Sadržaj

Predgovor

Svrha je ove knjige na sažet i pregledan način izložiti osnovne ideje i metode poslovne statistike studentima biotehničkih znanosti. Pisana je prema bilješkama s predavanja koje je prvi autor držao proteklih nekoliko godina za studente Agroekonomskog smjera na Agronomskom fakultetu u Zagrebu.

Prvi dio knjige posvećen je uvodu u deskriptivnu statistiku. Naglasak je stavljen na razvijanje osnovne statističke kulture kroz razumijevanje osnovnih pojmova, prepoznavanje i klasificiranje različitih tipova podataka, te njihovo prikazivanje. Obrađene su i osnovne numeričke deskriptivne mjere te ukazano na razne mogućnosti manipulacije pri prikazivanju i interpretaciji statističkih podataka.

Drugi dio knjige bavi se inferencijalnom statistikom. Počinje poglavljem u kojem su neformalno izloženi osnovni elementi teorije vjerojatnosti. Sljedeća dva poglavlja obrađuju najznačajnije i najzastupljenije tipove diskretnih i kontinuiranih slučajnih varijabli. Poglavlje u kojem se obrađuje problem izbora uzorka te odnosa između parametara populacije i statistika uzorka ključno je za razumijevanje i pravilnu primjenu metoda izloženih u poglavljima o intervalnim procjenama i o testiranju hipoteza.

Inferencijalna statistika je u biti matematička disciplina. Unatoč tome, sveli smo matematičke formalnosti u njenom izlaganju na najmanju moguću mjeru. Učinili smo to kako bi se skriptom mogli služiti i čitatelji sa srednjoškolskim predznanjem.

Iako je knjiga ponajviše namijenjena studentima agronomije, nismo se posebno trudili zaodjenuti sve primjere u agronomsko ruho. Primjerima iz drugih područja željeli smo naglasiti da su koncepti izloženi u ovoj knjizi općeniti i da se na njih mogu svesti naizgled vrlo različiti problemi.

Bit ćemo zahvalni svim čitateljima koji nam ukažu na propuste i pogreške u knjizi.

U Zagrebu, rujna 2008

T. Došlić
D. Vrgoč

Poglavlje 1

Osnovni pojmovi

Statistika je skup metoda koje se koriste za prikupljanje, analizu, prikaz i interpretaciju podataka, te za donošenje odluka.

Donošenje odluka uz neodređenosti.

Statistika:
- teorijska, matematička: razvoj, izvod i dokaz statističkih formula, teorema i metoda
- primijenjena: primjena ovog gore na probleme iz života

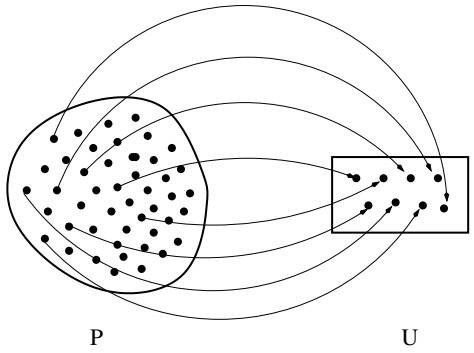
Statistika: - deskriptivna (opisna)
- inferencijalna

Deskriptivna (opisna) statistika je skup metoda za organiziranje, predočavanje i opis podataka koristeći tablice, grafikone i zbirne (skupovne) mjere.

Inferencijalna statistika se sastoji od metoda koje na temelju uzorka pomažu u donošenju odluka ili u izvođenju zaključaka o cijeloj populaciji.

Populacija (ciljna populacija) je skup svih elemenata - osoba, objekata ili stvari čija svojstva proučavamo.

Uzorak je dio populacije odabran za proučavanje.



Pregled je prikupljanje informacija o elementima populacije. Za žive ljude se zove i **anketa**. Ako pregled dohvaća sve elemente populacije, onda je to **popis** ili **cenzus**.

Uzorak je **reprezentativan** ako vjerno predstavlja karakteristike populacije.

Uzorak je **slučajan** ako je izabran na takav način da je svaki element populacije imao šansu biti izabran.

Izbor uzorka: - s vraćanjem
- bez vraćanja

Element ili **član** populacije ili uzorka je subjekt ili objekt o kojem se prikuplja informacija.

Varijabla je svojstvo koje se proučava i koje poprima razne vrijednosti za razne elemente.

Opažanje (mjerenje) je vrijednost varijable za pojedini element.

Skup podataka je kolekcija opažanja jedne ili više varijabli.

Primjer :

Krava	Količina mlijeka	Starost	Težina	Boja
Šarenka	3600	4	620	Bijela
Malenka	4200	5	660	Crna
Goluba	3800	4	712	Žuta
Rumenka	3200	6	580	Crvena

1.1 Tipovi varijabli

Varijable: - kvalitativne (kategoričke)
- kvantitativne (količinske): - diskretne
- kontinuirane

Kvalitativna varijabla je varijabla koja ne može poprimati brojčane vrijednosti, no može biti razvrstana u dvije ili više nebrojčanih kategorija. Podatci o takvoj varijabli su **kvalitativni podatci**.

Primjeri: Rod, boja dlake, marka traktora,...

Kvantitativna varijabla je varijabla koja poprima brojčanu (numeričku) vrijednost. Podatci o takvoj varijabli su kvantitativni podatci.

Kvantitativna varijabla čije vrijednosti su prebrojive, tj. koja može poprimiti strogo odvojene vrijednosti (bez međuvrijednosti) zove se **diskretna**. Primjeri su broj osoba, kuća, životinja,...

Kvantitativna varijabla koja može (načelno) poprimiti svaku vrijednost iz nekog intervala zove se **kontinuirana** ili **neprekidna**. Primjeri su masa, temperatura, zarada...

1.2 Skale mjerenja

1. **Nominalna skala**

   Primjenjuje se na podatke koji su razvrstani u različite kategorije u svrhu identifikacije.

   Primjeri su imena, marke, bračno stanje,... Tim se kategorijama mogu pridružiti numeričke vrijednosti, no njih nema smisla uspoređivati, niti s njima računati.

2. **Ordinalna skala**

   Primjenjuje se na podatke koji se razvrstavaju u kategorije koje se mogu rangirati, tj. poredati.

   Primjer - rangiranje izvrstan, zadovoljavajući, loš. Mogu im se pridijeliti brojčane vrijednosti koje ima smisla uspoređivati, no ne i računati s njima.

3. **Intervalna skala**

   Primjenjuje se na podatke koji se mogu rangirati i za koje se može izračunati i interpretirati razlika.

   Primjer - temperatura, IQ,...

4. **Omjerna skala**

   Primjenjuje se na podatke koji se mogu rangirati i za koje imaju smisla sve osnovne aritmetičke operacije.

   Primjeri su masa, cijena, zarada,...

Poglavlje 2

Organizacija i prikazivanje podataka

2.1 Sirovi podatci

Podatke zabilježene onako kako su prikupljeni i prije bilo kakve obrade zovemo **sirovi podatci**.

Primjer:

Student	Županija rođenja
A.B	Zagrebačka
C.A	Koprivničko-križevačka
A.A	Splitsko-dalmatinska
B.M	Koprivničko-križevačka
Ž.K	Krapinsko-zagorska
C.D	Koprivničko-križevačka
M.O	Ličko-senjska

2.2 Kvalitativni podatci

2.2.1 Frekvencijska raspodjela

Lista svih kategorija s brojem elemenata koji pripadaju pojedinoj kategoriji zove se frekvencijska raspodjela.

Primjer od gore: Varijable - županija rođenja
Kategorija - Koprivničko-križevačka
Frekvencija - 3

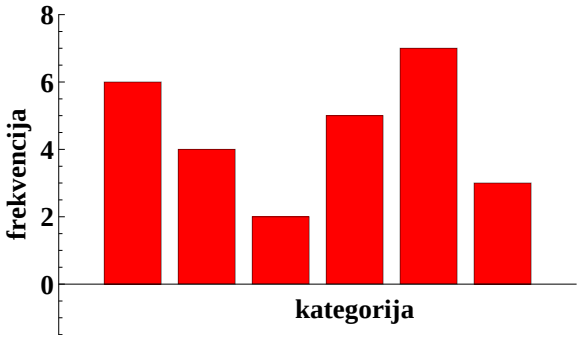
Kako se radi za male populacije/uzorke: 𝍸 𝍸 |||

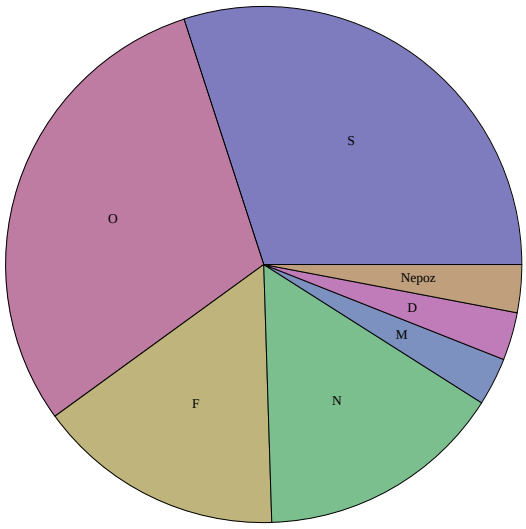
2.2.2 Relativna frekvencija i postotna raspodjela

Relativna frekvencija (učestalost) neke kategorije je omjer učestalosti te kategorije i zbroja svih frekvencija.

Postotak je relativna frekvencija puta 100.

2.2.3 Grafički prikaz kvalitativnih podataka





Prikaz na prvoj slici nazivamo **stupčasti grafikon** (engleski bar chart), a prikaz na drugoj slici **kružni grafikon** (engleski pie chart).

2.3 Kvantitativni podatci

2.3.1 Frekvencijska raspodjela

Kvantitativni podatci mogu biti grupirani ili negrupirani.

Grupirati - kada, kako i zašto?

Primjer:

broj automobila po kućanstvu - ne grupiramo
ukupni prihod kućanstva - grupiramo da ne bi sve vrijednosti imale učestalost 1
starost stanovništva - grupiramo jer nema velike razlike između 44 i 45 godina

---

Primjer : Mjesečna zarada zaposlenika u nekom poduzeću:

Mjesečna zarada	Broj radnika
1 - 2000	16
2001 - 3000	38
3001 - 4000	66
4001 - 5000	70
5001 - 6000	41
6001 - 7000	18
7001 - 8000	1

U ovom primjeru varijabla je mjesečna zarada, dok je učestalost broj radnika.

Treću kategoriju (3001 - 4000) nazivamo treća grupa ili treći razred, te možemo iščitati da je učestalost treće grupe 66.

U našem primjeru su granice sedme grupe 7001 (donja granica) i 8000 (gornja granica).

Rubovi trećeg razreda: 3000.5 - 4000.5

Sredina trećeg razreda: 3500.5

Širina trećeg razreda: 1000

Izračunavanje širine razreda:

$$\frac{\text{najveća vrijednost - najmanja vrijednost}}{\text{broj radnika}}$$

(Ovo je približna širina).

Svaki zgodni broj manji ili jednak od najmanje vrijednosti može poslužiti kao polazna vrijednost raspodjele.

2.3.2 Relativna frekvencija i postotna raspodjela

Relativna frekvencija i postotak definiraju se slično kao i za kvalitativne podatke. Za grupirane podatke je

$$\text{relativna frekvencija razreda} = \frac{\text{frekvencija razreda}}{\text{zbroj svih frekvencija}}.$$

Postotak razreda je njegova relativna frekvencija puta sto.

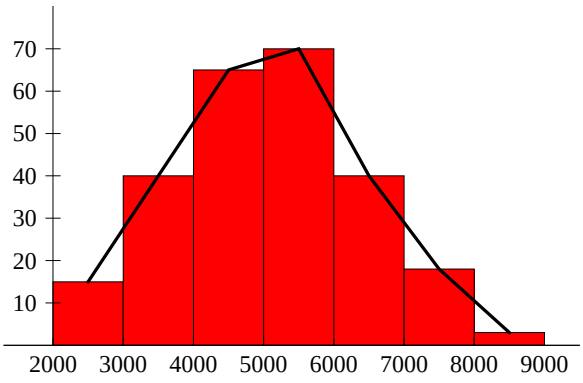
Mjesečna zarada	Broj radnika	Relativna frekvencija	Postotak
1 - 2000	16	0.064	6.4
2001 - 3000	38	0.152	15.2
3001 - 4000	66	0.264	26.4
4001 - 5000	70	0.280	28.0
5001 - 6000	41	0.164	16.4
6001 - 7000	18	0.072	7.2
7001 - 8000	1	0.004	0.4

2.3.3 Grafički prikaz grupiranih podataka

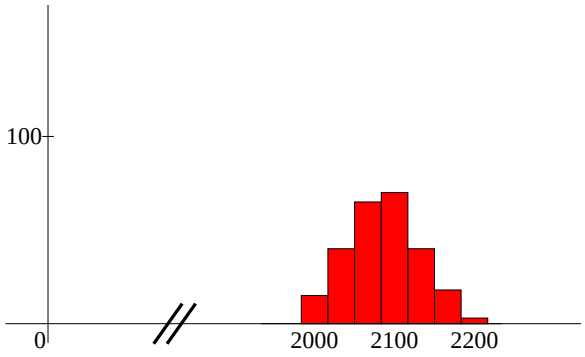
2.3.3.1 Histogram

Histogram može biti frekvencijski, relativno frekvencijski ili postotni.

Frekvencije (relativne frekvencije, postotci) su prikazani kao visine pravokutnika čije su osnovice na horizontalnoj osi određene granicama ili rubovima razreda. Pravokutnike crtamo jedan uz drugoga.



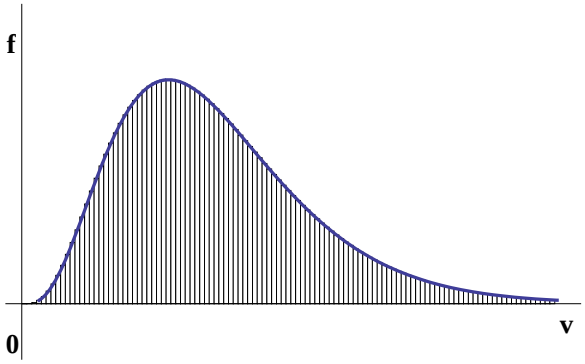
Mogli smo stvar nacrtati i s ishodištem u $(0, 0)$.



2.3.3.2 Poligon

Spojimo sredine (polovišta) gornjih stranica uzastopnih pravokutnika u histogramu ravnim crtama i dobijemo **poligon**. Crna crta na gornjoj slici.

Za velike skupove podataka poligon prelazi u krivulju frekvencijske raspodjele ili u **frekvencijsku krivulju**.



