Primijenjena statistika

Nikola Koceić Bilan

Prirodoslovno-matematički fakultet u Splitu, 2011.

Sadržaj

Uvod

Početci statistike (ako ne računamo statističke preglede podataka o broju umrlih, rođenih, zaraženih od neke bolesti, a koji se vode unatrag nekoliko stoljeća) su usko vezani uz početke Teorije vjerojatnosti, a možemo ih pratiti od pojave hazardnih igara sredinom 17.st. Pojam vjerojatnosti se od početaka vezivao uz intuitivno bliskiji pojam relativne frekvencije. Unatoč nepostojanju odgovarajućega matematičkog modela i neuspješnih pokušaja uvođenja aksioma, "naivnim pristupom" postignuti su značajni rezultati povezani s imenima Pascala, Fermata, Bernoullija i Laplacea. A.N. Kolmogorov 1933. uvodi opće prihvaćenu aksiomatiku u teoriju vjerojatnosti koja adekvatno reprezentira našu prirodnu predodžbu o vjerojatnosti nekoga događaja kao broju kojemu konvergiraju relativne frekvencije toga događaja kad se broj pokusa neograničeno (beskonačno) ponavlja. Statistika je grana primijenjene matematike koja se počelo masovno primjenjivati u različitim prirodnim i društvenim znanostima tek početkom 2. svjetskog rata. Danas je statistika postala dio općeg obrazovanja jer je čovjek izložen situacijama u kojima mu je potrebno poznavanje nekih osnovnih statističkih pojmova i statističkog načina razmišljanja i to zbog praćenja stručne literature i medija, zbog deskripcije i analize podataka prikupljenih nekim istraživanjem, zaključivanja iz konkretnog slučaja na opći zakon, zbog planiranja istraživanja i eksperimenta, kao i zbog mnogih drugih zahtjeva kako svakodnevnog života tako i gotovo svih zanimanja današnjice.

Ovaj nastavni materijal je namijenjen prvenstveno studentima informatike Prirodoslovno-matematičkog fakultetu u Splitu ali, isto tako, i studentima svih studijskih programa kojima je potrebna primijenjena statistika. Upravo zbog toga se u tekstu ne koristi prejaki matematički aparat kojim se inače služi Teorija vjero-

jatnosti i statistike i koji je objektivno razumljiv samo profesionalnim matematičarima, već je upotrebljen matematički jezik primjereniji studentima nematematičkih studija koji imaju solidno predznanje iz elementarne matematike. Cijeli materijal je potkrijepljen s mnogo primjera iz struke i svakodnevnog života putem kojih se ilustriraju prethodno teoretski razrađene statističke metode i zorno se prikazuje njihova primjena. Iako u tekstu nema strogih matematičkih dokaza, obrada pojedinih statističkih metoda nije svedena na obično posluživanje recepata i uputa za pojedine statističke postupke.

Tekst se sastoji od 6 poglavlja. U prvom poglavlju se obrađuje deskriptivna statistika i to zbog cjelovitosti ovoga nastavnog teksta kao i zbog onih studenata koji nisu tijekom obrazovanja slušali neki uvodni statistički kolegij. Drugo i treće poglavlje se odnose na vjerojatnost i slučajnu varijablu. Kod ovih poglavlja se pazilo da i oni čitatelji koji nisu upoznati s osnovnim kombinatornim metodama kao i značenjem i tehnikom integriranja mogu s lakoćom pratiti navedene teme. U četvrtom poglavlju se obrađuje dvodimenzionalna slučajna varijabla, te pojmovi korelacije i neovisnosti slučajnih varijabli. U petom i šestom poglavlju se obrađuju osnovne teme iz inferencijalne statistike koje se odnose na statističku obradu uzoraka uzetih iz promatrane populacije. U ovim poglavljima su detaljno obrađeni intervali očekivanja kao i najosnovniji testovi za testiranje hipoteza.

Napomenimo da je u većini numeričkih postupaka korišten znak jednakosti, iako su numeričke vrijednosti uglavnom zaokružene na 2 ili više decimala. Ipak, znak $\approx$, kojim označujemo približnu vrijednost, je korišten kad god se željela naglasiti razlika između aproksimacije i prave vrijednosti.

Cijeli tekst može dati dobru osnovu za detaljnije proučavanje i razumijevanje ostalih statističkih tema koje se mogu pronaći u priloženoj literaturi, kao i dovoljno predznanje za neka početna samostalna statistička istraživanja.

Poglavlje 1

Deskriptivna statistika

1.1 Populacije i varijable

Statistički skup ili **populacija** je svaki skup čiji su elementi jedinice kojima mjerimo (ispitujemo) jedno ili više obilježja (svojstava). Kardinalni broj (broj elemenata) populacije nazivamo **opsegom**. Populacija može imati konačan ili beskonačan opseg. Ako je svakom elementu populacije S pridruženo jedno obilježje iz skupa obilježja O, onda je definirana jedna funkcija $X : S \to O$ koju nazivamo **statističkom varijablom**. Koji put se i skup $X(S) \subseteq O$ svih obilježja elemenata statističkog skupa naziva populacija. Skup $X(S)$ ćemo nazivati **skupom obilježja populacije**. Svaki podskup populacije nazivamo **uzorkom**. Statističke varijable dijelimo na **numeričke** i **kvalitativne**. Kvalitativne varijable su: **nominalne** i **ordinalne**.

Nominalna varijabla pridružuje svakom članu populacije neki atribut. Između takvih atributa nema uređaja (redoslijeda), čak ni u slučaju kada su atributi brojčani, jer služe kao brojčani identifikatori. Primjerice, svima onima koji odgovore na referendumsko pitanje sa "DA" pridružimo broj 1, u protivnom broj 0. S takvim brojevima nema smisla raditi računske operacije.

Ordinalna varijabla pridružuje članovima populacije simbol ili broj prema intezitetu mjernog svojstva pri čemu je određen njihov redoslijed prema stupnju inteziteta. Primjerice, varijabla koja studentima nekog fakulteta pridružuje oc-

jenu iz nekog kolegija je ordinalna.

Numeričke varijable dijelimo na: **intervalne** i **omjerne**.
Intervalna varijabla pridružuje svakom članu populacije realan broj, sukladno intezitetu mjernog svojstva, pri čemu uređaj brojeva definira i redoslijed obilježja, te je definirana mjerna jedinica i dogovorna nula. Primjerice, varijabla koja svakom danu pridružuje temperaturu zraka u isto vrijeme na istom mjestu je intervalna, a temperatura od 0^0 ne znači da temperature nema.
Omjerna varijabla je numerička varijabla koja ima iste karakteristike kao i intervalna samo što nula nije dogovorno utvrđena, već znači nepostojanje svojstva na promatranom elementu. Primjerice, varijabla koja mjeri visinu neke ljudske populacije je omjerna. Kod omjerne ima smisla upotrebljavati omjere vrijednosti (npr. duplo veća visina) za razliku od intervalne varijable (temperatura od -2^0 nije duplo veća od temperature -1^0).

Ako numerička varijabla može poprimiti najviše konačno ili prebrojivo elemenata tj. ako elemenata skupa $X(S)$ ima najviše koliko i elemenata skupa $\mathbb{N}$ nazivamo ju **diskretnom**. Ako varijabla X poprima sve vrijednosti iz nekog intervala $\langle a, b\rangle \subseteq \mathbb{R}$, tj. ako je $\langle a, b\rangle \subseteq X(S)$, za neke $a, b \in \mathbb{R}$, $a < b$, varijablu nazivamo **kontinuiranom**.

