti testa, moramo izabrati **testnu statistiku** T. Ona je izvedena iz uzorka, a odabiremo je tako da su ekstremne vrijednosti od T relativno malo vjerojatne pod uvjetom da vrijedi nulhipoteza. Nakon nje određujemo i **kritično područje** C_α. Ono ovisi o α, i određujemo ga (prije nego što smo vidjeli podatke) tako da vrijedi

$$P(T \in C_\alpha, \text{ a } H_0 \text{ je istina}) \leq \alpha.$$

Kako je α tipično mala vrijednosti, kritično područje izabiremo tako da je vjerojatnost da testna statistika upadne u njega mala pod pretpostavkom da vrijedi nulhipoteza.

Nakon prikupljanja podataka, iz uzorka izračunamo vrijednost **testne statistike** $T = t$, te donosimo odluku na sljedeći način

$$t \in C_\alpha, \quad \text{odbacujemo } H_0,$$

ili

$$t \notin C_\alpha, \quad \text{ne odbacujemo } H_0.$$

Vjerojatnost pogreške 1. vrste sada možemo zapisati kao

$$\alpha_\theta = P\left(T \in C_\alpha\right), \qquad \text{za } \theta \in H_0,$$

dok je vjerojatnost pogreške 2. vrste

$$\beta_\theta = P\left(T \notin C_\alpha\right) \qquad \text{za } \theta \in H_A.$$

Uočimo obje vjerojatnosti α_θ i β_θ ovise o θ, no naglasimo da smo kritično područje C_α izabrali tako da je za sve $\theta \in \Theta_0$

$$\alpha_\theta = P\left(T \in C_\alpha, \text{ a } H_0 \text{ istinita}\right) \leq \alpha.$$

6.3 Testovi o parametrima razdiobe

Testiranje hipoteza o parametru p binomne razdiobe

U primjenama ove procedure tipično želimo testirati na osnovi uzorka da li je učestalost jedinki sa zadanim karakteristikama u nekoj populaciji u skladu s našim ili tuđim pretpostavkama.

Primjer 6.3.1 (nastavak primjera 6.1.1 dio iii)

Ako su članovi porote izabrani iz populacije na slučajan način, tj. tako da svi građani u toj populaciji s pravom glasa imaju jednaku vjerojatnost biti članovima porote, prirodno je pretpostaviti da je broj manjinskih članova porote S binomna sl. varijabla s parametrima 80 i p. Ovdje p predstavlja vjerojatnost da je pojedini sl. odabrani glasač pripadnik manjine. Kako nas brine mogućnost da su članovi manjine nedovoljno reprezentirani u poroti, možemo postaviti hipoteze na sljedeći način:

$$H_0 : p = 0.5 \quad \text{i} \quad H_A : p < 0.5.$$

⌟

U gornjem primjeru dakle, želimo detektirati da li se vjerojatnost p razlikuje od pretpostavljene vrijednosti 0.5 i to tako da je strogo manja. Za ovakav tip alternativne hipoteze kažemo da je jednostrana. Pretpostavimo da postavimo nivo značajnosti testa $\alpha = 0.05$.

U sljedećem koraku moramo odabrati testnu statistiku. Mi ćemo uzeti već poznatu vrijednost

$$Z = \frac{S - n/2}{\sqrt{n\frac{1}{2}\frac{1}{2}}},$$

jer znamo da je pod pretpostavkom da vrijedi nul-hipoteza Z približno $N(0,1)$ distribuirana.

Moramo odrediti i C_α. Uočite da ako vrijedi H_A očekivana vrijednost testne statistike bit će manja od 0, zato postavljamo

$$C_\alpha = (-\infty, -z_\alpha].$$

Jasno je da je pod hipotezom $H_0 : p = \frac{1}{2}$

$$P(Z \in C_\alpha) = P(Z \leq -z_\alpha) \approx \Phi(-z_\alpha) = \alpha.$$

U primjeru 6.3.1 možemo postaviti $\alpha = 0.05$. Tada je $z_\alpha = 1.65$, a kako je $n = 80$ i $S = 4$, dobijemo

$$Z = \frac{4 - 40}{\sqrt{20}} = -8.05 < -1.65,$$

pa odbacujemo nulhipotezu.

Općenito, razmatramo 3 mogućnosti za alternativnu hipotezu o parametru p ako je želimo usporediti s nekom fiksnom vrijednošću p_0:

i) $H_0 : p = p_0$ i $H_A : p > p_0$.

ii) $H_0 : p = p_0$ i $H_A : p \neq p_0$.

iii) $H_0 : p = p_0$ i $H_A : p < p_0$.

Prva i treća od alternativnih hipoteza su **jednostrane**, dok je druga **dvostrana**. Za testnu statistiku u sva tri slučaja možemo uzeti

$$Z = \frac{S - np_0}{\sqrt{np_0(1-p_0)}},$$

koja pod H_0 ima približno $N(0,1)$ razdiobu za velike uzorke.

Iako je testna statistika ista, kritična područja za tri testa su redom:

i) $C_\alpha = [z_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-z_{\alpha/2}, z_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -z_\alpha]$.

Ako dakle Z padne u neko od njih, odbacili bismo H_0 u korist odgovarajuće alternative H_A na nivou značajnosti α.

Test i p-vrijednost testa

Kod odlučivanja o ne/odbacivanju nul-hipoteze na osnovu zadane testne statistike za koju znamo konstruirati kritično područje C_α za svaki α, moguće je postupiti i na sljedeći način:

▷ Za dane H_0 i H_A, te testnu statistiku T odredimo iz podataka vrijednost testne statistike $T = t$, a zatim

▷ Nađemo najmanji p tako da je $t \in C_p$. Ovakav p ovisi o uzorku i zove se **p-vrijednost** testa. Uočite, p-vrijednost je najmanji broj p takav da bismo na osnovu $T = t$ uz nivo značajnosti p odbacili nulhipotezu,

▷ Nul-hipotezu odbacujemo na nivou značajnosti α ako je p-vrijednost manja od α.

Intuitivno, možemo reći da p-vrijednost zapravo govori **koliko je vjerojatno toliko ekstremno (ili još ekstremnije) odstupanje testne statistike** pod pretpostavkom da vrijedi H_0.

Primjer 6.3.2 Promotrimo problem iii) iz primjera 6.1.1 još jednom. Naša testna statistika iznosi $Z = -8.04$, a p-vrijednost u ovom slučaju je

$$P(Z < -8.04) \approx \Phi(-8.04) = 4.5 \cdot 10^{-16}.$$

Mi bismo naravno odbacili nulhipotezu o nepristranosti izbora porote uz bilo koji tipično korišteni nivo značajnosti α. No p-vrijednost nam govori i koliko su malo vjerojatni podaci uz pretpostavku da H_0 zaista vrijedi. U ovom slučaju dobivena p-vrijednost je manja od vjerojatnosti da ćete bacajući novčić 50 puta zaredom dobiti sama pisma. Dakle, ekstremno mala. ⌟

Primjer 6.3.3 Mendelovi zakoni u genetici se također mogu izreći korištenjem vjerojatnosnih modela, ali i provjeriti statističkim testom. Prisjetimo se prvi Mendelov zakon o segregaciji tvrdi da aleli, koji određuju neko svojstvo organizma, dolaze u parovim te da se slučajno razdvajaju pri razmnožavanju. Preciznije, potomci naslijeđuju od svakog roditelja po jedan alel, i to slučajno s jednakim vjerojatnostima.

Da bi provjerio ovu tvrdnju Mendel je proučavao populaciju u kojoj su postojala dva alela T i t, te je križao parove heterozigota, odn. organizama čiji su aleli bili Tt, međusobno. Razlikujući pateralni i materalni alel (s obzirom od koje su strane naslijeđeni), Mendel je mogao očekivati 4 vrste jedinki koje su sve bile jednako vjerojatne prema gornjem principu

$$(T,T),(T,t),(t,T),(t,t).$$

U njegovom je slučaju alel T dominantno određivao visinu biljke graška, pa je $3/4$ biljaka dobivenih na ovaj način bilo visoko, a $1/4$ nisko. No on je naslućivao da je među visokim biljkama odnos heterozigota i homozigota 2:1, tako je za te biljke nastavio postupak samooprašivanja još 10 puta da bi odredio koje od njih su heterozigoti, a koje homozigoti. Konačno je na 1073 visoke biljke dobio odnos 720:353 odn. 2.04:1 što je gotovo savršeno odgovaralo njegovom predviđanju. Mi možemo provjeriviti da li je postotak homozigota među visokim biljkama zaista jedna trećina postavljajući $H_0 : p = 1/3$ nasuprot $H_A : p \neq 1/3$. Izračunamo li uobičajenu testnu

$$Z = \frac{353 - 1073/3}{\sqrt{1073\frac{1}{3}\frac{2}{3}}} = -0.302213,$$

dobili bismo p-vrijednost 0.7625. Stoga ne bismo odbacili Mendelov 1. zakon ni na jednom uobičajenom nivou značajnosti.

Ovako dobra podudarnost podataka za teorijskom pretpostavkom je izuzetna, pogotovo stoga što Mendel, ni nakon 10 samooprašivanja, nije mogao biti siguran da je pronašao sve heterozigote. On je zapravo samo mogao usporediti broj pronađenih heterozigota s brojem preostalih biljaka. Taj bi omjer naime trebao biti

$$2(1-(3/4)^{10}) : (1+2(3/4)^{10}) = 1.7 : 1\,.$$

Zbog ovoga je slavni statističar R.A.Fisher posumnjao u potpunu autentičnost Mendelovih podataka. Zanimljivo je da diskusija o tim podacima do kojih he Mendel stigao ⊖ nevjerojatno ustrajnim traje i danas.

⌟

Složena nul-hipoteza

Do sada smo pretpostavljali da je nul-hipoteza uvijek **jednostavna**, tj. oblika $H_0 : p = 0.5$, no ponekad bismo željeli testirati i složenu nulhipotezu, npr. $H_0 : p_0 \geq 0.5$. Općenito, pretpostavimo da želimo testirati:

$H_0 : p \leq p_0$ i $H_A : p > p_0$,

ili

$H_0 : p \geq p_0$ i $H_A : p < p_0$.