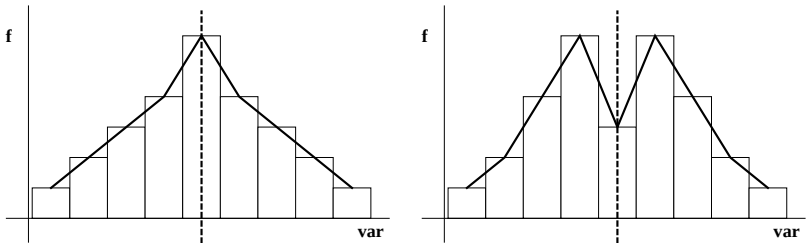
2.4 Oblici raspodjela

Pod oblikom raspodjele podrazumijevamo oblik pripadnog histograma, poligona ili frekvencijske krivulje.

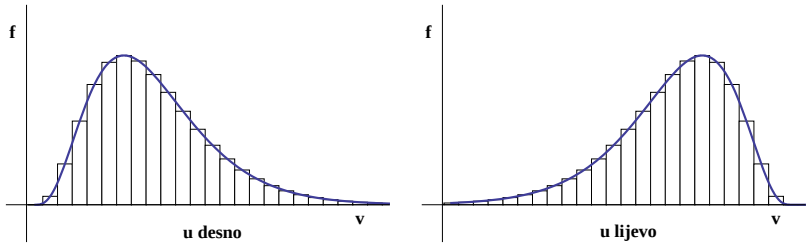
Najčešći oblici su:

- simetričan
- ukošen (nagnut)
- pravokutan ili uniforman.

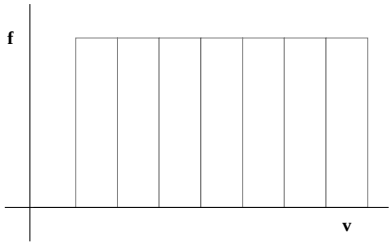
Simetričan oblik



Kosi oblik: rep na jednu stranu je dulji od repa na drugu.

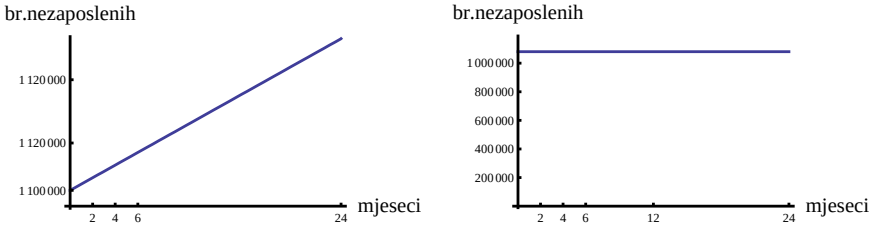


Uniformni oblik



2.5 Skale na osima i manipulacija

Primjer: Rast nezaposlenosti za vrijeme mandata neke vlade.



Isti podatci prikazani su u pro-vladinim i oporbenim novinama. Koje su koje?

2.6 Kumulativna frekvencijska raspodjela

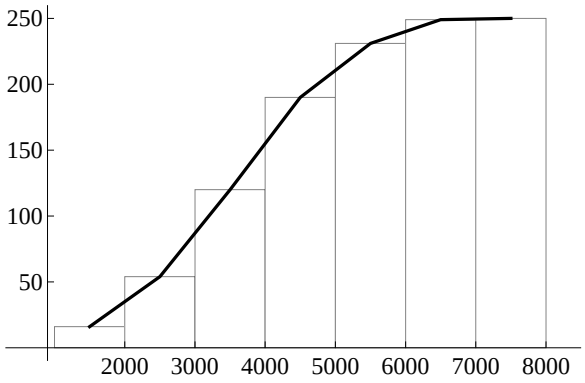
Kumulativna frekvencijska raspodjela daje ukupan broj vrijednosti koje padaju ispod gornjeg ruba (granice) svake klase.

Slično se definira i kumulativna relativna frekvencija i kumulativni postotak.

Primjer s poduzećem

Mjesečna zarada	Broj radnika	kum. r. f.	kum. %
≤ 2000	16	0.056	6.4
≤ 3000	54	0.216	21.6
≤ 4000	120	0.480	48.0
≤ 5000	190	0.760	76.0
≤ 6000	231	0.924	92.4
≤ 7000	249	0.996	99.6
≤ 8000	250	1.00.	100.0

Histogram, poligon i krivulja kumulativne frekvencijske distribucije definiraju se na analogan način.



2.7 Vremenski nizovi (slijedovi)

Podatci skupljeni za više elemenata u isto vrijeme ili za isto vremensko razdoblje zovu se **presječni podatci** (engleski cross-section data).

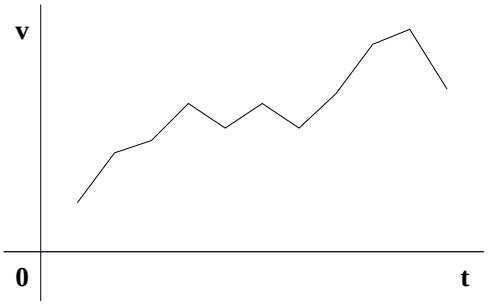
Primjer: Broj članova poljoprivrednih zadruga u Zagrebačkoj županiji.

Podatci prikupljeni za isti element i istu varijablu u različita vremena čine **vremenski niz**.

Primjer: Broj članova jedne te iste poljoprivredne zadruge u svakoj od 30 godina 1976. - 2005. (recimo 31. XII. svake godine).

U vremenskom slijedu imamo dvije relevantne varijable: broj članova i vrijeme.

Grafički podatci vremenskih slijedova mogu biti manipulirani kao grafički podatci drugih tipova varijabli. Za prikaz vremenskog slijeda obično rabimo po dijelovim linearnu funkciju.



Na horizontalnoj osi je vrijeme, na vertikalnoj je vrijednost varijable, a ne njena frekvencija.

Primjeri: kretanje cijena dionica,...

Poglavlje 3

Numeričke deskriptivne mjere

3.1 Mjerenje centralne tendencije za negrupirane podatke

3.1.1 Srednja vrijednost

Srednja vrijednost, prosjek ili aritmetička sredina definira se izrazom

$$\text{srednja vrijednost} = \frac{\text{zbroj svih vrijednosti}}{\text{broj vrijednosti}}.$$

Razlikujemo srednju vrijednost za populaciju i za uzorak.

$$\mu = \frac{\sum x}{N} \text{ za } \underline{\text{populaciju}} \text{ s N elemenata}$$

$$\overline{x} = \frac{\sum x}{n} \text{ za } \underline{\text{uzorak}} \text{ s n elemenata}$$

Kada bi histogram bio materijalni objekt, srednja vrijednost bi bila apscisa njegovog težišta.

Srednja vrijednost je osjetljiva na ekstremne vrijednosti - engleski termin je **outlier**. U standardnoj literaturi nismo našli dobar hrvatski prijevod, pa predlažemo ovdje riječ **odskočnik**. To ukazuje na podatak čija vrijednost odskače od ostalih.

Primjer: Na otoku živi 100 seljaka s po 1 jutrom zemlje i 1 veleposjednik s 1000 jutara. Prosječna veličina posjeda na tom otoku je nešto manja od 11

jutara. (Točna vrijednost je $\frac{1100}{101}$ jutara.) Agronomska savjetodavna služba koja bi raspolagala samo tim podatkom mogla bi donijeti preporuke koje ne bi bile prikladne niti za posjede od 1 jutra, niti za onaj od 1000 jutara.

U gornjem primjeru veleposjednik bi bio odskočnik.

3.1.2 Medijan

Medijan je vrijednost srednjeg člana u skupu podataka svrstanom po uzlaznim (rastućim) vrijednostima.

Ako je broj podataka neparan, imamo srednji član. Ako je broj podataka paran, izračunamo prosjek vrijednosti dva srednja člana.

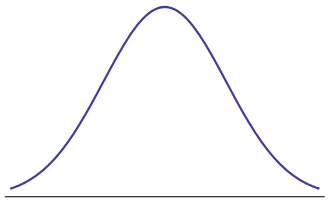
Medijan nije osjetljiv na odskočnike. Medijan daje položaj centra histograma, s polovicom vrijednosti lijevo i polovicom vrijenosti podataka desno od medijana.

3.1.3 Mod

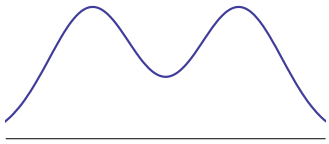
Mod je vrijednost koja se najčešće pojavljuje u skupu podataka.

Primjer : Ocjene studenta iz UPM.

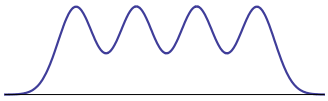
Glavni problem je da raspodjela ne mora imati mod, ili da ih može imati više.



Unimodalna



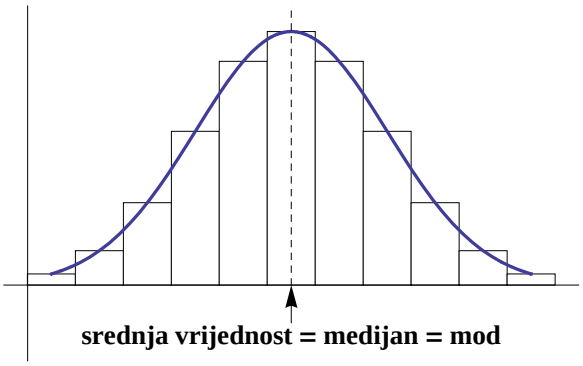
Bimodalna



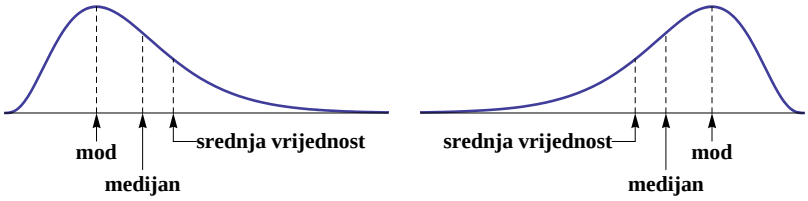
Multimodalna

3.1.4 Odnos među mjerama centralne tendencije

1. Za simetričnu i unimodalnu raspodjelu srednja vrijednost, medijan i mod se podudaraju i padaju u centar distribucije.



2. Za kosu unimodalnu raspodjelu medijan je uvijek između srednje vrijednosti i moda. Mod vidimo gdje je, a srednja vrijednost je uvijek na onoj strani moda na kojoj je dulji rep.



3.2 Mjere disperzije za negrupirane podatke

3.2.1 Raspon

Raspon je razlika najveće i najmanje vrijednosti u skupu podataka. Raspon je jako osjetljiv na odskočnike. Osim toga, daje informaciju izvučenu samo iz dviju krajnjih vrijednosti.

3.2.2 Varijanca i standardna devijacija

Standardna devijacija je najčešće korištena mjera disperzije. Ona kaže koliko su blizu vrijednosti skupljene oko srednje vrijednosti. Definira se kao korijen iz **varijance**.

Razlikujemo varijancu (pa i standardnu devijaciju) za populaciju i za uzorak.

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \text{ varijanca populacije}$$

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1} \text{ varijanca uzorka}$$

Standardna devijacija je korijen iz varijance:

$$\sigma = \sqrt{\sigma^2},\ s = \sqrt{s^2}.$$

Varijanca se računa kao srednja vrijednost kvadrata minus kvadrat srednje vrijednosti; za uzorak je malo drukčije:

$$\sigma^2 = \frac{\sum x^2}{N} - (\frac{\sum x}{N})^2; \qquad s^2 = \frac{\sum x^2}{n-1} - \frac{(\sum x)^2}{n(n-1)}$$

Razlog za drukčiju formulu za s^2 je da varijanca uzorka podcjenjuje varijancu populacije ako se uzme n^2 u nazivniku, a ne podcjenjuje ako se uzme $n(n-1)$.

Varijanca je srednja vrijednost kvadrata odstupanja. Uzima se kvadrat jer bi se za simetrične raspodjele odstupanja poništavala.

Zato su varijance i standardne devijacije uvijek nenegativne.

Varijanca se izražava u kvadratima jedinica varijable - pitanje: što je to kvadratna kuna? Zato se rabi standardna devijacija.

3.2.3 Koeficijent varijacije

Koeficijent varijacije izražava standardnu devijaciju kao postotak srednje vrijednosti.

$$CV = \frac{\sigma}{\mu} \cdot 100\% \text{ za populaciju}$$

$$CV = \frac{s}{\overline{x}} \cdot 100\% \text{ za uzorak}$$

3.2.4 Terminologija - parametar i statistika

σ i μ su **parametri** populacije; $\overline{x}$ i s su **statistike** uzorka.

3.3 Srednja vrijednost, varijanca i standardna devijacija za grupirane podatke

3.3.1 Srednja vrijednost za grupirane podatke

Ako su podatci grupirani, ne znamo njihove pojedinačne vrijednosti.

$$\mu = \frac{\sum mf}{N} \text{ za populaciju}$$

$$\overline{x} = \frac{\sum mf}{n} \text{ za uzorak}$$

Ovdje je m je sredina, a f frekvencija razreda. Zbraja se po razredima.
Sada možemo izračunati srednju vrijednost plaće za poduzeće iz 2.3.1.

Mjesečna zarada	Broj radnika (f)	Sredina razreda (m)
1 - 2000	16	1000.5
2001 - 3000	38	2500.5
3001 - 4000	66	3500.5
4001 - 5000	70	4500.5
5001 - 6000	41	5500.5
6001 - 7000	18	6500.5
7001 - 8000	1	7500.5

$$\Longrightarrow \overline{x} = 4028.5$$

3.3.2 Varijanca i standardna devijacija za grupirane podatke

$$\sigma^2 = \frac{\sum f(m-\mu^2)}{N} \text{ za populaciju}$$

$$s^2 = \frac{\sum f(m-\overline{x})^2}{n-1} \text{ za uzorak}$$

$$\sigma^2 = \frac{\sum m^2 f}{N} - (\frac{mf}{N})^2; \qquad s^2 = \frac{\sum m^2 f}{n-1} - \frac{(\sum mf)^2}{n(n-1)}$$

Ovo gore je formula za računanje.
Ako se ponovo vratimo na primjer s poduzećem dobivamo sljedeće:

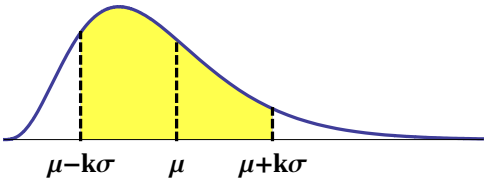
Mjesečna zarada	Broj radnika (f)	Sredina razreda (m)	$m - \overline{x}$
1 - 2000	16	1000.5	-3028
2001 - 3000	38	2500.5	-1528
3001 - 4000	66	3500.5	-528
4001 - 5000	70	4500.5	472
5001 - 6000	41	5500.5	1472
6001 - 7000	18	6500.5	2472
7001 - 8000	1	7500.5	3472

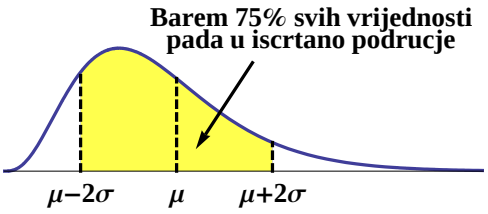
$$s^2 = 1928932.0996$$
$$s = 1388.86$$

3.4 Uporaba standardne devijacije

3.4.1 Čebiševljev teorem

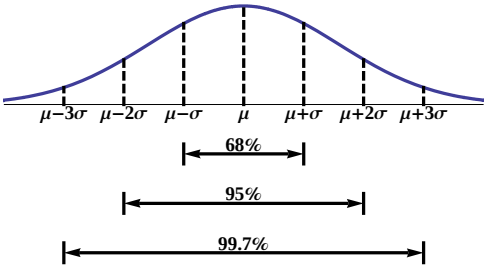
Čebiševljev teorem: Za svaki broj $k > 1$ barem $(1 - \frac{1}{k^2})$ svih vrijednosti podataka leži unutar k standardnih devijacija od srednje vrijednosti. ■





Čebiševljev teorem vrijedi i za populacije i za uzorke.