Primjer 1.1 *Broj svih pravnih osoba u Republici Hrvatskoj na dan 31.03.2001. je bio* 189 576*. Ako svakoj pravnoj osobi ispitujemo broj zaposlenika na taj dan definirali smo jednu diskretnu numeričku omjernu varijablu* $X : S \to \mathbb{N}$*, gdje je populacija* S *konačan skup svih pravnih osoba na određeni dan. Ako svakoj pravnoj osobi ispitujemo najmanju isplaćenu plaću za taj mjesec definirali smo drugu diskretnu numeričku omjernu varijablu* $X' : S \to \mathbb{R}$*. Ako pak svakoj pravnoj osobi pridružimo njezino sjedište, definiramo jednu kvalitativnu nominalnu varijablu, dok bi mjesta na rang listi HGK-e po godišnjoj bilanci definirala jednu ordinalnu varijablu.*

Više različitih varijabli koje djeluju na istoj populaciji možemo promatrati kao jednu *višedimenzionalnu varijablu* koju je najprikladnije prikazati matrično. Na primjer, ako je $X = (X_1, ..., X_k)$, gdje su X_i, $i = 1, ..., k$, statističke varijable

definirane na istoj populaciji S, onda se odgovarajuća matrica

$X_1(S)$ $\downarrow$	$X_2(S)$ $\downarrow$	$\cdots$	$X_k(S)$ $\downarrow$
a_{11}	a_{12}	$\cdots$	a_{1k}
$\vdots$	$\vdots$		$\vdots$
a_{i1}	a_{i2}	$\cdots$	a_{ik}
$\vdots$	$\vdots$		$\vdots$

sastoji od vrijednosti a_{ij} koje predstavljaju vrijednosti j-te varijable X_j u i-tom populacijskom članu.
Podaci o građanima koje popisivač stanovništva uzima predstavljaju jednu višedimenzionalnu varijablu (spol, dob, mjesto rođenja, broj članova kućanstva...).

Beskonačna populacija je na neki način teorijska tvorevina. Populacija će biti beskonačna ako je statistički skup hipotetski skup i vezan je nekim *stohastičkim procesom*. U tom slučaju su elementi populacije neki slučajni pokusi, eksperimenti koji se beskonačno puta nastavljaju, a numerička varijabla bilježi njihove ishode. Ishodi tih slučajnih procesa se ravnaju po zakonima vjerojatnosti, odnosno nisu unaprijed poznati. Navedimo neke primjere beskonačnih populacija i odgovarajućih varijabli.

Bacanje novčića je pokus koji se može ponavljati beskonačno puta. Takav hipotetski skup kojega tvore svi mogući pokusi bacanja novčića je beskonačna populacija, a nominalna varijabla koja ishodu glava pridruži 1, a pismo 0 je diskretna.

Ispitivanje broja kvarova mobitela nekog proizvođača u jamstvenom roku je jedna diskretna numerička varijabla definirana na hipotetskoj i beskonačnoj populaciji svih mobitela koji su proizvedeni i koji će se tek proizvesti u budućnosti u neprekidnoj proizvodnji (iako ih u realnom svijetu uvijek ima samo konačno) koja svakom mobitelu pripisuje broj kvarova $0, 1, 2, ...$ u jamstvenom roku.

Visina svih ljudi je jedna kontinuirana varijabla definirana na populaciji svih ljudi koji su se rodili ili koji će se roditi u neprekidnosti postojanja čovječanstva.

Temperatura zraka na nekom mjestu unutar 24 sata je kontinuirana varijabla koja mjeri temperaturu u svakom djeliću vremena (iako se u stvarnosti temperatura registrira samo nekoliko puta na dan).

Vrijeme potrebno za opsluživanje neke stranke na jednom šalteru određene banke za vrijeme smjene određene djelatnice je jedna kontinuirana varijabla. Populacija je hipotetski beskonačan skup koji se sastoji od svih stranki koje su se pojavile i koje će se pojaviti na tom šalteru u kontinuitetu, a varijabla bilježi vrijeme opsluživanja jedne stranke.

Primijetimo da varijabla može biti kontinuirana samo ako je populacija beskonačna.

Osobitosti populacija i numeričkih varijabli zadanih na njima se iskazuju različitim brojčanim veličinama koje nazivamo **parametrima**. Parametar ovisi o svim vrijednostima varijable X tj. o svim elementima $s \in S$ populacije i njihovim obilježjima $X(s)$. Ako se parametar računa samo na uzorku $S_0 \subseteq S$ populacije onda tu vrijednost dobivenu temeljem vrijednosti varijable $X|_{S_0}$ samo na uzorku S_0 nazivamo **procjenom parametra**, a analitički izraz tj. formulu kojom je izražena funkcijska veza između uzorka i vrijednosti varijable X na njemu nazivamo **procjeniteljem**.

Zadaća **deskriptivne statistike** je uređivanje, grupiranje, tabeliranje, grafičko prikazivanje dostupnih podataka i izračunavanje parametara varijabli zadanih na konačnoj populaciji čije su sve vrijednosti (obilježja) poznate. Pri tome se ne razmatra priroda procesa koji generira te podatke, a dobiveni parametri i zaključci o obilježjima se ne poopćavaju, već se odnose isključivo na dani empirijski materijal.

Zadaća **inferencijalne statistike** je donošenje zaključaka o parametrima varijabli koje su zadane na beskonačnoj populaciji ili su zadane na konačnom ali prevelikom skupu tako da nisu poznate sve vrijednosti varijable tj. obilježja svakog elementa populacije. U oba slučaja zaključci o cijeloj populaciji s odgovarajućom varijablom se donose temeljem dostupnih podataka na uzorku i oni predstavljaju procjenu parametra s određenom vjerojatnošću.

1.1.1 Frekvencija i proporcija

Neka je $S = \{s_1, ..., s_N\}$ konačna populacija opsega N i $X : S \to O$ varijabla. Neka je skup obilježja populacije (vrijednosti varijable) $X(S) = \{x_1, x_2, ..., x_k\}$. Označimo sa $y_i = X(s_i)$, $i = 1, ..., N$, vrijednosti članova populacije. Očito je

$y_i \in \{x_1, x_2, ..., x_k\}$, $i = 1, ..., N$, a neki različiti elementi populacije $s_i \neq s_j$ mogu imati iste vrijednosti, $y_i = y_j$, odnosno mogu imati isto obilježje. Budući da u statistici nisu bitni pojedinačni članovi populacije već samo ukupan broj populacijskih elemenata s istim obilježjem to se numerička statistička varijabla često zadaje konačnim nizom svojih vrijednosti $y_1, ..., y_N$ koji nazivamo **statističkim nizom**.

Definicija 1.2 *Broj elemenata skupa* $X^{-1}(x_i)$, $i = 1, ..., k$, *odnosno broj svih članova populacije koji imaju isto obilježje* x_i *nazivamo* ***frekvencijom obilježja*** x_i *i označujemo sa* f_i, *a broj* $p_i = \frac{f_i}{N}$ *nazivamo njegovom* ***relativnom frekvencijom*** *ili* ***proporcijom.***

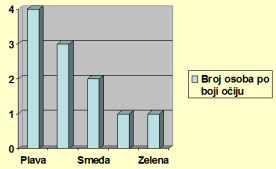
$$\text{Očito vrijedi } N = \sum_{i=1}^{k} f_i \qquad \text{i} \qquad \sum_{i=1}^{k} p_i = 1,\ p_i = \frac{f_i}{\sum_{i=1}^{k} f_i}.$$

Primjerice, ako osobe $s_1, s_2, ..., s_{10}$ tvore neku promatranu populaciju i imaju redom $20, 18, 18, 30, 25, 20, 20, 18, 25, 20$ godina, onda obilježja (godine) $18, 20, 25$ i 30 imaju redom frekvencije $3, 4, 2, 1$.