U oba slučaja, **test provodimo jednako** kao i da je $H_0 : p = p_0$, što intuitivno argumentiramo činjenicom da ako su podaci ekstremni u odn. na $p = p_0$ onda su još ekstremniji ako je p još manja odn. veća za ove primjere hipoteza.

Testiranje hipoteza o parametru μ normalne razdiobe

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine slučajni uzorak iz normalne razdiobe s parametrima μ i σ^2, te neka su nam oba parametra nepoznata. Pretpostavimo da želimo testirati:

i) $H_0 : \mu = \mu_0$ i $H_A : \mu > \mu_0$.

ii) $H_0 : \mu = \mu_0$ i $H_A : \mu \neq \mu_0$.

iii) $H_0 : \mu = \mu_0$ i $H_A : \mu < \mu_0$.

Za testnu statistiku odabiremo

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{\hat{s}}.$$

Za T je poznato da ima Studentovu t-razdiobu s $n-1$-im stupnjem slobode, ako vrijedi H_0 . Stoga testove provodimo određujući kritično područje na sljedeći način za tri gornja slučaja

i) $C_\alpha = [t_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-t_{\alpha/2}, t_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -t_\alpha]$.

Ovdje je t_α kao i prije $(1-\alpha)$-kvantil t razdiobe s $n-1$ stupnjem slobode.

U slučaju jednostranih alternativnih hipoteza i) odn. iii), postupak testiranja bismo izveli jednako i da je H_0 složena hipoteza oblika $H_0 : \mu \leq \mu_0$ u i) ili $H_0 : \mu \geq \mu_0$ u iii). Naime, ako imamo dovoljno razloga za odbaciti $H_0 : \mu = \mu_0$ u korist $H_A : \mu > \mu_0$, tada je jasno da imamo još jače razloge za odbaciti $H_0 : \mu < \mu_0$ npr.

Primjer 6.3.4 Na uzorku od 37 studenata, nađena je aritmetička sredina trajanja njihovih tipičnih putovanja do fakulteta u iznosu $\bar{X} = 39.03$ min te standardna devijacija $\hat{s} = 24.14$ min. Pretpostavimo da su podaci normalno distribuirani, iako bi to ovdje moglo biti upitno, te da želimo testirati da li je očekivano trajanje putovanja jednako

30 min, kako bi možda netko mogao tvrditi. Postavimo $\alpha = 0.01$ i

$$H_0 : \mu = 30 \text{ nasuprot } H_A : \mu \neq 30,$$

te odredimo

$$T = \sqrt{37}\frac{\bar{X} - 30}{\hat{s}} = -2.27$$

Kritično područje je

$$C_{0.01} = \mathbb{R} \setminus (-t_{0.005}, t_{0.005}) = (-2.71, 2.71)^c.$$

Kako T ne leži u njemu ne možemo odbaciti nul-hipotezu na ovom nivou značajnosti. Uvjerite se ipak da bismo je odbacili da smo postavili $\alpha = 0.05$. Možemo izačunati i p vrijednosti. Ona je u ovom slučaju 0.0289, pa je jasno da za $\alpha = 0.01$ ne možemo odbaciti nul-hipotezu. ⌟

Testiranje hipoteza o očekivanju proizvoljne razdiobe na osnovu velikog uzorka

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine sl. uzorak iz proizvoljne razdiobe s očekivanjem μ i varijancom $0 < \sigma^2 < \infty$, te neka su nam oba ova broja nepoznata. Pretpostavimo da želimo testirati:

i) $H_0 : \mu \leq \mu_0$ i $H_A : \mu > \mu_0$.

ii) $H_0 : \mu = \mu_0$ i $H_A : \mu \neq \mu_0$.

iii) $H_0 : \mu \geq \mu_0$ i $H_A : \mu < \mu_0$.

Za testnu statistiku ponovo odabiremo

$$T = \sqrt{n}\frac{\bar{X} - \mu_0}{\hat{s}}.$$

Ako je n dovoljno velik, iz centralnog graničnog teorema slijedi da T ima približno $N(0, 1)$ razdiobu ako vrijedi $\mu = \mu_0$. Kritična područja su sada:

i) $C_\alpha = [z_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-z_{\alpha/2}, z_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -z_\alpha]$.

Ako dakle Z padne u neko od njih, odbacili bismo H_0 u korist odgovarajuće alternative H_A.

Ovaj test možemo koristiti i za parametar μ normalne razdiobe uz uvjet da nam je poznata standardna devijacija σ i to tako da njenu vrijednost uvrstimo u formulu za T umjesto procjene $\hat{s}$.

Snaga testa

Primjetite da za zadane hipoteze i nivo značajnosti možemo potencijalno pronaći i više testnih statistika T i pripadnih kritičnih područja C_α. Pitanje kako ih odabrati što optimalnije s obzirom na vjerojatnosti pogreške 1. odn. 2. vrste.

Kao što smo istakli, za test koji nikad ne odbacuje H_0, vjerojatnost pogreške 1. vrste je 0. No takav test nije koristan naravno. Za ilustraciju, neka je naš statistički test ugradjen u protupožarni alarm baziran na detektoru dima. Hipoteze su ovdje: H_0 : nema požara, i H_A : u prostoriji je požar. Ako nikad ne odbacujemo H_0 u ovom primjeru, ovakav alarm se nikada neće oglasiti, stoga ga nema smisla ni postavljati. Naš bi cilj mogao biti: uz malu vjerojatnost da će se alarm oglasiti u slučaju da požara nema, dobiti i što veću vjerojatnost da će se oglasiti u slučaju kad nastupi požar. Tu vjerojatnost možemo zapisati kao

$$P\,(\text{test odbacuje } H_0 | H_A \text{ je istinita}) = 1 - \beta.$$

Ova vjerojatnost se u statistici zove **snaga testa**.

U parametarskim vjerojatnosnim modelima kakve smo susretali do sada, H_0 odn. H_A se daju izraziti preko vrijednosti parametra. Ako je npr. u binomnom modelu $H_0 : p \leq p_0$, tada će gornja vjerojatnost ovisiti o pravoj vrijednosti parametra p, stoga definiramo **funkciju snage** testa

$$\gamma(p) = P\,(\text{test odbacuje } H_0 | p \text{ je prava vrijednost parametra}) = 1 - \beta(p).$$

Dakako, idealno bi bilo da za par hipoteza $H_0 : p \leq p_0$ i $H_A : p > p_0$ vrijedi

$$\gamma(p) = 0 \text{ za } p \leq p_0 \text{ i } \gamma(p) = 1 \text{ za sve } p > p_0,$$

takvi idealni testovi u našim primjerima ne postoje. Mi smo stoga koristeći nivo značajnosti α, ograničili $\gamma(p) \leq \alpha$, za $p \leq p_0$, a razne testove s tim svojstvom uspoređujemo gledajući funkciju γ za $p > p_0$.

Primjer 6.3.5 (snaga jednostranog testa za parametar p) Pretpostavite da na osnovu uzorka duljine n, od kojih je S broj jedinki s izvjesnim karakteristikama, želimo testirati sljedeće hipoteze o postotku takvih jedinki u cijeloj populaciji

$$H_0 : p \leq p_0 \quad \text{nasuprot} \quad H_A : p > p_0.$$

Koristimo uobičajenu testnu statistiku

$$Z = \frac{S - np_0}{\sqrt{np_0(1-p_0)}}.$$

Kako smo objasnili, H_0 ćemo odbaciti za $Z \geq z_\alpha$ u ovom slučaju. Dakle funkcija γ se može izračunati kao

$$\begin{aligned}
\gamma(p) &= P\left(Z \geq z_\alpha, \text{ a prava vrijednost parametra je } p\right) \\
&= P\left(\frac{S - np_0}{\sqrt{np_0(1-p_0)}} > z_\alpha\right) \\
&= P\left(\frac{S - np}{\sqrt{np(1-p)}} > z_\alpha\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \sqrt{n}\frac{(p-p_0)}{p(1-p)}\right) \\
&\approx 1 - \Phi\left(z_\alpha\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \sqrt{n}\frac{(p-p_0)}{p(1-p)}\right).
\end{aligned}$$

Dakle, test ima veću snagu ako imamo veći uzorak ili veći nivo značajnosti, ali i ako je razlika $p - p_0$ veća. Što je dakako intuitivno jasno, veću razliku između stvarnog i pretpostavljenog parametra ćemo lakše uočiti.

Neka je zadano $\alpha = 0.05$. U slučaju testa

$$H_0 : p \leq 1/2 \;\text{ nasuprot }\; H_A : p > 1/2,$$

već znamo da je

$$\gamma(0.5) \approx \alpha = 0.05\,.$$

Snagu testa u ovom slučaju možemo vidjeti iz grafa funkcije γ. Što funkcija γ brže raste prema 1 za $p > 0.5$ test bismo smatrali snažnijim. Ovisnost funkcije snage o stvarnom parametru p ovdje ilustrira Slika 6.3.