3.4.2 Empiričko pravilo za zvonolike raspodjele



Udruženje *Mensa* prima članove čiji je IQ veći od 130. To znači da su za više od 3σ udaljeni (u desno) od prosjeka. (Testovi inteligencije se normiraju tako da bude $\mu = 100$ i $\sigma = 30$.)

3.5 Mjere položaja

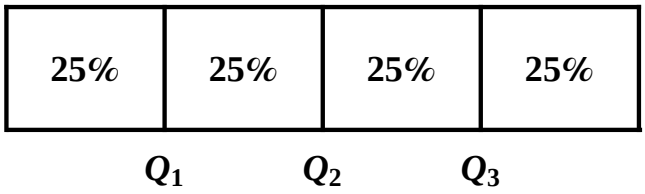
3.5.1 Kvartili

Kvartili su mjere koje dijele sortirane podatke u 4 jednaka dijela.

Označavamo ih s Q_1, Q_2, Q_3. Drugi kvartil je medijan, prvi kvartil je medijan onih koji su manji od medijana, treći kvartil je medijan onih koji su veći od medijana.

Kvartili su profinjenja medijana.

Razlika trećeg i prvog kvartila je međukvartilni raspon, $\sum QR = Q_3 - Q_1$.



3.5.2 Centili

Centili dijele sortirani skup podataka u 100 jednakih dijelova. Oznaka za k-ti centil je P_k, $k = 1, \ldots 99$.

$P_{25} = Q_1$, $P_{50} = Q_2$ =medijan.

P_k = vrijednost $\frac{kn}{100}$tog elementa u sortiranom skupu podataka.

Centilni rang vrijednosti x_i se računa kao omjer broja vrijednosti manjih od x_i i ukupnog broja vrijednosti, pomnoženog sa 100.

Poglavlje 4

Osnovni elementi teorije vjerojatnosti

4.1 Pokus, ishod i prostor događaja

Pokus je proces koji rezultira jednim i samo jednim od mnogo (više) mogućih opažanja. Ta opažanja zovemo **ishodima** pokusa, ili **događajima**. Prostor (skup) svih mogućih ishoda pokusa zovemo **prostor događaja**.

Primjer: Pokus - bacanje novčića
Ishod - kuna
Prostor događaja ={kuna, slavuj}

Primjer: Pokus - određivanje mjeseca rođenja
Ishod - listopad
Prostor događaja ={Si, Ve, Ož, Tr, Sv, Lip, Ko, Ru, Lis, St, Pr}

Primjer: Pokus - spol sudionika dvočlanog predstavništva
Ishod - MŽ
Prostor događaja ={ŽŽ, ŽM, MŽ, MM}

4.2 Jednostavni i složeni događaji

Jednostavni događaj se sastoji od jednog i samo jednog ishoda slučajnog pokusa.

Složeni događaj je kolekcija od više ishoda slučajnog pokusa.

Primjer: Bacanje kocke. "Pala je šestica" je jednostavan događaj. "Pao je paran broj" je složeni događaj. Drugi složeni događaj je "Pao je broj manji od 4".

4.3 Vjerojatnost; svojstva i pristupi

Vjerojatnost je numerička mjera "pojavljivosti" pojedinog događaja.

Promatramo slučajni pokus koji ima elementarne ishode E_i, $i = 1, \ldots, n$.

Vrijedi: $0 \leq P(E_i) \leq 1$

Nadalje, $\sum_{i=1}^{n} P(E_i) = 1$.

Za vjerojatnost događaja $\triangle$ vrijedi $0 \leq P(\triangle) \leq 1$.

Primjer: Bacanje kocke, bacanje novčića,...

Definicija vjerojatnosti je zadovoljavajuće formalno riješena tek 1930-ih godina - Kolmogorov.

Klasična definicija vjerojatnosti: $P(E_i) = \frac{1}{n}$

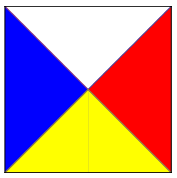
(geometrijska)

$$P(\triangle) = \frac{\text{broj ishoda povoljnih za } \triangle}{\text{broj mogućih ishoda}}.$$

Primjer: Bacanje kocke, $\mathrm{P}(1)=\mathrm{P}(2)=\ldots=\mathrm{P}(6)=\frac{1}{6}$

$$\mathrm{P}(\text{paran broj}) = \frac{1}{2}.$$

Primjer: Vjerojatnost padanja strelice pikada ispod obiju dijagonala kvadrata.



$$P(\triangle) = \frac{1}{4}.$$

$$P(\triangle) = \frac{\text{mjera područja povoljnog za događaj}}{\text{mjera cijelog područja}}.$$

Relativna frekvencija kao aproksimacija

$P(\triangle) = \frac{f}{n}$ relativna frekvencija je aproksimacija vjerojatnosti.

Zakon velikih brojeva: Ponavljanjem pokusa vjerojatnost iz relativne frekvencije teži prema teorijskoj (pravoj) vjerojatnosti.

Subjektivna vjerojatnost

Vjerojatnost pridjeljena događaju na temelju subjektivne procjene, iskustva, informacija, vjerovanja.

Mislim da je vjerojatnost da student agroekonomije položi statistiku na prvom roku jednaka 0.9.

4.4 Šanse ili izgledi

Pokušaj prijevoda engleskog termina **odds**.

"Šanse su jedan naprama milijun." Tisuću prema 1, itd.

4.5 Pravila prebrojavanja

Ako se pokus sastoji iz k koraka, i ako je broj mogućih ishoda u i-tom koraku jednak n_i, onda je ukupni broj mogućih ishoda jednak $n = \prod_{i=1}^{k} n_i$.

Primjer: Biramo između 2 juhe, 3 glavna jela i 4 deserta. Koliko različitih jelovnika možemo složiti? $2 \cdot 3 \cdot 4 = 24$.

Primjer: Na listiću sportske prognoze je 13 parova. Za svaki par postoje tri moguća ishoda. Na koliko se različitih načina može popuniti listić? $3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot = 3^{13}$.

Vjerojatnost dobitne kombinacije je $\frac{1}{3^{13}} = 3^{-13}$.

U slučaju da imamo n različitih objekata možemo ih poredati (razmjestiti) na $n! = 1 \cdot 2 \cdot \ldots \cdot n$ različitih načina.

Primjer: Koja je vjerojatnost da se slučajnim izvlačenjem iz šešira u kojem su papirići sa slovima A, B, E, G, R i Z dobije riječ ZAGREB?

P(ZAGREB)=$\frac{1}{720}$.

Iz n-članog skupa možemo izabrati k-člani podskup na $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ načina.

Ako je bitan i poredak, onda se to može na $n \cdot (n-1) \cdot, \ldots \cdot (n-k+1)$ načina.

Primjer: Obično izaslanstvo, izaslanstvo sa strukturom.

4.6 Uvjetna vjerojatnost

Primjer: Imamo uzorak od 100 ljudi, 60 muških i 40 ženskih. Od njih se traži odgovor na pitanje vole li ili ne vole određeni proizvod. Tablica s rezultatima.

S \ P	V	N	$\sum$
M	15	45	60
Ž	4	36	40
$\sum$	19	81	100

Uvjetna vjerojatnost događaja A uz uvjet B je vjerojatnost da se dogodi A ako se već dogodio B. Oznaka $P(A|B)$.

Iz tablice gore imamo: $P(V) = \frac{19}{100} = 0.19$

$$P(V|M) = \frac{15}{60} = 0.25$$

$$P(\check{Z}) = \frac{40}{100} = 0.4$$

$$P(\check{Z}|N) = \frac{36}{81} = \frac{4}{9} = 0.4444$$

4.7 Disjunktni događaji

Događaji su disjunktni (uzajamno se isključuju) ako ne mogu nastupiti istovremeno.

Primjer: Bacanje kocke, A=paran broj, B=neparan broj.

Primjer : $\left.\begin{array}{l} \text{A = slučajno odabrani student je muško} \\ \text{B = slučajno odabrani student je žensko} \end{array}\right\}$ isključuju se

Primjer: 30 studenata, 25 uči engleski, 12 uči njemački, 9 ih uči i engleski i njemački. Događaj A = student uči engleski i događaj B = student uči njemački nisu uzajmno isključivi; postoje studenti koji uče i engleski i njemački.

4.8 Nezavisni događaji

Događaji A i B su **nezavisni** ako jedan ne utječe na drugog, tj. $P(A|B) = P(A)$, ili $P(B|A) = P(B)$. Ako nisu nezavisni događaji su zavisni.

Iz tablice kod uvjetne vjerojatnosti vidimo da događaji Ž i V nisu nezavisni. Naime, P(Ž)=0.4, $P(\check{Z}|V)=\frac{4}{19} = 0.211$.

Važne napomene:

1. Događaji s pozitivnim vjerojatnostima su ili disjunktni ili nezavisni:
   a) Disjunktni događaji su uvijek zavisni.
   b) Nezavisni događaji nisu nikad disjunktni.
2. Zavisni događaji mogu i ne moraju biti disjunktni.

4.9 Komplementarni (suprotni) događaji

Suprotni događaj događaja A uključuje sve ishode koji nisu u A. Oznaka $\overline{A}$.

$$P(\overline{A}) = 1 - P(A)$$

Primjer: A = slučajno odabrani nazočni student je muško. Onda je $\overline{A}$ = slučajno odabrani nazočni student je žensko.

4.10 Presjek događaja

Događaj $A \cap B$ se sastoji od svih elementarnih događaja koji su i u A i u B.

Primjer: Bacanje kocke. A = paran broj, B = broj manji od 5. $A \cap B = \{2, 4\}$.

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

Za nezavisne događaje je $P(B|A) = P(B)$, pa je

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

Odavde,

$$P(A|B) = \frac{P(A \cap B)}{P(A)}, \text{ uz } P(A) \neq 0.$$

Za disjunktne događaje je $P(A \cap B) = 0$.

4.11 Unija događaja

Događaj $A \cup B$ uključuje sve ishode koji su u barem jednom od događaja A i B.

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

Primjer s jezicima: $P(A \cup B) = \frac{28}{30} = \frac{25 + 12 - 9}{30}$.

Za disjunktne događaje je

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

4.12 Bayesovi teoremi

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}.$$

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(A_1 \cap B) + P(A_2 \cap B)}.$$

Poglavlje 5

Diskretne slučajne varijable

Slučajna varijabla je varijabla čija je vrijednost određena ishodom slučajnog pokusa.

Slučajne varijable: - Diskretne - broj ovog ili onog: automobila, ljudi,...
- Kontinuirane - količina (mjera) vode, vremena,...

5.1 Vjerojatnostna raspodjela diskretne slučajne varijable

Vjerojatnostna rapodjela diskretne slučajne varijable je lista svih mogućih vrijednosti koje ta varijabla može poprimiti s pripadajućim vjerojatnostima.

Primjer: U mjestu s 2000 obitelji ispitano je koliko koja obitelj posjeduje mobitela. Frekvencijska raspodjela je dana sljedećom tablicom.

Broj mobitela	Frekvencija	Relativna frekvencija
0	30	0.015
1	470	0.235
2	850	0.425
3	490	0.245
4	160	0.080

Nasumično biranje jedne od 2000 obitelji iz tog mjesta je slučajni pokus. Vjerojatnost da slučajna varijabla "broj mobitela u obitelji" poprimi vrijednost $0 \leq i \leq 4$ je dana relativnom frekvencijom.

Broj mobitela	Vjerojatnost
0	0.015
1	0.235
2	0.425
3	0.245
4	0.080

Dakle vjerojatnostna raspodjela je pravilo koje svakoj vrijednosti koju slučajna varijabla može poprimati pridružuje vjerojatnost da se ta vrijednost i poprimi.

Imamo tablicu. Označimo li vjerojatnost da slučajna varijabla X poprimi vrijednost x_i s p_i, onda je svaki redak u toj tablici oblika x_i, p_i.

Neka slučajna varijabla X može poprimiti n vrijednosti, $x_1, x_2, \ldots, x_n$. Tada vrijedi:

$$0 \leq p_i \leq 1, i = 1, \ldots, n$$
$$\sum_{i=1}^{n} p_i = 1$$

$P(X = x_i) = p_i$.

U našem primjeru $P(X = 1) = 0.235$.

Iz vjerojatnostne raspodjele se mogu očitati i druge informacije. Zanima li nas vjerojatnost da slučajno odabrana obitelj posjeduje barem dva mobitela, pišemo to kao $P(X \geq 2)$. Iz raspodjele vidimo da je to jednako $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) = 0.425 + 0.245 + 0.08 = 0.75$.

U općem slučaju vjerojatnostna raspodjela može biti zadana **tablicom**, **grafom** ili **formulom**. To je stoga što je vjerojatnostna raspodjela funkcija koja vrijednostima slučajne varijable pridružuje njihove vjerojatnosti.

5.2 Funkcija distribucije

Vidjeli smo na primjeru da se vjerojatnostna raspodjela može izvesti iz tablice relativnih frekvencija. Alternativna karakterizacija diskretne slučajne varijable može se dati pomoću tablice kumulativnih relativnih frekvencija. Na taj način dobivamo funkciju distribucije.

Funkcija distribucije F diskretne slučajne varijable X dana je formulom $F(a) = P(x \leq a)$.