Primjer 1.3

Osoba O	Boja očijuX(O)
A	Plava
B	Crna
C	Zelena
D	Plava
E	Siva
F	Siva
G	Plava
H	Smeđa
I	Siva
J	Plava
K	Smeđa

Boja očiju	Frekvencija
Plava	4
Siva	3
Smeđa	2
Crna	1
Zelena	1



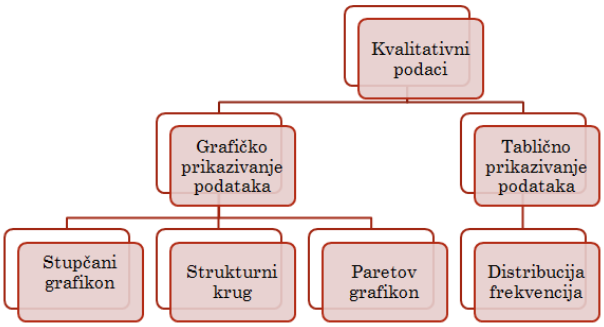
Definicija 1.4 *Funkciju koja svakom obilježju* x_i *pridružuje odgovarajuću (relativnu) frekvenciju* f_i (p_i), $i = 1, ..., k$, *nazivamo* ***funkcijom distribucije ili razdiobe (relativne) frekvencije varijable****, a skup točaka* $\{(x_i, f_i), i = 1, ..., k\}$

($\{(x_i, p_i), i = 1, ..., k\}$) nazivamo ***grafom*** *te* ***distribucije****. Spajanjem točaka grafa dobivamo* ***poligon distribucije varijable****.*

Ako po nekom kriteriju poredamo obilježja (numerička obilježja poredamo po uređaju $\leqslant$, a nenumerička najčešće po uređaju između njihovih frekvencija), tj. elemente statističkog niza, dobivamo **grupirani statistički niz**. Tada uz **niz frekvencija** $f_1, ..., f_k$, od odgovarajućih međusobno različitih elemenata toga niza definiramo i **kumulativni niz** $F(x_1) = f_1$, $F(x_2) = f_1 + f_2, ...$, $F(x_i) = f_1 + \cdots + f_i, ..., F(x_k) = f_1 + \cdots + f_k = N$.

1.1.2 Uređivanje kvalitativnih podataka

Kvalitativni podaci, tj. vrijednosti kvalitativne varijable zadane na konačnoj populaciji, se prikazuju grafički ili tablično.



Najčešći način zadavanja kvalitativne varijable je tablični prikaz distribucije frekvencija.

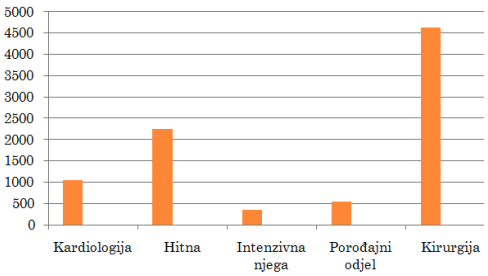
PRIMJER: BOLNIČKI PACIJENTI PO ODJELIMA

BOLNIČKI ODJEL	BROJ PACIJENATA (frekvencija)	RELATIVNA FREKVENCIJA
Kardiologija	1052	0.1192 = 11.92%
Hitna	2245	0.2545 = 24.45%
Intenzivna njega	340	0.0385 = 3.85%
Porođajni odjel	552	0.0625 = 6.25%
Kirurgija	4630	0.525 = 52.50%

Prethodnom tablicom je zadana distribucija frekvencija nominalne varijable koja djeluje na populaciji pacijenata koji leže u bolnici tako da svakom pacijentu pridruži ime odjela na kojem leži.

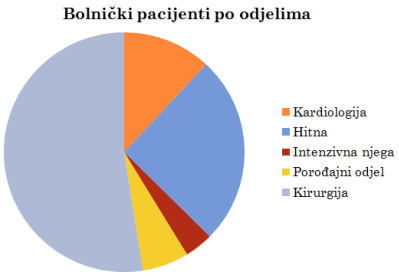
Grafički se kvalitativna varijabla najčešće zadaje pomoću stupčanog grafikona

PRIMJER STUPČANOG GRAFIKONA



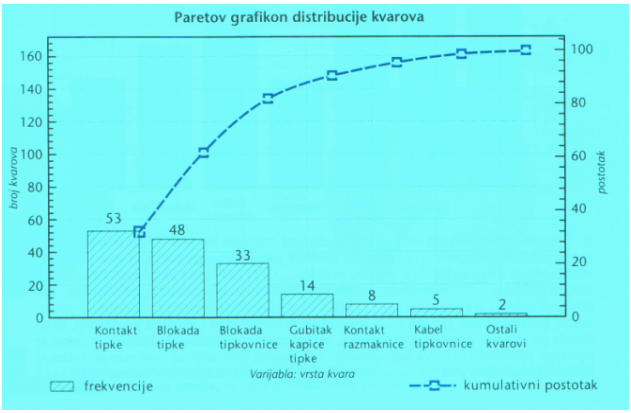
ili strukturnog kruga. Pripadni kut kružnog isječka u strukturnom krugu koji odgovara frekvenciji f_i iznosi $\frac{f_i}{N}360^0$.

PRIMJER – STRUKTURNI KRUG



Paretov dijagram prikazuje istovremeno i frekvencije (poredane od veće prema manjoj) i kumulativne frekvencije, iz čega se lakše uočava ako manji broj obilježja

(u ovom slučaju kvarova) daje veći dio ukupne distribucije.



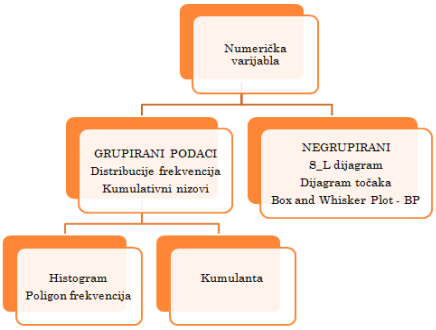
Za prikazivanje višedimenzionalne kvalitativne varijable najprikladnija je kombinirana tablica.

Vrsta hotela	2 zvjezdice	3 zvjezdice	4 zvjezdice	5 zvjezdica	Ukupno
Regija A	5	6	3	1	15
Regija B	7	6	4	0	17
Regija C	2	4	4	2	12
Ukupno	14	16	11	3	44

Prethodnom tablicom je zadana distribucija frekvencija dvodimenzionalne nominalne varijable koja djeluje na populaciji svih hotela tako da svakom hotelu pridruži broj zvjezdica kojima je kategoriziran i šifru regije u kojoj se nalazi. Svaku od tih varijabli možemo promatrati i zasebno. Nadalje, ova tablica prikazuje i distribuciju frekvencija induciranih nominalnih varijabli zadanih na suženim populacijama (hoteli samo određene regije ili samo određene kategorizacije).

1.1.3 Uređivanje numeričkih podataka

Numerička varijabla zadana na konačnoj populaciji se zadaje u obliku grupiranih ili negrupiranih podataka.



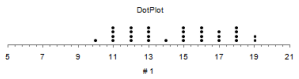
U donjoj tablici su prikazani neki načini zadavanja numeričke varijable u obliku negrupiranih podataka.

STEAM AND LEAF PLOT
S-L
ZADAN JE NIZ DULJINE 31
12 11 13 10 14 12 15 16 17 18 17 18 18 19 13 11 15 16 13 13 11
11 12 18 19 17 15 12 16 16 15

S-L dijagram

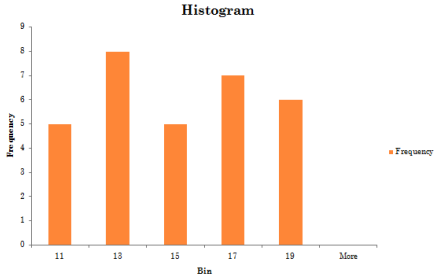
1	10	0
5	11	0000
9	12	0000
13	13	0000
14	14	0
(4)	15	0000
8	16	0000
11	17	000
(4)	18	0000
6	19	00

Dot plot
(dijagram točaka)



Kvalitativne podatke možemo grupirano prikazati pomoću: histograma u kojem frekvencije odgovaraju površini pravokutnika (ili visini pravokutnika, ako je osnovica jedinična), poligona frekvencija koji odgovara poligonu distribucije varijable ili kumulante koja odgovara grafičkom prikazu kumulativnog niza.

