Uvod u Bayesovu statistiku

Horvat, Tamara Anna

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tamara Anna Horvat

UVOD U BAYESOVU STATISTIKU

Diplomski rad

Voditelj rada:
doc. dr. sc. Ivan Ivec

Zagreb, rujan 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Mami i Tomislavu, hvala što ste mi sve ove godine bili kontinuirana podrška u svakom pogledu. Marku, hvala što si kroz sve stresne trenutke bio uz mene kao moj štit. Znaš i sam da je teško naći dobrog supporta, ali ja sam svog pronašla upravo u tebi. Lokiju, hvala što si uvijek znao kada je vrijeme za pauzu, maženje, igranje i neizostavne keksiće. To myself - jer zaslužila si. Remember, wherever life plants you - bloom with grace.

Sadržaj

Uvod

Korištenje riječi statistika u svakodnevnom životu najčešće je povezano s brojčanim vrijednostima kojima pokušavamo opisati bitne karakteristike nekog skupa podataka. Statistika kao znanstvena disciplina bavi se razvojem metoda prikupljanja, opisivanja i analiziranja podataka, ali i primjenom tih metoda u procesu donošenja zaključaka na temelju prikupljenih podataka. Danas razlikujemo matematičku statistiku, koja razvija i usavršava nove metode, te primijenjenu statistiku koja se bavi primjenom tako razvijenih metoda na stvarne podatke odnosno stvarnu problematiku.

Svako statističko istraživanje fokusirano je na skupu objekata (jedinki) i skupu odabranih veličina koje se na njima promatraju. Veličine koje se na jedinkama promatraju zovemo varijablama, a sve jedinke koje se žele danim istraživanjem obuhvatiti čine populaciju. Svako obilježje možemo mjeriti ovisno o mjernom instrumentu i njegovoj preciznosti. Svakoj varijabli pridružujemo mjernu vrijednost, a koliko točno ćemo nešto izmjeriti ovisi o našim potrebama, ali i mjernom instrumentu. Podaci se prikupljaju da bi se zaključivalo o varijablama na populaciji, no naravno, prije samog provođenja istraživanja populacija mora biti precizno opisana. Iz definirane populacije zatim se uzima uzorak, a podaci moraju biti prikupljeni na uzorku koji je reprezentativan za opisanu populaciju. Da bi uzorak bio reprezentativan, on mora odražavati populaciju tj. u njemu trebaju biti zastupljene sve tipične karakteristike populacije.

Najčešći način odabira jedinki iz populacije u reprezentativan uzorak je tzv. slučajni uzorak, tj. to je takav izbor u kojemu svaka jedinka ima jednaku vjerojatnost biti izabrana u uzorak. Upravo ta slučajnost, koja na svojevrsni način implicira određenu neizvjenost, stavlja naglasak na korištenje statističkih alata pri stvaranju, procjeni i testiranju određenih modela.

Općenito, postoje dva pristupa statističkoj analizi: frekvencionistički te Bayesovski. Osnovna razlika među njima leži u načinu na koji se interpretira osnovni pojam vjerojatnosti. Kao što samo ime kaže, frekvencionistička statistika, poznata kao i klasična statistika, vjerojatnost tretira kao limes relativnih frekvencija. Primijetimo da pridržavanje ovog pristupa nije uvijek moguće u praksi, primjerice kada proučavamo određene događaje koji se rijetko pojavljuju. U tom smo slučaju prisiljeni oslanjati se na teorijske rezultate, a upravo zbog nedovoljne veličine uzorka. Zagovornici Bayesovskog pristupa smatraju pak

da je vjerojatnost subjektivna, tj. da je ona stupanj uvjerenja koji se mijenja dolaskom novih informacija. Uz interpretaciju pojma vjerojatnosti vežemo i pojam neizvjesnosti. Sljedbenici klasičnog pristupa vjeruju kako je izvor neizvjesnosti isključivo realizacija slučajnih varijabli, dok Bayesova statistika ide korak dalje i smatra tu slučajnost nedovoljnom.

Do naglog rasta popularnosti Bayesove statistike dolazi 1980-ih kada ona postaje sve učestalija tehnika pri izgradnji statističkih modela za rješavanje svakodnevnih problema. Naziv je dobila po engleskom statističaru Thomasu Bayesu koji je prvi dokazao specijalni slučaj teorema kojeg danas nazivamo Bayesov teorem ili češće, Bayesova formula. Njegov rad "Essay Towards Solving a Problem in the Doctrine of Chances" objavio je Royal Society 1763. godine, dvije godine nakon njegove smrti. Zanimljivo je da je bez znanja o Bayesovom radu, francuski matematičar i astronom Pierre-Simon Laplace otkrio Bayesov teorem 1774. godine, ali u općenitijem i jasnijem obliku. Sljedećih 40 godina primjenjivao ga je na području statistike, meteorologije, astronomije pa čak i geodezije.

Iako je klasična statistika preuzela svojevrsnu vodeću poziciju među statističkim metodama, neki statističari nastavljali su širiti Laplaceove ideje. Tako su se stvorile dvije struje u Bayesovoj statistici: subjektivistička i objektivistička. Osnovne razlike ova dva pristupa su u interpretaciji apriornih distribucija. Danas Bayesova statistika ima širok spektar primjene u znanstvenim disciplinama, a Bayesovske metode modeliranja problema pružaju znanstvenicima mogućnost da na prirodan način oblikuju svoje podatke i znanja te da dobiju iskustvene i direktne odgovore na svoja pitanja. Cilj ovog diplomskog rada je dati jasan i sažet pregled teorije koja stoji iza Bayesove statistike te prikazati njenu uporabu kroz različite primjere.

Poglavlje 1

Osnovni pojmovi statistike

1.1 Definicija

Statistika je grana matematike koja obuhvaća skupljanje, analizu, interpretaciju i prezentaciju podataka te izradu predviđanja koja se temelje na tim podacima. Veliku važnost u korištenju statistike imaju i planiranje te provođenje pokusa, tj. skupljanje podataka koji će se analizirati, te interpretacija istih. Počeci statistike javili su se još u 5. stoljeću p. n. e., a najstariji zapisi o korištenju statistike potiču iz 9. stoljeća. Napisao ih je arapski znanstvenik Al-Kindi u svrhu proučavanja kodiranih poruka. Pojam statistika je prvobitno izveden iz latinskog izraza statisticum collegium (vijeće država) te talijanske riječi statista (državnik ili političar). Značenje sakupljanja i analize podataka statistika je dobila početkom 19. stoljeća, a riječ „statistika“ je u engleski jezik uveo Sir John Sinclair.

1.2 Osnovni pojmovi

Temeljni pojmovi koji se koriste u statistici su populacija i uzorak.

Definicija 1.2.1. ***Populacija*** *ili osnovni skup je skup podataka svih jedinki ili objekata od interesa.*

Primjerice populacija je skup svih biljaka koje jedna cvjećarna uzgaja.

Definicija 1.2.2. *Uzorak je podskup populacije za koji se mjere ili skupljaju podaci koji nas zanimaju.*

Obilježje možemo mjeriti ovisno o mjernom instrumentu i njegovoj preciznosti (metar, vaga, tlakomjer, brojač) ili opažati, prebrojavati (oči). Tako svako obilježje poprima neku mjernu vrijednost.

Primjer 1.2.3. *Masu možemo izmjeriti precizno; na kilogram točno (17 kg), na dekagram točno (17, 05 kg) ili do na gram točno (17, 054 kg).*

Naravno, koliko točno ćemo nešto izmjeriti ovisi o našim potrebama, ali i mjernom instrumentu pomoću kojeg vršimo mjerenje. Populacije koje promatramo mogu biti konačne i beskonačne, no uzorak koji promatramo uvijek je konačan.

Primjer 1.2.4. *Želimo provjeriti je li određena kocka pravedna. Populacija koju promatramo sastojat će se od beskonačno mnogo bacanja dane kocke, a uzorak će se sastojati od konačno mnogo bacanja te kocke, recimo njih 10 000.*

Dva su osnovna tipa statistike: deskriptivna te inferencijalna. Deskriptivna statistika samo opisuje dane podatke, a zapravo je uzorak jednak populaciji. Deskriptivna statistika obuhvaća postupke uređivanja, tabličnog i grafičkog prikazivanja podataka te izračunavanje opisnih statističkih pokazatelja.

Primjer 1.2.5. *Analiza rezultata kolokvija iz metodike nastave matematike, iz kojih se ne pokušava donijeti zaključak kakvi bi oni mogli biti za studente koji ispit nisu pisali.*

Inferencijalna statistika donosi zaključke o osnovnom skupu tj. populaciji na bazi promatranog uzorka.

Primjer 1.2.6. *Devedeset studenata pristupa ispitu kolegija metodike nastave matematike. Na temelju ostvarenog broja bodova studenata, možemo li zaključiti da će prosječno student imati više od 80 bodova na ispitu?*

Poglavlje 2

Deskriptivna statistika

U statističkim istraživanjima razlikujemo nekoliko osnovnih tipova varijabli koje se međusobno razlikuju po svojstvima vrijednosti koje mogu poprimiti.

2.1 Varijable

2.1.1 Kvalitativne varijable

Karakteristika kvalitativnih varijabli je da njihove vrijednosti nisu, po svojim svojstvima korištenim u istraživanju, realni brojevi. Tipičan primjer takve varijable je spol osobe. Vrijednosti kvalitativne varijable uobičajeno nazivamo kategorijama, a one mogu biti definirane u skladu s potrebama statističkog istraživanja.

Primjer 2.1.1. *Sljedeće su varijable kvalitativnog tipa:*

- *radna mjesta u školi (spremačica, domar, tajnik, nastavnik, pedagog, ravnatelj),*
- *boja očiju (plava, smeđa, zelena),*
- *krvne grupe (A, B, AB, 0),*
- *spol (m ili ž).*

2.1.2 Numeričke varijable

Numeričke varijable prirodno primaju vrijednosti iz skupa realnih brojeva. Tipični su primjeri numeričkih varijabli masa i visina osobe. Međutim, treba naglasiti da se i kategorije kvalitativnih varijabli mogu izražavati brojevima što ih ne čini numeričkim varijablama. Primjerice, spol osobe jedna je kvalitativna varijabla. Tako kategoriju "ženski spol"

možemo označiti s ”1”, a kategoriju ”muški spol” s ”2”, što može biti korisno prilikom unošenja podataka u bazu. Time smo kategorijama kvalitativne varijable pridružili numeričke vrijednosti, ali samu varijablu nismo učinili numeričkom po njezinim svojstvima.

Primjer 2.1.2. *Sljedeće su varijable numeričkog tipa:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine,*
- *broj bodova na državnoj maturi iz matematike,*
- *broj zaraženih COVID-19 virusom,*
- *temperatura mora,*
- *koncentracija soli u morskoj vodi.*

Među numeričkim varijablama razlikujemo diskretne i kontinuirane varijable. Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti.

Primjer 2.1.3. *Sljedeće su numeričke varijable diskretne:*

- *broj bodova na državnoj maturi iz matematike,*
- *broj ulovljenih komaraca u klopku,*
- *broj dana u godini s temperaturom zraka većom od 35°C.*

Skup je mogućih vrijednosti kontinuiranih numeričkih varijabli cijeli skup realnih brojeva ili neki interval.

Primjer 2.1.4. *Sljedeće su numeričke varijable kontinuirane:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine,*
- *temperatura mora,*
- *vodostaj neke rijeke.*

U svrhu prikaza podataka i nekih statističkih analiza, vrijednosti se numeričke varijable također mogu svrstati u kategorije. Za razliku od kategorija kvalitativnih varijabli, među kategorijama se numeričke varijable uvijek može prepoznati prirodan poredak.

U sljedećem primjeru prikazat ćemo mogućnost kategorizacije numeričke varijable. To je postupak koji se najčešće provodi stvaranjem nove varijable čije su vrijednosti kategorije kojih je (znatno) manje nego svih mogućih vrijednosti odgovarajuće numeričke varijable.

Primjer 2.1.5. *Dana je baza podataka koja se sastoji od sljedećih varijabli:*

– *automobili* - diskretna numerička varijabla koja sadrži podatke o broju prodanih automobila u jednom danu za sto promatranih dana

– *kategorija* - kvalitativna varijabla koja podatke iz varijable automobili svrstava u pet kategorija (prema kriteriju prikazanom u tablici)

S obzirom na to da broj prodanih automobila u jednom danu može biti vrlo malen (primjerice samo nekoliko osobnih automobila), ali i vrlo velik (npr. narudžbe automobila za veliko poduzeće), zaključujemo da diskretna numerička varijabla *automobili* može poprimiti velik broj različitih vrijednosti skupa prirodnih brojeva. Stoga je u nekim situacijama korisno kategorizirati vrijednosti te varijable prema nekom točno, unaprijed određenom kriteriju. Kategorizaciju broja prodanih automobila u jednome danu možemo napraviti kao što je prikazano varijablom kategorija.

broj prodanih automobila	kategorija
0 - 5	E
6 i 7	D
8 i 9	C
10 i 11	B
12 i više	A

Tablica 2.1: Primjer kategorizacije numeričke varijable automobili

2.1.3 Ordinalne varijable

Karakteristika ordinalnih varijabli jest da su one po svom karakteru kvalitativne, no među kategorijama se ipak može uspostaviti prirodan poredak. Tipični primjeri ordinalnih varijabli su stručna sprema osobe ili ocjene u školi.

Primjer 2.1.6. *Dana je baza podataka koja sadrži podatke prikupljene anonimnim anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz kolegija metodika nastave matematike 4. Prikupljeni podaci organizirani su na sljedeći način:*

– *prosjek* - varijabla koja sadrži podatke o prosječnoj ocjeni studiranja za 49 anketiranih studenata

– *položeno* - varijabla koja studente svrstava u dvije kategorije s obzirom na to jesu li položili ispit

položen/nepoložen ispit	kategorija
položen ispit	1
nepoložen ispit	0

Tablica 2.2: Kategorizacija studenata s obzirom na položenost ispita

– *predavanja / vježbe* - dvije varijable koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije

prisutnost studenta na p/v	kategorija
student nikada nije izostao sa p/v	1
student je samo jednom izostao sa p/v	2
student je s p/v izostao barem 2 puta	3

Tablica 2.3: Kategorizacija studenata s obzirom na broj izostanaka s predavanja/vježbi

– *težina kolegija / materijali* - dvije varijable koje sadrže subjektivne ocjene studenata o težini kolegija i količini dostupnih materijala za pripremanje ispita iz promatranog kolegija (u standardnoj skali ocjena 1-5).

Uočimo da se varijabla prosjek može promatrati kao neprekidna numerička varijabla, varijabla položeno jest kvalitativna, dok se varijable predavanja/vježbe, težina kolegija/materijali mogu svrstati u ordinalne varijable.

2.2 Metode opisivanja varijabli

2.2.1 Metode opisivanja kvalitativnih varijabli

Očito je kako su grafički prikazi vrlo važan dio deskriptivne statistike. Postoje različiti načini za grafičko prikazivanje podataka, a važno je napomenuti kako grafički prikaz mora biti čitljiv sam po sebi – bez obzira na tekst koji se nalazi uz njega. Potrebno je voditi brigu o ispravno označenim osima (naslovi, mjerne jedinice) te umetnutnim legendama (ukoliko za njima postoji potreba). Vrijednosti kvalitativne varijable jesu kategorije, a mjere kojima opisujemo zastupljenost pojedine kategorije u uzorku jesu frekvencija kategorije te relativna frenkvencija kategorije.

Definicija 2.2.1. ***Frekvencija kategorije*** *je broj izmjerenih vrijednosti varijable koje pripadaju danoj kategoriji.* ***Relativna frekvencija kategorije*** *je broj izmjerenih vrijednosti varijable koje pripadaju danoj kategoriji, podijeljen ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu.*

Pretpostavimo da varijabla može primiti vrijednost k različitih kategorija, s time da se u podacima nalazi N izmjerenih vrijednosti za tu varijablu. Frekvenciju i-te kategorije označit ćemo s f_i, a relativnu frekvenciju dobijemo kao $\frac{f_i}{N}$. Relativna frekvencija daje nam informaciju o udjelu kategorije u uzorku poznate veličine i najčešće se izražava kao postotak. Dobivene podatke pojedinih kategorija možemo prikazivati tablično i grafički, a najčešće se koriste stupčasti dijagrami frekvencija te relativnih frekvencija, ali i kružni dijagrami. Podatke iz sljedećeg primjera prikazat ćemo grafički, na nekoliko načina.