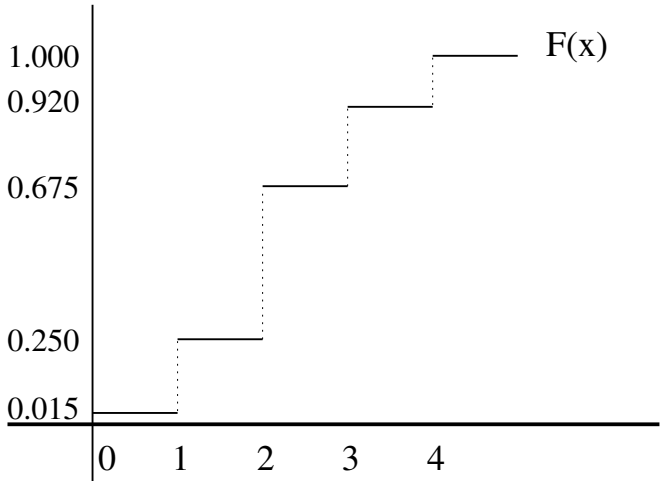
Za sve a vrijedi $0 \leq F(a) \leq 1$. Funkcija F je neopadajuća. Računa se iz vjerojatnostne raspodjele po formuli

$$F(a) = \sum_{x_i \leq a} p_i.$$

Za primjer s mobitelima imamo tablicu

x_i	p_i	$F(x_i)$
0	0.015	0.015
1	0.235	0.250
2	0.425	0.675
3	0.245	0.920
4	0.080	1.000

Za razliku od vjerojatnostne raspodjele koja je definirana samo za vrijednosti $a = x_i$, funkcija distribucije je definirana na cijelom skupu R. Između vrijednosti x_i i x_{i+1} funkcija f je konstantna i ima vrijednost $F(x_i)$. Za vrijednosti x_i funkcija distribucije ima skok koji iznosi p_i. Lijevo od najmanjeg x_i funkcija distribucije je jednaka nuli, a desno od najvećeg x_i ima vrijednost 1. Funkcija distribucije za primjer s mobitelima je prikazana na slici.



Ako je vjerojatnostna raspodjela diskretne slučajne varijable dana formulom,onda se i za funkciju distribucije te slučajne varijable često može naći formula kojom je zadana.

5.3 Očekivanje diskretne slučajne varijable

Očekivanje disketne slučajne varijable je očekivanje njene vjerojatnostne raspodjele. Za slučajnu varijablu X pišemo $E(X)$ i govorimo o **očekivanoj vrijednosti**. To je vrijednost koju dobijemo uprosječavanjem vrijednosti varijable opaženih u puno ponavljanja pokusa.

U primjeru s mobitelima dobije se da je očekivana vrijednost broja mobitela po obitelji jednaka 2.14. Što to znači? Sigurno ne da obitelji posjeduju sedmine mobitela. To znači da razne obitelji posjeduju razni broj mobitela, neke više, neke manje od očekivanog broja.

$\mu = E(X) = \sum x_i p_i$ - formula za računanje očekivane vrijednosti.

U primjeru s mobitelima, $E(X) = 0{\cdot}0.015+1{\cdot}0.235+2{\cdot}0.425+3{\cdot}0.245+4{\cdot}0.08 = 2.14$.

5.4 Standardna devijacija diskretne slučajne varijable

Standardna devijacija mjeri raspršenje vjerojatnostne raspodjele.

Označavamo ju sa σ. Mala vrijednost σ znači da je većina vrijednosti slučajne varijable grupirana oko očekivanja. Velika vrijednost σ znači da su vrijednosti razvučene preko većeg raspona.

Osnovna formula je:

$$\begin{aligned}\sigma = \sqrt{\sum[(x_i - \mu)^2 p_i]} &= \sqrt{\text{očekivanje kvadrata odstupanja}} \\ &= \sqrt{\text{prosječno kvadratno odstupanje}}\end{aligned}$$

Praktičnija za računanje je formula

$$\sigma = \sqrt{\sum x_i^2 p_i - \mu^2} = \sqrt{\text{očekivanje kvadrata - kvadrat očekivanja}}.$$

Varijanca σ^2 diskretne slučajne varijable je kvadrat standardne devijacije.

I ovdje vrijedi Čebiševljev teorem.

5.5 Binomna vjerojatnostna raspodjela

Binomna raspodjela se primjenjuje za opis situacija u kojima se pokus s dva moguća ishoda ponavlja više puta.

Binomni pokus:

1. Pokus se ponavlja n puta u identičnim uvjetima.
2. Svaki pokus ima točno 2 moguća ishoda (uspjeh ili neuspjeh).
3. Vjerojatnosti uspjeha (p) i neuspjeha (q) su konstantne.
4. Pokusi su nezavisni (ishod jednog ne utječe na druge).

"Uspjeh" i "neuspjeh" su termini koji ne znače da je jedan od ishoda nužno poželjan, a drugi nepoželjan.

Primjeri: Bacanje novčića, ponovljeno 10 puta.

Definicija: Slučajna varijabla koja predstavlja broj uspjeha u ponavljanju binomnog pokusa se zove **binomna slučajna varijabla** ili jednostavno binomna

varijabla. Za n ponavljanja pokusa i vjerojatnost uspjeha p, kažemo da je slučajna varijabla X binomna slučajna varijabla s parametrima n i p i pišemo $X \sim B(n,p)$.

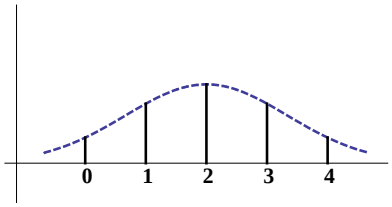
Vjerojatnost da od n ponavljanja k puta ishod bude "uspjeh" dana je **binomnom formulom**

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

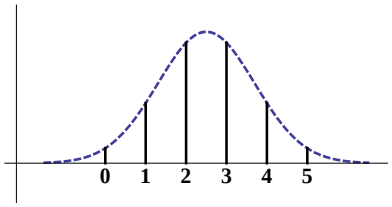
Ovdje je $q = 1-p$.

Binomni koeficijent $\binom{n}{k}$ broji načine kako je k uspješnih pokusa raspoređeno u ukupno n ponavljanja. Množenje vjerojatnosti je posljedica nezavisnosti (pretpostavljene u uvjetima) događaja, tj. ishoda.

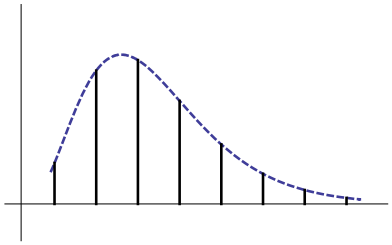
Oblik binomne distribucije je simetričan za $p = q = 0.5$. Za $p < 0.5$ raspodjela je ukošena u desno, a za $p > 0.5$ raspodjela je ukošena u lijevo.

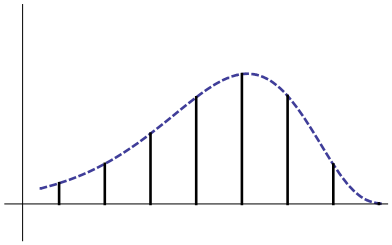


broj pojavljivanja kune u 4 bacanja simetričnog novčića



broj pojavljivanja kune u 5 bacanja simetričnog novčića





Digresija: Koliko puta treba ponoviti pokus da vjerojatnost uspjeha bude barem p_0?

Očekivanje i standardna devijacija binomne raspodjele dani su formulama

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

Primjer: Vjerojatnost da jaje nije mućak je 0.95. Koliki je očekivani broj pilića koji će se izleći u inkubatoru od 400 jaja? Kolika je standardna devijacija?

$\mu = 400 \cdot 0.95 = 380.$

$\sigma = \sqrt{400 \cdot 0.95 \cdot 0.05} = 20\sqrt{0.0475} = 20 \cdot 0.22 = 4.4.$

5.6 Geometrijska raspodjela

Promatramo binomni pokus s vjerojatnošću uspjeha p. Neka je X slučajna varijabla koja broji broj pokusa do prvog uspjeha. Kakva je njena raspodjela?

Iz pretpostavljene nezavisnosti pokusa slijedi da je vjerojatnost prvog uspjeha u k-tom izvođenju pokusa dana formulom $P(X = k) = (1 - p)^{k-1}p$. To je vjerojatnost da prvih $k-1$ izvođenja pokusa bude neuspješno i da k-to izvođenje bude uspješno.

Za takvu slučajnu varijablu X kažemo da ima **geometrijsku raspodjelu** s parametrom p i pišemo $X \sim G(p)$.

Očekivanje i standardna devijacija geometrijske slučajne varijable $X \sim G(p)$ dani su formulama

$$E(X) = \frac{1}{p}, \qquad \sigma(X) = \frac{\sqrt{1-p}}{p}.$$

Geometrijska slučajna varijabla ima svojstvo $P(X > k) = (1 - p)^k$. Odatle slijedi da je njena funkcija distribucije zadana s $F(a) = 1 - (1 - p)^{\lfloor a \rfloor}$, pri čemu $\lfloor a \rfloor$ označava najveći cijeli broj manji ili jednak od a.

Primjer Kolika je vjerojatnost da će cilj biti pogođen iz drugog pokušaja ako je vjerojatnost pogađanja cilja iz jednog pokušaja $p = 0.3$? Kolika je vjerojatnost da broj pokušaja ne će biti veci od 3?

5.7 Hipergeometrijska raspodjela

U slučajevima kada se iz konačne populacije bira uzorak bez vraćanja više ne vrijedi zahtjev nezavisnosti pokusa. Ako iz košare u kojoj je 20 ispravnih i 5 trulih jabuka izvučemo jednu, pa ju ili pojedemo ili bacimo, vjerojatnost izvlačenja trule jabuke u drugom pokušaju nije ista. U takvim se situacijama koristi hipergeometrijska raspodjela.

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

N - broj elemenata u populaciji
n - broj elemenata u uzorku
r - broj "uspjeha" u populaciji
k - broj "uspjeha" u uzorku

Očekivanje i varijanca dani su formulama

$$\mu = n\frac{r}{N},\ \sigma^2 = \frac{N-n}{N-1}\cdot\mu\frac{r}{N}(1-\frac{r}{N}).$$

Primjer Iz kokošinjca sa 7 bijelih i 3 crne kokoši lisica odnese 4. Kolika je vjerojatnost da je odnijela 2 crne kokoši? Najviše 2 crne?

5.8 Poissonova raspodjela

Znamo da se stroj kvari u prosjeku tri puta mjesečno. Želimo naći vjerojatnost da će se pokvariti točno dvaput tijekom sljedećeg mjeseca. Slično, znamo da u prodavaonicu dolazi u prosjeku 7 ljudi dnevno reklamirati kupljenu robu. Kolika je vjerojatnost da će ih sutra doći 5?

U ovakvim situacijama primjenjujemo Poissonovu raspodjelu.

Uvjeti za primjenu Poissonove raspodjele:

1. X mora biti slučajna varijabla.
2. Dešavanja su slučajna.
3. Dešavanja su nezavisna.

Stvari se uvijek gledaju u intervalima. Intervali mogu biti vremenski (broj pacijenata u satu), prostorni (broj prometnih nezgoda na određenoj cestovnoj dionici) ili volumni (broj neispravnih proizvoda među idućih 100 koji izlaze iz stroja).

Prosječan broj dešavanja u promatranom intervalu je **poznat**.

Ta se veličina označava s λ.

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ - vjerojatnost da se u promatranom intervalu desilo } k \text{ događaja.}$$

Očekivanje i varijanca Poissonove raspodjele su jednaki λ. Standardna devijacija je jednaka $\sqrt{\lambda}$.

Kao i za druge raspodjele, može se računati $P(X<k)$, $P(X\geq k)$, $P(k_1\leq X\leq k_2)$ i slično.

Intervali za λ i X moraju biti jednaki. Ako nisu, moramo redefinirati λ.

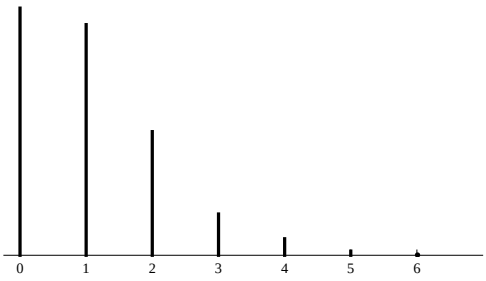
Primjer : Poznato je da, pri telefonskoj anketi, u prosjeku 2 od 10 nazvanih odbija odgovoriti na upit. Kolika je vjerojatnost da če od 20 nazvanih osoba njih 5 odbiti odgovoriti?

Želimo naći $P(X = 5)$. Znamo da je $\lambda = 4$, jer ako od 10 nazvanih u prosjeku 2 ne će odgovoriti, onda od 20 nazvanih u prosjeku 4 ne će odgovoriti. Dakle je $\lambda = 4$. Uvrštavanjem u formulu dobivamo $P(X = 5) = \frac{4^5 e^{-4}}{5!} = \frac{1024 \cdot 0.018316}{120} = 0.1563$.

Jedina informacija koja ulazi u račun je λ; to je jedini parametar raspodjele. Treba provjeriti uvjete primjenjivosti. Npr., znamo da na aerodrom slijeće u prosjeku 120 aviona dnevno, no njihovi dolasci nisu slučajni, već su po redu letenja. Može se desiti da ih je utorkom uvijek 75, a petkom 180, pa ne možemo primijeniti Poissonovu raspodjelu.

Primjer: Prodavač proda u prosjeku 0.9 automobila dnevno. Treba napisati vjerojatnostnu raspodjelu za broj prodanih automobila u danu, ako su zadovoljeni uvjeti Poissonove raspodjele.

X	$P(X = k)$
0	0.4066
1	0.3659
2	0.1647
3	0.0494
4	0.0111
5	0.0020
6	0.0003



Poglavlje 6

Kontinuirane slučajne varijable

6.1 Funkcija gustoće vjerojatnosti

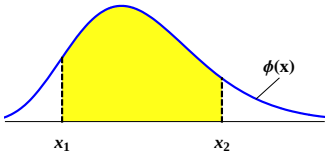
Kod kontinuirane slučajne varijable ne govori se o vjerojatnosti da ona poprimi točno određenu vrijednost. Govori se o vjerojatnosti da ona poprimi vijednost između određenih granica. Možemo reći da za kontinuiranu varijablu izraz $P(X = x_0)$ nema puno smisla.

Vjerojatnostna raspodjela kontinuirane slučajne varijable ima sljedeća svojstva:

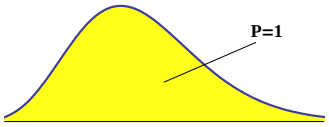
1. Vjerojatnost da X poprimi vrijednost u bilo kojem intervalu je između 0 i 1.