POVRŠINSKI GRAFIKONI
HISTOGRAM I BAR-CHART



Osim histograma koristi se i trodimenzionalni hologram u kojemu pojedinačne frekvencije odgovaraju volumenu kvadra. Budući je u dvodimenzionalnom prikazu teško uspoređivati volumene, to je ovakav prikaz pogodan za stvaranje pogrešne slike o pravim podacima.

Za prikazivanje dvodimenzionalne numeričke varijable najprikladniji je **dijagram rasipanja** (Scatterplot) u kojemu se sve vrijednosti y_i i y_i', $i = 1, ...N$, varijabli X i X' zadanima na istoj populaciji opsega N prikazuju u obliku uređenog para (y_i, y_i') u Kartezijevom koordinatnom sustavu.

Ako je numerička varijabla kontinuirana ili ako je diskretna s velikim brojem različitih vrijednosti, onda se distribucija frekvencija formira prema razredima, tj. podaci se grupiraju u disjunktne podintervale. U tom slučaju svaki razred je interval $[L_{1i}, L_{2i}\rangle$ a njegova frekvencija f_i je ukupan broj vrijednosti varijable X koji se nalaze u tom intervalu. Ako ima ukupno k razreda, onda skup $\{([L_{1i}, L_{2i}\rangle, f_i), i = 1, ..., k\}$ svih uređenih parova $([L_{1i}, L_{2i}\rangle, f_i)$ predstavlja distribuciju frekvencija te numeričke varijable grupirane u razrede. Umjesto f_i, u distribuciji mogu biti zastupljene i relativne frekvencije $p_i = \frac{f_i}{N}$, a možemo govoriti i o kumulativnom nizu gdje su razredi poredani po uređaju na $\mathbb{R}$. Broj razreda k za grupiranje n različitih vrijednosti numeričke varijable se računa izrazom $k \approx 1 + 3.3 \log n$. Razredi su jednakih veličina kad god su podaci približno simetrično raspoređeni, no općenito su uži tamo gdje je veća koncentracija podataka.

Pri brojčanoj analizi distribucije frekvencija s razredima bitno je da vrijedi $L_{2i} =$

L_{1i+1}, za svaki $i = 1, ..., k-1$. Takve granice nazivamo pravim granicama što se uvijek može postići zajedničkim izjednačavanjem gornje granice i -tog razreda i donje granice $i+1$-og razreda s brojem $\frac{L_{2i}+L_{1i+1}}{2}$. Nadalje, važno je izračunati *razrednu sredinu* $\frac{L_{1i}+L_{2i}}{2}$, te *veličinu razreda* $L_{2i} - L_{1i}$.

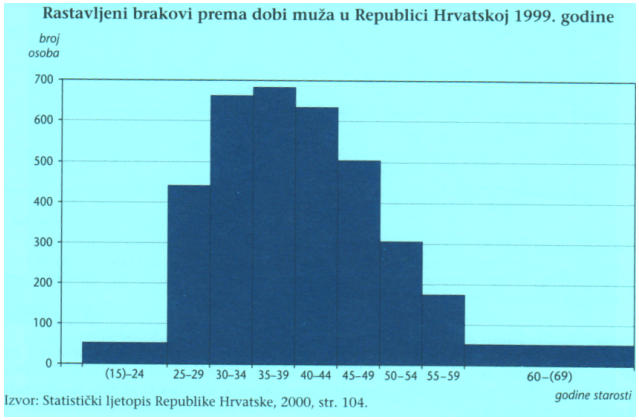
Ako se histogramom prikazuje distribucija frekvencija s razredima i ako su svi razredi jednake veličine, onda osnovica pravokutnika predstavlja veličinu razreda, a visina odgovara razrednoj frekvenciji. No, ako su razredi nejednakih veličina, potrebno je jednu veličinu uzeti za jediničnu, a frekvencije l puta većih ili l puta manjih razreda od jedinične duljine razreda treba dijeliti ili množiti s faktorom da bismo dobili *korigiranu frekvenciju* koja nam služi kao visina pravokutnika u histogramu. Korigiranim frekvencijama se koristimo i u poligonu frekvencija, dok u kumulanti uzimamo originalne frekvencije.

Primjer 1.5 *U tablici su dani podaci o rastavljenim brakovima prema dobi muža u R.H.1999.*

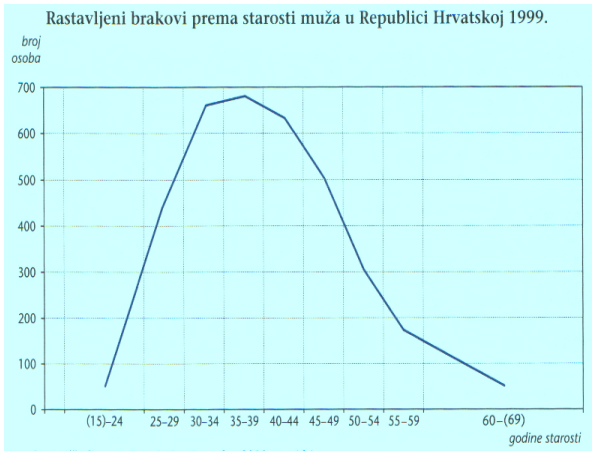
Godine života	Broj osoba	Prave granice	Razredne sredine	Veličina razreda	Korigirane frekvencije
	f_i		x_i	i_i	f_{ci}
(15)–24	105	(14.5)–24.5	19.5	10	52.5
25–29	439	24.5–29.5	27	5	439
30–34	662	29.5–34.5	32	5	662
35–39	683	34.5–39.5	37	5	683
40–44	635	39.5–44.5	42	5	635
45–49	503	44.5–49.5	47	5	503
50–54	305	49.5–54.5	52	5	305
55–59	174	54.5–59.5	57	5	174
60–(79)	208	59.5–(79.5)	69.5	20	52
ukupno	3714	–	–	–	–

(U zagradama su procijenjene granice prvog i zadnjeg razreda, odnosno njihove prave granice pri pretpostavci da bi gornja granica razreda prije prvoga bila 14, a

donja granica razreda poslije zadnjega 80).



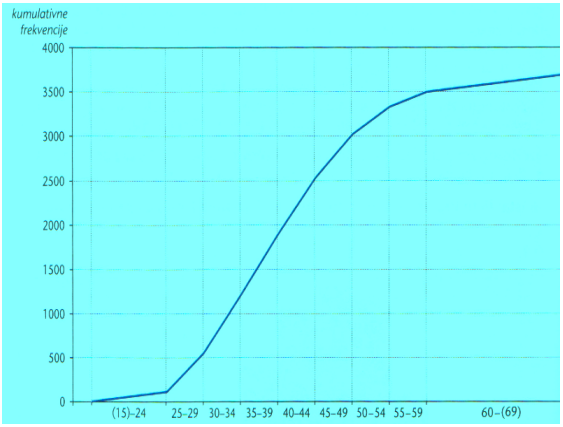
Izvor: Statistički ljetopis Republike Hrvatske, 2000, str. 104.



Poligon frekvencija je poligonalna crta koju tvore dužine koje spajaju točke čije apscise su jednake razrednim sredinama, a ordinate su im (korigirane) frekvencije.