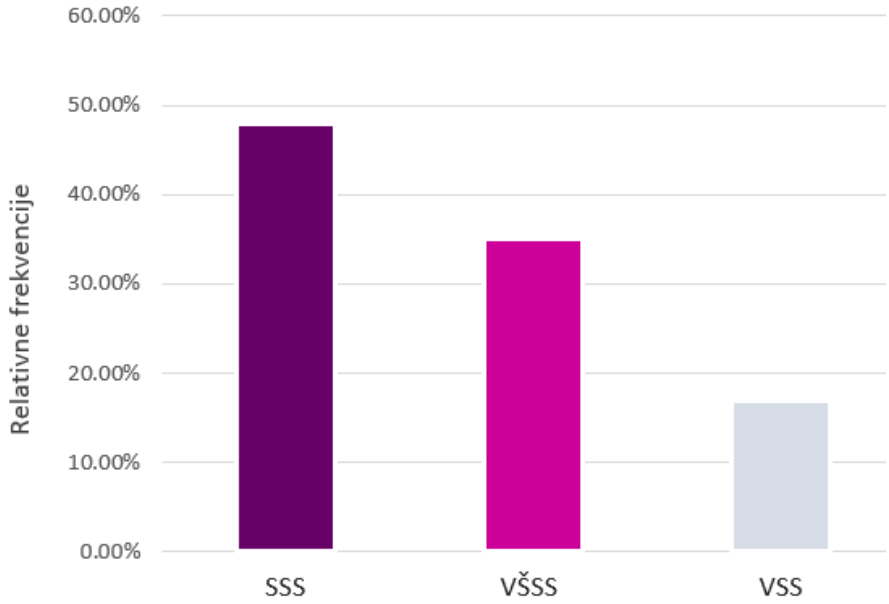
Primjer 2.2.2. *Dana je baza podataka o uzorku od 100 djelatnika tvornice Podravka u Koprivnici. Promotrimo kvalitativnu varijablu obrazovanja čije su vrijednosti svrstane u tri kategorije: srednja stručna sprema (SSS), viša stručna sprema (VŠSS) te visoka stručna sprema (VSS).*

Podatke ćemo prikazati pomoću tablice frekvencija i relativnih frekvencija te stupčastog i kružnog dijagrama relativnih frekvencija svih kategorija varijable obrazovanje.

kategorija	frekvencija	relativna frekvencija
SSS	48	48/100 = 0.48
VŠSS	35	35/100 = 0.35
VSS	17	17/100 = 0.17

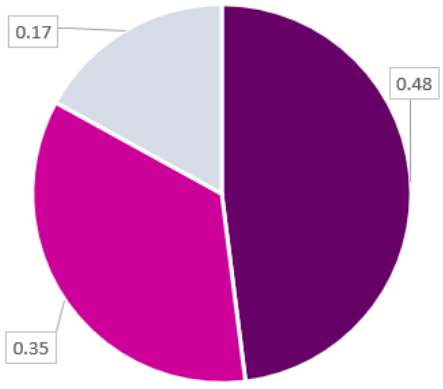
Tablica 2.4: Tablica frekvencija i relativnih frekvencija varijable obrazovanje

Stupčasti dijagram je grafički prikaz pomoću stupaca gdje je visina svakog stupca proporcionalna frekvenciji kategorije koju predstavlja.



Slika 2.1: Stupčasti dijagram relativnih frekvencija varijable obrazovanje

Kružni dijagram ili strukturni krug također se koristi ukoliko se prikazuju udjeli nekih kategorija u sveukupnom iznosu. Čitljiv je i pregledan samo ukoliko je broj ”sektora” kruga malen, a u suprotnom je potrebno izabrati prikladniji grafički prikaz.



Slika 2.2: Kružni dijagram relativnih frekvencija varijable obrazovanje

2.2.2 Metode prikazivanja numeričkih varijabli

Po svojoj prirodi numeričke varijable mogu biti diskretne i neprekidne, a u oba slučaja se može dogoditi da u prikupljenim zadacima postoji veliki broj međusobno različitih vrijednosti. U takvim slučajevima tablični i grafički prikazi uvedeni za dane kvalitativne varijable mogu biti nedovoljno informativni. Ukoliko su numeričke varijable diskretne s malo mogućih vrijednosti, tada za opis podataka možemo koristiti iste metode kao pri opisivanju kvalitativnih podataka, a koje su opisane u prethodnom poglavlju. U suprotnom, ako numerička varijabla prima mnogo međusobno različitih vrijednosti, za prikazivanje skupa izmjerenih vrijednosti tablice frekvencija (napravljenih na osnovu svake pojedine izmjerene vrijednosti) nam neće biti korisne. U svrhu dobivanja korisnih, ali i preglednih stupčastih dijagrama za podatke iz kontinuiranih numeričkih varijabli, potrebno je izmjerene vrijednosti kategorizirati na određeni način. Veliki, polazni skup podataka podijelit ćemo u nekoliko disjunktnih intervala po proizvoljnom kriteriju koji smo unaprijed odredili. Crtat ćemo posebnu vrstu stupčastog dijagrama – histogram.

Definicija 2.2.3. ***Histogram*** *je grafički prikaz nekog skupa podataka koji se sastoji od međusobno susjednih pravokutnika s po jednom stranicom na osi apscisa. Pritom se površine dijelova odnose kao (relativne) frekvencije podataka, a visine iznad pojedine apscise su te (relativne) frekvencije po jednoj jedinici apscise, tj. površine podijeljene sa širinom pojedine skupine podataka.*

Dakle, u histogramu odnose među frekvencijama ne pokazuju ordinate već površine, a ukupna površina histograma jednaka je N (odnosno 1 ukoliko su ordinate relativne frekvencije). Ukoliko su širine skupina podataka jednake, onda su i visine pravokutnika u histogramu također proporcionalne frekvencijama.

Strogo matematički, histogram je funkcija koja broji koliko podataka pripada određenoj klasi. Ako je k broj klasa na koje smo rasporedili N podataka, tada mora vrijediti:

$$\sum_{i=1}^{k} m_i = N$$

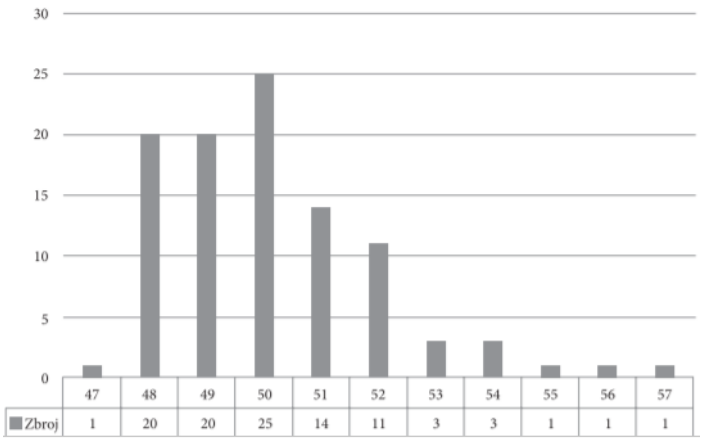
gdje je m_i broj podataka u i-toj klasi. Grafički prikaz tako definiranog histograma sastoji se od pravokutnika čije se širine odnose kao širine klasa, a površine kao brojevi m_i.

Primjer 2.2.4. *Sljedeći niz podataka predstavlja 100 mjerenja prosječnog broja otkucaja mog srca u minuti tijekom sna (prikupljenih od 1. 5. 2020. do 23. 7. 2020. pomoću uređaja FitBit):*

51	49	52	53	52	51	55	52	52	57	50	54
50	54	51	50	52	52	48	50	49	49	49	49
50	50	51	49	50	50	48	47	49	48	50	48
49	51	52	50	53	53	50	52	50	50	49	52
48	48	52	51	50	48	56	48	51	49	48	50
50	50	49	49	50	48	51	49	48	48	54	50
50	48	48	50	49	49	49	51	51	48	50	50
51	50	48	48	48	49	51	49	48	51	49	49
48	50	51	52								

Slika 2.3: Niz podataka

Ovako poslagani brojevi slabo nam govore o otkucajima srca, stoga ćemo ih pretvoriti u pregledniju formu. Odredit ćemo frekvencije pojedinih brojeva, a zatim ćemo nacrtati histogram koristeći se Excel programom.



Slika 2.4: Histogram

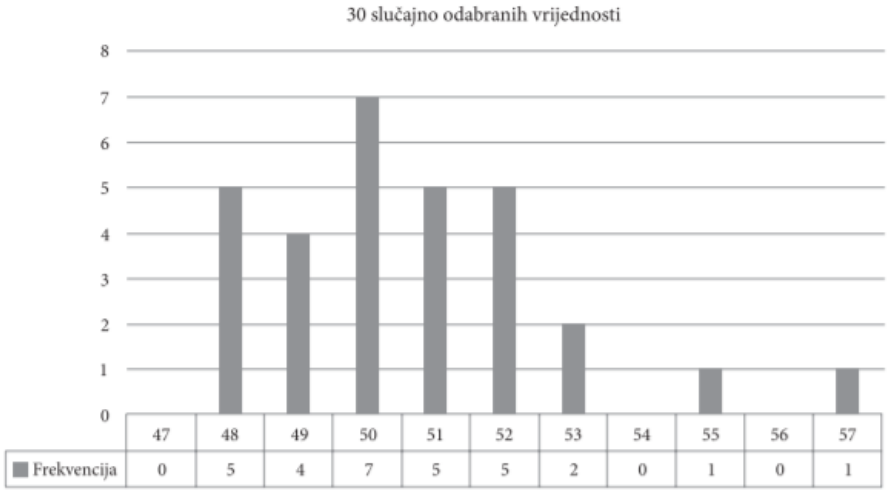
Iz histograma i tablice frekvencija sada preglednije vidimo neke informacije. Bez dubljeg ulaženja u analizu sa slike možemo odmah reći da tijekom sna imam prosječno od 47 do 57 otkucaja srca u minuti, a možemo očitati i određene odnose. Primjerice:

- u prosjeku rijetko imam više od 54 otkucaja i manje od 48 po minuti;
- prosjek broja otkucaja srca se u 90% slučajeva nalazi u intervalu $[48, 52]$;

– u najviše slučajeva imala sam prosjek od 50 otkucaja srca u minuti.

Vrlo često neće nam biti dostupni svi podatci već ćemo imati samo uzorak koji će nam isto tako često biti dovoljan kako bi došli do određenih zaključaka o svim podatcima. Promatrajući podatke možemo vidjeti da će, ako uzmemo manji uzorak, ponašanje podataka biti slično kao da smo ih uzeli sve.

Pokažimo to ponovno na podacima o prosječnom broju otkucaja srca. Slučajno odaberimo 30 vrijednosti. Jedna od mogućih realizacija ovakvog odabira podataka dana je histogramom i tablicom frekvencija prikazanom na slici 2.5.



Slika 2.5: Histogram uzorka

Uočimo sličnosti s originalnim podatcima i prikazom podataka na slici 2.4

– brojka 50 pojavljuje se najviše puta;

– 86.66% podataka se i dalje nalazi u intervalu $[48, 52]$;

– vrijednosti iznad 54 i ispod 48 su rijetke.

Uočimo kako smo slične zaključke imali kod svih podataka.

2.3 Numeričke karakteristike

2.3.1 Mjere centralne tendencije

Za numeričke varijable možemo definirati i numeričke karakteristike koje imaju logičnu interpretaciju, a mogu se iskoristiti s ciljem prikazivanja skupa podataka. Definirajmo neke

od najčešće korištenih numeričkih karakteristika skupa podataka.

Definicija 2.3.1. ***Aritmetička sredina*** *niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X je*

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Definicija 2.3.2. ***Medijan*** *je broj koji se nalazi u sredini sortirane liste podataka (ili je aritmetička sredina srednjih dvaju podataka). Malo preciznije, to je broj sa svojstvom da 50% svih podataka ima vrijednost bar koliko on iznosi. Vertikala povučena u medijanu dijeli histogram na dva dijela jednake površine.*

Definicija 2.3.3. ***Mod*** *je iznos koji se najčešće pojavljuje, tj. to je vrijednost s najvećom frekvencijom. On ne mora postojati, a ne mora biti ni jedinstveno određen. Može se opisati i kao najtipičnija vrijednost mjerene varijable.*

Primjer 2.3.4. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Izračunajte aritmetičku sredinu, medijan te mod ovog skupa vrijednosti.

Aritmetička sredina iznosi 3.25, medijan je 2.5, a mod je 2.

2.3.2 Mjere raspršenja podataka

Osim podatka koji opisuje prosječni ili tipični podatak, bitan statistički podatak je i unutar kojeg intervala se nalazi većina podataka. Najjednostavnija mjera raspršenja je raspon ili rang, a bitni su nam i prvi (donji) kvartil te treći (gornji) kvartil. Koristit ćemo i varijancu te standardnu devijaciju, koje karakteriziraju raspršenost podataka oko aritmetičke sredine.

Definicija 2.3.5. ***Raspon*** *podataka $x_1, x_2, \ldots, x_n$ poredanih po veličini je razlika najvećeg i najmanjeg podatka.*

Primjerice, raspon podataka $1, 1, 2, 2, 3, 11, 64$ je $64 - 1 = 63$.

Definicija 2.3.6. ***Prvi (donji) kvartil*** *je broj od kojega je 25% podataka manje ili je njemu jednako.* ***Treći (gornji) kvartil*** *je broj od kojega je 75% podataka manje ili je njemu jednako.*

Slično kvartilima definiraju se i druge podjele, npr. na percentile, kojima se histogram dijeli na 100 dijelova, svaki od kojih ima odprilike 1% površine. Mogu se razmatrati i dijelovi s drugim postocima površina, a općenito govorimo o kvantilima.

Definicija 2.3.7. ***Varijanca*** *niza izmjerenih vrijednosti $x_1, x_2, x_3, ..., x_n$ varijable X definirana je izrazom:*

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

Definicija 2.3.8. ***Standardna devijacija*** *jest kvadratni korijen varijance, tj.:*

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

Primjer 2.3.9. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

Odredite varijancu te standardnu devijaciju skupa podataka.

Prvo, odredimo aritmetičku sredinu - ona približno iznosi 5.42. Zatim, odredimo varijancu te standardnu devijaciju:

$$\overline{s}_n^2 = \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87$$

$$\overline{s}_n = \sqrt{\overline{s}_n^2} = \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81.$$

2.3.3 Čebiševljev teorem

Teorem 2.3.10 (Čebiševljev teorem [7])**.** *Neka je x aritmetička sredina i $\overline{s}_n$ standardna devijacija uzorka $x_1, x_2, \ldots, x_n$. Tada u intervalu $\langle \overline{x} - 2 \cdot \overline{s}_n, \overline{x} + 2 \cdot \overline{s}_n \rangle$ ima barem 75% podataka, a u intervalu $\langle \overline{x} - 3 \cdot \overline{s}_n, \overline{x} + 3 \cdot \overline{s}_n \rangle$ ima barem 88% podataka.*

Primjer 2.3.11. *Kontrolom slučajno odabranih 20 staklenki s kemikalijom, punjenih od jednog proizvođača, dobiveni su sljedeći podatci (u litrama):*

$$1.97, 1.95, 2.02, 1.99, 1.95, 2.03, 2.00, 1.96, 1.98, 2.00$$

$$2.01, 1.99, 1.98, 1.97, 1.97, 1.94, 1.94, 2.04, 2.02, 1.93.$$

$\overline{x} = 1.982$
$\overline{s}_n = 0.031$ (zaokružena na tri decimale), pa je stoga $2\overline{s}_n = 0.062$.
Zato je $\overline{x} + 2 \cdot \overline{s}_n = 2.044$ i $\overline{x} - 2 \cdot \overline{s}_n = 1.920$. Vidimo da su svi zadani podatci između 1.920 i 2.044 (a teorem garantira bar 75%).