2. Ukupna vjerojatnost svih (disjunktnih) intervala u kojima se može naći X je jednaka 1.

Krivulja vjerojatnostne raspodjele kontinuirane slučajne varijable zove se još i **funkcija gustoće vjerojatnosti**.



$$0 \leq P(x_1 \leq X \leq x_2) \leq 1$$
$$P(x_1 < X < x_2)$$



$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} \varphi(x)dx$$

Za funkciju gustoće $\varphi(x)$ zadanu na intervalu $\langle a, b\rangle$, svojstvo 2 pišemo kao

$$\int_a^b \varphi(x)dx = 1.$$

6.2 Funkcija distribucije

Funkcija distribucije kontinuirane slučajne varijable s funkcijom gustoće $\varphi(x)$ definira se formulom

$$F(x) = \int_a^x \varphi(t)dt.$$

Funkcija $F(x)$ je rastuća funkcija s vrijednostima u intervalu $[0, 1]$.

Ako znamo $F(x)$, funkciju gustoće vjerojatnosti dobijemo deriviranjem, $\varphi(x) = F'(x)$. Nadalje, $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$.

6.3 Očekivanje i varijanca kontinuirane slučajne varijable

Za kontinuiranu slučajnu varijablu x s funkcijom gustoće $\varphi(x)$ definiramo njeno očekivanje μ i varijancu formulama

$$\mu = \int_a^b x\varphi(x)dx \quad \text{i} \quad \sigma^2 = \int_a^b (x-\mu)^2\varphi(x)dx.$$

Integrira se po cijelom intervalu $\langle a, b\rangle$ na kojem je zadana funkcija gustoće φ. Taj interval može biti i cijeli $\mathbb{R}$, tj. $\langle -\infty, \infty\rangle$.

Standardna devijacija σ definira se kao korijen iz varijance.

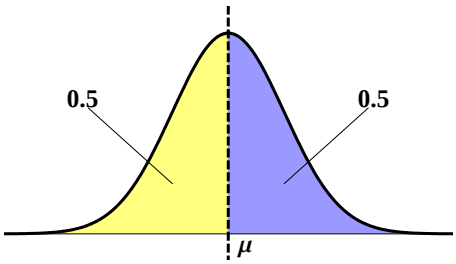
6.4 Normalna raspodjela

Normalna raspodjela je jedna od najvažnijih i najčešćih vjerojatnostnih raspodjela koje kontinuirana slučajna varijabla može imati. Po njoj se ravnaju mnoge slučajne varijable u prirodi i društvu: visina i težina ljudi, rezultati na ispitima, životni vijek uređaja, količina mlijeka u boci, vrijeme obavljanja nekog posla, itd.

Slučajna varijabla koja se ravna po normalnoj vjerojatnostnoj raspodjeli se zove **normalna slučajna varijabla**.

Svojstva normalne raspodjela:

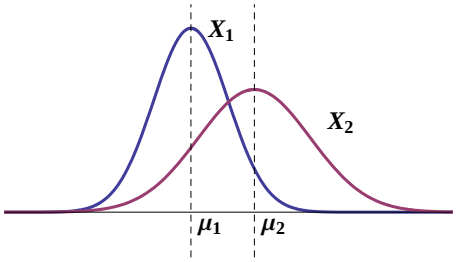
1. Karakterizirana je očekivanjem (srednjom vrijednošću) μ i standardnom devijacijom σ.
2. Krivulja je zvonolika i
   - površina pod njom je 1;
   - simetrična je oko μ;
   - repovi krivulje odlaze u beskonačnost.



Veličine μ i σ su **parametri** normalne raspodjele $X \sim N(\mu, \sigma)$

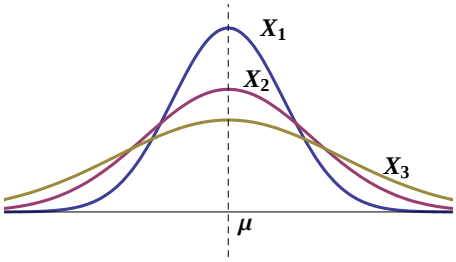
Ne postoji samo jedna normalna raspodjela, postoji mnoštvo normalnih raspodjela.

Vrijednost μ određuje apscisu maksimuma krivulje.



$\mu_2 > \mu_1$

Vrijedost σ određuje (mjeri) "razmazanost" krivulje.



$\sigma_3 > \sigma_2 > \sigma_1$

6.4.1 Standardna normalna raspodjela

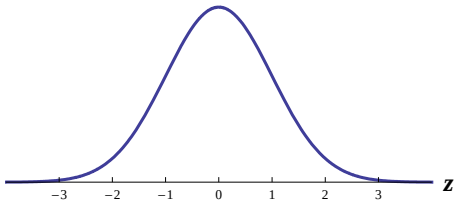
Normalna raspodjela s parametrima $\mu = 0$ i $\sigma = 1$ se zove **standardna normaln raspodjela**.

$$X \sim N(0,1).$$

Funkcija gustoće standardne normalne raspodjele dana je formulom

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

Funkcija distribucije standardne normalne varijable nije elementarna funkcija i ne može se izraziti jednostavnom formulom. Zove se **integral vjerojatnosti** i obično označava s $\Phi(x)$. Njene su vrijednosti tabelirane u statističkim tablicama.

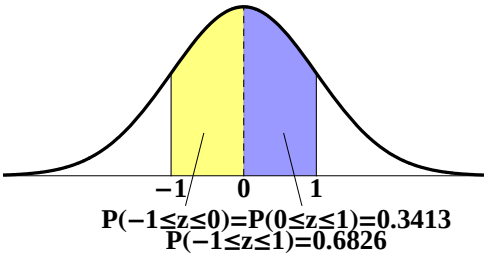


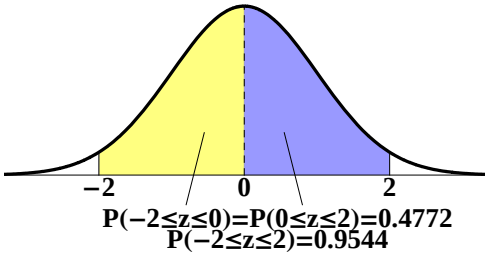
Za normalnu raspodjelu s parametrima μ i σ funkcija gustoće je

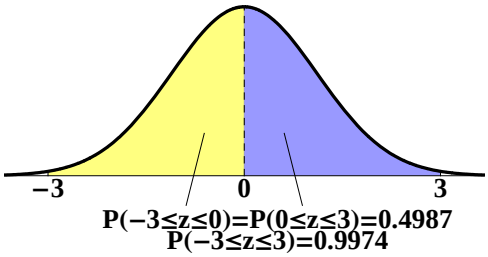
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x-\mu}{\sigma}^2}$$

Vrijednosti na osi apscisa standardne normalne raspodjele se obično označavaju sa z. Izražavaju se u jedinicama standardnih devijacija.

Izraz $z = 2$ označava da je točka na osi apscise za 2 standardne devijacije desno, tj. veća, od srednje vrijednosti.







Standardna normalna distribucija je tabelirana u statističkim tablicama. Njima se treba znati služiti.

Sve normalne raspodjele mogu se svesti na standardnu raspodjelu.

Postupak se zove **standardizacija**.

$$X \sim N(\mu, \sigma) \longrightarrow N(0,1)$$
$$z = \frac{x-\mu}{\sigma} \longleftrightarrow x = \mu + \sigma z$$

Primjer: $X \sim N(50, 10)$ $x = 35 \Longrightarrow$ $z = \frac{35-50}{10} = -1.5$

$x = 55 \Longrightarrow$ $z = \frac{55-50}{10} = 0.5$

Svi problemi s $X \sim N(\mu, \sigma)$ se prvo svode na $Z \sim N(0,1)$, a onda se primjenjuju tablice.

Problem: - iz poznatih x ili z vrijednosti odrediti površine pod krivuljom
- iz poznatih površina pod krivuljom odrediti x ili z vrijednosti

$$x \to z \longrightarrow \text{Tablice} \longrightarrow z \to x$$

6.4.2 Normalna aproksimacija binomne raspodjele

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Za velike n formule vezane uz $B(n,p)$ postaju nespretne. Ako je n velik i $p \sim 0.5$, onda je binomna raspodjela simetrična i liči na normalnu. U tim je slučajevima aproksimacija dobivena korištenjem normalne umjesto binomne raspodjele dobra. To vrijedi i za vjerojatnosti uspjeha koje nisu jako blizu 0.5 ako su i np i nq dovoljno veliki.

Empiričko pravilo:

Normalna raspodjela je dobra aproksimacija binarne ako je $np > 5$ i $nq > 5$.

Postupak:

1. Izračunati μ i σ za binomnu distribuciju.
2. Pretvoriti diskretnu slučajnu varijablu u kontinuiranu - napraviti **korekciju zbog neprekidnosti**.
3. Izračunati vjerojatnost koristeći normalnu raspodjelu (standardizacija i sve što već treba).

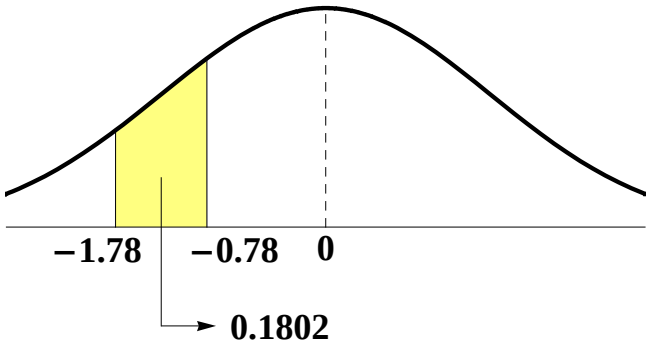
Primjer : $n = 50$, $p = 0.237$, $X \sim B(n,p)$, $P(7 \leq n \leq 9)$.

$np > 5$, $nq > 5 \Rightarrow$ možemo zamijeniti $B(n,p)$ s $N(\mu, \sigma)$

$\mu = np = 11.85$

$\sigma = \sqrt{50 \cdot 0.237 \cdot 0.763} = 3.007$

$P(7 \leq n \leq 9) \rightarrow P(6.5 \leq X \leq 9.5) \rightarrow P(-1.78 \leq Z \leq -0.78)$

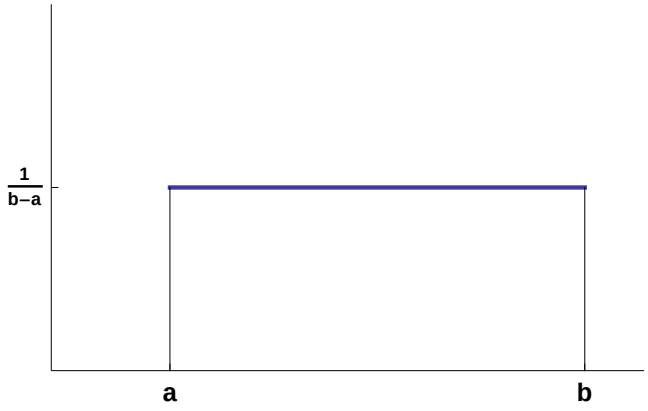


$P(7 \leq X \leq 9) = 0.1802$

6.5 Uniformna vjerojatnostna raspodjela

Slučajna varijabla koja se ravna po uniformnoj raspodjeli zove se **uniformna slučajna varijabla**.

Za uniformnu slučajnu varijablu na intervalu $[a, b]$ funkcija gustoće vjerojatnosti dana je formulom $f(x) = \dfrac{1}{b-a}$. Funkcija distribucije $F(x)$ je 0 za $x < a$, 1 za $x > b$ i $F(x) = \frac{x-a}{b-a}$ za $a \leq x \leq b$.



Primjer: Raspodjela oštećenja na komadu autoceste.

Vrijedi:

$$\mu = \frac{a+b}{2},\ \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Temeljna formula je

$$P(x \leq x \leq d) = \frac{d-c}{b-a}.$$

Ona odražava temeljno svojstvo, a to je da vjerojatnost da slučajna varijabla poprimi vrijednost u nekom intervalu ne ovisi o položaju tog intervala, već samo o njegovoj duljini.

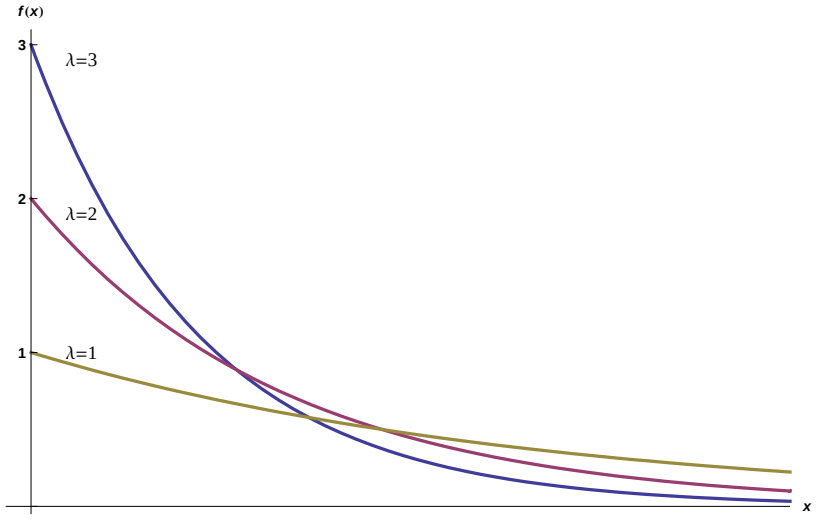
6.6 Eksponencijalna raspodjela

Eksponencijalna raspodjela je bliska Poissonovoj raspodjeli. Kod Poissonove raspodjele je slučajna varijabla poprimala vrijednosti koje su brojile dešavanje određenog događaja u intervalu, vremenskom, prostornom ili volumnom. Kod eksponencijalne raspodjele gledamo vrijeme između dvaju uzastopnih dešavanja događaja.

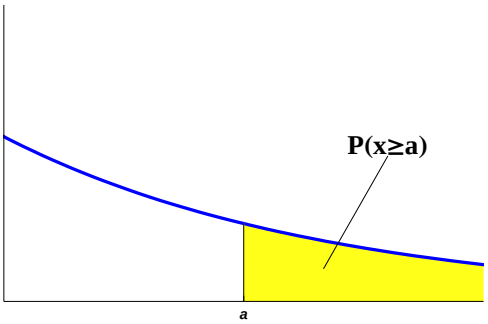
Eksponencijalna raspodjela ima samo jedan parametar, λ, čije je značenje isto kao kod Poissonove raspodjele - prosječan broj pojavljivanja događaja u jedinici vremena.