Kumulativne frekvencije razreda tj. vrijednosti kumulativnog niza računamo pomoću originalnih frekvencija. Naime, te vrijednosti se odnose na zbroj frekvencija svih obilježja (godina) do godine s kojom završava taj razred: za 1. razred $F(24.5) = f_1 = 105$, *za 2. razred* $F(29.5) = f_1 + f_2 = 544$, *za 3. razred* $F(34.5) = f_1 + f_2 + f_3 = 1206\ldots$ *za 9. razred* $F(79.5) = f_1 + \cdots + f_9 = 3714 = N$.

Kumulanta distribucije je poligonalna crta koju tvore dužine koje spajaju točke čije apscise su jednake gornjim granicama razreda, a ordinate su im vrijednosti kumulativnog niza. Početna točka ima apscisu jednaku donjoj granici prvoga razreda i ordinatu 0.



1.2 Populacijski parametri

Populacijske parametre (parametre neke varijable ili statističkog niza) dijelimo na: **srednje vrijednosti** i **mjere disperzije**. Srednje vrijednosti dijelimo na **potpune srednje vrijednosti** i **položajne srednje vrijednosti**. Potpune srednje vrijednosti su: aritmetička sredina, harmonijska sredina, geometrijska sredina i moment. Položajne srednje vrijednosti (određene položajem unutar danog niza) su: mod, medijan i kvantil. Mjere disperzije dijelimo na: **nepotpune mjere disperzije** i **potpune mjere disperzije.** Nepotpune mjere disperzije su: raspon varijacije, interkvartil i koeficijent kvartilne devijacije. Potpune mjere disperzije su: varijanca, standardna devijacija i koeficijent varijacije.

1.2.1 Aritmetička sredina

Neka je $X : S \to X(S) = \{x_1, x_2, ..., x_k\} \subseteq \mathbb{R}$ numerička varijabla zadana na konačnoj populaciji opsega N i neka je $\{(x_i, f_i), i = 1, ..., k\}$ njezina distribucija frekvencije.

Definicija 1.6 ***Aritmetička sredina*** μ *numeričke varijable* X *zbroj umnoška*

vrijednosti svakog obilježja i njegove frekvencije podijeljen s opsegom populacije tj.

$$\mu = \frac{\sum_{i=1}^{k} f_i x_i}{N}.$$

Ako je numerička varijabla X zadana statističkim nizom $y_1, ..., y_N$, onda je njezina aritmetička sredina očigledno jednaka

$$\mu\left(y_1, ..., y_N\right) = \frac{\sum_{i=1}^{N} y_i}{N}.$$

Ako je distribucija numeričke varijable grupirana u (prave) razrede, tada se čitav razred identificira s razrednom sredinom, tj. $x_i f_i$ predstavlja umnožak frekvencije f_i razreda $[L_{1i}, L_{2i}\rangle$ i razredne sredine $x_i = \frac{L_{1i}+L_{2i}}{2}$. Takvu aritmetičku sredinu nazivamo **vaganom**.

Primjer 1.7 *Proizvodnja deterdženta Lahor u tijeku jedne dekade iznosila je u tonama* $105, 100, 110, 112, 108, 100, 104, 115, 96, 120$. *Prosječna proizvodnja je* $\mu = \frac{1}{10}\left(105 + 100 + 110 + 112 + 108 + 100 + 104 + 115 + 96 + 120\right) = 107$.

Primjer 1.8 *Dana je distribucija broja dana prema broju odsutnih zaposlenika nekog poduzeća u fiksnom periodu:*

Broj dana	0	1	2	3	4	5	6	7	8	9	10
Broj odsutnih	4	10	20	27	17	8	8	6	5	3	2 .

Aritmetička sredina numeričke varijable koja svakom zaposleniku pridružuje broj dana u kojima je izostao s posla u fiksnom periodu je $\mu = \frac{4\cdot 0+10\cdot 1+\cdots+3\cdot 9+2\cdot 10}{4+10+\cdots+2} = \frac{416}{110} = 3.78182$.

Zadatak 1.9 *Izračunajte prosječnu mjesečnu plaću u djelatnostima prijevoza, skladištenja i veza u R.H. u kolovozu 2000.*

djelatnost	*broj zaposlenih u tisućama*	*prosječna plaća*
kopneni prijevoz	*28.3*	*3115*
vodeni prijevoz	*2.7*	*3430*
zračni prijevoz	*0.7*	*5914*
pomoćne djelatnosti	*24.5*	*3360*
pošta i telekomunik.	*23.3*	*4560*
ukupno	*79.5*	

Rješenje: $\mu = \frac{\sum_{i=1}^{k} \mu_i N_i}{\sum_{i=1}^{k} N_i} = \frac{3115 \cdot 28.3 + 3430 \cdot 2.7 + \cdots + 4560 \cdot 23.3}{79.5} = \frac{290123.3}{79.5} = 3649.35$

Zadatak 1.10 *Odredite prosječan promet trgovina ako su zadani sljedeći podaci:*

Promet u 000 kn	Broj trgovina
30-40	2
40-50	5
50-60	10
60-70	12
70-90	10
90-110	9
110-150	2
Ukupno	50

.

Rješenje:

Promet u 000 kn	Broj trgovina (f_i)	Razredne sredine (x_i)	Podtotali ($x_i f_i$)
30-40	2	35	70
40-50	5	45	225
50-60	10	55	550
60-70	12	65	780
70-90	10	80	800
90-110	9	100	900
110-150	2	130	260
Ukupno	50		3585

$$\mu = \frac{\sum_{i=1}^{7} x_i f_i}{50} = \frac{3585}{50} = 71.7 \text{ } tisuća \text{ } kuna$$

Primjer 1.11 *Dana je udaljenost u kilometrima od mjesta stanovanja do radnog mjesta djelatnika nekog poduzeća:*

udaljenost u km (x_i)	*broj djelatnika* (f_i)
10	*1*
15	*2*
20	*5*
30	*3*
40	*1*

.

Prosječna udaljenost je $\mu = \frac{\sum_{i=1}^{5} f_i x_i}{\sum_{i=1}^{5} f_i} = \frac{1\cdot 10+2\cdot 15+\cdots+1\cdot 40}{1+2+\cdots+1} = \frac{270}{12} = 22.5$ *km. Primijetimo da je zbroj svih linearnih individualnih odstupanja* $x_i - \mu$ *od srednje vrijednosti jednak* 0*, tj.* $\sum_{i=1}^{5} f_i (x_i - \mu) = 1\cdot(10-22.5)+2\cdot(15-22.5)+\cdots+1\cdot(40-22.5) = 0.$

1.2.2 Standardna devijacija i varijanca

Neka je $X : S \to X(S) = \{x_1, x_2, ..., x_k\} \subseteq \mathbb{R}$ numerička varijabla zadana na konačnoj populaciji opsega N i neka je $\{(x_i, f_i), i = 1, ..., k\}$ njezina distribucija frekvencije. Primijetimo da je zbroj svih linearnih individualnih odstupanja vrijednosti varijable y_i, $i = 1, ..., N$, od njezine srednje vrijednosti μ uvijek jednak 0, tj. $\sum_{i=1}^{N} (y_i - \mu) = \sum_{i=1}^{k} f_i (x_i - \mu) = 0$. Najpodobniji parametar za mjeru odstupanja (raspršenosti, disperzije) je srednja vrijednost kvadratnih odstupanja.