⌟

Primjer 6.3.6 (snaga testa za parametar μ) Pretpostavite da na osnovu normalno distribuiranog uzorka $X_1, \ldots X_n$, želimo testirati sljedeće hipoteze

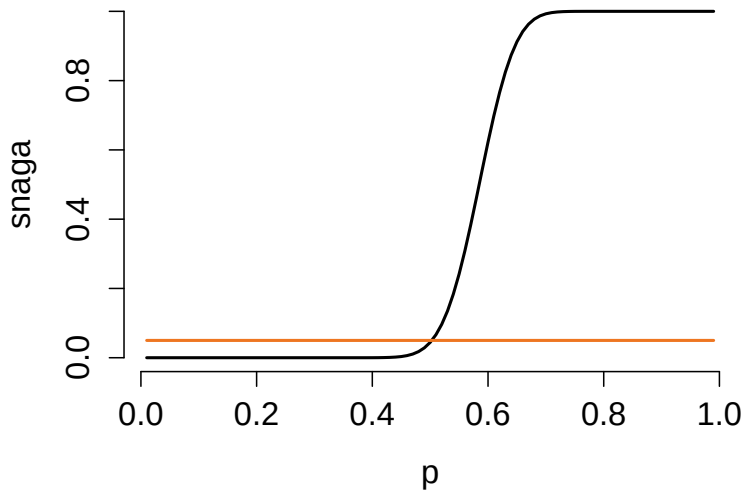
$$H_0 : \mu \geq \mu_0 \;\text{ nasuprot }\; H_A : \mu < \mu_0.$$

Neka nam je poznata varijanca uzorka σ^2 i neka je zadan nivo značajnosti α. Ako koristimo uobičajenu testnu statistiku

$$Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma},$$

kako smo objasnili, H_0 ćemo odbaciti za $Z \leq -z_\alpha$ u ovom slučaju. Dakle funkcija snage se može izračunati kao

$$\begin{aligned}
\gamma(\mu) &= P\left(Z \leq -z_\alpha, \text{ a prava vrijednost parametra je } \mu\right) \\
&= P\left(\sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \leq -z_\alpha\right) \\
&= P\left(\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq -z_\alpha - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right) \\
&= \Phi\left(-z_\alpha + \sqrt{n}\frac{\mu_0 - \mu}{\sigma}\right).
\end{aligned}$$



Slika 6.1: Funkcija snage $\gamma(p)$ za test $H_0 \leq \frac{1}{2}$.

Ponovo se možemo uvjeriti da test ima veću snagu ako imamo veći uzorak ili veći nivo značajnosti, ali i ako je razlika $\mu_0 - \mu$ veća. S druge strane velika varijanca σ^2 smanjuje snagu testa.

Formula za snagu može biti od velike koristi kod dizajna samog pokusa. Pretpostavite da smatramo kako je u populaciji koju studiramo očekivanje neke veličine manje od 38, te da ga želimo usporediti s nulhipotezom koja tvrdi da je ono veće ili jednako od 40. Neka je $\alpha = 0.05$, a standardna devijacija neka je poznata i iznosi 4. Pretpostavite nadalje da nakon pokusa želimo imati vjerojatnost 99% odbacivanja u ovom slučaju pogrešne nulhipoteze. Mi možemo odrediti kolika nam je veličina uzorka potrebna da bismo to postigli. Naime trebalo bi vrijediti

$$\Phi\left(-1.65 + \sqrt{n}\frac{40-38}{4}\right) > 0.99.$$

Kako je $\Phi(2.33) \approx 0.99$, trebamo postići

$$-1.65 + \sqrt{n}\frac{40-38}{4} > 2.33.$$

Odavde slijedi da nam treba $n \geq 64$.

⌟

Testiranje hipoteza o očekivanjima dvije normalne razdiobe

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine sl. uzorak iz normalne razdiobe s parametrima μ_1 i σ^2, a $Y_1, \ldots, Y_m$, $m \geq 1$, neka čine sl. uzorak iz normalne razdiobe s parametrima μ_2 i σ^2

(nezavisan od X_i), te neka su nam sva 3 parametra nepoznata. Dakle, test provodimo uz pretpostavku da su **varijance jednake** u oba uzorka. Pretpostavimo da želimo testirati:

i) $H_0 : \mu_1 \leq \mu_2$ i $H_A : \mu_1 > \mu_2$.

ii) $H_0 : \mu_1 = \mu_2$ i $H_A : \mu_1 \neq \mu_2$.

iii) $H_0 : \mu_1 \geq \mu_2$ i $H_A : \mu_1 < \mu_2$.

Iz svakog uzorka posebno izračunamo njihove aritmetičke sredine $\bar{X}$ i $\bar{Y}$, te uzoračke varijance $\hat{s}_X^2$ odn. $\hat{s}_Y^2$. Za očekivati je da ćemo razliku izmedju μ_1 i μ_2 moći uočiti izmedju razlike $\bar{X}$ i $\bar{Y}$, no pitanje je kako?

Za testnu statistiku odabiremo

$$T = \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \frac{\bar{X} - \bar{Y}}{\hat{s}},$$

gdje je

$$\hat{s}^2 = \frac{1}{n+m-2} \left((n-1)\hat{s}_X^2 + (m-1)\hat{s}_Y^2 \right).$$

Može se pokazati da, uz uvjet $\mu_1 = \mu_2$, statistika T ima Studentovu t razdiobu s $n + m - 2$ stupnja slobode. Stoga, kritično područje za svaki od tri testa odredjujemo na sljedeći način

i) $C_\alpha = [t_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-t_{\alpha/2}, t_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -t_\alpha]$.

Gdje je t_α $(1-\alpha)$-kvantil t razdiobe s $n + m - 2$ stupnja slobode.

Na ovaj način možemo usporediti npr. rezultate kolokvija za one studente koji su posjetili predavanja i one koji to nisu činili. Pretpostavka da su podaci normalno distribuirani uz jednake varijance je katkad upitna u praksi. Ako ona nije zadovoljena testnu statistiku za 3 hipoteze o odnosu očekivanja dvije razdiobe možemo malo promijeniti, i ponovo kontruirati test o odnosu μ_1 i μ_2, uz uvjet da imamo veliki uzorak (v. Bhattacharyya i Johnson [**?**] npr.).

Usporedba varijanci dva normalno distribuirana uzorka

Neka $X_1, \ldots, X_n$, $n \geq 1$, čine sl. uzorak iz normalne razdiobe s parametrima μ_1 i σ_1^2, a $Y_1, \ldots, Y_m$, $m \geq 1$, neka čine sl. uzorak iz normalne razdiobe s parametrima μ_2 i σ_2^2, te neka su nam sva 4 parametra nepoznata. Pretpostavimo da želimo testirati:

$H_0 : \sigma_1 = \sigma_2$ nasuprot $H_A : \sigma_1 \neq \sigma_2$.

ili što je potpuno ekvivalentno

$H_0 : \frac{\sigma_1}{\sigma_2} = 1$ nasuprot $H_A : \frac{\sigma_1}{\sigma_2} \neq 1$.

Izračunajmo uzoračke varijance $\hat{s}_X^2$ odn. $\hat{s}_Y^2$, a za testnu statistiku postavimo

$$F = \frac{\hat{s}_X^2}{\hat{s}_Y^2}.$$

Za ovu testnu statistiku je poznato da pod nul hipotezom H_0 ima takozvanu Fisherovu F razdiobu s $(n-1, m-1)$ stupnjeva slobode. Sada kritično podučje za nivo značajnosti α možemo konstruirati koristeći tablice za F razdiobu.

Usporedba očekivanja dva sparena normalno distribuirana uzorka

Primjer 6.3.7 Pretpostavimo da želimo usporediti razliku izmedju krvnog tlaka za osobe koje uzimaju i one koje ne uzimaju odredjeni lijek. Mogli bismo postupiti kao i do sada kad smo željeli usporediti da li je očekivana vrijednost neke varijable u dvije populacije jednaka, npr. mogli bismo koristiti t-test.

Pretpostavite, medjutim, da je varijanca mjerenja σ^2 vrlo velika, tada t-test može detektirati samo vrlo značajne razlike izmedju μ_1 i μ_2. Ovakav test ima malu snagu ako se μ_1 i μ_2 razlikuju za relativno malu vrijednost. Imate li bolju ideju?

Bolja ideja je da pokušamo eliminirati dio varijabilnosti izmedju podataka tako da isto mjerenje napravimo na jednoj osobi po dva puta, tj. jednom u periodu uzimanja lijeka i jednom u periodu kada ne uzima lijek. Zbog istog razloga, tj. smanjenja varijabilnosti, u medicini (ali i biologiji) istraživanja se često oslanjaju na podatke dobivene na parovima jedinki iz iste obitelji, ili još bolje na parovima jednojajčanih blizanaca.

Nakon što odlučimo kako i što spariti u ovakvom pokusu, poželjno je provesti i **randomizaciju**, tj. svaku jedinku iz para treba dodijeliti 1. ili 2. tretmanu u ovisnosti o ishodu nezavisnih bacanja novčića. Npr. ako bacimo pismo, prvo mjerimo tlak osobe uz uzimanje lijek, a zatim bez uzimanja lijeka, a ako padne glava napravimo obrnuto. Na ovaj način smanjujemo prostor za druge nekontrolirane izvore varijabilnosti.

Dakle ovdje na svakoj jedinki napravimo dva mjerenja. Tako da slučajni uzorak možemo označiti kao niz parova $(X_1, Y_1), \ldots, (X_n, Y_n)$. Ako želimo testirati razliku izmedju očekivanja pod jednim odn. drugim tremanom, za sve jedinke možemo promatrati razlike $D_i = X_i - Y_i$, $i = 1, \ldots, n$. Uočite da hipoteze o μ_1 i μ_2 možemo prevesti u hipoteze o očekivanju razlike, npr.

$$\mu_1 > \mu_2 \text{ je ekvivalentno } ED_i > 0.$$

Stoga $D_1, \ldots, D_n$ tretiramo kao uzorak, pri tom je često je razumno pretpostaviti da su D_i normalno distribuirane s očekivanjem $d = \mu_1 - \mu_2$ i varijancom σ_d^2.

Pod gornjim uvjetima za hipoteze

i) $H_0 : \mu_1 \leq \mu_2$ i $H_A : \mu_1 > \mu_2$,

ii) $H_0 : \mu_1 = \mu_2$ i $H_A : \mu_1 \neq \mu_2$,

iii) $H_0 : \mu_1 \geq \mu_2$ i $H_A : \mu_1 < \mu_2$,

definiramo testnu statistiku

$$T = \sqrt{n}\frac{\bar{D}}{\hat{s}_d},$$

gdje je $\hat{s}_d^2$ uzoračka varijanca, a $\bar{D}$ aritmetička sredina uzorka $D_1, \ldots, D_n$. Uočite da se tri alternativne hipoteze mogu redom izreći i kao: $H_A : d > 0,\ d \neq 0$ odn. $d < 0$

Za ovakvu statistiku T znamo da ima Studentovu t-razdiobu sa $n-1$-im stupnjem slobode ako vrijedi $ED_i = 0$, tj. $\mu_1 = \mu_2$. Stoga testove provodimo odredjujući kritično područje na sljedeći način

i) $C_\alpha = [t_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-t_{\alpha/2}, t_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -t_\alpha]$.