Funkcija gustoće vjerojatnosti dana je formulom

$$f(x) = \lambda e^{-\lambda x}.$$



Zanima nas $P(X \geq a)$, $P(X \leq a)$, $P(a \leq X \leq b)$.



$$P(X \geq a) = \int_a^{\infty} \lambda e^{-\lambda x} dx = -e^{-\lambda x}|_a^{\infty}.$$

$$P(X \geq a) = e^{-\lambda a}, \lambda > 0, a > 0$$

Odatle i iz činjenice da je ukupna površina ispod krivulje jednaka 1 slijedi

$$P(X \leq a) = 1 - e^{-\lambda a}.$$

Dakle je funkcija distribucije $F(x)$ dana formulom $F(x) = 1 - e^{-\lambda x}$. Odatle i iz aditivnosti vjerojatnosti slijedi i formula

$$P(a \leq X \leq b) = e^{-\lambda a} - e^{-\lambda b}.$$

Eksponencijalnu raspodjelu možemo smatrati i graničnim slučajem geometrijske raspdjele za $n \to \infty$. Za slučajnu varijablu X koja se ravna po eksponencijalnoj raspodjeli s parametrom λ imamo

$$E(X) = \sigma(X) = \frac{1}{\lambda}.$$

Primjer : Službenik na šalteru posluži u prosjeku 30 stranaka na sat.

Ako je vrijeme potrebno za poslužiti jednu stranku slučajna varijabla s eksponencijalnom raspodjelom, koja je vjerojatnost da će iduća stranka potrošiti više od 5 minuta? A da će biti poslužena u manje od dvije?

Jedinica vremena s kojom radimo je minuta. Treba preračunati λ u minute. Dobije se $\lambda = \frac{30}{60} = 0.5$. Sad je

$P(X \geq 5) = e^{-0.5 \cdot 5} = e^{-2.5} = 0.0821.$

$P(X \leq 2) = 1 - e^{-0.5 \cdot 2} = 1 - e^{-1} = 0.6321.$

6.7 Paretova raspodjela

Krajem XIX stoljeća talijanski ekonomist Vilfredo Pareto uočio je da se broj ljudi čiji su prihodi veći od X može dobro aproksimirati funkcijom Cx^{α} za neke pozitivne konstante C i α. Te se konstante razlikuju od države do države, no α je obično oko 1.5. Po sličnim se zakonima ravnaju i populacije gradova, veličine tvrtki, isplate osiguranja, magnitude potresa i mnoge druge pojave. Promatramo li te veličine kao slučajne varijable, njihova bi funkcija gustoće bila oblika $F(x) = \frac{\alpha}{x^{\alpha}}$ za $x \geq 1$. Takve slučajne varijable imaju **Paretovu raspodjelu** s parametrom α. Pišemo $X \sim Par(\alpha)$.

Očekivanje i standardna devijacija Paretove slučajne varijable s parametrom α dani su, za $\alpha > 1$, formulama

$$E(X) = \frac{\alpha}{\alpha - 1}, \qquad \sigma(X) = \frac{1}{\alpha - 1}\sqrt{\frac{\alpha}{\alpha - 2}}.$$

Za $0 < x \leq 1$, očekivanje i standardna devijacija Paretove slučajne varijable s takvim parametrom su beskonačno veliki. To znači da za takve slučajne varijable ne postoje "tipične" vrijednosti.

Funkcija distribucije dana je formulom $F(x) = 1 - x^{-\alpha}$ za $x \geq 1$.

Poglavlje 7

Populacije i uzorci

7.1 Anketa

Anketa (proba, pregled uzorka) je tehnika prikupljanja informacija koja uključuje samo dio populacije. Kad je riječ o ljudskoj populaciji govorimo o anketi.

Anketa
- osobno - najkvalitetnije, najskuplje, najviše odgovora i najdugoročniji utjecaj na ispitanika.
- telefon - pristojan odaziv, nije jako skupo ni dugotrajno. Ljudi ne vole biti gnjavljeni kod kuće, a neki i nemaju telefon.
- pošta - mali odaziv, ali isto jeftino.

Najsloženiji dio je priprema upitnika. Formulacija pitanja može znatno utjecati na odgovore.

7.2 Slučajni i neslučajni uzorci

Uzorak je **slučajan** ako svaki član populacije ima neke izglede biti izabran u uzorak. U neslučajnom uzorku neki članovi populacije nemaju nikakve izglede biti izabrani.

Za slučajnost uzorka nije nužno da svi članovi imaju jednake izglede za uključivanje u uzorak.

Izvlačenje iz šešira, loto, slučajni brojevi,... daju slučajne uzorke.

Slučajni uzorak je obično i reprezentativan.

Pogodnostni uzorak uključuje najdostupnije članove populacije (u danom trenutku ili na danom mjestu). Prvi koji dođu pod ruku.

Ekspertni uzorak je odabran iz populacije na temelju prosudbe i apriornog znanja o populaciji. Takav uzorak može biti reprezentativan, no za veliku populaciju izgledi su obično mali.

Pogodnostni i ekspertni uzorci su neslučajni.

Pseudoankete su nereprezentativni uzorci. Anketa u časopisima, televizijske ankete na 060 broj i slične stvari su beskorisne za procjenu situacije u općoj populaciji jer uključuju samo čitatelje dotičnog časopisa / gledatelje te emisije i/ili one koji ne žale novac za poziv, te su emocionalno angažirani.

7.3 Jednostavni slučajni uzorak

Jednostavni slučajni uzorak je uzorak pri čijem je izboru svaki član populacije imao iste izglede biti uključen.

Izvlačenje, lutrija, slučajni brojevi iz tablica ili računala,...

7.4 Raspodjele populacije i uzorka

7.4.1 Populacijska raspodjela

Svaka populacija ima točno jednu vrijednost parametra μ. S druge strane, svaki uzorak daje svoju aritmetičku sredinu, tj. srednju vrijenost $\overline{x}$. Dakle je $\overline{x}$ slučajna varijabla, pa kao i svaka druga slučajna varijabla i $\overline{x}$ ima svoju vjerojatnostnu raspodjelu.

Uzorkovna raspodjela od $\overline{x}$ je vjerojatnostna raspodjela slučajne varijable $\overline{x}$. Ona svakoj vrijednosti koju $\overline{x}$ može poprimiti pridružuje odgovarajuću vjerojatnost.

Primjer: Populacija od pet članova, slučajna varijabla je godišnja plaća u tisućama kuna: 17, 24, 35, 35 i 43. Za tu populaciju je $\mu = \frac{154}{5} = 30.8$, $\sigma = 9.174$ (u tisućama kuna).

Različiti uzorci (ima ih $\binom{5}{k}$) daju različite vrijednosti za $\overline{x}$. Uzmemo li $k = 3$, imamo 10 različitih uzoraka.

Osoba	Plaća
A	17
B	24
C	35
D	35
E	43

Uzorak	$\overline{x}$
ABC	25.33
ABD	25.33
ABE	28.00
ACD	29.00
ACE	31.67
ADE	31.67
BCD	31.33
BCE	34.00
BDE	34.00
CDE	37.67

$\Longrightarrow$

$\overline{x}$	$P(X = \overline{x})$
25.33	0.2
28.00	0.1
29.00	0.1
31.33	0.1
31.67	0.2
34.00	0.2
37.67	0.1

$P(\overline{x} = 31.67) = 0.2$

7.5 Pogreške uzorka i ostale pogreške

Prilikom prikupljanja i obrade podataka dolazi do raznih pogrešaka. Dio je uzrokovan strukturom uzorka, a dio potječe od prikupljanja i obrade podataka.

Pogreška uzorka je razlika između μ i $\overline{x}$, tj. $\overline{x} - \mu$.

To vrijedi za slučajne uzorke i ako nema ostalih pogrešaka.

Razlozi ostalih pogrešaka:

- neslučajnost uzorka;
- netočni i/ili neiskreni ogovori;
- nejasna pitanja;
- pogrešan unos/kopiranje podataka.

U prijašnjem primjeru, za uzorak ABE pogreška uzorka je $\overline{x}-\mu = 31.67-30.80 = 0.87$ (tisuća kuna). Ako smo još krivo zapisali plaću od E kao 45, onda je $\overline{x} = 32.33$. Sada je $\overline{x} - \mu = 1.53$. Od toga je 0.87 pogreška uzorka, a 0.66 su ostale pogreške.

U realnom životu nikad ne znamo pogrešku uzorka.

7.6 Očekivanje i standardna devijacija od $\overline{x}$

Gledamo li $\overline{x}$ kao slučajnu varijablu, onda njeno očekivanje i standardnu devijaciju računamo iz vjerojatnostne raspodjele.

Označavamo ih s $\mu_{\overline{x}}$ i $\sigma_{\overline{x}}$. Standardna devijacija od $\overline{x}$ se još zove i **standardna pogreška** od $\overline{x}$.

Vrijedi: $\mu_{\overline{x}} = \mu$ - statistika uzorka $\overline{x}$ je procjena očekivanja.

Ako je očekivanje statistike uzorka jednako vrijednosti odgovarajućeg parametra populacije, kažemo da je statistika uzorka **nepristrana procjena** parametra populacije.

Može se pokazati da je $\overline{x}$ nepristrana procjena od μ.

Za standardnu devijaciju imamo drukčiju situaciju.

$x = \sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$ - Ovo vrijedi za beskonačne populacije ili za konačne populacije i uzorak biran s vraćanjem. Praktični kriterij je $\frac{n}{N} \leq 0.05$. Ako to nije zadovoljeno, koristimo formulu:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad , \quad (\frac{n}{N} \geq 0.05)$$

Faktor $\sqrt{\dfrac{N-n}{N-1}}$ je korekcija zbog konačnosti populacije. Imamo lijep primjer toga kod varijance i standardne devijacije za hipergeometrijsku raspodjelu.

1. $\sigma_{\overline{x}} < \sigma$ - logično. Na uzorku se ne može manifestirati sav raspon populacije. Broj u nazivniku je veći od 1, $\sigma_{\overline{x}} < \sigma$.

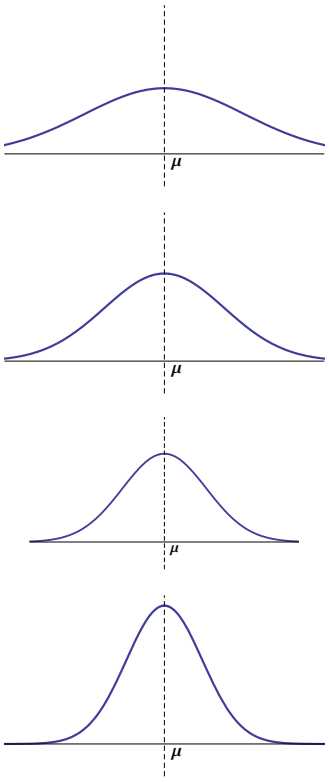
2. $\sigma_{\overline{x}}$ pada s porastom veličine uzorka. To je isto očito iz $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Ako standardna devijacija statistike uzorka pada s porastom veličine uzorka kažemo da je ta statistika **konzistentna procjena**. Dakle je $\overline{x}$ konzistentna procjena parametra μ.

7.7 Oblik uzorkovne raspodjele od $\overline{x}$

7.7.1 Populacija ima normalnu raspodjelu

1. $\mu_{\overline{x}} = \mu$
2. $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$, za $\dfrac{n}{N} < 0.05$
3. Raspodjela od $\overline{x}$ je normalna, za sve n



populacija

$\overline{x}$, $n = 5$

$\overline{x}$, $n = 20$

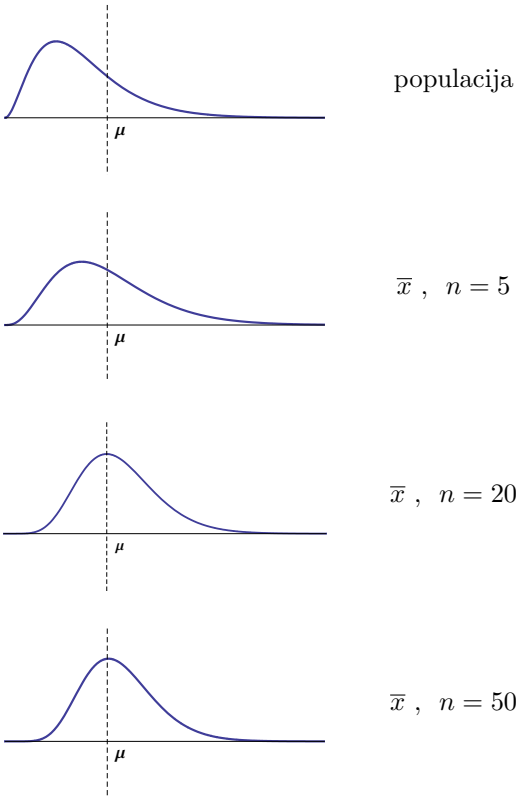
$\overline{x}$, $n = 50$

sve raspodjele su normalne

7.7.2 Populacija nema normalnu raspodjelu

Centralni granični teorem: Ako je uzorak dovoljno velik, raspodjela od $\overline{x}$ je približno normalna, bez obzira na oblik raspodjele populacije. Očekivanje i standardna devijacija su dani s $\mu_{\overline{x}} = \mu$, $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Uzorak se obično smatra velikim za $n \geq 30$.



populacija

$\overline{x}$, $n = 5$

$\overline{x}$, $n = 20$

$\overline{x}$, $n = 50$

Posljednje dvije distribucije su približno normalne raspodjele (i dalje vrijedi $\dfrac{n}{N} \leq 0.05$).

7.8 Primjene raspodjele od $\overline{x}$

Kolika je vjerojatnost da $\overline{x}$ za neki uzorak padne između nekih vrijednosti?

$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}}$, ako znamo μ i σ. A obratno?

7.9 Vjerojatnost (proporcija, udio) u populaciji i uzorku

Populacijski udio (vjerojatnost) je omjer broja članova populacije sa svojstvom koje zas zanima i ukupnog broja članova populacije. Označavamo ga s $p = \frac{X}{N}$.

Isti takav omjer za uzorak zovemo **uzorkovni udio** i označavamo s $\hat{p} = \frac{x}{n}$.

Primjer: Udio pušača u ukupnoj populaciji i u uzorku.

$$p = \frac{X}{N} \;\; ; \;\; \hat{p} = \frac{x}{n}$$

7.10 Vjerojatnostna raspodjela od $\hat{p}$

Primjer : Populacija od 5 studenata, troje voli statistiku. Gledamo dvočlane uzorke.

A	da
B	ne
C	ne
D	da
E	da

AB	0.5
AC	0.5
AD	1.0
AE	1.0
BC	0.0
BD	0.5
BE	0.5
CD	0.5
CE	0.5
DE	1.0

	$P(\hat{p})$
0.0	0.1
0.5	0.6
1.0	0.3

Očekivanje od $\hat{p}$ je jednako populacijskom udjelu p: $\mu_{\hat{p}} = p$.

Na našem primjeru $\mu_{\hat{p}} = 0.0 \cdot 0.1 + 0.5 \cdot 0.6 + 1.0 \cdot 0.3 = 0.6 = p$.

Vidimo da je $\hat{p}$ nepristrana procjena od p.

Za standardnu devijaciju vrijedi:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \quad , \quad q = 1 - p \quad , \quad \frac{n}{N} \leq 0.05$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad , \quad \frac{n}{N} \leq 0.05$$

Ponovo imamo faktor korekcije za konačnu populaciju oblika $\sqrt{\frac{N-n}{N-1}}$.