Definicija 1.12 ***Varijanca*** σ^2 *numeričke varijable* X *zadane na konačnoj populaciji je zbroj svih umnožaka između kvadrata razlike vrijednosti obilježja* x_i *i aritmetičke sredine* μ *te varijable i frekvencije* f_i *toga obilježja podijeljen s opsegom populacije, tj.*

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{\sum_{i=1}^{k} f_i}.$$

Ako je varijabla zadana statističkim nizom $y_1, ..., y_N$ onda je njezina varijanca očigledno jednaka

$$\sigma^2 (y_1, ..., y_N) = \frac{\sum_{i=1}^{N} (y_i - \mu (y_1, ..., y_N))^2}{N}.$$

Varijanca je minimum svih srednjih vrijednosti kvadratnih odstupanja $\frac{\sum_{i=1}^{k} f_i(x_i-a)^2}{\sum_{i=1}^{k} f_i}$ od nekog broja a.

Lako se pokaže da vrijedi i sljedeća formula za varijancu

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i {x_i}^2 - \frac{1}{N}\left(\sum_{i=1}^{k} f_i x_i\right)^2}{N}.$$

Definicija 1.13 ***Standardna devijacija*** *je drugi korijen iz varijance tj.* $\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i-\mu)^2}{\sum_{i=1}^{k} f_i}}$.

Standardna devijacija se tumači kao prosječno odstupanje vrijednosti numeričke varijable od njezine aritmetičke sredine. Devijaciju je potrebno uvijek promatrati skupa sa sredinom μ ili u omjeru $V = \frac{\sigma}{\mu}100\%$ kojega nazivamo **koeficijentom varijacije** a kojega tumačimo kao prosječno odstupanje u jedinicama sredine. Primjerice, ako je prosječno pakiranje kutije šećera 750 g, a devijacija 5 grama, tada je prosječno odstupanje $V = \frac{5}{750}100 = 0.66\%$ od prosječne težine pakiranja.

Primjer 1.14 *U tablici je dano stanovništvo R.H. po starosti iz 1991.*

Godine života	Broj stanovn. u tis. f	Sredina razreda x
0-5	280,1	2,5
5-10	314,7	7,5
10-20	657,7	15,0
20-40	1.403,7	30,0
40-60	1.221,1	50
60-75	618,1	67,5
75 i više	216,9	82,5
Ukupno	4.712,3	

Vrijedi: $\mu = \frac{\sum_{i=1}^{7} f_i x_i}{\sum_{i=1}^{7} f_i} = \frac{175708}{4712,3} = 37,29$ *godine je bila prosječna starost stanovništva i* $\sigma = \sqrt{\frac{\sum_{i=1}^{7} f_i(x_i-\mu)^2}{4712,3}} = \sqrt{\frac{2224367.07}{4712.3}} \approx 21.7$ *je prosječno odstupanje od te vrijednosti.*

Teorem 1.15 (Čebišev) *Neka su μ i σ aritmetička sredina i standardna devijacija bilo koje numeričke varijable $X : S \to \mathbb{R}$ zadane na konačnoj populaciji S, te $k \in \mathbb{R}$, $k > 1$. Tada barem $\left(1 - \frac{1}{k^2}\right) 100\%$ članova ima obilježje koje se nalazi u intervalu $\langle \mu - k\sigma, \mu + k\sigma \rangle$, tj. zbroj relativnih frekvencija svih obilježja koji pripadaju tom intervalu je najmanje $\left(1 - \frac{1}{k^2}\right)$.*

Posljedica Čebiševa teorema je da barem 75% elemenata populacije ima numeričko obilježje u intervalu $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$, barem 89% elemenata je u intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$, a barem 93% elemenata je u intervalu $\langle \mu - 4\sigma, \mu + 4\sigma \rangle$.

Outlieri su ekstremne vrijednosti varijable koje znatno više od ostalih vrijednosti varijable odstupaju od prosjeka. Obično je outlier ona vrijednost koja odstupa od μ za više od 4σ. Često se takve vrijednosti izuzimaju iz analize skupa prikupljenih podataka jer nisu reprezentativne i mogu biti pogrešne.

Primjer 1.16 *U Primjeru 1.11 varijanca je* $\sigma^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i (x_i - 22.5)^2}{12} = \frac{1 \cdot (10-22.5)^2 + 2 \cdot (15-22.5)^2 + \cdots + 1 \cdot (40-22.5)^2}{12} = \frac{775}{12} = 64.58$*, a standardna devijacija je* $\sigma = \sqrt{64.58} = 8.04$*. Primijetimo da svi djelatnici osim jednoga, dakle 91.67% njih, imaju obilježje u intervalu* $\langle \mu - 2\sigma, \mu + 2\sigma \rangle = \langle 6.42, 38.58 \rangle$*. Bez poznavanje pojedinačnih obilježja, po Čebiševom teoremu, možemo samo zaključiti da se u istom intervalu nalazi barem 75% podataka.*

1.2.3 Standardizirana varijabla

Za ocjenu veličine individualnog odstupanja numeričkog obilježja x_i od aritmetičke sredine μ u jedinicama standardne devijacije koristi se relativna mjera odstupanja koju nazivamo **standardizirano obilježje** $z_i = \frac{x_i - \mu}{\sigma}$ i ona mjeri "koliko standardnih devijacija obilježje x_i odstupa od μ". Standardizirano obilježje z_i je pridruženo onom elementu populacije koji ima obilježje $x_i = \mu + z_i\sigma$. Stoga je standardizirano obilježje nova numerička varijabla Z koja djeluje na istoj populaciji kao i X tako što svakom elementu s populacije S pridruži odstupanje njegova obilježja $X(s)$ od μ izraženo u jedinicama σ tj. $Z(s) = \frac{X(s) - \mu}{\sigma}$. Nazivamo ju **standardiziranom varijablom**.

Propozicija 1.17 *Aritmetička sredina standardizirane varijable je uvijek* 0*, a standardna devijacija je uvijek* 1.

Primjer 1.18 *Prosječan broj bodova na 1. kolokviju iznosi 50 bodova, a prosječno odstupanje od prosjeka je 10. Na 2. kolokviju prosječan broj bodova je 90, a standardna devijacija je 20. Ako je student na 1. kolokviju postigao 62 boda, a na 2. 105 bodova, što možemo zaključiti o njegovom uspjehu?*
Budući su rasponi bodovnih skala nepoznati, možemo pretpostaviti da su različiti, pa prosudbu o uspjehu, iako se uspjeh iskazuje u istim mjernim jedinicama-bodovima, možemo donijeti jedino temeljem standardiziranog obilježja koje eliminira problem raspona skale. Zaključujemo: $z_1 = \frac{x_1-\mu_1}{\sigma_1} = \frac{62-50}{10} = 1.2$ *(uspjeh na 1. kolokviju je za 1.2 devijacije bolji od prosjeka) i* $z_2 = \frac{x_2-\mu_2}{\sigma_2} = \frac{105-90}{20} = 0.75$ *(uspjeh na 2. kolokviju je za* 0.75σ *bolji od prosjeka).*

Zadatak 1.19 *Skupina od 100 mladića natječe se u trčanju na 100 m i skoku u dalj. U prvoj disciplini je* $\mu_1 = 12.8$ *s i* $\sigma_1 = 2s$*, a u drugoj je* $\mu_2 = 485$ *cm* $\sigma_2 = 50$ *cm. Ako mladić A ima rezultat* 12.2 *s u 1. disciplini i* 490 *cm u 2., a mladić B trči* 13 *s na* 100 *m i skače* 580 *cm u dalj, koji je mladić uspješniji?*
Rješenje: *Budući su mjerne jedinice ovih disciplina različite moramo koristiti standardiziranu varijablu. Vrijedi:*

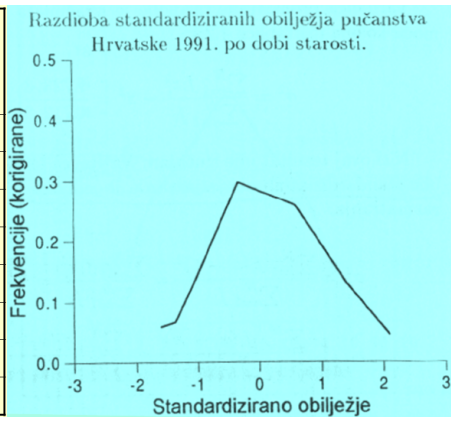
$z_{A1} = \frac{12.2-12.8}{2} = -0.3$ *(što je zapravo* $+0.3$ *jer je to za* 0.3σ *brže od prosjeka),*
$z_{B1} = \frac{13-12.8}{2} = 0.1$ *(što je zapravo* -0.1 *jer je to za* 0.1σ *sporije od prosjeka),*
$z_{A2} = \frac{490-485}{50} = 0.1$*,*
$z_{B2} = \frac{580-485}{50} = 1.9$*.*

Prosjek mladića A je $\frac{0.3+0.1}{2} = 0.2$ *(ukupno je približno prosječan) a mladića B je* $\frac{-0.1+1.9}{2} = 0.9$ *(ukupno je iznadprosječan).*

Primjer 1.20 *Iz podataka u Primjeru 1.14 izračunata su standardizirana obilježja*

i dan je graf distribucije frekvencija standardizirane varijable.