Usporedba parametra p za dvije binomne razdiobe

Na dva uzorka iz dvije različite populacije provodimo istraživanje kako bismo utvrdili postoji li razlika u postotku jedinki u jednoj odn. drugoj populaciji sa zadanim obilježjem.

Primjer 6.3.8 Na uzorku od 40 studentica matematike 2. godine utvrđeno je da je 25 njih zadovoljilo sve svoje obaveze iz prve godine, na uzorku od 40 studenata taj broj je nešto manji i iznosi 21. Možemo li tvrditi da postoji razlika izmedju uspješnosti muških odn. ženskih studenata matematike? ⌟

Podatke ovdje općenito reprezentiramo s dvije nezavisne binomne slučajne varijable $S_1 \sim B(n_1, p_1)$ i $S_2 \sim B(n_2, p_2)$, koje označavaju broj jedinki sa traženim krakteristikama u 1. odn. 2. uzorku. O odnosu p_1 i p_2 možemo postavite sljedeća tri para hipoteza

i) $H_0 : p_1 \leq p_2$ i $H_A : p_1 > p_2$.

ii) $H_0 : p_1 = p_2$ i $H_A : p_1 \neq p_2$.

iii) $H_0 : p_1 \geq p_2$ i $H_A : p_1 < p_2$.

Za testnu statistiku izabiremo

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

gdje su naravno

$$\hat{p}_1 = \frac{S_1}{n_1} \text{ i } \hat{p}_2 = \frac{S_2}{n_2},$$

dok je

$$\hat{p} = \frac{S_1 + S_2}{n_1 + n_2}.$$

Centralni granični teorem nam omogućuje da za velike uzorke koristimo sljedeća tri već dobro poznata kritična područja:

i) $C_\alpha = [z_\alpha, \infty)$.

ii) $C_\alpha = \mathbb{R} \setminus (-z_{\alpha/2}, z_{\alpha/2})$.

iii) $C_\alpha = (-\infty, -z_\alpha]$.

6.4 Testovi prilagodbe

Bez obzira da li podaci dolaze iz diskretne ili neprekidne razdiobe, zanimljivo je znati da li im dani teorijski model odgovara. Mi smo već vidjeli kako na grafu kvantila uzorak možemo usporediti podatke i teorijsku razdiobu. Ipak u nekim situacijama je poželjno obaviti i formalni test.

χ^2-test

Pretpostavimo da slučajni uzorak $X_1, \ldots, X_n$ grupiramo u k razreda. Pretpostavimo da postoji i teorijski model za podatke, tako da možemo odrediti vjerojatnosti oblika $P(X_i \in A)$ za svaki skup A. Koliko dobro model odgovara podacima možemo naslutiti uspoređujući za $i = 1, \ldots, k$ sljedeće brojeve

$$O_i = \text{ broj opažanja u } i\text{-tom razredu}$$

i

$$E_i = \text{ očekivani broj opažanja u } i\text{-tom razredu prema modelu}\,.$$

Razrede određujemo tako da je očekivani broj podataka po svim razredima ukupno n. Ukoliko želimo testirati H_0 : podaci dolaze iz danog teorijskog modela, nasuprot H_A : podaci imaju neku drugu razdiobu, često korištena statistika je

$$Hi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}\,. \tag{6.4.1}$$

Egzaktnu razdiobu ove testne statistike nije lako naći, no za velike uzorke, ona ima približno χ^2 razdiobu s $k-p-1$ stupnjem slobode, gdje p u ovom slučaju označava broj parametara koje smo procijenili da bismo teorijsku razdiobu prilagodili danim podacima. Konačno na nivou značajnosti α nulhipotezu odbacujemo ako testna statistika pripada skupu

$$C_\alpha = [\chi^2_\alpha, \infty)\,.$$

Pri tom χ^2_α označava $(1-\alpha)$-kvantil χ^2 razdiobe s $k-p-1$ stupnjeva slobode.

U praktičnim situacijama postavlja se pitanje – koliki n je dovoljno velik za primjenu ovog testa? Uobičajeni je odgovor da je nužno da su svi $E_i \geq 5$ ili barem 3. O tome je važno voditi računa kod grupiranja podataka u razrede.

Primjer 6.4.1 Pretpostavimo da želimo odrediti vjerojatnosni model za broj rekombinacija tijekom mejoze na jednom kromosomu neke biljne vrste. Broj rekombinacija Y namjeravamo modelirati Poissonovom razdiobom, tako da vrijedi

$$P(Y=j) = \frac{\lambda^j}{j!}e^{-\lambda}\,,$$

za $j = 0, 1, 2, \ldots$ Pretpostavljamo da imamo podatke nakon $n = 60$ mejoza. Da bismo procijenili parametar λ, prisjetimo se da je očekivanje Poissonove razdiobe upravo jednako λ, stoga je prirodan procjenitelj upravo aritmetička sredina uzorka. Neka su dobiveni sljedeći podaci

Broj rekombinacija	O_i	E_i
0	32	60 P(Y=0)=28.34
1	15	60 P(Y=1)= 21.26
2	9	60 P(Y=2)= 7.97
3	4	60 P(Y=3)= 1.99
Ukupno	60	59.56

Da bismo izračunali E_i u trećem stupcu iskoristili smo procjenitelj za λ

$$\hat{\lambda} = \overline{X}_n = \frac{32\cdot 0 + 15\cdot 1 + 9\cdot 2 + 4\cdot 3}{60} = \frac{3}{4}.$$

Zbroj brojeva u trećem stupcu ja manji od 60, naime prema Poissonovoj razdiobi postoji i striktno pozitivna vjerojatnost da ćemo opaziti 4 ili čak i više grešaka. Također posljednji broj u tom stupcu je manji od 5, pa je u ovakoj situaciji razumno združiti razrede sa 2,3 ili više grešaka u jedan razred, odn. kreirati novu tablicu

Ako želimo testirati da li podaci slijede Poissonovu razdiobu, možemo postaviti nivo značajnosti na $\alpha = 0.05$. Nakon toga se možemo uvjeriti da je u našem slučaju je $Hi^2 = 2.94$, što moramo usporediti sa $\chi^2_\alpha(1) = 3.84$. Naime, kako imamo samo 3 razreda u drugoj tablici, broj stupnjeva slobode je 3-1-1=1. Dakle, možemo zaključiti da na

Broj rekombinacija	O_i	E_i
0	32	$60P(Y=0)=28.34$
1	15	$60P(Y=1)=21.26$
2 ili više	13	$60P(Y\geq 2)=10.40$
Ukupno	60	60

nivou značajnosti 0.05, ne možemo odbaciti nulhipotezu o Poissonovoj distribuiranosti ovih podataka.

⌟

Primjer 6.4.2 Mendelov drugi zakon govori o nezavisnoj segregaciji različitih alela. Zakon tvrdi da se dva alela za dva odvojena svojstva nasljeđuju nezavisno. Mendel je konkretno proučavao dva svojstva graška za koje je pronašao da postoji dominantni alel: žutu boju i zaobljenost zrna. Označimo li alele za prvo odn. drugo svojstvo sa Yy odn Rr, ako samooprašujemo biljke koje su heterozigoti za oba svojstva, tj. nose alele $rRyY$ lako je izračunati vjerojatnosti različitih fenotipova. Uočimo prvo da je po prvom Mendelovom zakonu vjerojatnost da ćemo samooprašivanjem dobiti biljku zaobljenog zrna 3/4, što je vjerojatnost da će ta biljka naslijediti bar jedan alel R, a slično vrijedi i za žutu boju v. primjer 6.3.3. Dakle, ako označimo sa X_R, odn. X_Y 0–1 varijable koje indiciraju prisutnost bar jednog alela R odn. Y koji nasljeđuje slučajno odabrana biljka potomak, razdioba ovih slučajnih varijabli je

$$X_R, X_Y \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

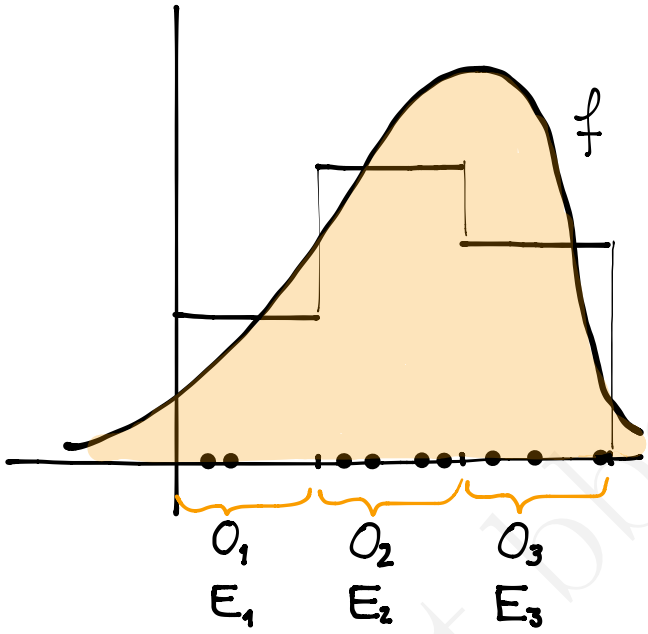
Drugi Mendelov zakon zapravo tvrdi da su slučajne varijable X_R i X_Y nezavisne. Stoga, ako vrijede oba zakona zajednička razdioba ovih slučajnih varijabli je

	0 (nezaobljeno)	1 (zaobljeno)
0 (zeleno)	$P(X_R=0,X_Y=0)=1/16$	$P(X_R=1,X_Y=0)=3/16$
1 (žuto)	$P(X_R=0,X_Y=1)=3/16$	$P(X_R=1,X_Y=1)=9/16$

Iako se radi o dvodimenzionalnoj slučajnoj varijabli, (X_R, X_Y) može kao par primiti jednu od 4 vrijednosti pa možemo koristeći χ^2–test testirati da li su neki podaci u skladu s ovako pretpostavljenom razdiobom.