Poglavlje 3

Osnovni pojmovi teorije vjerojatnosti

3.1 Kombinatorika

Definicija 3.1.1. *Neka je S skup od n elemenata.* ***Permutacija*** *je uređena n-torka međusobno različitih elemenata skupa S. Permutacija skupa S je bijekcija tog skupa na samoga sebe.*

Broj permutacija od n različitih elemenata:

$$P_n^n = n! = n \cdot (n-1) \cdot \ldots \cdot 1.$$

Primjer 3.1.2. *Na koliko različitih načina možemo poredati tri različita slova, primjerice A, B i C?*

$$ABC, ACB, BAC$$
$$BCA, CAB, CBA$$

Vidimo, ukupno je $3! = 3 \cdot 2 \cdot 1 = 6$ takvih načina.

Definicija 3.1.3. ***Permutacija duljine r ($r \leqslant n$)*** *ili r-permutacija skupa S je uređena r-torka čije su komponente različiti elementi skupa S.*

Broj permutacija duljine r koje se mogu složiti od n različitih elemenata:

$$P_r^n = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdot \ldots \cdot (n-r+1).$$

Primjer 3.1.4. *Na koliko načina možemo poredati* 2 *od ukupno* 3 *različita slova, primjerice A, B i C?*

$$AB, AC, BA$$

$$BC, CA, CB$$

Vidimo, ukupno je $\frac{3!}{(3-2)!} = \frac{6}{1} = 6$ takvih načina.

Definicija 3.1.5. *Permutacija od n elemenata među kojima ima n_1 elemenata prve vrste, n_2 elemenata druge vrste, ..., n_r elemenata r-te vrste i vrijedi $\sum_{i=1}^{r} n_i = n$ je uređena n-torka od čega je n_i broj elemenata i-te vrste; $i = 1, 2, \ldots, r$. Zovemo je* ***permutacija s ponavljanjem****.*

Broj permutacija od n elemenata od kojih je
n_1 jedne vrste,
n_2 druge vrste,
n_3 treće vrste,
n_r r-te vrste,
$n_1 + n_2 + n_3 + \ldots + n_r = n$,
jednak je:

$$P^{n}_{n_1,n_2,\ldots,n_r} = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_r!}.$$

Primjer 3.1.6. *Na koliko načina 4 knjige, od kojih su dvije plavih korica, jedna zelenih korica, a jedna crvenih korica, možemo složiti na policu?*

Označimo knjige s plavim koricama slovom P, knjigu zelenih korica sa Z te knjigu crvenih korica sa C.

$$PPZC, PPCZ, PZCP, PCZP, ZCPP, CZPP$$

$$CPPZ, ZPPC, CPZP, ZPCP, PZPC, PCPZ$$

Vidimo, ukupno je $\frac{4!}{2!\cdot 1!\cdot 1!} = \frac{4\cdot 3\cdot 2\cdot 1}{2\cdot 1} = 12$ takvih načina.

Definicija 3.1.7. ***r-kombinacija od*** *S je neuređena r-torka ($r \leqslant n$) različitih elemenata iz S.*

Broj r-članih kombinacija skupa od n različitih elemenata je

$$C_n^r = \frac{n!}{r! \cdot (n-r)!} = \binom{n}{r}.$$

Napomena: $\binom{n}{0} = \binom{n}{n} = 1$.

Primjer 3.1.8. *Na koliko se načina, u skupu od* 5 *različitih čokolada, može formirati uzorak veličine* 3 *različite čokolade?*

Označimo čokolade slovima A, B, C, D i E.

$$ABC, ABD, ABE, BCD, BCE$$

$$CDE, ACD, ACE, ADE, DBE$$

Vidimo, ukupno je $\frac{5!}{3!\cdot(5-3)!} = \frac{120}{12} = \binom{5}{3} = 10$ takvih načina.

3.2 Prostor elementarnih događaja

Teorija vjerojatnosti je matematička disciplina čiji je zadatak formirati i proučavati matematički model nekog danog slučajnog pokusa. U 17. stoljeću, Blaise Pascal te Pierre de Fermat bili su prvi koji su se bavili vjerojatnosnim problemima, a inspiraciju su pronašli u raznim kockarskim igrama. Svaki pokus (eksperiment) definiran je odnosom uzorka i posljedica, a pretpostavke za realizaciju pokusa su ponavljanje pokusa proizvoljno konačno mnogo puta te poznavanje mogućih ishoda. Ishodi pokusa jedini su objekti koji nam služe za izgradnju matematičkog modela.

Dvije su vrste pokusa:

– deterministički pokus – ishod je jednoznačno određen uvjetima pokusa

– slučajni pokus – ishod nije jednoznačno određen uvjetima pokusa

Osnovna pretpostavka slučajnog pokusa je da svako izvođenje pokusa mora dati ishod tj. događaj koji odgovara jednom i samo jednom elementarnom događaju. Slučajni pokus je definiran svojim osnovnim ishodima koji se međusobno isključuju i zovu se elementarni događaji, a označavaju se malim grčkim slovima ω_1, ω_2, ω_3...

Definicija 3.2.1. ***Slučajni pokus*** *je definiran svojim osnovnim ishodima koji se međusobno isključuju i zovu se elementarni događaji. Označavaju se malim grčkim slovima* $\omega_1, \omega_2, \ldots$. *Skup* $\Omega = \{\omega_i : \omega_i = \text{elementarni događaji}, i = 1, 2, \ldots, n, \ldots\}$ *je neprazan skup i zove se* ***prostor elementarnih događaja****.*

Primjer 3.2.2. *Bacimo simetrični novčić jednom te napišimo prostor elementarnih događaja.*

$\Omega = \{\omega_1, \omega_2\}$ gdje je ω_1 = palo je pismo, a ω_2 = pala je glava.

Definicija 3.2.3. ***Slučajni događaj*** *je podskup prostora elementarnih događaja. Slučajni događaji označavaju se velikim tiskanim slovima latinice A, B,* $\ldots \subseteq \Omega$.

Definicija 3.2.4. *Cijeli prostor elementarnih događaja* Ω *je* ***siguran događaj*** *koji se mora dogoditi u svakom izvođenju pokusa.*

Definicija 3.2.5. *Prazan skup* $\emptyset$ *je* ***nemoguć događaj*** *koji se nikada neće dogoditi.*

Definicija 3.2.6. *Elementarni događaj koji pripada događaju A zove se* ***povoljan događaj*** *za A. Pojavljivanje tog elementarnog događaja u pokusu povlači da se dogodio događaj A.*

Primjer 3.2.7. *Neka je dan slučajan pokus bacanja igraće kocke. Odredite prostor elementarnih događaja te neka A označava da je pao paran broj. Odredite elementarne događaje povoljne za A.*

Prvo, označimo elementarne događaje.

$$\begin{gathered}\omega_1 = \text{pala je jedinica} \\ \omega_2 = \text{pala je dvojka} \\ \vdots \\ \omega_6 = \text{pala je šestica}\end{gathered}$$

Zatim, odredimo prostor svih elementarnih događaja.

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\} = \{1, 2, 3, 4, 5, 6\}$$

A = pao je paran broj
A =$\{\ \omega_2, \omega_4, \omega_6\ \} = \{2, 4, 6\}$.

Vidimo, elementarni događaji $\omega_2, \omega_4, \omega_6$ povoljni su za događaj A.

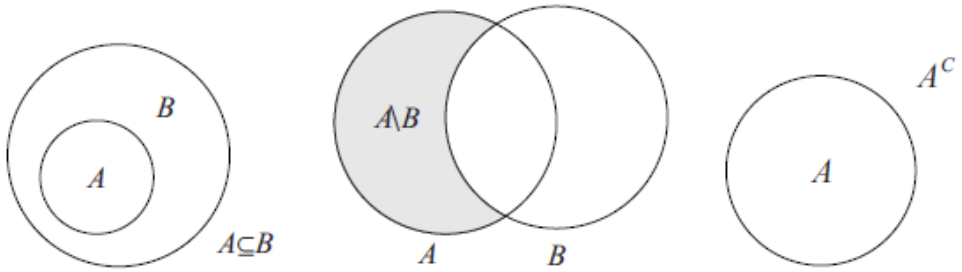
3.3 Operacije sa skupovima

Slučajni događaji su podskupovi od Ω. Operacije s događajima definiramo koristeći se operacijama sa skupovima.

Definicija 3.3.1. ***Podskup događaja*** *$A \subseteq B$: (dogodi se $A \Rightarrow$ dogodi se B)*

Definicija 3.3.2. ***Razlika događaja*** *$A \setminus B$: (događaj $A \setminus B$ se dogodi ako se dogodi A i ne dogodi se B)*

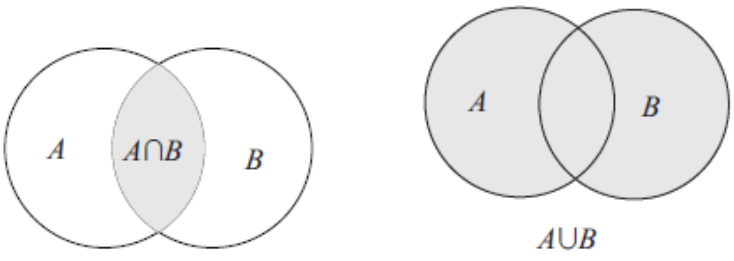
Definicija 3.3.3. ***Suprotan događaj*** *$A^c = \Omega \setminus A$: (A^c se dogodi $\iff$ A se ne dogodi)*



Slika 3.1: Vizualni prikaz podkupa, razlike te komplementa dvaju skupova

Definicija 3.3.4. ***Presjek događaja*** *$A \cap B$: (događaj $A \cap B$ se dogodi $\iff$ dogode se i A i B)*

Definicija 3.3.5. ***Unija događaja*** *$A \cup B$: (događaj $A \cup B$ se dogodi $\iff$ dogodi se ili A ili B)*



Slika 3.2: Vizualni prikaz presjeka te unije dvaju skupova

Također, vrijede i de Morganova pravila:

$$\left(\bigcup_k A_k\right)^c = \bigcap_k A_k^c, \quad \left(\bigcap_k A_k\right)^c = \bigcup_k A_k^c.$$

Definicija 3.3.6. *Za događaje A i B kažemo da se međusobno* ***isključuju*** *ako je njihov presjek jednak* $\emptyset$.

Definicija 3.3.7. *Skupovi* $A_1, A_2, \ldots, A_n$ *čine* ***potpun sistem događaja*** *ako se svi međusobno isključuju i ako im je unija cijeli prostor elementarnih događaja:*

$$A_i \cap A_j = \emptyset, \forall i \neq j, i, j = 1, \ldots, n; \bigcup_{i=1}^{n} A_i = \Omega.$$

3.4 Definicije vjerojatnosti

3.4.1 Klasična definicija vjerojatnosti

Definicija 3.4.1. *Neka je prostor elementarnih događaja konačan skup* $|\Omega| = n$*, te neka je* $\Omega = \{\omega_1, \omega_2, \ldots \omega_n\}$*. Neka su svi elementarni događaji jednako mogući. Neka događaj A ima m povoljnih elementarnih događaja,* $A \subseteq \Omega, |A| = m$*. Vjerojatnost svakog elementarnog događaja je* $P(\omega_i) = \frac{1}{|\Omega|}$*, a vjerojatnost događaja A definira se kao broj:* $P(A) = \frac{|A|}{|\Omega|}$.

Vrijede i sljedeća svojstva:

(1) $P(\Omega) = 1$,

(2) $P(A^c) = 1 - P(A)$,

(3) $P(\emptyset) = 0$,

(4) $0 \leq P(A) \leq 1$,

(5) $A \subseteq B \Rightarrow P(A) \leq P(B)$,

(6) $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$,

(7) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Definicija 3.4.2. *Neka je* Ω *prostor elementarnih događaja.* ***Partitivni skup*** *ili skup svih podskupova od* Ω *zovemo skup svih mogućih događaja slučajnog pokusa. Podskup* $\mathcal{A} \subseteq P(\Omega)$ *zovemo* ***familija događaja*** *iz* Ω.

Definicija 3.4.3. *Neka familija događaja $\mathcal{F} \subseteq P(\Omega)$ ima svojstva:*

(i) $\emptyset \in \mathcal{F}$,

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,

(iii) Ako je $A_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

*Takvu familiju skupova $\mathcal{F}$ zovemo **sigma algebra** događaja (σ-algebra). Ako je Ω konačan skup, onda je i svaka σ-algebra $A \subseteq P(\Omega)$ konačna i naziva se algebra događaja.*

3.4.2 Aksiomatska definicija vjerojatnosti

Definicija 3.4.4. *Neka je Ω prostor elementarnih događaja slučajnog pokusa i neka je $\mathcal{F}$ σ-algebra skupova na Ω. Funkcija $P : \mathcal{F} \to \mathbb{R}$ zove se **vjerojatnost** na $\mathcal{F}$ ako vrijedi:*

(P1) $P(A) \geq 0, A \subseteq \mathcal{F}$ *(svojstvo nenegativnosti);*

(P2) $P(\Omega) = 1$ *(svojstvo normiranosti);*

(P3) $A_i \in \mathcal{F}, i \in \mathbb{N}, A_i \cap A_j = \emptyset, i \neq j \Rightarrow P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ *(svojstvo prebrojive aditivnosti).*

Definicija 3.4.5. ***Vjerojatnosnim prostorom** zovemo uređenu trojku $(\Omega, \mathcal{F}, P)$, gdje je $\mathcal{F}$ σ-algebra na Ω, a P vjerojatnost na Ω.*

Definicija 3.4.6. *Ako je Ω prebrojiv ili konačan skup elementarnih događaja, onda $(\Omega, \mathcal{F}, P)$ zovemo **diskretni vjerojatnosni prostor**.*

Definicija 3.4.7. *Ako je Ω konačan skup elementarnih događaja, $|\Omega| = n$, onda $(\Omega, \mathcal{F}, P)$ zovemo **n-dimenzionalni diskretni vjerojatnosni prostor**.*

Teorem 3.4.8. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor. Tada za funkciju vjerojatnosti P vrijedi:*

(a) $P(\emptyset) = 0$;

(b) $A_i \in \mathcal{F}, i \in \{1, \ldots, n\}, A_i \cap A_j = \emptyset, i \neq j \Rightarrow P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ *(svojstvo konačne aditivnosti);*

(c) $A, B \in \mathcal{F}, A \subseteq B \Rightarrow P(A) \leq P(B)$ *(svojstvo monotonosti;*

(d) $A \in \mathcal{F} \Rightarrow P(A^c) = 1 - P(A)$;

(e) $A, B \in \mathcal{F} \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Dokaz. (a) Neka je $A_1 = \Omega$, $A_i = \emptyset, i \geq 2$. Tada prema uvjetima ($P2$) i ($P3$) vrijedi

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \Rightarrow P(\Omega) = P(\Omega) + \sum_{i=2}^{\infty} P(\emptyset) \Rightarrow 1 = 1 + \sum_{i=2}^{\infty} P(\emptyset) \Rightarrow P(\emptyset) = 0.$$

(b) Neka je $A_i = \emptyset, i > n$. Prema svojstvu (a) je $P(A_i) = 0$. Koristeći definiciju vjerojatnosti ($P2$) dobivamo

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \Rightarrow P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

(c) $A \subseteq B \Rightarrow B = A \cup (B \setminus A)$. Prema svojstvu ($b$) vrijedi

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

Prema uvjetu ($P1$) $P(B \setminus A) \geq 0 \Rightarrow P(B) \geq P(A)$.

(d) $\Omega = A \cup A^c \Rightarrow 1 = P(A \cup A^c) = P(A) + P(A^c)$.

(e) Uočimo sljedeće relacije: $A \cup B = A \cup (B \setminus A)$, $B = (A \cap B) \cup (B \setminus A)$. Prema svojstvu ($b$) računamo:
$P(A \cup B) = P(A \cup (B \setminus A)) = P(A) + P(B \setminus A)$ i
$P(B) = P((A \cap B) \cup (B \setminus A)) = P(A \cap B) + P(B \setminus A)$.
Zaključujemo da je $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

□

Poglavlje 4

Nezavisnost i uvjetna vjerojatnost

Thomas Bayes (1702. - 1762.) uvodi pojam uvjetne vjerojatnosti: vjerojatnost da se dogodi događaj B ako se dogodio događaj A jednaka je kvocijentu vjerojatnosti da se dogode događaji i A i B i vjerojatnosti događaja A.