Godine života	Broj stanovn. u tis. f	Sredina razreda x	$z = \frac{x-\mu}{\sigma}$
0-5	280,1	2,5	-1,603
5-10	314,7	7,5	-1,373
10-20	657,7	15,0	-1,027
20-40	1.403,7	30,0	-0,336
40-60	1.221,1	50	0,586
60-75	618,1	67,5	1,392
75 i više	216,9	82,5	2,083
Ukupno	4.712,3		



Zadatak 1.21 *Tečaj jedne dionice zabilježen u \$ na jednoj burzi 25 dana uzastopno je:* $125, 127, 132, 127, 122, 129, 121, 124, 128, 132, 126, 125, 129, 128, 132, 122, 121, 120, 125, 133, 127, 125, 127, 134, 134$. *Jedan broker predviđa da će u vremenu predviđenom za prodaju dionica tečaj iznositi* 139 *\$, a drugi da će u tom istom vremenu iznositi* 133 *\$. Prosudite rizik procjene za oba brokera.*

Rješenje: *Moramo pretpostaviti da je tržište stabilno tj. da je kretanje tečaja "normalno", isto kao i za zabilježenih 25 dana. Prosječna cijena za navedeno razdoblje je* $\frac{3175}{25} = 127$ *\$, a standardna devijacija je* $\frac{416}{45} = 4.079$ *\$. Prognozirani tečaj od 139 \$ odstupa od prosjeka za* $\frac{139-127}{4.079} = 2.94$ *standardne devijacije, a tečaj od 133 odstupa* 1.47σ. *To znači da prva prognoza nije u intervalu* $\langle 127 - 2.94\sigma, 127 + 2.94\sigma\rangle$ *što povlači (sve uz pretpostavku da je kretanje tečaja uobičajno) da ta cijena nije i neće biti među* $\left(1 - \frac{1}{2.94^2}\right) 100\% = 88,43\,\%$ *svih (i budućih) cijena dionica.Druga cijena nije u intervalu* $\langle 127 - 1.47\sigma, 127 + 1.47\sigma\rangle$, *tj. nije i neće biti među* $\left(1 - \frac{1}{1.47^2}\right) 100\% = 53,7\%$ *svih cijena dionica. Očito je da prvi broker više riskira.*

Zadatak 1.22 *Prosječna godišnja plaća zaposlenika s određenom kvalifikacijom u jednom poduzeću iznosi* 57345 *kn. Prosječno odstupanje od toga prosjeka je* 7540 *kn. Može li se reći da je zaposlenik istih kvalifikacija koji ima godišnju plaću od* 34000 *kn diskriminiran?*

Rješenje: *Odredimo relativan polažaj osobe s godišnjom plaćom* 34000 *prema prosječnoj godišnjoj plaći:* $z = \frac{x_i-\mu}{\sigma} = \frac{34000-57345}{7540} = -3.09615$. *Plaća te osobe*

ne pripada intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ *u kojemu se, po Čebiševom teoremu, nalazi barem* 89% *plaća. Dakle, može se reći da je ta osoba diskriminirana, jer njezina plaća spada u najviše* 11% *najlošijih i najboljih plaća. Točnije, za* $k = 3.09615$, *interval* $\langle \mu - 3.09615\sigma, \mu + 3.09615\sigma \rangle$, *kojemu ne pripada promatrana plaća, sadrži* $\left(1 - \frac{1}{k^2}\right) 100\%$ *plaća, što je* $89,57\%$ *ukupnih plaća.*

1.2.4 Geometrijska sredina

Neka je $\{(x_1, f_1), ..., (x_k, f_k)\}$ distribucija numeričke varijable zadane na konačnoj populaciji opsega N koja poprima same pozitivne vrijednosti. **Geometrijskom sredinom** te numeričke varijable nazivamo broj

$$G = \sqrt[N]{x_1^{f_1} \cdot \dots \cdot x_k^{f_k}}.$$

Primijetimo da vrijedi

$$G = x_1^{\frac{f_1}{N}} \cdot \dots \cdot x_k^{\frac{f_k}{N}} = x_1^{p_1} \cdot \dots \cdot x_k^{p_k},$$

gdje su $p_1, ..., p_k$ relativne frekvencije obilježja $x_1, ..., x_k$ redom.
Ako je varijabla zadana statističkim nizom $y_1, ..., y_N$ (s pozitivnim vrijednostima) onda je njezina geometrijska sredina očigledno jednaka

$$G(y_1, ..., y_N) = \sqrt[N]{y_1 \cdot \dots \cdot y_N}$$

i vrijedi $\log G = \frac{\log y_1 + \dots + \log y_N}{N}$.

Geometrijska sredina se upotrebljava kao mjera prosječne brzine nekih promjena. Primjerice, ako je neko mjesto 2000. godine imalo 2000 stanovnika, 2005. godine 9000 stanovnika, a 2010. godine 18000, onda se broj stanovnika prvo povećao za 4.5 puta, a u drugom razdoblju 2 puta. Prosječna promjena stanovništva po razdobljima nije $\frac{4.5+2}{2} = 3.25$, već $\sqrt[2]{4.5 \cdot 2} = 3$. Zaista, $2000 \cdot 3 \cdot 3 = 18000$.

1.2.5 Harmonijska sredina

Harmonijskom sredinom numeričke varijable zadane na konačnoj populaciji opsega N distribucije $\{(x_1, f_1), ..., (x_k, f_k)\}$ koja poprima same pozitivne vrijed-

nosti nazivamo broj

$$H = \frac{N}{\frac{f_1}{x_1} + \cdots + \frac{f_k}{x_k}}.$$

Primijetimo da vrijedi

$$H = \frac{1}{\frac{\frac{f_1}{x_1}}{N} + \cdots + \frac{\frac{f_k}{x_k}}{N}} = \frac{1}{\frac{\frac{f_1}{N}}{x_1} + \cdots + \frac{\frac{f_k}{N}}{x_k}} = \frac{1}{\frac{p_1}{x_1} + \cdots + \frac{p_k}{x_k}},$$

gdje su $p_1, ..., p_k$ relativne frekvencije obilježja $x_1, ..., x_k$ redom.
Ako je varijabla zadana statističkim nizom $y_1, ..., y_N$ (s pozitivnim vrijednostima) onda je njezina harmonijska sredina očigledno jednaka

$$H(y_1, ..., y_N) = \frac{N}{\frac{1}{y_1} + \cdots + \frac{1}{y_N}}.$$

Ova veličina se primjenjuje kao adekvatna srednja vrijednost nekih srednjih vrijednosti, tj. omjera istih brojnika.