Mendel je zapisao da je od 556 ovakvih potomaka dobio sljedeće frekvencije potomaka

	0 (nezaobljeno)	1 (zaobljeno)
0 (zeleno)	32	108
1 (žuto)	101	315



Slika 6.2: Opažene i očekivane frekvencije razreda

Ako bismo brojeve u prethodnoj tablici pomnožili sa $n = 556$ dobili bismo sljedeće očekivane frekvencije: 34.75, 104.25, 104.25 i 317.75. Ako ih označimo sa E_i, a brojeve u prethodnoj tablici sa O_i, dobit ćemo 0.47 za vrijednost χ^2-statistike. Kako imamo 4 grupe podataka koristili smo 3 stupnja slobode, što za ovu statistiku daje veliku p-vrijednost od 0.9254. Posebno, možemo zaključiti da su podaci u skladu s Mendelovim zakonima.

⌟

Za neprekidne razdiobe uobičajeno je razrede formirati u obliku intervala od kojih neki mogu biti i neograničeni. Ako je X neprekidna slučajna varijabla s gustoćom f i funkcijom distribucije F prisjetimo se da vrijedi

$$P(X \in (a, b]) = \int_a^b f(s)\mathrm{d}s = F(b) - F(a).$$

Stoga je očekivani broj podataka u uzorku duljine n

$$E_i = n[F(b_i) - F(a_i)]$$

ako je i-ti razred oblika $(a_i, b_i]$. Usporedbu ovih veličina ilustrira slika 6.4

Primjer 6.4.3 Prisjetimo se podataka o visinama studenata sa slike 2.4 (v. str. 26). Njihova aritmetičku sredina je $\bar{X} = 179.15$, a standardna devijacija im je $s = 10.594$. Usporedimo njihovu razdiobu s normalnom. Za očekivanje odn. standardnu devijaciju te normalne razdiobe uzimamo upravo ove dvije procjenjene vrijednosti. Podatke ćemo grupirati u intervale: $(-\infty, 170.15), [170.15, 179.15), [179.15, 188.15)$ i $[188.15, \infty)$.

interval	O_i	E_i
$(-\infty, 170.15)$	11	7.912
$[170.15, 179.15)$	9	12.088
$[179.15, 188.15)$	15	12.088
$[188.15, \infty)$	5	7.912
Ukupno	40	40

Iskoristimo li sada formulu (6.4.1) da bismo izračunali testnu statistiku, dobijemo $Hi^2 = 5.001$ te p-vrijednost 0.02533044. Pri tome koristimo χ^2 razdiobu s 1 stupnjem slobode, jer smo imali 4 razreda, ali smo procijenili dva parametra. Možemo zaključiti da bismo nulhipotezu o normalnosti odbacili na nivou značajnosti 0.05, ali ne i na nivou značajnosti 0.01 npr. Također, lako je zamisliti da bi se vrijednost testne statistike, a onda i naš zaključak mogao promijeniti drugačijim izborom razreda.

⌟

Kako prethodni primjer ilustrira kod usporedbe uzorka s nekom teorijskom neprekidnom razdiobom zaključak χ^2 testa može ovisiti o izboru razreda. To je naravno vrlo nezgodno, zbog toga u tim situacijama postoje i koriste se drugi testovi prilagodbe (engl. goodness of fit). Najopćenitiji od njih je Kolmogorov–Smirnovljev test. U slučaju da nas zanima samo test normalnosti važno je znati da postoje posebni testovi, npr. Lillieforsov, Anderson-–Darlingov, itd. ⊖

6.5 Testovi nezavisnosti

Procedura koju ćemo opisati može se smatrati posebnim slučajem testa prilagodbe, a zasniva se na podacima koji su vrlo jednostavni. Svaku od n jedinki u uzorku klasificiramo u odn. na dva (kvalitativna i binarna) obilježja. Dakle, za svaku jedinku imamo dvije sl. varijable s vrijednostima 0 i 1, koje indiciraju ima li jedinka obilježje x i/ili obilježje y. Mogli bismo dakle reprezentirati podatke kao niz parova Bernoullijevih sl. varijabli (X_i, Y_i), $i = 1, \ldots, n$. Ono što želimo provjeriti je da li je prisutnost ovih obilježja kod pojedine sl. odabrane jedinke nezavisno. U jeziku vjerojatnosti to bi značilo da su nezavisne sl. varijable X_i i Y_i. Mogućih testova je i ovdje više.

Primjetite da bismo sakupljene podatke moglu prikazati u frekvencijskoj tablici oblika 2×2:

	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

gdje n_{kl} označava broj jedinki u uzorku (ili parova (X_i, Y_i)) za koje je $X_i = k$, a $Y_i = l$. Ovakve tablice se zovu i kontingencijske tablice (engl. contingency tables), a mogu biti

i reda $m \times n$, $n, m \geq 2$ općenito. Mi ćemo zbog jednostavnosti promatrati samo slučaj $m = n = 2$.

Primjer 6.5.1 Godine 1889. u istraživačkoj stanici u Rothamstedu (UK), za vrijeme popodnevne pauze za čaj, algolog dr. B.M. Bristol odbila je šalicu čaja uz napomenu kako pije čaj isključivo ako se prvo ulije miljeko u šalicu, a tek zatim čaj. Nazočan je bio i statističar R.A. Fisher, koji je ustvrdio da je to besmisleno, jer nitko ne može osjetiti razliku izmedju šalica čaja u kojima je mlijeko uliveno prije ili nakon čaja. Tada je dr. Bristol uzdahnula
"Oh, ali razlika stvarno postoji!"
Nakon čega je netko uzviknuo (kasnije će se ispostaviti, bio je to budući suprug gospođice Bristol)
"Testirajmo je!"
A Fisher je ubrzo izašao i s idejom o tome kako provesti ovaj test. ⌟

Postavimo u gornjem primjeru $X_i = 1$ ako je u i-toj šalici uliveno prvo mlijeko, a 0 inače. Slično, nakon što dr. Bristol proba i-ti čaj ne znajući o kakvoj se šalici radi, postavimo $Y_i = 1$ ako je ustvrdila da je u i-tu šalicu uliveno prvo mlijeko, a 0 inače.

Ako su slučajne varijable X_i i Y_i zaista nezavisne, trebalo bi vrijediti

$$P(X_i = k, Y_i = l) = P(X_i = k)P(Y_i = l) \tag{6.5.1}$$

za sve $k, l = 0, 1$.

Iako ne znamo pravu razdiobu sl. vektora (X_i, Y_i) prirodno je možemo aproksimirati na sljedeći način

	0	1
0	$\hat{p}_{00} = n_{00}/n$	$\hat{p}_{01} = n_{01}/n$
1	$\hat{p}_{10} = n_{10}/n$	$\hat{p}_{11} = n_{11}/n$

gdje je $\hat{p}_{kl}$ dakako procjena za $p_{kl} = P(X_i = k, Y_i = l)$.

Procjena za vjerojatnosti $P(X_i = k)$ i $P(Y_i = l)$ se sada može lako dobiti iz tablice. Tako $p_k = P(X_i = k)$ procjenjujemo sa

$$\hat{p}_k = \hat{p}_{k0} + \hat{p}_{k1},$$

a $q_l = P(Y_i = l)$ procjenjujemo sa

$$\hat{q}_l = \hat{p}_{0l} + \hat{p}_{1l}.$$

Uz pretpostavku nezavisnosti, na osnovu (6.5.1) očekivali bismo

$$\hat{p}_{kl} \approx \hat{p}_k \hat{q}_l \,.$$

Dakle uz pretpostavku nezavisnosti očekivali bismo sljedeću tablicu frekvencija

	0	1
0	$n\hat{p}_0\hat{q}_0$	$n\hat{p}_0\hat{q}_1$
1	$n\hat{p}_1\hat{q}_0$	$n\hat{p}_1\hat{q}_1$

Stoga za suprotstavljene hipoteze

$H_0:$ X_i i Y_i su nezavisne i $H_A:$ X_i i Y_i nisu nezavisne,

testnu statistiku možemo izabrati kao

$$C^2 = \sum_{k,l=0,1} \frac{(n_{kl} - n\hat{p}_k\hat{q}_l)^2}{n\hat{p}_k\hat{q}_l} = n \sum_{k,l=0,1} \frac{(\hat{p}_{kl} - \hat{p}_k\hat{q}_l)^2}{\hat{p}_k\hat{q}_l}.$$

Iz centralnog graničnog teorema slijedi da ova statistika ima približno χ^2 razdiobu s 1 stupnjem slobode. Zbog toga je kritično područje ovog testa

$$C_\alpha = [\chi^2_\alpha, \infty),$$

gdje je χ^2_α $(1-\alpha)$-kvantil χ^2 razdiobe s jednim stupnjem slobode.

Ako želimo na sličan način testirati zavisnost kategorijalnih varijabli s više od dva moguća ishoda, to možemo učiniti na potpuno analogan način: ako npr. varijabla X ima r mogućih ishoda $\{1, 2, \ldots, r\}$ a varijabla Y ima c mogućih ishoda $\{1, 2, \ldots, c\}$, dobili bismo tablicu dimenzija $r \times c$. Očekivane frekvencije bismo našli na potpuno jednak način kao i slučaju 2×2 i testna statistika bi imala sličan oblik

$$C^2 = \sum_{k=1}^{r} \sum_{l=1}^{c} \frac{(n_{kl} - n\hat{p}_k\hat{q}_l)^2}{n\hat{p}_k\hat{q}_l},$$

no kritične vrijednosti bismo tražili koristeći $(1-\alpha)$-kvantil χ^2 razdiobe s $(r-1)\cdot(c-1)$ stupnjeva slobode. Napomenimo na kraju da je Fisher pronašao i tzv. egzaktnu verziju testa nezavisnost, odn. način određivanja kritičnog područja koji se ne oslanja na aproksimativnu χ^2 razdiobu. Taj test se može naći između ostalog implementiran i u R-u.