Definicija 4.0.1. *Neka je* $(\Omega, \mathcal{F}, P)$ *vjerojatnosni prostor i* $A \in \mathcal{F}$ *takav da je* $P(A) > 0$. *Funkcija* $P_A : \mathcal{F} \to [0,1]$ *definirana s*

$$P_A(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad B \in \mathcal{F},$$

je vjerojatnost na $\mathcal{F}$ *i zovemo je* ***uvjetna vjerojatnost uz uvjet A****. Broj* $P(B|A)$ *zovemo vjerojatnost od B uz uvjet da se A dogodio.*

Primjer 4.0.2. *Bačene su dvije kockice. Kolika je vjerojatnost da se pojavio broj 5, ako znamo da je zbroj znamenaka jednak 9?*

S A i B označimo događaje:

$$A = \{\text{zbroj znamenaka jednak je 9}\}$$

$$B = \{\text{pojavio se broj 5}\}.$$

Očito je tražena vjerojatnost $P(B|A)$

$$A = \{(3,6), (4,5), (5,4), (6,3)\}.$$

Vidimo da je vjerojatnost događaja A jednaka :

$$P(A) = \frac{4}{36},$$

i vjerojatnost događaja $A \cap B$:

$$P(A \cap B) = \frac{2}{36},$$

jer su nam povoljni događaji samo {(5,4), (4,5)}.

Sada možemo izračunati vjerojatnost događaja $P(B|A)$ uvrštavajući poznate podatke:

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

$$P(B|A) = \frac{\frac{2}{36}}{\frac{4}{36}},$$

$$P(B|A) = \frac{1}{2}.$$

Dakle, vjerojatnost da na kockici padne broj 5, ako znamo da je zbroj na kockicama jednaka 9, jednaka je $\frac{1}{2}$.

Primjer 4.0.3. *U kutiji se nalazi 8 bijelih i 4 crne kuglice. Kuglice izvlačimo jednu po jednu bez vraćanja. Kolika je vjerojatnost da će prve dvije kuglice koje izvučemo biti bijele?*

S A i B označimo događaje:

$$A = \{\text{prva kuglica je bijela}\}$$

$$B = \{\text{druga kuglica je bijela}\}$$

Tada je $A \cap B$ događaj čiju vjerojatnost tražimo. Ukupno imamo 12 kuglica, od kojih je 8 bijelih pa lako zaključujemo kolika je vjerojatnost događaja A:

$$P(A) = \frac{8}{12}$$

Ako znamo da je prva izvučena kuglica bijela, tada nam u kutiji preostaje 7 bijelih kuglica od ukupno 11 kuglica. Pa je vjerojatnost događaja B uz uvjet A:

$$P(B|A) = \frac{7}{11}$$

Sada možemo uvrstiti u formulu podatke koje imamo:

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

$$P(A \cap B) = P(B|A) \cdot P(A),$$
$$P(A \cap B) = \frac{7}{11} \cdot \frac{8}{12},$$
$$P(A \cap B) = \frac{14}{33}.$$

Dakle, vjerojatnost da će prve dvije kuglice koje izvučemo biti bijele jednaka je $\frac{14}{33}$.

Definicija 4.0.4. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i neka su $A, B \in \mathcal{F}$. Za događaje A i B kažemo da su nezavisni ako vrijedi*

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

Primjer 4.0.5. *U kutiji se nalaze 4 plave, 5 bijelih i 6 crnih kuglica. Na sreću odabiremo 3 kuglice. Izvlačimo jednu po jednu kuglicu bez vraćanja u kutiju. Označimo događaje:*

A={*sve su tri izvučene kuglice različitih boja*}

B={*prva izvučena kuglica je bijela*}.

Izračunaj $P(A)$ i $P(B)$. Jesu li događaji A i B nezavisni?

Prvo, izračunajmo vjerojatnost događaja A. Pogledajmo primjerice vjerojatnost da smo izabrali redom plavu, pa bijelu, pa crnu kuglicu. Kako sve skupa ima 15 kuglica, vjerojatnost da su izvučene plava - bijela - crna je

$$\frac{4}{15} \cdot \frac{5}{14} \cdot \frac{6}{13}.$$

Postoji 3! permutacija te tri boje, ali uočimo da ćemo za svaku permutaciju dobiti ovu gornju vjerojatnost (samo se permutiraju brojnici), pa je

$$P(A) = 3! \cdot \frac{4 \cdot 5 \cdot 6}{15 \cdot 14 \cdot 13}.$$

Uočimo da bi dobili jednaku vjerojatnost da nismo pazili na redoslijed,

$$P(A) = \frac{\binom{4}{1} \cdot \binom{5}{1} \cdot \binom{6}{1}}{\binom{15}{3}} = \frac{24}{91}$$

(biramo po jednu kuglicu svake boje, a sve skupa 3 kuglice od 15 možemo odabrati na $\binom{15}{3}$ načina).

Vjerojatnost da je prva izvučena kuglica bijele boje lako izračunamo. Od 15 kuglica, 5 ih je bijele boje pa je

$$P(B) = \frac{5}{15} = \frac{1}{3}.$$

Kako bismo provjerili jesu li događaji A i B nezavisni, koristit ćemo nužan i dovoljan uvjet nezavisnosti, odnosno ako su događaji A i B nezavisni mora vrijediti $P(A \cap B) = P(A) \cdot P(B)$. Preostaje nam izračunati $P(A \cap B)$. Vjerojatnost $P(A \cap B)$ jest vjerojatnost da su sve tri izvučene kuglice različitih boja, ali da je pritom prva izvučena kuglica bijela.
Moguća su 2 slučaja: BCP ili BPC, odnosno da je prvo izvučena bijela pa crna pa plava ili bijela pa plava pa crna. U prvom slučaju vjerojatnost za bijelu kuglicu je $\frac{5}{15}$, za crnu kuglicu $\frac{6}{14}$ jer je nakon jedne izvučene ostalo 14 kuglica i vjerojatnost za plavu kuglicu iznosi $\frac{4}{13}$ jer nakon izvučene dvije kuglice u kutiji preostaje još 13 kuglica. Po principu produkta, vjerojatnost za prvi slučaj iznosi

$$\frac{5}{15} \cdot \frac{6}{14} \cdot \frac{4}{13} = \frac{4}{91}.$$

Analogno, vjerojatnost za drugi slučaj iznosi

$$\frac{5}{15} \cdot \frac{4}{14} \cdot \frac{6}{13} = \frac{4}{91}.$$

Budući da su ova 2 slučaja disjunktna, da bismo izračunali vjerojatnost $P(A \cap B)$ potrebno je zbrojiti vjerojatnosti prvog i drugog slučaja. Stoga vrijedi:

$$P(A \cap B) = \frac{4}{91} + \frac{4}{91} = \frac{8}{91} = \frac{24}{91} \cdot \frac{1}{3} = P(A) \cdot P(B).$$

Zaključujemo, događaji A i B su nezavisni.

Prije sljedeće definicije, promotrimo primjer.

Primjer 4.0.6. *Voćarnica se opskrbljuje jabukama iz triju voćnjaka, i to 35% potrebne količine iz prvog, 45% potrebne količine iz drugog i 20% potrebne količine iz trećeg voćnjaka. 10% jabuka prvog voćnjaka prve su kvalitete, dok to vrijedi za 15% jabuka drugog voćnjaka i 20% jabuka trećeg voćnjaka. Kolika je vjerojatnost da na sreću odabrana jabuka bude prve kvalitete?*

Ponekad kada računamo vjerojatnost potrebno je sve moguće ishode podijeliti u različite klase. U ovom primjeru, odaberimo na sreću jednu jabuku u voćarnici. Tri su mogućnosti:

$$H_1 = \{\text{odabrana je jabuka iz prvog voćnjaka}\};$$

$$H_2 = \{\text{odabrana je jabuka iz drugog voćnjaka}\}.$$

$$H_3 = \{\text{odabrana je jabuka iz trećeg voćnjaka}\}.$$

Vjerojatnost da se ostvari neki od ovih događaja su:

$$P(H_1) = 0.35\ ,\ P(H_2) = 0.45\ i\ P(H_3) = 0.2.$$

Sa A označimo traženi događaj:

$$A = \{\text{odabrana je jabuka prve kvalitete}\}$$

Događaj A "razbili" smo na tri disjunktna događaja:

$$A \cap H_1 = \{\text{odabrana jabuka prve kvalitete je iz prvog voćnjaka}\}$$

$$A \cap H_2 = \{\text{odabrana jabuka prve kvalitete je iz drugog voćnjaka}\}.$$

$$A \cap H_3 = \{\text{odabrana jabuka prve kvalitete je iz trećeg voćnjaka}\}.$$

Zato je vjerojatnost događaja A zbroj vjerojatnosti događaja $A \cap H_1$, $A \cap H_2$ i $A \cap H_3$:

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + P(A \cap H_3).$$

Vjerojatnosti presjeka računamo na poznati način:

$$P(A \cap H_1) = P(H_1) \cdot P(A|H_1)$$

$$P(A \cap H_2) = P(H_2) \cdot P(A|H_2)$$

$$P(A \cap H_3) = P(H_3) \cdot P(A|H_3).$$

Iz danih podataka u zadatku znamo da je vjerojatnost da je jabuka prve kvalitete, ako je poznato da potječe iz prvog voćnjaka jednaka:

$$P(A|H_1) = 0.1 \Rightarrow P(A \cap H_1) = 0.35 \cdot 0.1 = 0.035.$$

Analogno tome imamo:

$$P(A|H_2) = 0.15 \Rightarrow P(A \cap H_2) = 0.45 \cdot 0.15 = 0.0675$$

i

$$P(A|H_3) = 0.2 \Rightarrow P(A \cap H_3) = 0.2 \cdot 0.2 = 0.04$$

Sada možemo izračunati traženu vjerojatnost događaja A:

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + P(A \cap H_3) = 0.035 + 0.0675 + 0.04 = 0.1425.$$

Ovaj slučaj možemo i generalizirati. Pretpostavimo da skup elementarnih događaja možemo rastaviti na n disjunktnih događaja:

$$\Omega = H_1 \cup H_2 \cup ... \cup H_n$$

pri čemu su događaji H_i, H_j disjunktni za $i \neq j$ i vrijedi $P(H_i) > 0$ za svaki i. Ovakav rastav nazivamo **particija vjerojatnostnog prostora**. Kažemo još da familija H_1,...,H_n čini **potpun sustav događaja**.
Sada promotrimo neki događaj $A \subset \Omega$. Familijom $H_1, H_2, ..., H_n$ on je podijeljen na međusobno disjunktne događaje:

$$A = (A \cap H_1) \cup (A \cap H_2) \cup ... \cup (A \cap H_n).$$

Znamo da su događaji $A \cap H_i$ međusobno disjunktni, pa vrijedi:

$$P(A) = P(A \cap H_1) + P(A \cap H_2) + ... + P(A \cap H_n),$$

$$P(A) = P(H_1)P(A|H_1) + ... + P(H_n)P(A|H_n).$$

Teorem 4.0.7. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnostni prostor i neka skupovi $H_1, H_2, \ldots, H_n \in \mathcal{F}$ čine potpun sistem događaja. Tada*

$$\forall A \in \mathcal{F} \Rightarrow P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i).$$

Teorem 4.0.8. *Vjerojatnost produkta (presjeka) dvaju događaja $P(A \cap B)$ jednaka je produktu vjerojatnosti jednog od njih i uvjetne vjerojatnosti drugog, pod uvjetom da se prvi dogodio.*

$$P(A \cap B) = P(A) \cdot P_A(B), \quad P(A \cap B) = P(B) \cdot P_B(A).$$

Poglavlje 5

Bayesov teorem

Iz poznatih relacija

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

možemo napisati

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

Ovu formulu najčešće koristimo onda kada je događaj B jedna od hipoteza $H_1, H_2, ..., H_n$ na koje je razbijen skup Ω.

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

Pritom se vjerojatnost $P(A)$ računa uglavnom pomoću formule potpune vjerojatnosti. Tako dobivamo **Bayesov teorem**. Dokazao ga je engleski statističar i filozof Thomas Bayes (1702. - 1762.) po kojemu je i dobio ime. Zapišimo i dokažimo teorem.

Teorem 5.0.1. *Neka je $(\Omega, \mathcal{F}, P)$ vjerojatnosni prostor i neka skupovi $H_1, H_2, \ldots, H_n \in \mathcal{F}$ čine potpun sistem događaja. Neka događaj $A \in \mathcal{F}$ ima pozitivnu vrijednosti $P(A) > 0$. Tada je $\forall i$*

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}.$$

Dokaz. Definicija uvjetne vjerojatnosti i formula produkta vjerojatnosti povlači da vrijedi:

$$P(H_i|A) = \frac{P(H_i \cap A)}{P(A)} = \frac{P(H_i) \cdot P(A|H_i)}{P(A)}$$

Zatim, prema formuli potpune vjerojatnosti dobivamo:

$$P(H_i|A) = \frac{P(H_i) \cdot P(A|H_i)}{\sum_{j=1}^{n} P(H_j) \cdot P(A|H_j)}$$

□

Primjer 5.0.2. *U ponoć su na parkiralištu bila dva bijela i jedan crveni Harley, tri bijela i četiri crvena Kawasaki-ja te tri bijele i jedna crvena Yamaha. Te noći kradljivac motocikala nasumce je odabrao motocikl i ukrao ga. Ako je ukradeni motocikl bijele boje, kolika je vjerojatnost da je to bio Kawasaki?*

Označimo tri događaja koja čine potpun sustav događaja,

$$H_1 = \{\text{motocikl je marke Harley}\},$$

$$H_2 = \{\text{motocikl je marke Kawasaki}\},$$

$$H_3 = \{\text{motocikl je marke Yamaha}\},$$

te njihove odgovarajuće vjerojatnosti:

$$P(H_1) = \tfrac{3}{14} \rightarrow \text{jer su 3 motocikla marke Harley, od njih 14;}$$

$$P(H_2) = \tfrac{1}{2} \rightarrow \text{jer je 7 motocikala marke Kawasaki, od njih 14;}$$

$$P(H_3) = \tfrac{2}{7} \rightarrow \text{jer je 4 motocikla marke Yamaha, od njih 14.}$$

S A označimo događaj

$$A = \{\text{ukradeni motocikl je bijele boje}\}.$$

Preostaje nam izračunati vjerojatnosti $P(A|H_1)$, $P(A|H_2)$ i $P(A|H_3)$ kako bi mogli iskoristiti Bayesovu formulu i dobiti traženu vjerojatnost događaja $H_2|A$.
Kako imamo 2 bijela motocikla marke Harley, 3 bijela motocikla marke Kawasaki i 3 bijela motocikla marke Yamaha, odgovarajuće vjerojatnosti iznose:

$$P(A|H_1) = \tfrac{2}{3}, P(A|H_2) = \tfrac{3}{7}, P(A|H_3) = \tfrac{3}{4}.$$

Sada možemo poznate podatke uvrstiti u Bayesovu formulu:

$$P(H_2|A) = \frac{P(H_2)P(A|H_2)}{P(H_1)P(A|H_1) + P(H_2)P(A|H_2) + P(H_3)P(A|H_3)}$$

$$P(H_2|A) = \frac{\frac{1}{2} \cdot \frac{3}{7}}{\frac{3}{14} \cdot \frac{2}{3} + \frac{1}{2} \cdot \frac{3}{7} + \frac{2}{7} \cdot \frac{3}{4}}$$

$$P(H_2|A) = 0.375.$$

Primjer 5.0.3. *Kako bi se što bolje pripremila za polaganje mature iz matematike, Ana je preko interneta naručila zbirku riješenih zadatka koji su bili na maturi prethodnih godina, ne gledajući pri tome koji je izdavač. Od prijatelja koji su već polagali maturu, čula je da su takvu zbirku objavile tri izdavačke kuće te da prva izdavačka kuća ima najbolju reklamu na internetu i ona tiska* 80% *zbirki iz kojih se budući studenti pripremaju, druga izdavačka kuća tiska* 15% *zbirki, dok je treća izdavačka kuća tek počela s radom i ona tiska* 5% *takvih zbirki. Prijatelji su također rekli Ani da ponekad naručene zbirke stignu na kućnu adresu nepotpune, odnosno u njima se nalaze samo tekstovi zadataka, a ne i rješenja. Naime, zbirke koje tiska prva izdavačka kuća tiskaju se bez rješenja u* 4% *slučajeva, u drugoj izdavačkoj kući taj je postotak* 6%*, dok posljednja izdavačka kuća tiska* 9% *zbirki koje ne sadrže rješenje.*

(a) *Ana se zabrinula jer nije pogledala izdavača kada je naručivala i pita se kolika je vjerojatnost da je zbirku naručila od prvog izdavača?*

(b) *Kada je zbirka stigla na kućnu adresu Ana je bila u školi, ali je njezin brat odmah otvorio paket i vidio da je zbirka stigla bez rješenja! Odmah je poslao poruku Ani. Sada kada je sigurna da zbirka nije stigla u obliku kom se nadala, Ana se ponovno pita kolika je vjerojatnost da se radi o prvom izdavaču. Razlikuje li se ta vjerojatnost od one dobivene u prethodnom slučaju?*

Možemo uočiti da potpun sustav događaja čine sljedeća tri skupa:

$$H_1 = \{\text{zbirka dolazi od prvog izdavača}\},$$

$$H_2 = \{\text{zbirka dolazi od drugog izdavača}\},$$

$$H_3 = \{\text{zbirka dolazi od trećeg izdavača}\},$$

Također, zanimaju nas i događaji:

$$N = \{\text{zbirka je nepotpuna (stigla je Ani bez rješenja)}\},$$

$$N^c = \{\text{zbirka je potpuna (stigla je Ani s rješenjima)}\},$$

(a) Ana je zbirku naručila slučajnim odabirom (nije gledala tko je izdavač) te je vjerojatnost da je ona izdana od strane prvog izdavača jednaka 0.8, jer prvi izdavač tiska 80% zbirki iz kojih budući studenti vježbaju.