Oblik distribucije od $\hat{p}$ slijedi iz centralnog graničnog teorema.

Uzorkovna raspodjela od $\hat{p}$ je približno normalna za dovoljno veliki uzorak. Uzorak se smatra dovoljno velikim ako je $np > 5$ i $nq > 5$.

To je isti uvjet kao i za aproksimaciju binomne raspodjele normalnom.

Želimo li računati vjerojatnost da je za određeni uzorak iz populacije s poznatim parametrima $\hat{p}$ između nekih granica, imamo $z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$.

To ne izgleda naročito zanimljivo. Više nas zanima obratno, kao i za $\overline{x}$.

Poglavlje 8

Procjene očekivanja i udjela

Dolazimo do dijela statistike koji se bavi zaključivanjem o populaciji na temelju informacije dobivene iz uzorka - inferencijalne statistike.

Inferencijalna statistika kojom ćemo se baviti obuhvaća procjene očekivanja i vjerojatnosti za jednu populaciju, te testiranje hipoteze o očekivanju i vjerojatnosti za jednu populaciju.

8.1 Procjenjivanje

Procjenjivanje je pridruživanje vrijednosti parametru populacije na temelju vijednosti statistike uzorka.

$\left.\begin{array}{l}\mu \text{ - pravo očekivanje} \\ p \text{ - prava vjerojatnost, pravi udio}\end{array}\right\}$ populacije

$\left.\begin{array}{l}\overline{x} \text{ - procjena očekivanja} \\ \hat{p} \text{ - procjena vjerojatnosti}\end{array}\right\}$ uzorka

Procjena je vrijednost pridružena parametru populacije na temelju vrijednosti statistike uzorka.

8.2 Točkovne i intervalne procjene

8.2.1 Točkovna procjena

Vrijednost statistike uzorka korištena kao procjena parametra populacije je **točkovna procjena**.

Točkovna procjena se obično daje s **marginom pogreške**, koja se uzima kao $\pm 1.96\sigma_{\overline{x}}$. Obično ne znamo $\sigma_{\overline{x}}$, pa se uzima $s_{\overline{x}}$ kao točkovna procjena od $\sigma_{\overline{x}}$.

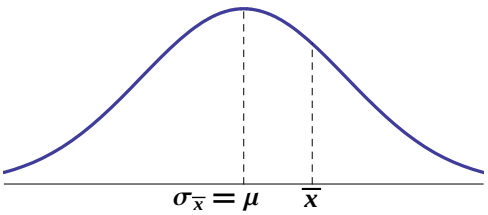
Svaki uzorak daje drugačiju točkovnu procjenu parametra populacije.

Točkovna procjena se gotovo uvijek razlikuje od prave vrijednosti.

8.2.2 Intervalna ocjena

Kod intervalne ocjene konstruira se interval oko točkovne ocjene i daje se vjerojatnost da taj interval sadrži pravu vrijednost parametra populacije.

Imamo neku informaciju o kvaliteti procjene.

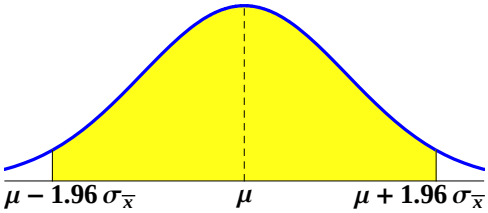


Znamo da se točkovna procjena gotovo nikada ne podudara s pravom vrijednošću. Oduzmemo li i dodamo isti iznos od točkovne procjen dobivamo krajeve intervala u kojem bi se mogla nalaziti i prava vrijednost promatranog parametra. Što više širimo taj interval, to je vjerojatnije da će prava vrijednost biti unutra. Treba nam kvantitativni odnos između širine intervala i vjerojatnosti da on sadrži pravu vrijednost. To ovisi o dvjema veličinama: o $\sigma_{\overline{x}}$ i o vjerojatnosti padanja u taj interval. Jasno je da veća standardna devijacija znači i širi interval. Jasno je i da veća vjerojatnost znači širi interval.

Koeficijent pouzdanosti je vjerojatnost koju pridjeljujemo događaju da promatrani interval sadrži pravu vrijednost parametra populacije. Obično želimo odrediti granice intervala tako da vjerojatnost sadržavanja prave vrijednosti bude 0.9, 0.95 ili 0.99.

Ovo treba malo pojasniti. Prava vrijednost parametra koji gledamo je neovisna o uzorku koji imamo. Za svaki interval konstruiran iz podataka dobivenih iz uzorka ta vrijednost ili pada unutra ili ne, dakle vjerojatnosti su 1 ili 0. Značenje broja koji predstavlja vjerojatnost, tj. pouzdanost, je da će 95% (recimo) od svih uzoraka iste veličine vrijednost $\overline{x}$ pasti između $\mu - 1.96\sigma_{\overline{x}}$ i $\mu + 1.96\sigma_{\overline{x}}$. Dakle $\mu - 1.96\sigma_{\overline{x}} \leq \overline{x} \leq \mu + 1.96\sigma_{\overline{x}}$ vrijedi za 95% svih uzoraka te veličine.

Odavde se dobije da vrijedi $\overline{x} - 1.96\sigma_{\overline{x}} \leq \mu \leq \overline{x} + 1.96\sigma_{\overline{x}}$ za 95% svih (velikih) uzoraka iste veličine. Slično se dobiva i za druge koeficijente pouzdanosti.



8.3 Procjene očekivanja - veliki uzorci

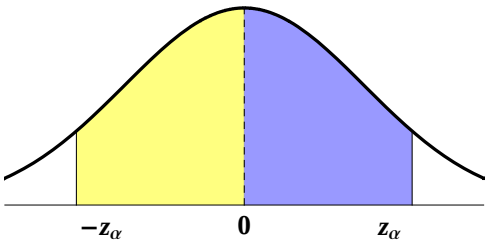
Vidjeli smo da je raspodjela veličine $\overline{x}$ približno normalna bez obzira na raspodjelu populacije ako je uzorak dovoljno velik. Za praktične potrebe uzimamo da je uzorak s više od 30 elemenata dovoljno velik.

Standardna devijacija populacije je najčešće nepoznata. Stoga za vrijednost $\sigma_{\overline{x}}$ koristimo točkovnu procjenu od $\sigma_{\overline{x}}$, dakle koristimo $s_{\overline{x}} = \dfrac{s}{\sqrt{n}}$ umjesto $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.
Formule za računanje veličine s imamo u točki 3.2.2.

Za zadanu pouzdanost α (obično 0.9, 0.95 ili 0.99, najčešće 0.95) intervalna procjena je dana s:

$$\begin{array}{ll} \langle \overline{x} - z_\alpha \sigma_{\overline{x}}, \overline{x} + z_\alpha \sigma_{\overline{x}} \rangle & \text{ako je } \sigma \text{ poznato} \\ \langle \overline{x} - z_\alpha s_{\overline{x}}, \overline{x} + z_\alpha s_{\overline{x}} \rangle & \text{ako } \sigma \text{ nije poznato.} \end{array}$$

Ovdje je z_α vrijednost koja se očitava iz tablice standardne devijacije normale raspodjele i odgovara vrijednosti standardne normalne varijable za koju je površina između 0 i z_α jednaka $\dfrac{\alpha}{2}$.



Najčešće korištene vrijednosti z_α su:

α	z_α
0.90	1.645
0.95	1.96
0.99	2.58

Vrijednost $E = z_\alpha \sigma_{\overline{x}}$ ili $E = z_\alpha s_{\overline{x}}$ je **maksimalna pogreška** procjene od μ. Širina intervala ovisi o z_α, σ i n. Na σ ne možemo utjecati, pa ako želimo smanjiti širinu intervala moramo ili smanjiti pouzdanost, ili povećati uzorak. Smanjenje pouzdanosti nije poželjno, dakle treba nastojati povećati uzorak.

8.4 Procjene očekivanja - mali uzorci

Ako je populacija normalna i ako je σ poznato, smislene intervalne procjene se mogu dobiti i s malim uzorcima. Ako to nije slučaj, ne možemo više koristiti normalnu raspodjelu. Koristimo tzv. Studentovu t-raspodjelu.

Uvjeti primjenjivosti t-raspodjele:

1. Populacija ima (približno) normalnu raspodjelu.
2. Uzorak je mali ($n < 30$).
3. Standardna devijacija populacije je nepoznata.

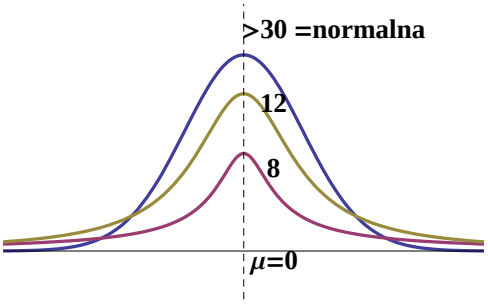
8.4.1 t-raspodjela

Uveo ju je W. S. Gossett 1908., pod pseudonimom Student.

t-distribucija ima samo jedan parametar, **stupanj slobode**.

$df = n - 1$. Za svaku vrijednost df imamo posebnu t-raspodjelu.

Sve one imaju $\mu = 0$. Za t-raspodjelu s df stupnjeva slobode standardna devijacija je dana s $\sqrt{\frac{df}{df-2}}$. Vidimo da je to uvijek veće od 1 i da pada prema 1 s porastom df, dakle s porastom n. Za dovoljno veliki n (≥ 30) t-distribuciju možemo zamijeniti standardnom normalnom raspodjelom.



Zašto stupnjevi slobode i zašto $n-1$? Ako imamo n opažanja, $n-1$ od njih uvijek možemo odabrati po volji, n-to je određeno njima.

Ako je sredina četiriju vrijednosti jednaka 20, tri od njih možemo odabrati po volji, četvrtu više ne. Odaberemo li te tri kao 17, 25 i 22, četvrta mora biti 16 da bi sredina bila 20.

8.4.2 Interval pouzdanosti za μ pomoću t-raspodjele

Za zadanu pouzdanost α interval pouzdanosti konstruiramo kao

$$\langle \overline{x} - t_\alpha s_{\overline{x}}, \overline{x} + t_\alpha s_{\overline{x}} \rangle,$$

gdje je $s_{\overline{x}} = \dfrac{s}{\sqrt{n}}$, a vrijednost t_α očitavamo iz tablice t-raspodjele za $n-1$ stupnjeva slobode tako da površina ispod krivulje između 0 i t_α bude $\dfrac{\alpha}{2}$.

8.5 Intervalne procjene vjerojatnosti - veliki uzorci

Zanima nas koji je udio članova populacije s promatranim svojstvom, tj. koja je vjerojatnost da će slučajno odabrani element populacije imati to svojstvo. Prava vrijednost tog parametra je p. Statistika uzorka je $\hat{p}$.

O raspodjeli veličine znamo:

1. Približno je normalna.
2. $\mu_{\hat{p}} = p$.
3. $\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}$, $q = 1 - p$.

Veličina $\sigma_{\hat{p}}$ nije poznata. Točkovno ju ocjenjujemo veličinom $s_{\hat{p}} = \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$.

Za zadanu pouzdanost α interval pouzdanosti konstruiramo kao

$$\langle \hat{p} - z_\alpha s_{\hat{p}}, \hat{p} + z_\alpha s_{\hat{p}} \rangle,$$

pri čemu se z_α određuje iz tablice normalne raspodjele kao i za intervalnu procjenu očekivanja.

8.6 Određivanje veličine uzorka

Često je bitno odrediti minimalnu veličinu uzorka koja će nam dati intervalnu procjenu željene pouzdanosti i s određenom maksimalnom pogreškom. Kako maksimalna pogreška ovisi o stndardnoj devijaciji populacije, a ona obično nije poznata, formule koje dajemo mogu biti dosta neprecizne.

8.6.1 Određivanje veličine uzorka za procjenu očekivanja

U 8.3 smo imali formulu $E = z_\alpha \sigma_{\overline{x}}$. Zbog $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, imamo $E = z_\alpha \frac{\sigma}{\sqrt{n}}$, pa odatle i formula $n = (\frac{z_\alpha \sigma}{E})^2$.

Problem je da obično ne znamo σ. Obično se rješava tako da se uzme promatrani uzorak proizvoljne veličine i da se iz njegove standardne devijacije s zaključi da je $\sigma = s$ i dalje se radi s tim. Kvaliteta ocjene minimalnog n tada ovisi o tome koliko je s blizu σ, tj. koliko smo imali sreće s izborom uzorka.

8.6.2 Određivanje veličine uzorka za procjenu vjerojatnosti

Iz 8.5 imamo da je $E = z_\alpha \sigma_{\hat{p}} = z_\alpha \sqrt{\frac{pq}{n}}$. Odatle slijedi formula $n = pq(\frac{z_\alpha}{E})^2$. Naravno, ne znamo ni p ni q. Možemo postupiti kao kod procjene očekivanja i uzeti $\hat{p}$ i $\hat{q}$ iz preliminarnog uzorka umjesto p i q. Druga mogućnost je napraviti **najkonzervativniju procjenu** izborom $p = q = 0.5$. Kako je $pq \leq \frac{1}{4}$ za sve p, q za koje je $p + q = 1$, tj. zbog $p - p^2 \leq \frac{1}{4}$ na $[0, 1]$, slijedi da izborom $p = q = 0.5$ povećavamo fakzor pq u našoj formuli i time povećavamo minimalni traženi n. Griješimo na sigurnu stranu.

$$n = (\frac{z_\alpha}{2E})^2.$$

(Ovdje bi svuda umjesto n trebalo stajati $n_{\min}$, no već imamo dosta kompliciranih oznaka.)

Poglavlje 9

Testiranje hipoteza

9.1 Motivacija i definicije

U bocama vina bi trebalo biti 750 ml vina. Na uzorku od 100 boca ustanovljeno je da je u njima prosječno 730 ml vina. Možemo li na temelju tog nalaza optužiti vinariju da vara kupce? Što ako neki drugi zorak da prosjek od 760 ml? Testiranje hipoteze je postupak kojim rješavamo takve dileme.

9.1.1 Dvije pretpostavke (hipoteze)

Promatrajmo situaciju u kojoj je osoba optužena za neko nedjelo i izvedena pred sud. Postoje samo dvije mogućnosti:

1. Osoba je kriva.
2. Osoba nije kriva.

Na početku suđenja se smatra da osoba nije kriva. Dužnost je tužitelja dokazati da je osoba kriva. Pretpostavka od koje polazimo je **nul-hipoteza**; ona druga je **alternativna hipoteza**. Nul hipotezu označavamo s H_0, alternativnu s H_1.