6.6 Testiranje hipoteza u R-u

Sažetak o testiranju

Procedura testiranja statističkih hipoteza može se podijeliti na sljedeće korake:

- Formuliramo H_0 i H_A, te odredimo nivo značajnosti α.
- Odredimo testnu statistiku T i kritično područje C_α tako da

$$P(T \in C_\alpha | H_0) \leq \alpha.$$

- Izračunamo vrijednost testne statistike za naše podatke, npr. $T = t$.
- Odbacujemo H_0 ako $t \in C_\alpha$, inače je ne odbacujemo na nivou značajnosti α.

Sjetimo se da posljednji korak ima i alternativni oblik ako možemo izračunati p-vrijednost dobivene testne statistike $T = t$, tada bismo odbacili H_0 ako je $p < \alpha$.

Testovi o parametrima p i μ

Neka u uzorku od 1000 mušica, njih 550 nosi određeni genotip, pretpostavite da na nivou značajnost α želimo testirati

$$H_0 : p = 0.4 \text{ i } H_A : p \neq 0.4,$$

gdje je p proporcija takvih mušica u populaciji. Test možemo provesti npr. već poznatim naredbama

```
> uzorak <- 550
> alfa <- 0.05
> prop.test(uzorak, 1000, p=0.4, conf.level = 1-alfa)
```

Uočite da smo dobili χ^2–testnu statistiku umjesto statistike Z koju smo mi koristili, no da bismo odlučili želimo li odbaciti H_0 ili ne, dovoljno je promotriti samo p–vrijednost odn. `p-value` koju nam daje R.

Ukoliko bismo imali jednostrane alternative, npr. $H_A : p < 0.4$ ili $H_A : p > 0.4$, test bismo proveli naredbama

```
> prop.test(uzorak, 1000, p=0.4, conf.level = 1-alfa,alternative="less")
> prop.test(uzorak, 1000, p=0.4, conf.level = 1-alfa,alternative="greater")
```

Slično za podatke o visinama iz prehodnog poglavlja, ako pretpostavimo normalnu $N(\mu, \sigma^2)$ razdiobu možemo testirati npr. $H_0 : \mu = 170$ nasuprot $H_A : \mu \neq 170$, naredbom

```
> t.test(visine,conf.level=1-alfa,mu=170,alternative="two.sided")
```

ili npr. $H_0 : \mu \leq 170$ nasuprot $H_A : \mu > 170$, naredbom

```
> t.test(visine,conf.level=1-alfa,mu=170,alternative="greater")
```

Testovi za usporedbu dva normalna uzorka

Ako podijelimo npr. uzorak visina na dva dijela naredbama

```
> visine1 <- visine[1:20]
> visine2 <- visine[21:40]
```

Uz pretpostavku da oba uzorka dolaze is normalne razdiobe s istom varijancom i očekivanjima μ_1 odn. μ_2 možemo testirati npr. da li $H_0 : \mu_1 \leq \mu_2$ nasuprot $H_A : \mu_1 > \mu_2$, naredbom

```
> t.test(visine1,visine2,conf.level=1-alfa,var.equal=T,alternative="greater")
```

Da bismo provjerili jednakost varijanci na raspolaganju nam je naredba

```
> var.test(visine1,visine2)
```

Ukoliko imamo sparene uzorke, $H_0 : \mu_1 = \mu_2$ nasuprot $H_A : \mu_1 \neq \mu_2$, možemo testirati jednostavno naredbom

```
> t.test(uzorak1-uzorak2,conf.level=1-alfa)
```

Testovi prilagodbe i kontingencijske tablice

Ako želimo testirati da li Mendelovi podaci odgovaraju razdiobi koju možemo izvesti iz njegovih zakona nasljeđivanja v. primjer 6.4.2

```
> x <- c(315,101,108,32)  ### podaci o boji i obliku graška
> p <- c(9,3,3,1)         ### teoretski odnos frekvencija prema Mendelovim principima
```

test provodi naredba

```
> chisq.test(x, p = p, rescale.p = TRUE)
```

Ukoliko imamo kontingencijsku tablicu u R je možemo učitati na više načina, npr. korištenjem naredbe `table` ili direktno ako smo prebrojali sami sve ishode kao u idućem primjeru

```
> razredi <- c(4,5,6)
> ocjene <- c(49,50,69)
> popularnost <- c(24,36,38)
> sport <- c(19,22,28)
```

Tablica je preuzeta iz članka Chase, M.A i Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," *Research Quarterly for Exercise and Sport*, 63, 418-424. U istraživanju su učenici 4., 5. i 6. razreda izrazili svoje preference prema tome što im je najbitnije u školi: ocjene, popularnost ili sportski uspjeh. Ako želimo usporediti zavisnost između dobi (odn. razreda) učenika i njihovog odabira možemo konstruirati tablicu i provesti test sljedećim naredbama

```
> tableSk<-rbind(ocjene,popularnost,sport)
> colnames(tableSk) <- razredi
> tableSk
> chisq.test(tableSk)
```

Uočite da se radi o tablici 3×3 tako da imamo 4 stupnja slobode za testnu statistiku.

Zadaci

Zadatak 1. Na uzorku od 100 ljudi iz neke populacije uočeno je da njih 17 ima plavu kosu. Neka je p vjerojatnost da slučajno odabrana osoba u toj populaciji ima plavu kosu. Testirajte da li vrijedi $H_0 : p \leq 0.15$ nasuprot alternative $H_a : p > 0.15$ na nivou značajnosti od 1%.

Zadatak 2. Uz pretpostavke primjera 6.3.6 postavite hipoteze

$$H_0 : \mu \leq \mu_0 \quad \text{nasuprot} \quad H_A : \mu > \mu_0.$$

Koristeći statistiku Z i argumente iz primjera 6.3.6, pokažite da se funkcija snage može izračunati kao

$$\gamma(\mu) = 1 - \Phi\left(z_\alpha - \sqrt{n}\frac{\mu - \mu_0}{\sigma}\right).$$

Zadatak 3. U laboratoriju je mjerena količine hrane u gramima koju zamorci dnevno jedu u normalnim uvjetima i uz slušanje pop glazbe s radija. Na uzorku od tri zamorca su dobiveni sljedeći rezultati

	1	2	3
bez glazbe	205	181	201
s glazbom	215	180	234

Procijenite aritmetičku sredinu i standardnu devijaciju razlika u ovim mjerenjima. Postavite H_0 i H_a ako želimo utvrditi postoji li razlika u očekivanoj dnevnoj konzumaciji hrane u različitim režimima. Provedite test za ove hipoteze uz razinu značajnosti od 5%.

Zadatak 4. Pretpostavite da je S broj jedinki s izvjesnom karakteristikama u uzorku duljine n, te da je S binomna slučajna varijabla s paremetrima n, p. Ako želimo testirati sljedeće hipoteze $H_0 : p = p_0$ nasuprot $H_A : p \neq p_0$, koristimo uobičajenu testnu statistiku $Z = (S - np_0)/\sqrt{np_0(1-p_0)}$ i kritično područje $C_\alpha = \mathbb{R} \setminus (-z_{\alpha/2}, z_{\alpha/2})$. Pokažite da je snaga ovog testa uz pravi parametar $p = p_S \neq p_0$ jednaka

$$\gamma(p_S) = P(|Z| \geq z_{\alpha/2}) \approx 1 - \Phi(g + k) + \Phi(g - k)$$

gdje je Φ funkcija distribucije $N(0, 1)$ razdiobe, te vrijedi

$$g = \frac{\sqrt{n}(p_0 - p_S)}{\sqrt{p_S(1-p_S)}} \quad \text{i} \quad k = z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{p_S(1-p_S)}}.$$

Zadatak 5. Dana je tablica studenata koji pohađaju nastavu i polažu ispit iz Statistike:

	nisu položili	položili
ne pohađaju	10	18
pohađaju	8	36

Možemo li zaključiti da ne postoji veza između dolaženja na nastavu i uspješnog polaganja ispita? Zaključak donesite na nivou značajnosti od 10%.

Zadatak 6. Pretpostavlja se da je težina (u gramima) plodova rajčice u nekom plasteniku normalno distribuirana s očekivanjem 65 g i standardnom devijacijom 25 g. Rezultati mjerenja 100 plodova prikazani su u tablici, no bez teorijskih frekvencija. Popunite tablicu do kraja i testirajte hipotezu da podaci dolaze iz ove normalne razdiobe na razini značajnosti 0.05.

interval	O_i	E_i
$(-\infty, 50)$	32	
$[50, 80]$	55	
$(80, \infty)$	13	
Ukupno	100	100

7

Linearni modeli

7.1 Jednostavna linearna regresija

Pretpostavljamo da za svaku od n jedinki u uzorku imamo dva numerička podatka. Tada se podaci daju reprezentirati kao niz parova numeričkih varijabli

$$(x_i, y_i),\ i = 1, \ldots, n.$$

Primjer 7.1.1 U jednoj studiji prikupljeni su podaci o biljkama kupusa. Za svaku od 60 biljaka izmjerena je veličina glavice u kilogramima (x_i) i sadržaj vitamina C (u nepoznatim jedinicama) (y_i).

⌟

Prisjetimo se definicije koeficijenta korelacije za dvije sl. varijable X i Y kao

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - EX)(Y - EY)]}{\sigma_X \sigma_Y}.$$

Ako $\text{cov}(X, Y) = 0$, slijedi $\text{corr}(X, Y) = 0$, u tom slučaju kažemo da su X i Y nekorelirane. Npr. nezavisne sl. varijable su uvijek nekorelirane, no obrnuto ne vrijedi.

Broj $\text{corr}(X, Y)$ je uvijek izmedju -1 i 1, što je bliži granicama tog intervala linearna veza izmedju sl. varijabli X i Y je jača.

Kako mi imamo samo podatke i ne znamo pravu (teorijsku) razdiobu para sl. varijabli (X, Y), mi možemo izračunati samo procjenu za koeficijent korelacije. Nju zovemo **korelacija slučajnog uzorka** i računamo je kao

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y},$$

uz uobičajene oznake, tako je npr.