(b) Kada je Ana saznala da je zbirka došla bez rješenja, koristeći tu dodatnu informaciju krenula je opet izračunati vjerojatnost da ona dolazi od prvog izdavača, odnosno vjerojatnost $P(H_1|N)$, za što su joj trebali i sljedeći podaci koje je prvo zapisala iz podataka koji su joj poznati:

$$P(H_1) = 0.80 \rightarrow \text{jer prvi izdavač tiska 80\% zbirki;}$$

$$P(H_2) = 0.15 \rightarrow \text{jer drugi izdavač tiska 15\% zbirki;}$$

$$P(H_3) = 0.05 \rightarrow \text{jer treći izdavač tiska 5\% zbirki.}$$

$$P(N|H_1) = 0.04 \rightarrow \text{jer 4\% zbirki bez rješenja dolazi od prvog izdavača;}$$

$$P(N|H_2) = 0.06 \rightarrow \text{jer 6\% zbirki bez rješenja dolazi od drugog izdavača;}$$

$$P(N|H_3) = 0.09 \rightarrow \text{jer 9\% zbirki bez rješenja dolazi od trećeg izdavača;}$$

Uvrštavanjem navedenih vrijednosti u Bayesovu formulu dobivamo

$$P(H_1|N) = \frac{0.80 \cdot 0.04}{0.80 \cdot 0.04 + 0.15 \cdot 0.06 + 0.05 \cdot 0.09} = 0.703.$$

Dakle, uz dani uvjet da je zbirka došla bez rješenja, vjerojatnost da ona dolazi od prvog izdavača se smanjila jer prvi izdavač tiska najmanji postotak nepotpunih zbirki (onih koje ne sadrže rješenja).

Bayesova formula ima vrlo široku primjenu, pa se tako osim u matematici često koristi i u medicinskoj dijagnostici. Pretpostavimo da se u nekoj bolnici liječe bolesnici od kojih svaki može imati jednu od bolesti $H_1, H_2, \ldots, H_n$ te da je kod slučajno odabranog bole-snika nakon pregleda ustanovljen skup od A simptoma. Uz pretpostavku da se u bolnici bilježe statistički podatci o broju bolesnika koji su imali bolest H_i, $i = 1, \ldots, n$ možemo izračunati pripadne vjerojatnosti $P(H_i)$ i $P(A|H_i)$, $i = 1, \ldots, n$, što su redom relativna frekvencija bolesti H_i i relativna frekvencija utvrđivanja skupa simptoma A među svim bolesnicima koji su bolovali od bolesti H_i (relativna frekvencija je približno jednaka odgovarajućoj vjerojatnosti te k njoj teži s povećanjem uzorka kojeg promatramo). Ovakvi podatci nam omogućavaju primjenu Bayesove formule koja će nam dati odgovor na pitanje kolika je vjerojatnost da osoba ima bolest H_i ako je kod nje ustanovljen skup simptoma A.

Ilustrirajmo primjenu Bayesove formule u medicini sljedećim primjerom.

Primjer 5.0.4. *Ani je teško vježbati zadatke za državnu maturu jer ima simptome alergije. Išla je na razna ispitivanja k liječniku, no svi su testovi bili negativni. Preostala je samo mogućnost da ima vrlo rijedak oblik alergije koji se nasumično pojavljuje kod jedne osobe u populaciji od* 10000 *ljudi. Ako se Ana odluči testirati, uz pretpostavku da test daje točan rezultat u* 99% *slučajeva, i ako rezultat testa bude pozitivan, kolika je zaista vjerojatnost da Ana ima spomenutu rijetku alergiju?*

Ako s A označimo događaj da osoba ima navedeni rijedak oblik alergije, a s B događaj da je rezultat testa pozitivan, možemo zaključiti da je $P(B|A) = 0.99$ i $P(A) = 0.0001$.

Sada primjenom formule $P(B) = \sum_{i=1}^{n} P(H_i) \cdot P(B|H_i)$, uz dani potpun sustav događaja $\{A,A^c\}$ dobivamo

$$P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c) = 0.99 \cdot 0.0001 + 0.01 \cdot 0.9999 = 0.01.$$

Primjenom Bayesove formule imamo da je

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} = 0.0099,$$

odnosno vjerojatnost da Ana ima alergiju, uz uvjet da će test biti pozitivan je manja od 1%. Razlog ovog zaključka je činjenica da je navedeni oblik alergije jako rijedak.

Poglavlje 6

Slučajne varijable

6.1 Diskretne slučajne varijable

Prisjetimo se da su elementarni događaji upravo oni događaji koji mogu biti rezultat nekog razmatranog pokusa. Nakon provođenja pokusa u pravilu nešto prebrojavamo ili mjerimo, a rezultat je neki broj, ovisno o tome koji se događaj pojavio. Matematički se to može predočiti funkcijom koja svakom događaju pridružuje neki broj, a vrijednosti te funkcije ovise o događaju koji se pojavio. Dakle, ta je funkcija definirana na skupu svih događaja u pokusu i slučajno (ovisno o događaju koji se dogodio) postiže određene vrijednosti i upravo zato se naziva slučajnom varijablom.

Neka je Ω skup događaja. Svakom događaju w_k iz skupa događaja na odgovarajući način pridružimo realan broj x_k. Na primjer:

Ω	$\mathbb{R}$
ω_1	x_1
ω_2	x_2
ω_3	x_3
ω_4	x_1
ω_5	x_2

Pogladamo li prethodnu tablicu, vidimo da smo nekim različitim događajima pridružili isti broj, npr. ω_2 i ω_5 smo pridružili x_2.

Ako s X označimo funkciju koja događajima iz Ω pridružuje neke x_k-ove tada se radi o funkciji koja je slučajna varijabla.

Definicija 6.1.1. *Neka je Ω vjerojatnosni prostor. Slučajna varijabla je funkcija*

$$X : \Omega \to \mathbb{R}.$$

Ako popišemo sve različite x_k i zbrojimo pripadne vjerojatnosti koje smo imali za eventualno iste x_k, onda smo dobili diskretnu slučajnu varijablu, preciznije, njezin zakon razdiobe. U literaturi se (diskretna) slučajna varijabla zajedno sa zakonom razdiobe često označava s

$$X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix}$$

Tablicu zovemo distribucija ili zakon razdiobe slučajne varijable.

Definicija 6.1.2. *Funkcija distrubucije slučajne varijable X je funkcija*

$$F_x(x) = P(X \leq x).$$

Vjerojatnost $P(X \leq x)$ je vjerojatnost da slučajna varijabla poprimi vrijednost manju ili jednaku vrijednosti x:

$$P(X \leq x) = P\left(\bigcup_{y \leq x} X^{-1}(y)\right)$$

Primjer 6.1.3. *Napišimo zakon razdiobe slučajne varijable X, gdje je X broj glava palih u 2 bacanja (simetričnog) novčića.*

Označimo događaje, njihove vjerojatnosti i broj palih glava za dva bacanja novčića.

P	Ω	$\mathbb{R}$
$\frac{1}{4}$	PP	0
$\frac{1}{4}$	PG	1
$\frac{1}{4}$	GP	1
$\frac{1}{4}$	GG	2

Sada vidimo da je glava mogla pasti samo $0, 1$ ili 2 puta i to su mogući x_k-ovi za našu slučajnu varijablu. Zakon razdiobe dobivamo zbrajanjem vjerojatnosti za svaki x_k, tj. imamo

X	P_X
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

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

Primjer 6.1.4. *Napišimo zakon razdiobe slučajne varijable Z, gdje je Z zbroj brojeva koji su pali na dvije kocke.*

Zapišimo tablicu mogućih zbrojeva na dvije kocke, ona će nam dati opis varijable Z.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Sada, zapišimo još iz prethodne tablice i zakon razdiobe slučajne varijable Z.

X	$\Pr_X$
2	$\frac{1}{36}$
3	$\frac{2}{36}$
4	$\frac{3}{36}$
5	$\frac{4}{36}$
6	$\frac{5}{36}$
7	$\frac{6}{36}$
8	$\frac{5}{36}$
9	$\frac{4}{36}$
10	$\frac{3}{36}$
11	$\frac{2}{36}$
12	$\frac{1}{36}$

Iz ove tablice se lako čitaju sljedeće vjerojatnosti:

$$\Pr(6 \leq Z \leq 10) = \frac{23}{36}$$

$$\Pr(Z > 3) = \frac{33}{36}$$

$$\Pr(Z < 8) = \frac{21}{36}$$

Očekivanje diskretne slučajne varijable

Definicija 6.1.5. *Očekivanu vrijednost slučajne varijable označavamo s $E(X)$, a definiramo je ovako:*

$$E(X) = \sum_{svim\, x_k} x_k P(x_k).$$

Dakle, očekivanje slučajne varijable je prosjek svih vrijednosti varijable X.

U primjeru 2.1.44. slučajna varijabla X bila je broj glava palih na dva novčića i njena razdioba bila je

X	$\Pr_X$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

Nađimo očekivanje te slučajne varijable.

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

Raspršenje diskretne slučajne varijable

Definicija 6.1.6. *Raspršenje X od $E(X)$ mjerimo varijancom $Var(X)$ ili $V(X)$ koja se definira kao*

$$V(X) = E(X - E(X))^2,$$

pri čemu s desne strane imamo prosječnu vrijednost kvadrata otklona.

Prema tome, za varijancu diskretne slučajne varijable vrijedi

$$V(X) = \sum_{\text{svi}\, x_k} (x_k - E(X))^2 \, \mathrm{P}(x_k).$$

Drugi korijen od varijance nazivamo standardna devijacija i označavamo s

$$\sigma_X = \sqrt{V(X)}.$$

6.1.1 Binomna distribucija

Slučajna varijabla X koja broji uspjehe u n nezavisnih izvođenja istog pokusa sa samo dva moguća ishoda (uspjeh i neuspjeh), pri čemu se u svakom izvođenju uspjeh realizira s vjerojatnošću p (neuspjeh sa q ili $1-p$), ima binomnu distribuciju s parametrima $n \in \mathbf{N}$ i $p \in \langle 0, 1\rangle$ te oznaku $X \sim B(n, p)$.
Slika $X \sim B(n, p)$ distribucije jest $R(X) = \{0, 1, 2, \ldots, n\}$, a vjerojatnosti realizacije elemenata slike jesu:

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, \quad x \in R(X).$$

Očekivanje i varijanca $X \sim B(n, p)$ distribucije jesu:

$$E[X] = n \cdot p,$$

$$Var(X) = n \cdot p \cdot (1-p).$$

Pretpostavimo da neki slučajni pokus, koji može imati samo dva ishoda: uspjeh ili neuspjeh, pri čemu je vjerojatnost uspjeha p, ponavljamo n puta. Pretpostavimo da ishod pojedinog pokusa ne utječe na ishode preostalih pokusa. Neka je X slučajna varijabla koja registrira broj uspjeha u n pokusa. Ako se nijednom kao ishod nije pojavio uspjeh, tada je

$$P(X = 0) = (1-p)^n = q^n.$$

Ako se jedanput pojavio uspjeh, tada je

$$P(X = 1) = p(1-p)^{n-1} + (1-p)p(1-p)^{n-2} + \cdots + (1-p)^{n-1}p = npq^{n-1}.$$

Ako se uspjeh pojavio k puta, vrijedi:

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

Zato je **funkcija vjerojatnosti binomne slučajne varijable**:

$$f(x) = \begin{cases} \binom{n}{k} \cdot p^x \cdot q^{n-x}, & x \in \{0, 1, \ldots, n\} \\ 0, & x \notin \{0, 1, \ldots, n\} \end{cases}$$

Funkcija distribucije računa se zbrajanjem:

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} P(X = k) = \sum_{k \leq x} \binom{n}{k} \cdot p^k \cdot q^{n-k}.$$

Primjer 6.1.7. *Nađite vjerojatnost da se u* 4 *bacanja simetričnog novčića pismo pojavi točno dva puta.*

Očito se radi o binomnoj slučajnoj varijabli kojoj je $n = 4$ i $p = \frac{1}{2}$, a promatramo vrijednost $x = 2$. Prema tome, uvrštavajući n, p, i x u

$$P(X = x) = f(x) = \binom{n}{k} \cdot p^x \cdot q^{n-x}$$

dobivamo

$$P(X = 2) = \binom{4}{2} \cdot \binom{1}{2}^2 \cdot \binom{1}{2}^2 = \frac{12}{32} = \frac{3}{8} = 0.375.$$

Dakle, vjerojatnost da se pismo pojavi točno dva puta u 4 bacanja simetričnog novčića jednaka je $\frac{3}{8}$.

Primjer 6.1.8. *Neka je* $X \sim B(6, 0.7)$*, gdje je* $n = 6$*,* $p = 0.7$*. Odredite vrijednosti funkcije vjerojatnosti i funkcije distribucije.*

Rješenje:
Za $X = 0$ imamo

$$f(0) = \binom{6}{0} \cdot (0.7)^0 \cdot (0.3)^{6-0} = 0.0007,$$
$$F(0) = p(X \leq 0) = \binom{6}{0} \cdot (0.7)^0 \cdot (0.3)^{6-0} = 0.0007.$$

Za $X = 1$ imamo

$$f(1) = \binom{6}{1} \cdot (0.7)^1 \cdot (0.3)^{6-1} = 0.0102,$$
$$F(1) = p(X \leq 1) = \binom{6}{0} \cdot (0.7)^0 \cdot (0.3)^{6-0} + \binom{6}{1} \cdot (0.7)^1 \cdot (0.3)^{6-1} = 0.0109.$$

Analogno za $X = \{2, 3, 4, 5, 6\}$ dobivamo vrijednosti zapisane u sljedećoj tablici:

X	0	1	2	3	4	5	6
$f(X)$	0.0007	0.0102	0.0595	0.1832	0.3241	0.3136	0.1187
$F(X)$	0.0007	0.0109	0.0704	0.2536	0.5777	0.8813	1

Primjer 6.1.9. *Jedna tvornica proizvodi čokoladice. Vjerojatnost da je jedna čokoladica defektna, tj. oštećena je 10%. Koja je vjerojatnost da ćemo, uzevši uzorak od 50 čokoladica, dobiti 3 oštećene čokoladice?*

Zadano je

$$n = \textit{broj proizvedenih čokoladica} = 50$$
$$x = \textit{broj oštećenih čokoladica} = 3$$
$$p = \text{vjerojatnost da je pojedina čokoladica oštećena} = 0.1$$

Uvrstimo li podatke u formulu

$$P(X = x) = f(x) = \binom{n}{x} \cdot p^x \cdot q^{n-x}$$

dobivamo

$$P(X = 3) = \binom{50}{3} \cdot (0.1)^3 \cdot (0.9)^{47} = 0.13857.$$

Vjerojatnost da ćemo dobiti 3 oštećene čokoladice jednaka je 0.13857.