U našem primjeru s vinom smatramo da je tvdnja o 750 ml vina u prosječnoj boci istinita. Štoviše, tvrdnja da vinarija ne vara kupce je istinita i ako je u boci više od 750 ml vina. To može i ne mora biti istinito, no mi uzimamo kao polaznu hipotezu da vinarija govori istinu, tj. da ne vara kupce. Dakle, H_0 je hipoteza koja kaže

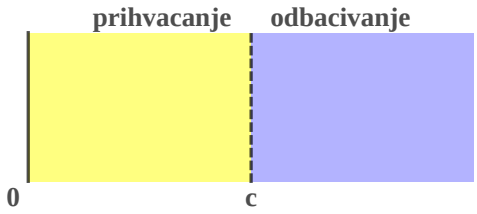
$$H_0 : \mu \geq 750\ ml.$$

Alternativna hipoteza je $H_1 < 750\ ml$.

Obje hipoteze odnose se na (nepoznati) parametar populacije, a ne na statistike uzorka.

9.1.2 Područja odbacivanja i prihvaćanja

Vratimo se usporedbi sa suđenjem. Ako tužitelj ne iznese nikakve dokaze, sud i/ili porota će optuženika osloboditi. Ako tužitelj iznese neke dokaze, i ako oni nisu brojni i/ili uvjerljivi, sud će smatrati da nema dovoljno razloga za proglasiti optuženika krivim, i da nema razloga odbaciti nul-hipotezu da je ovaj nevin. Ako, pak, "količina" dokaza prijeđe neku kritičnu točku C, sud će zaključiti da je optuženi kriv i odbaciti će nul-hipotezu o njegovoj nevinosti. Na gornjoj slici područje između 0 i C na "osi dokaza" je **područje prihvaćanja** nul-hipoteze, a područje desno od C je **područje odbacivanja** nul-hipoteze.



(Pravi sudski postupci se ne daju ovako jednostavno provesti jer je teško kvantificirati dokaze na jednodimenzionalnoj skali. Vrijednost C je također teško egzaktno definirati.)

Vrijednost C je **kritična točka**.

9.1.3 Dva tipa pogreške

Odluka suda / Činjenice	nevin	kriv
nevin	ispravna odluka	Tip II (β)
kriv	Tip I (α)	ispravna odluka

Vrijednost α, **signifikantnost**, je vjerojatnost pogreške tipa I, tj. odbacivanja istinite nul-hipoteze. Parametru α pridružujemo vrijednost **prije** testiranja hipoteze. Želimo imati mali α, obično se uzima $\alpha = 0.01, 0.025, 0.05$ ili 0.1. U analogiji sa sudom, želimo imati malu vjerojatnost osude nevine osobe.

Odluka / Stanje	H_0 istinita	H_0 lažna
H_0 se prihvaća	$\surd$	β
H_0 se odbacije	α	$\surd$

U prvom retku nevin/H_0 se prihvaća, može se desiti da je optuženik počinio nedjelo, ali nema dovoljno dokaza ili da je zaista nevin. U prvom slučaju smo učinili pogrešku tzv. tipa II, oslobodili smo krivca. Ako je β vjerojatnost takve

pogreške, onda se veličina $1-\beta$ zove **snaga testa**. To je vjerojatnost nečinjenja pogreške tipa II.

$\alpha = P\{H_0$ se odbacuje $|H_0$ je istinita $\}$ - pouzdanost, signifikantnost (značajnost)

$\beta = P\{H_0$ se prihvaća $|H_0$ nije istinita $\}$ - $1-\beta =$ snaga

Vrijednosti α i β nisu međusobno nezavisne, i ne mogu se obje smanjiti uz istu veličinu uzorka. Želimo li smanjiti obje, moramo povećati veličinu uzorka.

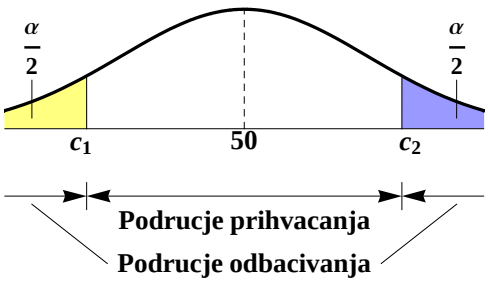
9.1.4 Repovi testa - jednostrani i dvostrani testovi

Vratimo se usporedbi sa sudom. U statistici, vrijednost kritične točke ne određujemo proizvoljno, nego iz unaprijed zadane vrijednosti α. Za razliku od suda, gdje gomilanje dokaza znači da imamo samo jedno područje prihvaćanja i odbacivanja, u statistici se može desiti da područje odbacivanja bude sastavljeno od dva komada. Čak i ako je od jednog komada, ono može biti lijevo ili desno od područja prihvaćanja, a ne samo desno, kao kod suda. U ovisnosti o tome koliko ima područja odbacivanja, govorimo o jednostranom ili dvostranom testu. Jednostrani testovi se dijele u lijevorepne i desnorepne, već prema tome je li područje odbacivanja u lijevom ili desnom repu krivulje gustoće raspodjele.

9.1.4.1 Dvostrani test

Nul hipoteza u dvostranom testu je formulirana u terminima jednakosti. Npr., prosječna duljina čavla je 50 mm.

$H_0 : \mu = 50$. Alternativna hipoteza je $H_1 : \mu \neq 50$.



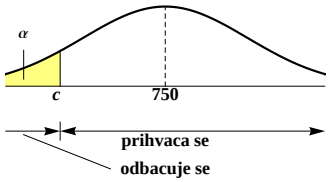
Dvostrani test ima 2 kritične vrijednosti, c_1 i c_2.

Odbacujemo H_0 ako vrijednost $\overline{x}$ padne bilo lijevo bilo desno izvan područja prihvaćanja. Odbacivanje H_0 znači da smo zaključili da je razlika između μ i $\overline{x}$ statistički signifikantna, tj. značajna. Prihvaćanje H_0 znači da smatramo da je razlika između μ i $\overline{x}$ uzrokovana izborom uzorka, da nije statistički značajna.

9.1.4.2 Jednostrani test

Lijevi rep

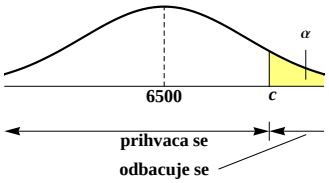
U takve testove spada naš primjer s vinom. Ako je u boci u prosjeku manje od 750 ml, možemo vinariju optužiti za varanje. Ako je barem 750 ml, onda ne. Polazimo od pretpostavljene nevinosti:



$$H_0 : \mu \geq 750ml$$
$$H_1 : \mu < 750ml$$

Desni rep

Imamo situaciju u kojoj vlada tvrdi da troškovi života nisu porasli u nekom razdoblju, a oporba da su porasli. Ako su pred tri mjeseca bili 6500 kn, onda se hipoteza formulira ovako:



$$H_0 : \mu \leq 6500kn$$
$$H_1 : \mu > 6500kn$$

	dvostrani test	lijevorepi test	desnorepi test
uvjet u H_0	$=$	$=$ ili $\geq$	$=$ ili $\leq$
uvjet u H_1	$\neq$	$<$	$>$
područje odbacivanja	u oba repa	u lijevom repu	u desnom repu

Postupak kod testiranja hipoteza:

1. Formulirati nul- i alternativnu hipotezu.
2. Odabrati raspodjelu.
3. Odrediti područja prihvaćanja i odbacivanja.
4. Izračunati vrijednosti statistike testa ($\overline{x}$ ili $\hat{p}$).
5. Donijeti odluku.

9.2 Testiranje hipoteza o očekivanju - veliki uzorci

Velikim uzorcima smatramo one koji imaju barem $n = 30$ elemenata.

Bez obzira znamo li σ ili ne, možemo se koristiti normalnom raspodjelom.

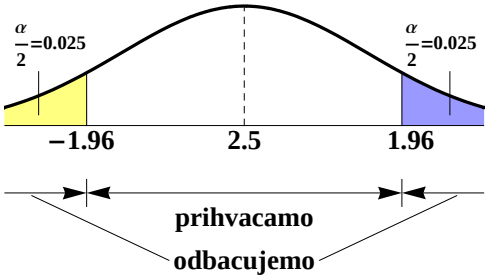
Slučajna varijabla $z = \dfrac{\overline{x} - \mu}{\sigma_{\overline{x}}}$ ili $z = \dfrac{\overline{x} - \mu}{s_{\overline{x}}}$ se zove **statistika testa**. Zove se još i **opažena vrijednost** od z. Slučajna varijabla $\overline{x}$ je statistika uzorka.

Primjer : Duljina osovina za neki fini mehanizam mora biti 2.5 cm. Proizvodi ih stroj. Ako prosječna duljina osovine odstupa od zadane, stroj se mora zaustaviti i podesiti. Slučajan uzorak od 49 osovina dao je $\overline{x} = 2.49$ cm i standardnu devijaciju od 0.021 cm. Uz $\alpha = 0.05$, treba li zaustaviti stroj ili ne?

Rješenje:

$n = 49$
$\overline{x} = 2.49$ cm
$s = 0.021$ cm

1. $H_0 : \mu = 2.5$
   $H_1 : \mu \neq 2.5$
2. Koristimo normalnu raspodjelu.
3. Test je dvostran zbog uvjeta $\neq$ u H_1.
4. Vrijednost μ nije poznata, pa z računamo pomoću $s_{\overline{x}}$
   $s_{\overline{x}} = \dfrac{s}{\sqrt{n}} = 0.003$
   $z = \dfrac{\overline{x} - \mu}{s_{\overline{x}}} = \dfrac{2.49 - 2.5}{0.003} = \dfrac{-0.01}{0.003} = -3.33$
5. Vrijednost $z = -3.33$ pada u područje odbacivanja nul-hipoteze, pa stroj treba zaustaviti i podesiti.



Odbacivanjem nul-hipoteze tvrdimo da je razlika između $\overline{x}$ i μ prevelika da bi bila posljedica samo pogreške uzorka, odnosno, da je vjerojatnost da ta razlika potječe od pogreške uzorka mala, manja od 0.05. To znači da nam samo 5% svih mogućih uzoraka daje tako veliku pogrešku. Donijeli smo pogrešnu odluku, tj. počinili smo pogrešku tipa I, ako nas je zapao baš jedan od tih 5% uzoraka.

Korite se još termini **statistički značajna razlika** i **razlika koja nije statistički značajna**. Ako je razlika statistički značajna, nul-hipoteza se odbacuje; ako odstupanje nije statistički značajno, onda se nul-hipoteza prihvaća.

9.3 Testiranje hipoteza o očekivanju - mali uzorci

Kao i za procjene, za male uzorke se služimo t-raspodjelom.

Uvjeti primjene t-raspodjele su:

1. Populacija je (približno) normalna.
2. Standardna devijacija populacije σ nije poznata.
3. Uzorak je mali, $n < 30$.

Statistika testa se sada označava s t i računa kao:

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}}, \text{ gdje je } s_{\overline{x}} = \frac{s}{\sqrt{n}}.$$

Ovo je opažena vrijednost od t.

Postupak je isti kao i za velike uzorke, samo se ne služimo tablicama normalne raspodjele već tablicama t-raspodjele s $n - 1$ stupnjeva slobode.

Primjer : Tvrtka tvrdi da njeni akumulatori traju u prosjeku barem 65 mjeseci. Na uzorku od 15 akumulatora nađeno je da u prosjeku traju 63 mjeseca, sa standardnom devijacijom s od 2 mjeseca. Uz razinu pouzdanosti od 5%, možemo li prihvatiti tvrdnju proizvođača?

Rješenje:

$n = 15$, $\overline{x} = 63$ mj., $\mu = 65$ mj.

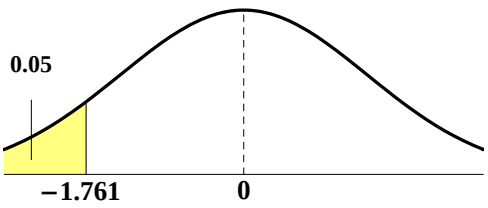
$s = 2$ mj.

1. $H_0 : \mu \geq 65$

   $H_1 : \mu < 65$

2. Uzorak je mali, ne znamo σ, trajanje je približno normalno raspodijeljeno. Uzimamo t-raspodjelu s 14 stupnjeva slobode.

3. $\alpha = 0.05$, test je jednostrani, lijevorepni. Područje odbacivanja je u lijevom repu, površina pod njim mora biti jednaka 0.05.



4. $s_{\overline{x}} = \dfrac{s}{\sqrt{n}} = \dfrac{2}{\sqrt{15}} = 0.5164$

$t = \dfrac{63-65}{0.5164} = -3.873$

5. Zbog opažene vrijedosti $t = -3.873$ koja je manja od kritične vrijednosti $t = -1.761$ odbacujemo H_0.

9.4 Testiranje hipoteza o vjerojatnosti - veliki uzorci

Postupak je sličan postupku za očekivanje. Razlikuje se samo računanje statistike testa. Za velike uzorke služimo se normalnom distribucijom.

Opažena vrijednost z se računa kao

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}, \text{ gdje je } \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}.$$

Primjer : U režimu ispravnog funkcioniranja stroj proizvodi najviše 4% neispravnih proizvoda. Ako proizvodi više, treba ga podesiti. Na slučajnom uzorku od 200 proizvoda nađeno je 11 neispravnih. Odredite, uz $\alpha = 0.05$, treba li zaustaviti stroj i podesiti ga.

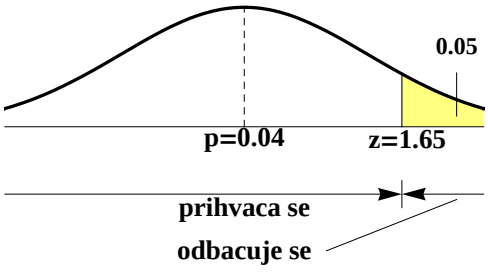
Rješenje: $n = 200, \hat{p} = \dfrac{11}{200} = 0.055, \alpha = 0.05$

1. $H_0 : p \leq 0.04$

   $H_1 : p > 0.04$

2. Uzorak je velik, $np = 8 > 5$, $nq = 192 > 5$, koristimo normalnu distribuciju.

3. Test je jednostrani, desnorepni.



4. $\sigma_{\hat{p}} = \sqrt{\dfrac{0.04 \cdot 0.96}{200}} = 0.0139$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.055 - 0.04}{0.0139} = \frac{0.015}{0.0139} = 1.079.$$

5. Zbog $1.079 < 1.65$ vidimo da opažena vrijednost od z pada u područje prihvaćanja hipoteze H_0, pa stroj ne treba zaustavljati.

Bibliografija

[1] Prem S. Mann, Statistics for Business and Economics, J. Wiley, New York, 1995.

[2] Boris Petz, Osnovne statističke metode za nematematičare, Naklada Slap, Jastrebarsko, 1997.

[3] Željko Pauše, Uvod u matematičku statistiku, Školska knjiga, Zagreb, 1993.

[4] Darrell Huff, How to lie with Statistics, Norton, New York, 1954.