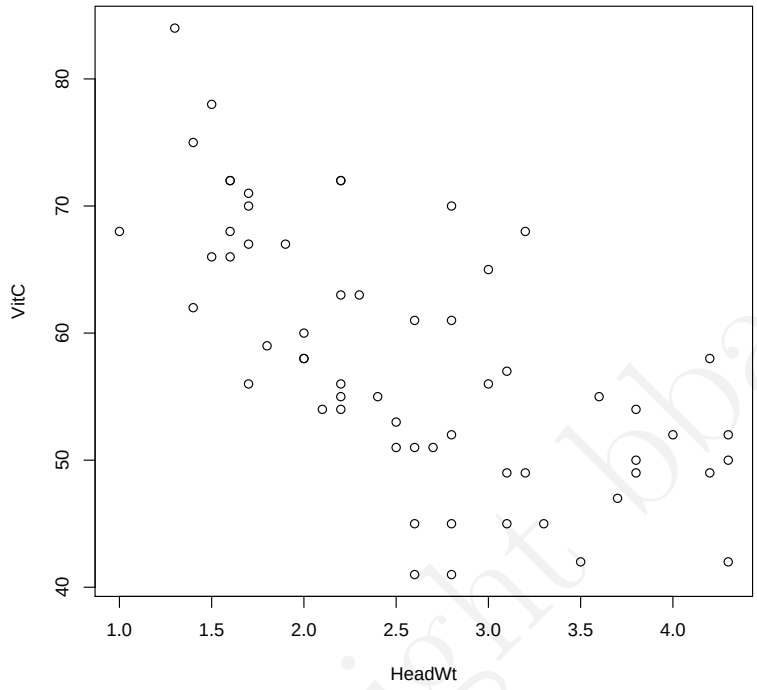
$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

varijanca uzorka $x_1, \ldots, x_n$. Uvedemo li

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

te

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{i} \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$



Slika 7.1: Scatterplot ukazuje da dvije varijable nisu nezavisne.

možemo pisati

$$r = \frac{n-1}{n} \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \approx \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} ,$$

za velike n.

Kod podataka o kupusu su npr

$$\bar{x} = 2.593,\ \bar{y} = 57.95,\ s_X = 0.8823825,\ s_Y = 10.11866 ,$$

a korelacija uzorka iznosi

$$r = -0.659892.$$

U modeliranju zavisnosti izmedju varijabli najjednostavnije je prepostaviti da postoji linearna veza izmedju njih, tj.

$$y_i = a + bx_i .$$

Takvu vezu zovemo determinističkom, a svi bi podaci u tom slučaju ležali točno na pravcu koji opisuje gornja jednadžba. Zbog raznih izvora varijabilnosti kod stvarno prikupljenih podataka, često je razumnije uvesti vjerojatnosni model za podatke

$$Y_i = a + bx_i + \varepsilon_i ,$$

gdje ε_i predstavlja slučajnu pogrešku.

Pitanje je kako bismo našli optimalne koeficijente pravca kroz točke podataka, recimo $\hat{a}$ i $\hat{b}$. Najčešće korištena metoda minimizira tzv. sumu kvadrata ostataka (eng. residual sum of squares–RSS ili sum of squares for error–SSE)

$$SSE = \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b}x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Ova metoda za pronalaženje optimalnih koeficijenata $\hat{a}$ i $\hat{b}$ naziva se **metoda najmanjih kvadrata**.

Termina regresija uveo je F. Galton, a R. Bošković je predložio minimizirati sumu apsolutnih pogrešaka $\sum |y_i - \hat{y}_i|$, no tada optimalne a i b nije lako naći.

Ispostavlja se da je optimalni koeficijent $\hat{b}$ zapravo

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \approx r\frac{s_Y}{s_X}$$

dok je

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

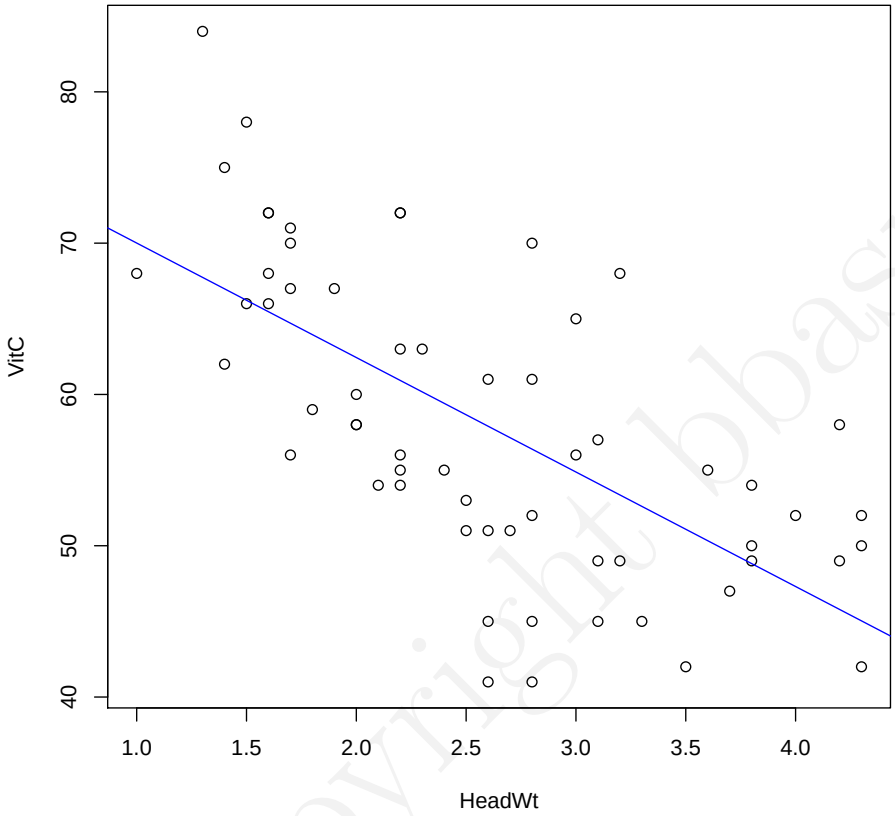
Veličine

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \ i = 1, \ldots, n$$

zovemo procjenama za vrijednosti y_i, a

$$y_i - \hat{y}_i, \ i = 1, \ldots, n,$$

nazivamo ostacima.



Procjenjeni koeficijenti za podatke o kupusu su $\hat{a} = 77.57$ i $\hat{b} = -7.57$.

Naravno nas interesira i neizvjesnost u procjeni ovih parametara ili čak statistički test. Ako želimo npr. napraviti interval pouzdanosti $(1-\alpha)100\%$ za parametar b to možemo učiniti ako vrijede dodatni (Gauss-Markovljevi) uvjeti na slučajne greške ε_i

- i) $E(\varepsilon_i) = 0$ za sve i
- ii) $\text{var}\,(\varepsilon_i) = \sigma^2 < \infty$ za sve i
- iii) $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ za sve $i \neq j$

Pod ovim prepostavkama $(1-\alpha)100\%$-tni interval pouzdanosti za parametar b je

$$\left(\hat{b} - t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}, \hat{b} + t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right),$$

gdje je kao gore $S_{xx} = (n-1)s_X^2$, a

$$s^2 = \frac{SSE}{n-2}.$$

Korisno je znati i da se i SSE može izračunati kao

$$SSE = (1 - r^2)S_{yy} = (1 - r^2)(n - 1)s_y^2 .$$

Ako želimo testirati sljedeće dvije hipoteze

$$H_0 : b = 0 \quad \text{nasuprot} \quad H_A : b \neq 0,$$

uz nivo značajnosti $\alpha > 0$ testna statistika je

$$T = \frac{\hat{b}\sqrt{S_{xx}}}{s}$$

a kritično područje je

$$C_\alpha = \mathbb{R} \setminus (-t_{\alpha/2}(n-2), t_{\alpha/2}(n-2)).$$

Dakle, koristimo Studentovu t-razdiobu s $n - 2$ stupnja slobode.

U primjeru s vrstama kupusa za test hipoteze $H_0 : b = 0$, testna statistika iznosi $T = -6.69$, a broj stupnjeva slobode jednak je 58. Tako da je $T \in C_{0.05}$, a i p vrijednost zanemarivo mala. Dakle odbacili bismo nul-hipotezu na nivou $\alpha = 0.05$.

Ako nam je potreban i $(1 - \alpha)100\%$-tni interval pouzdanosti za vrijednost Y kada znamo vrijednost kontrolirane varijable $x = x_0$, on se može dobiti kao

$$\hat{a} + \hat{b}x_0 \pm t_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

Ponekad se u softverskim paketima regresijska analiza provodi koristeći tzv. ANOVA (analysis of variance) tablice. U njima se varijabilnost podataka y_i razlaže na razne komponente.

izvor varijabilnosti	greška
varijabilnost zbog regresije	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
varijabilnost zbog greške	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
ukupno	$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$

Grubo govoreći, kvocijent $r^2 = SSR/S_{yy}$ mjeri koliki udio u ukupnoj varijabilnosti varijable y, možemo objasniti linearnom regresijom. Primjetimo na kraju, ako imamo više od dvije numeričke varijable, i tada možemo napraviti regresiju jedne od njih u odnosu na ostale, takav model se naziva višestruka regresija (multiple regression).

7.2 Jednofaktorska analiza varijance

Ponekad o svakoj jedinki u uzorku, osim jednog numeričkog opažanja, imamo i jedno kvalitativno koje indicira pripadnost jedinke nekoj od kategorija ili vrsti tretmana pod kojim je izvršeno opažanje. Prva je dakle varijabla numerička, a drugu u ovom kontekstu nazivamo **faktorom**.