Koja je vjerojatnost da ćemo, uzevši uzorak od 1 000 000 čokoladica dobiti 50 000 oštećenih čokoladica? Uočavamo da ukoliko bismo pristupili kao u prethodnom primjeru, trebali bismo izračunati binomni koeficijent $\binom{1\,000\,000}{50\,000}$. No, to nije najbolje rješenje jer se radi o računu s veoma velikim brojevima. Poissonova razdioba će biti bolji izbor.

Teorem 6.1.10. *Kod binomne razdiobe $X \sim B(n, p)$ najveća vjerojatnost pripada onoj vrijednosti slučajne varijable x_0 za koju vrijedi:*

$$np - q \leq x_0 \leq np + p.$$

Dokaz. Odredimo najprije rekurzivnu formulu za $P(x) = P(X = x)$.
Dijeljenjem izraza

$$P(x) = \binom{n}{k} \cdot p^x \cdot q^{n-x}$$

i

$$P(x-1) = \binom{n}{x-1} \cdot p^{x-1} \cdot q^{n-x+1}$$

dobivamo

$$\frac{P(x)}{P(x-1)} = \frac{\binom{n}{x} \cdot p^x \cdot q^{n-x}}{\binom{n}{x-1} \cdot p^{x-1} \cdot q^{n-x+1}} = \frac{\frac{n!}{(x-1)!x(n-x)!}}{\frac{n!}{(x-1)!(n-x)!(n-x+1)}} \cdot \frac{p^x q^{n-x}}{p^x p^{-1} q^{n-x} q} = \frac{n-x+1}{x} \cdot \frac{p}{q}$$

Dakle vrijedi $P(x) = \dfrac{n-x+1}{x} \cdot \dfrac{p}{q} \cdot P(x-1)$

Ako je x_0 ona vrijednost varijable x kojoj pripada najveća vjerojatnost, onda je:

$$P(x_0 - 1) \leq P(x_0) \text{ i } P(x_0) \geq P(x_0 + 1).$$

Najprije razmotrimo nejednakost $P(x_0 - 1) \leq P(x_0)$. Izrazimo li $P(x_0)$ pomoću rekurzivne formule preko $P(x_0 - 1)$, dobivamo

$$P(x_0 - 1) \leq \frac{n - x_0 + 1}{x_0} \cdot \frac{p}{q} \cdot P(x_0 - 1), \text{ tj.}$$

$$1 \leq \frac{n - x_0 + 1}{x_0} \cdot \frac{p}{q}$$

Nakon sređivanja dobivamo

$$x_0 \leq np + p.$$

Nadalje, razmotrimo nejednakost $P(x_0) \geq P(x_0+1)$ te izrazimo $P(x_0+1)$ pomoću rekurzivne formule preko $P(x_0)$

$$P(x_0) \geq \frac{n - x_0}{x_0 + 1} \cdot \frac{p}{q} \cdot P(x_0),$$

odnosno

$$1 \geq \frac{n - x_0}{x_0 + 1} \cdot \frac{p}{q},$$

odakle slijedi

$$x_0 \geq np - q.$$

□

Pogledajmo sada na konkretnom primjeru primjenu prethodno dokazanog teorema.

Primjer 6.1.11. *Automat izrađuje proizvod i daje 8% defektnih proizvoda. Proizvodi se bez kontrole pakiraju u kutije od po 30 komada. Koliko će neispravnih proizvoda biti najčešće u kutiji?*

U našem primjeru je $n = 30$, $p = 0.08$. Označimo sa x broj defektnih proizvoda.
Dalje nastavljamo direktnom primjenom teorema 6.1.10.:

$$30 \cdot 0.08 - 0.92 \leq x_0 \leq 30 \cdot 0.08 + 0.08$$
$$1.48 \leq x_0 \leq 2.48$$

Budući da je x, tj. broj defektnih proizvoda, prirodan broj, slijedi da je $x_0 = 2$.

6.1.2 Poissonova distribucija

Poissonova razdioba daje model vjerojatnosti „rijetkih" događaja (ponekad se naziva i „zakon rijetkih događaja") koji se događaju u jedinici vremena, površine, volumena i slično. Koristi se u slučajevima kada je vjerojatnost nekog događaja jako mala, a broj uzoraka veoma velik. Neki od primjera su:

- broj prometnih nesreća na određenoj dionici autoceste u jednom danu,
- telefonski pozivi na centrali u jednoj minuti,
- broj mjesečnih nesreća u tvornici,
- broj oboljelih stabala po aru šume,
- broj vidljivih grešaka na dijamantu.

U ovim primjerima možemo uočiti da je npr. veoma mala vjerojatnost da će se u nekom kratkom vremenskom periodu dogoditi prometna nesreća na promatranoj dionici ili da će se pojaviti defekt na određenom dijelu bakrene žice.

Definicija 6.1.12. *Slučajna varijabla X s parametrom $\lambda > 0$ i funkcijom vjerojatnosti te slučajne varijable*

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \ldots$$

naziva se Poissonova slučajna varijabla.

Poissonova razdioba je zapravo granični prijelaz binomne u kojoj parametar n raste u beskonačnost, ali pod uvjetom da produkt np ostane konstantan. Zbog toga slijedi da će parametar p težiti k nuli. Produkt np označimo s $\lambda : \lambda := np$.
Ako formulu za binomnu razdiobu

$$P(x) = \binom{n}{k} \cdot p^x \cdot q^{n-x}$$

zapišemo kao

$$P(x) = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} p^x q^{n-x}$$

te p zamijenimo s $\frac{\lambda}{n}$, tada dobivamo

$$P(x) = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \cdot \frac{\lambda^x}{n^x} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

Iz toga slijedi

$$P(x) = \frac{(1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{x-1}{n})}{x!} \cdot \lambda^x \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-x}.$$

Kada n teži u beskonačnost vrijedi

$$\lim_{n\to\infty,\, np=\lambda} P(x) = \lim_{n\to\infty,\, np=\lambda} \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n = \frac{\lambda^x}{x!}\cdot e^{-\lambda}.$$

Time smo dobili formulu za Poissonovu razdiobu

$$P(x) = \frac{\lambda^x}{x!}\, e^{-\lambda},\ x = 0, 1, 2, \ldots$$

Riješimo primjer 6.1.9. pomoću Poissonove formule:

Primjer 6.1.13. *Jedna tvornica proizvodi čokoladice. Vjerojatnost da je jedna čokoladica defektna, tj. oštećena je 10%. Koja je vjerojatnost da ćemo, uzevši uzorak od 50 čokoladica, dobiti 3 oštećene čokoladice?*

U našem primjeru je

$$n = \textit{veličina uzorka} = 50,$$

$$\lambda = \textit{prosječan broj oštećenih čokoladica u uzorku} = 0.1 \cdot 50 = 5,$$

$$x = \text{broj oštećenih čokoladica u uzorku} = 3,$$

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

$$P(3) = \frac{5^3}{3!}\, e^{-5} = 0.1425 = 14.25\%.$$

Dakle, pomoću Poissonove razdiobe dobili smo vjerojatnost 14.25% da ćemo, uzevši uzorak od 50 čokoladica, dobiti 3 oštećene, dok smo pomoću binomne razdiobe dobili vjerojatnost od 13.85%. Razlika je mala. Možemo zaključiti da što je manji p, a n veći, aproksimacija je bolja. Sada možemo riješiti drugi dio zadatka, tj. izračunati vjerojatnost da ćemo, uzevši uzorak od 1 000 000 čokoladica, dobiti 50 000 oštećenih.

U ovom slučaju je

$$n = \textit{veličina uzorka} = 1\,000\,000,$$

$$\lambda = \textit{prosječan broj oštećenih čokoladica u uzorku} = 0.1 \cdot 1\,000\,000 = 100\,000,$$

$$x = \text{broj oštećenih čokoladica u uzorku} = 50\,000,$$

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

$$P(50\,000) = \frac{100\,000^{50\,000}}{50\,000!}\, e^{-100\,000} \approx 1.6667 \cdot 10^{17} \cdot 1.0733 \cdot 10^{-433} \approx 1.79 \cdot 10^{-416}.$$

Primjer 6.1.14. *Pretpostavimo da je* 220 *grešaka raspoređeneo slučajno unutar knjige od 200 stranica. Odredite vjerojatnost da dana stranica knjige sadrži:*
a) niti jednu grešku,
b) točno jednu grešku,
c) barem dvije greške.

Definirajmo slučajnu varijablu X koja broji greške na pojedinoj stranici. Ona ima Poissonovu razdiobu. Kako bismo odredili njenu funkciju vjerojatnosti, potreban nam je parametar λ. Znamo da je taj parametar jednak prosječnom broju događaja (broj grešaka) koji se dogode u jednoj jedinici vremena (na jednoj stranici).

Stoga je

$$\lambda = \frac{220}{200} = 1.1,$$

$$P(x) = P(X = x) = \frac{\lambda^x}{x!}\, e^{-\lambda} = \frac{1.1^x}{x!}\, e^{-1.1},\ k = 0, 1, 2, \ldots$$

a) $P(X = 0) = \dfrac{1.1^0}{0!}\, e^{-1.1} \approx 0.333,$

b) $P(X = 1) = \dfrac{1.1^1}{1!}\, e^{-1.1} \approx 0.366,$

c) $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) \approx 1 - 0.333 - 0.366 = 0.301.$

6.2 Kontinuirane slučajne varijable

Do sada smo govorili samo o slučajnim varijablama koje su diskretne, a naravno postoje i kontinuirane, tj. neprekidne slučajne varijable. Kod njih nećemo moći samo popisati x_k-ove, nego ćemo morati definirati funkciju gustoće slučajne varijable. Objasnimo to kroz primjer.

Promatrajmo minutnu kazaljku analognog sata. Prepostavimo da se kazaljka kontinuirano miče. Neka je položaj te kazaljke slučajna varijabla X.
Očito je da je

$$0 \leq X < 60,$$

ali isto tako i da vjerojatnost za bilo koji $X = x_k$, $0 \leq x_k < 60$ je

$$P(X = x_k) = 0.$$

Zbog toga ima smisla promatrati samo intervale u kojima se X može nalaziti. Tako ima smisla promatrati na primjer

$$P(15 \leq X \leq 30) = \frac{1}{4}, \quad P(32 \leq X \leq 37) = \frac{1}{12}, \quad P(X < 1) = \frac{1}{60}.$$

Zaključujemo da funkcija vjerojatnosti kontinuirane slučajne varijable nema smisla, ali ima smisla funkcija gustoće takve slučajne varijable.

Za svaki $a, b \in \mathbb{R}$, $\mathrm{P}\,(a \leq X \leq b)$ je površina ispod funkcije gustoće slučajne varijable $f(x)$ pa je

$$\mathrm{P}\,(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Tako je u prethodnom primjeru

$$\mathrm{P}\,(15 \leq X \leq 30) = \int_{15}^{30} \frac{1}{60}\,dx = \frac{1}{60}\,x\Big|_{15}^{30} = \frac{30-15}{60} = \frac{1}{4}.$$

Za funkciju gustoće slučajne varijable moraju vrijediti i neka svojstva:

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Za neprekidne slučajne varijable možemo definirati očekivanje i varijancu na sličan način kao za diskretne slučajne varijable:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

6.2.1 Normalna distribucija

U 18. stoljeću, De Moivre je pokazao da se binomna slučajna varijabla dobro aproksimira jednom neprekidnom slučajnom varijablom te da je dana aproksimacija bolja što je broj ponavljanja pokusa veći. Ta se nova slučajna (neprekidna) varijabla naziva normalna slučajna varijabla.

Definicija 6.2.1. *Za kontinuiranu slučajnu varijablu X kažemo da ima normalnu ili Gaussovu razdiobu s parametrima μ i σ^2 i pišemo $X \sim N(\mu, \sigma^2)$, ako je njezina funkcija gustoće vrijednosti zadana formulom:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

gdje je: σ - standardna devijacija, σ^2 - varijanca, μ očekivanje.

Takvu slučajnu varijablu X zovemo normalna slučajna varijabla. Specijalno, ako je $\mu = 0$, $\sigma^2 = 1$, normalnu slučajnu varijablu zovemo standardna normalna slučajna varijabla.

Želimo li izračunati vjerojatnost da je normalna slučajna varijabla $N(\mu, \sigma^2)$ manja ili jednaka nekom broju a, onda ćemo morati poznavati funkciju distribucije slučajne varijable N, tj. morat ćemo izračunati integral

$$\mathrm{P}(X \leq a) = \int_{-\infty}^{a} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,dx,$$

što je netrivijalni integral koji ovisi o čak tri parametra (a, μ i σ) pa bi tablice za to bile prekomplicirane. Umjesto toga se tabelira samo funkcija distribucije standardne normalne slučajne varijable $N(0, 1)$, a sve ostale normalne slučajne varijable se jednostavnom transformacijom svode na tu normalnu slučajnu varijablu.

Postupak standardizacije

Neka je X normalna slučajna varijabla $X \sim N(\mu, \sigma^2)$. Tada je slučajna varijabla $Z = \frac{X-\mu}{\sigma}$ standardna normalna slučajna varijabla, tj. normalna slučajna varijabla s očekivanjem 0 i varijancom 1. Vrijednost varijable Z zapravo predstavlja udaljenost od očekivanja izraženu u dijelovima standardne devijacije. Njezine vrijednosti čitamo iz tablice (slika 6.1).

	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Slika 6.1: Tablica vrijednosti funkcije distribucije standardne normalne razdiobe

Funkciju distribucije od Z označavamo s Φ i vrijedi:

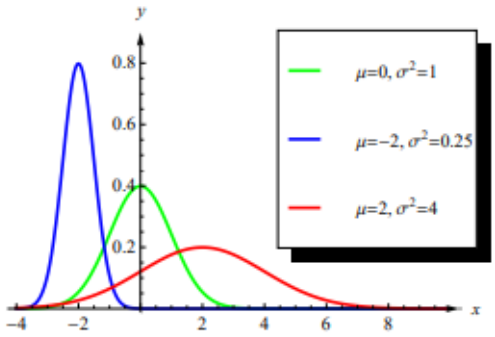
$$\Phi(x) = P(Z < x).$$

Za $x \leq 0$ vrijedi:

$$\Phi(x) = 1 - \Phi(-x).$$

Krivulja normalne razdiobe:

- je simetrična u odnosu na pravac $x = \mu$
- ima maksimum $\frac{1}{\sigma\sqrt{2\pi}}$ za $x = \mu$
- ima točke infleksije su $x = \mu - \sigma$ i $x = \mu + \sigma$.



Slika 6.2: Krivulja normalne razdiobe

Promatrajuću sliku 6.2 vidimo da vrh krivulje leži na samoj očekivanoj vrijednosti μ. Krivulja je simetrična s obje strane, a njeni krajevi padaju u zvonoliki oblik te se asimptotski približavaju x osi, a to znači da se dodiruju u beskonačnosti. Normalna krivulja ima dvije točke infleksije, koje su od očekivane vrijednosti udaljene za iznos standardne devijacije σ. Stoga je međusobna udaljenost ovih točaka 2σ.

Za manje iznose standardne devijacije krivulja je strmija od, primjerice, krivulja s većim iznosom devijacije. Moguće je donijeti zaključak da se širina razdiobe povećava kako se povećava i vrijednost σ.

Položaj bilo kojeg podatka u razdiobi je moguće odrediti pomoću takozvane Z-vrijednosti koja predstavlja udaljenost nekog rezultata mjerenja od sredine iskazana kao dio standardne devijacije skupa pdataka. Izračunava se kao $Z = \frac{X-\mu}{\sigma}$, pri čemu je X rezultat mjerenja. Ako za neki podatak izračunamo vrijednost, na primjer, 1.25, to znači da se taj podatak nalazi na 1.25 standardne devijacije desno od srednje vrijednosti. Uočimo kako je predznak Z vrijednosti bitan. Kako očitavamo vrijednosti iz tablice dane na slici 6.1?

U prvom stupcu tablice nalaze se Z-vrijednosti izražene s jednom decimalom, a drugu decimalu pronalazimo u prvom retku na vrhu tablice. Tako, npr. za $Z = 1.25$ očitamo vrijednost na mjestu gdje se sijeku redak koji počinje s 1.2 i stupac 5, što je 0.8944. To znači da je 89.44% manjih od podataka čija je Z-vrijednost 1.25.

Primjer 6.2.2. *Vijek trajanja neke automobilske gume je normalno distribuiran s očekivanjem* 34000 *km i standardnom devijacijom od* 4000 *km. Izračunajte vjerojatnost da guma traje više od* 40000 *km.*

Neka je X vijek trajanja gume.

$$X \sim N(34000, 4000^2)$$
$$Z = \frac{X - 34000}{4000} \sim N(0, 1)$$

Tražimo $P(X > 40000) = ?$

$$P\left(\frac{X - 34000}{4000} > \frac{40000 - 34000}{4000}\right) = P(Z > 1.5) = 1 - P(Z < 1.5)$$
$$= 1 - \Phi(1.5)$$
$$= 1 - 0.9332 = 0.0668.$$

Vjerojatnost da će guma trajati više od 40000 km je 6.68%.

Primjer 6.2.3. *Debljina željeznih ploča je slučajna varijabla. Možemo pretpostaviti da je to kontinuirana slučajna varijabla koja ima normalnu distribuciju s očekivanjem* 10 *mm i standardnom devijacijom* 0.02 *mm. Kolika je vjerojatnost defektne ploče ako je kontrola dala sljedeći kriterij:*

a) ploča mora biti tanja od 9.97*mm*
b) ploča mora biti deblja od 10.05*mm?*

X nam označava debljinu željeznih ploča. Za slučajnu varijablu $X \sim N(10, 0.02^2)$ računamo:

a) $P(X < 9.97) = ?$

$$Z = \frac{X - 10}{0.02} \sim N(0, 1)$$
$$P(X < 9.97) = P(\frac{X - 10}{0.02} < \frac{9.97 - 10}{0.02}) = P(Z < -1.5) = \Phi(-1.5)$$

$$= 1 - \Phi(1.5) = 1 - 0.9332 = 0.0668$$

Uz ovaj kriterij očekuje se 6.68% oštećenih ploča.

b) $Z = \frac{X-10}{0.02} \sim N(0, 1)$

$$P(X > 10.05) = P(\tfrac{X-10}{0.02} > \tfrac{10.05-10}{0.02}) = P(Z > 2.5) = 1 - P(Z < 2.5)$$
$$1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.$$

Uz ovaj kriterij vjerojatnost oštećenih ploča je 0.62%.

Poglavlje 7

Osnovni elementi Bayesovskog zaključivanja

U prethodnim poglavljima dane su osnove statistike i vjerojatnosti, naveden je i dokazan Bayesov teorem, a navedeni su i primjeri nekih slučajnih varijabli te njihove definicije. Nakon tih postavljenih matematičkih temelja na kojima ćemo graditi daljnju teoriju, uvedimo i ključne principe Bayesovske analize.

Znamo, Bayesov teorem jedan je od osnovnih rezultata u teoriji vjerojatnosti i temelj je Bayesove statistike, a kao što je već naglašeno, u Bayesovoj statistici parametri se promatraju kao slučajne varijable. Taj se postupak provodi na sljedeći način: odaberemo distribuciju koja najbolje opisuje naša vjerovanja o parametru (ova distribucija se naziva apriorna distribucija), a zatim iskoristimo podatke kako bismo ažurirali ta vjerovanja, i to korištenjem Bayesovog teorema. Prvo, definirajmo funkciju vjerodostojnosti.

7.1 Funkcija vjerodostojnosti

Koristimo notaciju $f(x|\theta)$ da bi predstavili zajedničku (uvjetnu) funkciju gustoće slučajnog uzorka uz parametar $\theta \in \Theta$. $f(x|\theta) = L(\theta)$ je funkcija od θ i zovemo ju **vjerodostojnost**.

Definicija 7.1.1. *Neka je $x = (x_1, x_2, \ldots, x_n)$ opaženi uzorak za varijablu X s populacijskom gustoćom $f(x|\theta)$, te neka je $\theta \in \Theta$ nepoznati parametar, gdje je Θ parametarski prostor. Tada je vjerodostojnost funkcija $L : \Theta \to \mathbb{R}$ definirana sa*

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta).$$

U slučaju diskretne slučajne varijable, umjesto $f(x|\theta)$ koristimo $P(x|\theta)$ - uvjetnu vjerojatnost danih podataka x uz uvjet da je vjerovanje u θ istinito.

Prema principu vjerodostojnosti sve informacije o parametru θ koje donosi opažanje x od X su sadržane u funkciji vjerodostojnosti $L(\theta)$. Štoviše, dvije funkcije vjerodostojnosti sadrže istu informaciju o θ ako su proporcionalne jedna drugoj.

Primjer 7.1.2. *Ispitanik je doznao da 9 osoba koristi, a 3 osobe ne koriste određeni proizvod, dok je* $\theta \in [0, 1]$ *vjerojatnost da osoba koristi proizvod.*

Ako su to jedine informacije s kojima raspolažemo, problem možemo modelirati na dva načina:

1. Ispitanik je ispitao 12 osoba i uočio da je broj osoba koje koriste proizvod X binomna slučajna varijabla $B(12, \theta)$, s opservacijom $X = 9$. Funkcija vjerojatnosti od X je jednaka
$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{12}{9} \theta^9 (1-\theta)^3 = 220\, \theta^9 (1-\theta)^3.$$
2. Ispitanik je ispitivao osobe sve dok nije dobio $\alpha = 3$ ispitanika koji su izjavili da ne koriste dani proizvod i uočio da je X, broj osoba koje koriste proizvod, negativna binomna varijabla $NB(3, 1-\theta)$. Funkcija vjerojatnosti od X je jednaka:
$$f(x|\theta) = \binom{\alpha + x - 1}{\alpha - 1} (1-\theta)^\alpha \, [1-(1-\theta)]^x = \binom{3+9-1}{3-1} (1-\theta)^3 \, \theta^9 = 55\, \theta^9 (1-\theta)^3.$$

Uočimo da je u oba slučaja vjerodostojnost proporcionalna s $\theta^9 (1-\theta)^3$, čime je zadovoljen princip vjerodostojnosti. Sve informacije o parametru θ su sadržane u funkciji vjerodostojnosti $L(\theta) \sim \theta^9 (1-\theta)^3$.

Ažuriranje vjerovanja o parametru

Iz Bayesove formule
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
očigledno slijedi
$$P(A|B) \sim P(B|A)P(A).$$

Objasnimo način na koji se vrši ažuriranje vjerovanja o parametru Θ. Pretpostavimo da je Θ diskretna slučajna varijabla. Nadalje, neka E označava skup nekih informacija o Θ. Ono što nas zanima je distribucija parametra Θ u svjetlu novih informacija E, tj. zanima nas vjerojatnost $P(\Theta = \theta|E)$. Koristimo Bayesov teorem:
$$P(\Theta = \theta|E) = P(\Theta = \theta) \frac{P(E|\Theta = \theta)}{P(E)}.$$

Primijetimo nekoliko činjenica:

1. $P(\Theta = \theta)$ je inicijalna funkcija vjerojatnosti od Θ, tj. funkcija vjerojatnosti koja se odnosi na vjerovanja o parametru Θ prije uzimanja u obzir novih infomacija E. Ovu funkciju vjerojatnosti nazivamo apriornom funkcijom vjerojatnosti.

2. Nakon uzimanja u obzir informacija E, ažuriramo naša dosadašnja vjerovanja o Θ koja su sada sadržana u funkciji vjerojatnosti $P(\Theta = \theta|E)$. Ovu funkciju nazivamo aposteriorna funkcija vjerojatnosti.

3. $\frac{P(E|\Theta=\theta)}{P(E)}$ je izraz koji pokazuje koliko će novi skup informacija E izmijeniti naša vjerovanja o Θ, odnosno koliko će promijeniti apriornu distribuciju. Primijetimo i da je $P(E|\Theta)$ vjerodostojnost za fiksnu vrijednost θ.

7.2 Apriorna i aposteriorna distribucija

Neka je Θ skup svih vrijednosti koje parametar θ može poprimiti. S π označimo apriornu distribuciju parametra θ. Primijetimo da je $\pi : \Theta \to \mathbb{R}$. Neka je X slučajna varijabla ili vektor iz kojeg dolaze podaci, a x njegova realizacija. Uvjetnu funkciju gustoće od X, uz danu vrijednost $\theta \in \Theta$ označimo s $f(x|\theta)$. Uvodimo sljedeće distribucije i oznake.

Definicija 7.2.1. *Distribuciju od θ, za dane $x_1, x_2, \ldots, x_n$ zovemo aposteriori distribucija i definiramo sa*

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{f(x)},$$

gdje je $f(x)$ marginalna distribucija od X.

Marginalna distribucija $f(x)$ se računa formulom:

$$f(x) = \begin{cases} \sum_\theta f(x|\theta)\pi(\theta), & \text{u diskretnom slučaju} \\ \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)d\theta, & \text{u neprekidnom slučaju} \end{cases},$$

gdje je $\pi(\theta)$ apriorna distribucija.

Dakle, vjerojatnost nekog događaja prije analize podataka, zovemo apriori vjerojatnost, a prilagođenu vjerojatnost aposteriori vjerojatnost.

Zamijetimo kako se ovim pristupom slučajnost opisuje kombinirajući informacije dobivene iz dva izvora: obrađenih podataka i apriori uvjerenja. Pogledajmo to na sljedećom primjeru.

Primjer 7.2.2. *Promatramo manadžera fonda koji testira strategiju pronalaženja akvizicijskih meta te ispitivanja njihove efikasnosti, preciznije pripadne cijene oslobođenja novčanog toka (eng. PFCF).*

Definiramo dva događaja:

D = Tvrtki X, u protekle 3 godine, PFCF je više od tri puta manji od prosjeka u tom sektoru,

E = Tvrtka X postala je akvizicijska meta u toku promatrane godine.

Na početku, nezavisno od pokazatelja, menadžer pretpostavljaa da je vjerojatnost da tvrtka postane meta 40%, što znači

$$P(E) = 0.4 \quad \text{odnosno} \quad P(E^c) = 0.6.$$

Nadalje pretpostavimo da nakon analize podataka menadžer ima vjerojatnost od 75% da tvrtka koja je postala meta, ima više od tri puta manji PFCF nego prosjek u tom sektoru. Dok je vjerojatnost da tvrtka koja nije meta, ima više od tri puta manji PFCF nego prosjek u tom sektoru jednaka 35%:

$$P(D|E) = 0.75 \quad \text{i} \quad P(D|E^c) = 0.35.$$

Sada je menadžer u stanju nadograditi apriori uvjerenja na način da, ako zna da tvrtka ima tri ili više puta manji PFCF nego prosjek u tom sektoru, može izračunati vjerojatnost da će postati akvizicijska meta. Koristeći Bayesov teorem dobije se:

$$P(E|D) = \frac{0.75 \times 0.4}{0.75 \times 0.4 + 0.35 \times 0.6} \approx 0.59.$$

Zaključujemo da, nakon što se uzme u obzir razina PFCF, vjerojatnost da tvrtka postane akvizicijska meta raste sa 40% na 59%.

Iz načina na koji je definirana aposteriorna distribucija parametra, zaključujemo da, ukoliko znamo funkciju gustoće $f(x|\theta)$ i apriornu funkciju gustoće parametra $\pi(\theta)$, možemo izračunati aposteriornu distribuciju parametra θ (odnosno imamo formulu za to, sam izračun je često problematičan), odnosno ažurirati naša vjerovanja o parametru θ uzimajući u obzir nove informacije $X = x$. Sada ćemo definirati Bayesovski statistički model.

Definicija 7.2.3. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i neka je $\mathcal{P}$ neka familija vjerojatnosti na $(\Omega, \mathcal{F})$. Uređena trojka $(\Omega, \mathcal{F}, \mathcal{P})$ naziva se statistička struktura. Familija $\mathcal{P}$ često je parametrizirana i zapisuje se u obliku:*

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

Definicija 7.2.4. *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je svaka slučajna varijabla (ili slučajni vektor za $k \geq 2$) $T : \Omega \to \mathbb{R}^k$ takva da za neki $n \in \mathbb{N}$ postoji n-dimenzionalni slučajni vektor $(X_1, X_2, \ldots, X_n)$ na $(\Omega, \mathcal{F}, \mathcal{P})$, te izmjeriva funkcija $t : R^n \to R^k$ takva da je $T = t(X_1, X_2, \ldots, X_n)$.*

Definicija 7.2.5. *Bayesovski statistički model slučajne varijable (ili slučajnog vektora)* $X : \Omega \to R^k$ *je familija* $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ *iz prethodne definicije pri čemu je parametru* θ *pridružena njegova apriorna distribucija s gustoćom* $\pi(\theta)$, $\theta \in \Theta$.

7.3 Bayesovsko zaključivanje za binomnu distribuciju

U ovom odjeljku gledat ćemo primjenu Bayesovskog zaključivanja danog formulom (vidi Definiciju 7.2.1.):

$$\pi(\theta|x) = \frac{f(\theta|x)\pi(\theta)}{f(x)},$$

u posebnom slučaju kada je

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad 0 \le \theta \le 1,$$

tj. u slučaju kada su rezultati opažanja opisani binomnom distribucijom. Podsjetimo se, to je uvijek slučaj kada radimo niz pokusa s dva moguća ishoda: uspjehom i neuspjehom. Pokazat ćemo primjer u kojem anketiramo 100 Zagrepčana s pitanjem podržavaju li izgradnju novog mosta preko rijeke Save.

Pritom u računu možemo izostaviti konstantu $\binom{n}{x}$ jer se ista konstanta javlja pri računanju marginalne distribucije $f(x)$ pa se u gornjoj (Bayesovoj) formuli ta konstanta pokrati. Štoviše, poželjno je u potpunosti izbjeći računanje marginalne distribucije jer se u slučaju kontinuirane apriorne distribucije to često svodi na numeričku integraciju koja nije praktična, pogotovo ako ju treba provesti puno puta.

To se može postići korištenjem konjugirane apriorne distribucije, tj. takve apriorne distribucije za koju znamo da će i aposteriorna distribucija biti iz iste familije distribucija. Definirat ćemo sada familiju $Be(\alpha, \beta)$ beta distribucija i pokazati da je ta familija konjugirana za binomnu raspodjelu.

Definicija 7.3.1. *Beta distribucija* $Be(\alpha, \beta)$ *definira se funkcijom gustoće vjerojatnosti*

$$g(x\,;\alpha,\beta) = \begin{cases} k \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1} & , \quad 0 \le x \le 1 \\ 0 & , \quad x \notin [0,1] \end{cases}$$

Konstanta k se, kao i kod svake druge funkcije gustoće vjerojatnosti, određuje tako da ukupni integral funkcije iznosi 1, i u ovom slučaju iznosi

$$k = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},$$

gdje je $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ poznata gama funkcija koja zadovoljava rekurzivnu relaciju $\Gamma(x) = (x-1)\cdot\Gamma(x-1)$, tj. za prirodan broj x je $\Gamma(x) = (x-1)!$

Ako beta distribuciju označimo s X, lako se koristeći spomenutu rekurzivnu relaciju, pokazuje da su njeno očekivanje i varijanca dani formulama

$$E(X) = \frac{\alpha}{\alpha+\beta}$$

i

$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Konačno, ako beta distribuciju uzmemo za apriornu distribuciju

$$\pi(\theta) \sim \theta^{\alpha-1}\cdot(1-\theta)^{\beta-1},$$

a funkcija vjerodostojnosti nam je istog oblika

$$f(x|\theta) \sim \theta^x\cdot(1-\theta)^{n-x},$$

Bayesovsko zaključivanje daje da je i aposteriorna distribucija istog oblika:

$$\pi(\theta|x) \sim f(x|\theta)\pi(\theta) = \theta^{x+\alpha-1}\cdot(1-\theta)^{n-x+\beta-1},$$

tj. Bayesovsko zaključivanje iz apriorne $Be(\alpha,\beta)$ distribucije daje aposteriornu distribuciju

$$Be(\alpha+x, \beta-x+n),$$

gdje je x broj uspjeha u n izvođenja pokusa, tj. dobivamo vrlo jednostavno pravilo za ažuriranje vjerovanja o parametru.