Analiza ovakvih podataka ovisi o samom dizajnu pokusa. Dizajn eksperimenta je izuzetno važan korak svakog istraživanja. Istraživač se mora odlučiti koje faktore želi proučavati, ali i kako će dodjeljivati različite pokusne jedinke različitim fakorima. Vrijednosti faktora su dakle razne vrste tretmana kojima podvrgavamo jedinke u pokusu, a ono što želimo tipično utvrditi je: postoji li zavisnost izmedju vrste tretmana i numeričke varijable.

Umjesto da usporedjujemo različite tretmane u parovima (npr. t-testovima) očito je efikasnije usporediti ih sve odjednom. Ako uzorke skupljamo u svakoj od k kategorija ili populacija (tj. za svaki tretman) na slučajan i nezavisan način govorimo o **potpuno randomiziranom dizajnu**.

Sve podatke sada dijelimo na k uzoraka ovisno o nivou faktora tj. vrsti tretmana

$$\begin{aligned} &x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}, \\ &x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}, \\ &\qquad \vdots \\ &x_{k,1}, x_{k,2}, \ldots, x_{k,n_k}, \end{aligned}$$

gdje je $n = n_1 + n_2 + \cdots + n_k$ ukupan broj jedinki u pokusu.

Za svaki od k uzoraka možemo izračunati aritmetičku sredinu

$$\bar{x}_i = \frac{1}{n_i}(x_{i,1} + x_{i,2} + \cdots + x_{i,n_i}) = \frac{\sum_{j=1}^{n_i} x_{i,j}}{n_i}$$

odn. varijancu

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2,$$

$i = 1, \ldots, k$.

Ukupna aritmetička sredina podataka je

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j}.$$

Ako postoji zavisnost izmedju faktora i numeričke varijable za očekivati je da se aritmetičke sredine raznih uzoraka bitno razlikuju, no nije apriori jasno kako odrediti ono što bismo zvali bitnim razlikama.

U tu svrhu provodimo **analizu varijance** (analysis of variance–ANOVA) razlažući varijabilnost u podacima. Definiramo sumu kvadrata u odn. na tretman (sum of squares

for treatments) kao

$$SST = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2,$$

te sumu kvadrata pogrešaka (sum of squares for error)

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2.$$

Na osnovu ovih veličina definiramo i pripadna srednjekvadratna odstupanja:

- MST (mean square for tretments)

$$MST = \frac{SST}{k-1}$$

- MSE (mean square for error)

$$MSE = \frac{SSE}{n-k}$$

Intuitivno je jasno da ako se značajan dio varijabilnosti dade objasniti različitim tretmanima tada očekujemo da je MST značajno veća od MSE.

Vjerojatnosni model za podatke možemo napisati na sljedeći način

$$X_{i,j} = \mu + \beta_i + \varepsilon_{i,j},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

gdje je μ ukupno očekivanje za numeričku varijablu, β_i predstavlja efekt i-tog tretmana, a $\varepsilon_{i,j}$ su nezavisne greške za koje pretpostavljamo da sve imaju $N(0, \sigma^2)$ razdiobu.

Ako želimo utvrditi utječu li tretmani na numeričku varijablu, možemo testirati

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

nasuprot alternative da je bar neka od β_i različita od 0.

Uz naše prepostavke testna statistika

$$F = \frac{MST}{MSE}$$

pod hipotezom H_0 ima Fisherovu razdiobu s $(k-1, n-k)$ stupnjeva slobode ili kraće $F(k-1, n-k)$ razdiobu.

Na nivou značajnosti $\alpha > 0$ odbacujemo nul-hipotezu ako je

$$F \geq F_\alpha(k-1, n-k),$$

tj. ako je F veća od $(1-\alpha)$-kvantila odgovarajuće F razdiobe.

Cijeli račun se tipično izvodi koristeći tzv. ANOVA tablice koje imaju sljedeći oblik

izvor varijabilnosti	stup. slob.	suma kvad.	srednjekv. ods.	F stat.
tretman	$k-1$	SST	MST	MST/MSE
greška	$n-k$	SSE	MSE	
ukupno	$n-1$	SS	s_X^2	

Pri tom je S_{xx} jednostavna suma kvadrata (sum of squares) tj.

$$S_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x})^2 = (n-1)s_X^2.$$

Uočite da vrijedi

$$S_{xx} = SST + SSE.$$

Postoje i mnoge druge ANOVA verzije. No postavlja se i pitanje ako je metoda izvedena pod prepostavkom da su greške normalno distribuirane, koliko je ona robusna na odstupanja od ove prepostavke, Ispostavlja se da je ANOVA nažalost osjetljiva na slučaj nejednakih varijanci, no taj je efekt mali u slučaju jednako velikih uzoraka za razne faktore (p. 574. McClave et al).

7.3 Linearni modeli u R-u

Jednostavna linearna regresija

Jednostavnu linearnu regresiju možemo prvo primijeniti na podatke *faithful* koji se tiču erupcije gejzira Old Faithful i standardni su dio R paketa. Skup podataka sadrži vremena trajanja erupcija i vremena čekanja između dvije erupcije. Procjenitelje za parametre a i b možemo lako naći naredbama

```
> procjena.lm <- lm(eruptions ~ waiting, data=faithful)
> summary(procjena.lm)
```

Primijetite da je ujedno proveden i test nulhipoteze $H_0 : b = 0$, te da je p– vrijednost ograničena odozgo vrlo malim brojem, tako da bismo na svim uobičajenim razinama značajnosti nulhipotezu odbacili. Naredba `coefficients(procjena.lm)` će nam dati samo procjenjene koeficijente. Grafički prikaz podataka i procjenjenog modela možemo dobiti naredbama

```
> attach(faithful)
> plot(waiting,eruptions)
> abline(procjena.lm)
```

Ukoliko želimo procjeniti interval pouzdanosti za vrijednost varijable *eruption* kada znamo vrijednost kontrolirane varijable *waiting* x_0, možemo pozvati naredbe

```
> x0 <- data.frame(waiting=75)
> predict(procjena.lm, x0, interval="confidence")
```

Slično, regresijski model bismo mogli uspostaviti i za vlastite podatke, recimo o ovisnosti broja sati učenja i broju bodova postignutih na nekom ispitu

```
> satiucenja<-c(37.3, 26.2, 33.0, 29.2, 28.9, 20.9, 23.2, 36.2,37.0, 37.1)
> brojbodova<-c( 107, 70,   58,   44,  57,  67,   77,  80,   71, 92 )
> plot(satiucenja,brojbodova)
> abline(lm(brojbodova~satiucenja),col="blue")
> summary(lm(brojbodova~satiucenja))
```

Jednofaktorska analiza varijance

Prtepostavimo da pratimo biljke čije visinu stabljike u cm mjerimo ovisno o tri tipa staništa

tip	visine
tip A	136.67, 95.71, 111.64, 110.68, 69.98
tip B	89.03, 91.32, 82.98, 121.06 , 79.73, 63.60
tip C	111.21, 95.77, 103.00, 89.30, 143.97, 103.68

Ako želimo testirati utjecaj tipa staništa na visinu biljaka podatke možemo pripremiti na sljedeći način

```
> fenotip <- scan()
136.67  95.71 111.64 110.68  69.98
89.03  91.32  82.98  121.06  79.73  63.60
111.21 95.77 103.00 89.30 143.97  103.68
```

Nakon toga moramo učitati i razine faktora npr. naredbama

```
> tipovi<-factor(c(rep("tip A",5),rep("tip B",6),rep("tip C",6))
```

Jenostavnije bi bilo `tipovi<-factor(c(1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3))`. Grafički bismo evt. razlike između tipovima najlakše mogli ilustrirati naredbom

```
>  boxplot(fenotip~tipovi)
```

Sad formalnu analizu varijance možemo provesti naredbama

```
>  anova.tab <- anova(lm(fenotip~tipovi))
>  anova.tab
```

Primijetite da na kraju dobijemo i uobičajenu ANOVA tablicu.

Zadaci

Zadatak 1. Na uzorku od 20 stabala oraha mjerena je njihova visina u metrima (varijabla x) i širina njihovog debla u centimetrima (varijabla y). Podaci su sažeti: $\sum_{i=1}^{20} x_i = 370$, $\sum_{i=1}^{20} y_i = 1020$, $\sum_{i=1}^{20}(x_i - \overline{x})(y_i - \overline{y}) = 310$, $\sum_{i=1}^{20}(x_i - \overline{x})^2 = 120$, $\sum_{i=1}^{20}(y_i - \overline{y})^2 = 220$, $\sum_{i=1}^{20}(y_i - \hat{y})^2 = 21$. Procijenite parametre a i b u linearnoj regresiji $Y = a + bx + \varepsilon$.

Zadatak 2. Za podatke iz prethodnog zadatka odredite koeficijent korelacije r te testirajte hipotezu $b = 0$ uz nivo značajnosti od 90%.

Zadatak 3. Na jednoj lokaciji mjerena je visina tri različite vrste kukuruza, i to na po 6 primjeraka svake vrste. Dobiveno je $MST = 0.5$ i $MSE = 0.2$. Testirajte da li postoji statistički značajna razlika u prosječnoj visini klasa kukuruza između tih vrsta uz razinu značajnosti 0.01. Sastavite pripadnu ANOVA tablicu.

Bibliografija

[1] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

[2] Dalgaard, P. (2008). *Introductory statistics with R.* Springer.

[3] Gonick, L. (1993). *Cartoon guide to statistics.* HarperCollins.

[4] McCandless, D. (2009) *Information is beautiful.* HarperCollins UK.

[5] Sarapa, N. (2003). *Teorija vjerojatnosti.* Školska knjiga, Zagreb.

[6] Sorić, B. (1981) Poboljšanje metode i kontrola ispravnosti statističkog zaključivanja. (In Croatian). *Zdravstvo*, **2**3, 154–170.

[7] Sorić, B. (1989) Statistical "discoveries" and effect size estimation. *J. Am. Statist. Ass.*, **8**4, 608–610.

[8] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer.

[9] Williams, D. (2001). *Weighing the odds: a course in probability and statistics.* Cambridge: Cambridge University Press.

[10] Zar, J. H. (1999). *Biostatistical analysis* 4 ed. Princeton-Hall, New Jersey.