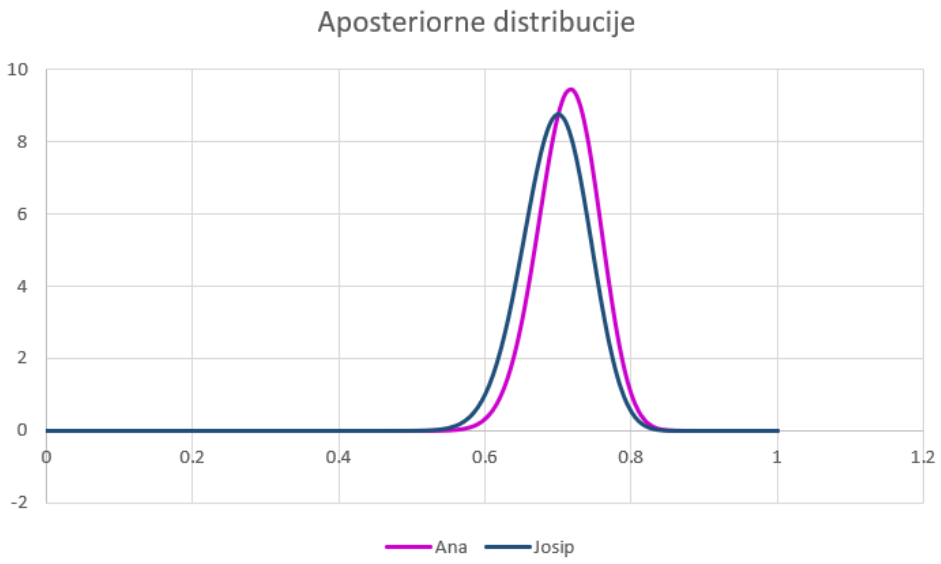
Opišimo sada primjer najavljen u uvodu.

Primjer 7.3.2. *Dvoje studenata, Ana i Josip, anketirali su* 100 *Zagrepčana s pitanjem podržavaju li izgradnju novog mosta preko rijeke Save. Prije anketiranja Ana je smatrala da* 80% *Zagrepčana podržava izgranju mosta s mogućim odstupanjem od* 10%*, te je podatke uvrstila u formule za očekivanje i varijancu beta distribucije:*

$$\frac{\alpha}{\alpha+\beta} = 0.8\,, \qquad \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.1^2\,.$$

Rješavanjem ovog sustava jednadžbi dobila je $\alpha = 12$, $\beta = 3$, tj. ona je izabrala $Be(12, 3)$ apriornu distribuciju. Josip je bio neodlučan pa je izabrao uniformnu, tj. $Be(1, 1)$ apriornu distribuciju prema kojoj su sve mogućnosti jednako vjerojatne.

Na anketi je 70 Zagrepčana podržalo izgradnju mosta, te su Ana i Josip odlučili ažurirati svoja uvjerenja Bayesovskim zaključivanjem: Ana je dobila $Be(12 + 70, 3 - 70 + 100) = Be(82, 33)$ aposteriornu distribuciju, a Josip $Be(1 + 70, 1 - 70 + 100) = Be(71, 31)$ aposteriornu distribuciju. Sa slike je vidljivo da su te dvije distribucije vrlo bliske, tj. Bayesovsko zaključivanje daje dobar rezultat čak i u slučaju loše apriorne distribucije (kao što je to bila Josipova).



Slika 7.1: Aposteriorne distribucije dvoje studenata

7.4 Bayesovsko zaključivanje za Poissonovu distribuciju

Sada proučavamo Bayesovsko zaključivanje u posebnom slučaju kada je

$$f(x|\theta) = \frac{\theta^x \cdot e^{-\theta}}{x!}, \qquad \theta > 0,$$

tj. kada promatramo rijetke događaje koji se u prosjeku događaju θ puta u jedinici vremena ili prostora pa znamo da je njihovo pojavljivanje dobro opisano Poissonovom distribucijom.

Distribucija čija gustoća ima isti oblik kao $f(x|\theta)$, i koja će prema tome biti konjugirana za Poissonovu distribuciju, je gama(α,β) distribucija s gustoćom

$$g(\theta;\alpha,\beta) = \frac{\beta^\alpha \theta^{\alpha-1}\cdot e^{-\beta\theta}}{\Gamma(\alpha)}, \qquad \theta \geq 0.$$

Preciznije, ako gama(α,β) distribuciju uzmemo za apriornu distribuciju

$$\pi(\theta) \sim \theta^{\alpha-1}\cdot e^{-\beta\theta},$$

Bayesovsko zaključivanje daje aposteriornu distribuciju

$$\begin{aligned}\pi(\theta|x) \sim f(x|\theta)\pi(\theta) &\sim \theta^x \cdot e^{-\theta}\cdot\theta^{\alpha-1}\cdot e^{-\beta\theta}\\ &= \theta^{\alpha+x-1}\cdot e^{-(\beta+1)\theta},\end{aligned}$$

tj. aposteriorna distribucija je gama$(\alpha+x,\beta+1)$ distribucija.

Za primjenu gama(α,β) distribucije važno je poznavati njezino očekivanje i varijancu:

$$\begin{aligned}E(X) = \int_0^\infty \theta\, g(\theta\,;\alpha,\beta)\, d\theta &= \int_0^\infty \theta\cdot\frac{\beta^\alpha\theta^{\alpha-1}e^{-\beta\theta}}{\Gamma(\alpha)}\, d\theta\\ &= \frac{\alpha}{\beta}\cdot\int_0^\infty \frac{\beta^{\alpha+1}\theta^{\alpha+1-1}e^{-\beta\theta}}{\Gamma(\alpha+1)}\, d\theta\\ &= \frac{\alpha}{\beta}\end{aligned}$$

(zadnji integral je jednak 1 jer pod integralom prepoznajemo gustoću gama$(\alpha+1,\beta)$ distribucije), a na sličan način možemo izračunati i $V(X) = \frac{\alpha}{\beta^2}$.

Primjer 7.4.1. *Dvoje studenata, Matea i Mario, žele procijeniti broj prometnih nesreća na novoobnovljenom remetinečkom rotoru u razdoblju od* 6 *mjeseci. Matea nema predodžbu o broju nesreća na prometnicama pa izabire uniformnu, tj. gama*$(1,0)$ *distribuciju. Mario zna da je u razdoblju od* 6 *mjeseci znalo biti i više od* 200 *prometnih nesreća na remetinečkom rotoru, te se nada da će nakon obnove to pasti barem na* 50*, s mogućim odstupanjem od* ± 20*.*

Te je podatke uvrstio u formule za očekivanje i varijancu gama distribucije:

$$\frac{\alpha}{\beta} = 50\,, \qquad \frac{\alpha}{\beta^2} = 20^2\,,$$

odakle je dobio $\beta = \frac{1}{8} = 0.125$ i $\alpha = \frac{50}{8} = 6.25$, tj. izabrao je gama$(6.25, 0.125)$ apriornu distribuciju.

Nakon što su saznali da se u prvih 6 mjeseci nakon obnove na rotoru dogodilo samo 7 prometnih nesreća, odlučili su ažurirati svoje uvjerenje Bayesovskim zaključivanjem.

Matea dobiva gama$(1 + 7, 0 + 1)$ = gama$(8, 1)$ aposteriornu distribuciju s očekivanjem $\frac{8}{1} = 8$ i standardnom devijacijom $\sqrt{\frac{8}{1^2}} \approx 2.83$.
Mario dobiva gama$(6.25 + 7, 0.125 + 1)$ = gama$(13.25, 1.125)$ aposteriornu distribuciju s očekivanjem $\frac{13.25}{1.125} \approx 11.78$ i standardnom devijacijom $\sqrt{\frac{13.25}{1.125^2}} \approx 3.24$.

Zaključujemo da se broj prometnih nesreća na remetinečkom rotoru drastično smanjio, ali da će trebati još vremena da dobijemo točnije podatke o očekivanju i standardnoj devijaciji za šestomjesečno razdoblje.

7.5 Bayesovsko zaključivanje za očekivanje normalne razdiobe

Središnji granični teorem nam kaže da suma velikog broja slučajnih varijabli često približno slijedi normalnu razdiobu. Posebno, ako neku veličinu promatramo na velikom broju uzoraka neke populacije, srednja vrijednost te veličine u uzorku će na uzorcima fiksne veličine biti približno normalno distribuirana, i aproksimacija će biti tim bolja, čim su uzorci veći. Ovo je razlog velike popularnosti i učestalosti korištenja normalne distribucije u praksi.

U ovom slučaju imamo

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2},$$

pa ako za apriornu distribuciju parametra θ (koji u ovom slučaju predstavlja očekivanu vrijednost srednje vrijednosti promatrane veličine u uzorku) uzmemo normalnu $N(m, s^2)$ distribuciju:

$$\pi(\theta) \sim e^{-\frac{1}{2}\left(\frac{\theta-m}{s}\right)^2},$$

Bayesovsko zaključivanje sada daje

$$\begin{aligned}\pi(\theta|x) \sim f(x|\theta)\, pi(\theta) &\sim e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2} \cdot e^{-\frac{1}{2}\left(\frac{\theta-m}{s}\right)^2} \\ &= e^{-\frac{1}{2}\left[\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta-m)^2}{s^2}\right]} \\ &\sim e^{-\frac{1}{2}\cdot\frac{s^2+\sigma^2}{s^2\sigma^2}\left[\theta-\frac{s^2x+\sigma^2m}{s^2+\sigma^2}\right]^2},\end{aligned}$$

tj. $N \sim \left(\frac{s^2x+\sigma^2m}{s^2+\sigma^2}, \frac{s^2\sigma^2}{s^2+\sigma^2}\right)$ aposteriornu distribuciju.

Valja još napomenuti da kod ovog zaključivanja standardnu devijaciju σ smatramo poznatom te nastojimo ažurirati naše uvjerenje o očekivanju promatrane normalne razdiobe. Također, ako umjesto jednog mjerenja/promatranja x, imamo cijeli uzorak $x_1, x_2, \cdots, x_n$, možemo uzorak zamijeniti s njegovom srednjom vrijednošću $\overline{x}$, uz efektivnu varijancu $\frac{\sigma^2}{n}$.

Npr. ako je naše uvjerenje o prosječnoj visini ljudi u nekoj državi 165 cm ± 5 cm, to možemo opisati apriornom $N(165, 5^2)$ distribucijom ($m = 165$, $s = 5$). Ako zatim izmjerimo 1000 slučajno odabranih stanovnika te zemlje, i na tom uzorku dobijemo srednju vrijednost $x = 168$, a poznato je da je standardna devijacija za visinu ljudi približno 9.5 cm pa za efektivnu varijancu srednje vrijednosti uzorka imamo $\sigma^2 = \frac{9.5^2}{1000} \approx 0.09$, računamo

$$\frac{s^2x + \sigma^2m}{s^2 + \sigma^2} = \frac{5^2 \cdot 168 + 0.09 \cdot 165}{5^2 + 0.09} \approx 167.99\,,$$

$$\frac{s^2\sigma^2}{s^2 + \sigma^2} = \frac{5^2 \cdot 0.09}{5^2 + 0.09} \approx 0.09\,,$$

čime smo dobili $N(167.99, 0.09)$ aposteriornu distribuciju, tj. ažurirali smo naše uvjerenje o prosječnoj visini ljudi u toj državi na 167.99 cm ± $\sqrt{0.09}$ cm = 167.99 cm ± 0.3 cm.

Poglavlje 8

Usporedba Bayesove i frekvencijske statistike

Kao što je navedeno i u uvodu, Bayesova statistika je zahvaljujući pojavi računala doživjela procvat krajem prošlog stoljeća. Drukčijeg je pristupa od onih koji se mogu naći u školskim i fakultetskim udžbenicima i zbog toga zaslužuje dodatnu pažnju. Razvila se toliko da na razini svjetske statistike danas razlikujemo dvije skupine zastupnika: Bayesovce i frekvencioniste (zastupnike standardne statistike). U standardnoj statistici vjerojatnost se računa samo na osnovi dobivenih podataka, dok je u Bayesovoj statistici kao takvoj pogled na vjerojatnosti subjektivan. Tako apriorne distribucije formalno izražavaju stupanj osobnog vjerovanja, a jedna od glavnih kritika Bayesove statistike je usmjerena upravo prema proizvoljnosti izbora apriorne distribucije. Kao alternativu subjektivnom pristupu imamo objektivna pravila za određivanje distribucija. Standardna statistika osmišljena je da se primijenjuje u izoliranom okruženju, dok s Bayesovom statistikom sami možemo kombinirati zaključke drugih s novim podacima, te na taj način dobivamo nove zaključke (objektivne u onoj mjeri u kojoj su i oni na koje se oslanjamo). S obzirom na eksploziju podataka i sve moćnija računala, vjeruje se kako je budućnost statistike u Bayesovom pristupu.

Konačno, navedimo prednosti koje sa sobom nosi Bayesovski pristup statističkog zaključivanja:

- Bayesovske metode omogućuju da se apriorna informacija formalno iskoristi pri zaključivanju,
- rezultati dobiveni Bayesovskom metodom lakši su za interpretaciju ljudima kojima statistika nije specijalnost, već se njome samo služe kao alatom,
- zaključak ovisi o odabranom uzorku,

- Bayesova analiza zasniva se isključivo na aposteriornoj raspodjeli,
- odgovori na postavljena pitanja slijede direktno iz osnova Bayesove analize.

Bibliografija

[1] M. Benšić, N. Šuvak, *Primijenjena statistika*, Osijek, 2013.

[2] W. M. Bolstad, *Introduction to Bayesian statistics*, A John Wiley Sons inc. publication, Hoboken, New Jersey, 2004.

[3] V. Čuljak, *Vjerojatnost i statistika (skripta)*, Građevinski fakultet Sveučilišta u Zagrebu, Zagreb, 2011.

[4] A. Jazbec, *Osnove statistike*, Šumarski fakultet Sveučilišta u Zagrebu, Zagreb, 2008.

[5] D. J. Maširević, *Primjena Bayesove formule i algoritamskog pristupa Bayesovoj formuli na situacijama iz svakodnevnog života*, Hrvatski matematički elektronički časopis, 2009.

[6] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[7] *Uvod u matematičku statistiku*, dostupno na `http://matematika.fkit.hr/novo/statistika_i_vjerojatnost/predavanja/1%20-%20Deskriptivna%20statistika.pdf`

Sažetak

Tema ovog rada uvod je u Bayesovsko statističko zaključivanje. Nakon definicija osnovnih pojmova iz područja statistike i vjerojatnosti, naveden je i dokazan Bayesov teorem te je prikazana njegova uporaba na nekoliko primjera. Navedeni su osnovni elementi Bayesovskog zaključivanja, dana su objašnjenja pojmova apriorne i aposteriorne distribucije te je dana definicija funkcije vjerodostojnosti, a objašnjena je i njezina važnost. Navedene su različite distribucije koje su potkrijepljene primjerima, a navedeni su i primjeri Bayesovskog zaključivanja za binomnu, Poissonovu te normalnu razdiobu. Konačno, navedena je i usporedba Bayesove i klasične frekvencijske statistike.

Summary

The topic of this graduate thesis is the introduction to Bayesian statistical inference. After the definitions of basic terms in the field of statistics and probability are given, the Bayesian theorem is proved and its use is shown on several examples. In the following chapters, the basic elements of Bayesian reasoning are given along with concept explanations of prior and posterior distributions. The credibility function is defined and its overall importance is explained. Different distributions are defined and explained along with examples, and different examples of Bayesian inference for binomial, Poisson and normal distribution are presented. The final chapter offers a comparison of Bayesian and classical frequency statistics.

Životopis

Rođena sam 22. siječnja 1993. godine u gradu Bad Hersfeldu u Njemačkoj. Osnovnu školu Tina Ujevića u Zagrebu završila sam 2007. godine nakon koje sam svoje obrazovanje nastavila u XVIII. gimnaziji u Zagrebu. Pohađala sam dvojezični program u kojemu sam nastavu pratila na njemačkom jeziku. Srednju sam školu završila s izvrsnim uspjehom, a položila sam i DSD ispite za C1 razinu njemačkog jezika. Prirodoslovno-matematički fakultet u Zagrebu upisujem 2011. godine te stječem razinu bakalara edukacije nastave matematike. Nakon preddiplomskog studija odlazim u Sjedinjene Američke Države gdje radim u struci, a zatim se 2017. godine vraćam u Osijek gdje polažem ispite za pedagoško-psihološko-didaktičko-metodičku izobrazbu na Filozofskom fakultetu. Iste godine upisujem diplomski studij na Prirodoslovno-matematičkom fakultetu u Zagrebu te stječem razinu magistre edukacije nastave matematike.