Primjerice ako se vozimo prosječnom brzinom 100 km/h u jednom smjeru, i 50 km/h u drugom, prosječna brzina nije 75 km/h već $H = \frac{2}{\frac{1}{100}+\frac{1}{50}} = 66.7$ km/h (razlomke $\frac{1}{100}$ i $\frac{1}{50}$ shvaćamo kao vremena potrebna za prevaliti jedinični dio puta, tj. 1 km, pri brzini 100 km/h odnosno 50 km/h).
Ili, ako domaćinstvu A litra mlijeka prosječno traje 5 dana, domaćinstvu B 10 dana, a domaćinstvu C 15 dana, onda prosječno trajanje litre mlijeka u ta 3 domaćinstva nije 10 dana, već $H = \frac{3}{\frac{1}{5}+\frac{1}{10}+\frac{1}{15}} = 8.2$ (razlomke $\frac{1}{5}$, $\frac{1}{10}$ i $\frac{1}{15}$ shvaćamo kao dio litre mlijeka u domaćinstvima A, B i C redom, koja se potroši za 1 dan).

Primjer 1.23 *Ako su zadani prosječni dnevni prometi (u stotinama kuna) u lancu supermarketa po regijama i struktura vrijednosti prometa u regijama u odnosu na cijelo područje, izračunajte prosječni dnevni promet za cijelo područje.*

i-ta regija	$\overline{x_i}$	$p_i 100\%$
1. sjever	*490*	*35%*
2. jug	*494*	*25%*
3. središnja regija	*500*	*40%*

Kad bi ovi postotci predstavljali ukupan udio broja supermarketa n_i i-te regije u njihovom ukupnom broju $N = n_1 + n_2 + n_3$, tj. $p_i = \frac{n_i}{N}$, onda bi ukupan prosječni

dnevni promet bio $\overline{x_1}p_1+\overline{x_2}p_2+\overline{x_3}p_3 = 490\cdot 0.35+494\cdot 0.25+500\cdot 0.4$. *No, postotci predstavljaju udio prometa* $\sum_{k=1}^{n_1} x_k^{(1)}$, $\sum_{k=1}^{n_2} x_k^{(2)}$, $\sum_{k=1}^{n_3} x_k^{(3)}$ *svih* n_1, n_2, n_3 *supermarketa 1., 2. odnosno 3. regije redom, u ukupnom prometu* $\sum_{k=1}^{N} x_k = \sum_{k=1}^{n_1} x_k^{(1)}+\sum_{k=1}^{n_2} x_k^{(2)}+\sum_{k=1}^{n_3} x_k^{(3)}$ *svih supermarketa. Dakle vrijedi* $p_i = \frac{\sum_{k=1}^{n_i} x_k^{(i)}}{\sum_{k=1}^{N} x_k}$, $i = 1,2,3$. *Stoga je prosjek jednak*

$$\frac{\sum_{k=1}^{n_1} x_k^{(1)}+\sum_{k=1}^{n_2} x_k^{(2)}+\sum_{k=1}^{n_3} x_k^{(3)}}{N} = \frac{\sum_{k=1}^{N} x_k}{N} = \frac{1}{\frac{n_1+n_2+n_3}{\sum_{k=1}^{N} x_k}} =$$

$$\frac{1}{\frac{n_1}{\sum_{k=1}^{N} x_k}+\frac{n_2}{\sum_{k=1}^{N} x_k}+\frac{n_3}{\sum_{k=1}^{N} x_k}} = \frac{1}{\frac{\sum_{k=1}^{n_1} x_k^1}{\frac{\sum_{k=1}^{N} x_k}{n_1}}+\frac{\sum_{k=1}^{n_2} x_k^2}{\frac{\sum_{k=1}^{N} x_k}{n_2}}+\frac{\sum_{k=1}^{n_3} x_k^3}{\frac{\sum_{k=1}^{N} x_k}{n_3}}} = \frac{1}{\frac{p_1}{\overline{x_1}}+\frac{p_2}{\overline{x_2}}+\frac{p_3}{\overline{x_3}}} = \frac{1}{\frac{0.35}{490}+\frac{0.25}{494}+\frac{0.40}{500}} = H =$$

494.96.

1.2.6 Momenti

Neka je $X : S \to X(S) = \{x_1, x_2, ..., x_k\} \subseteq \mathbb{R}$ numerička varijabla zadana na konačnom skupu, opsega N, s vrijednostima $y_1, ..., y_N$, distribucijom

$$\{(x_1, f_1), ..., (x_k, f_k)\}$$

i aritmetičkom sredinom μ. **Moment** od X je aritmetička sredina niza odstupanja vrijednosti numeričke varijable od njezine aritmetičke sredine μ (**centralni moment**) ili neke druge vrijednosti (**pomoćni moment**) podignuta na neku potenciju $r \in \mathbb{N}_0$.

Tako je je r-ti centralni moment definiran sa: $\mu_r = \frac{\sum_{i=1}^{k} f_i(x_i-\mu)^r}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{N}(y_i-\mu)^r}{N}$.

Očito je $\mu_1 = 0$, $\mu_2 = \sigma^2$. Moment μ_3 izražava asimetriju podataka u odnosu na μ, dok μ_4 izražava "zaobljenost" distribucije.

Slično, r-ti pomoćni moment oko nule definiramo kao: $m_r = \frac{\sum_{i=1}^{k} f_i x_i^r}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{N} y_i^r}{N}$.

Primijetimo da je $m_1 = \mu$.

1.2.7 Mod

Mod, u oznaci M_o, je ono obilježje statističke varijable koje ima najveću frekvenciju. Možemo reći da je to obilježje koje se najčešće javlja. Određuje se za kvalitativna i kvantitativna obilježja.

Ako je distribucija numeričke varijable grupirana u (prave) razrede, onda se razred s najvećom korigiranom frekvencijom b naziva **modalni razred**. Ako je L_1 donja granica toga razreda, a l njegova večina, te a frekvencija razreda koji prethodi modalnom, a c frekvencija razreda koji slijedi iza modalnoga, onda se mod aproksimira formulom $M_o = L_1 + \frac{(b-a)}{(b-a)+(b-c)} l$.

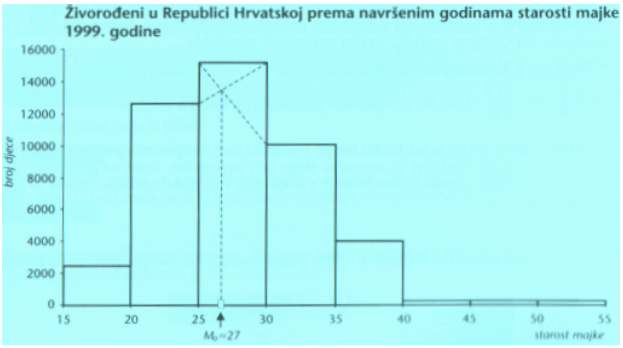
Primijetimo da mod općenito nije jedinstven, odnosno više različitih vrijednosti neke statističke varijable čije frekvencije su maksimalne i međusobno jednake mogu biti njezin mod.

Zadatak 1.24 *U tablici su zadani podaci o dobi majki živorođene djece u Republici Hrvatskoj u 1999. godini. Odredite mod (dob majki s najvećim brojem živorođene djece).*

Dob majke	*Broj živorođene djece*	*Korigirane frekvencije*	*Veličina razreda*
$15-20$		2436	5
$20-25$	12 613	12 613	5
$25-30$	15 183	15 183	5
$30-35$	10 046	10 046	5
$35-40$	4001	4001	5
$40-(55)$	815	271.67	(15)

Rješenje: *Očito je modalni razred* $25-30$*, a mod je*

$M_o = 25 + \frac{(15\,183-12\,613)}{(15\,183-12\,613)+(15\,183-10\,046)} \cdot 5 = 26.67.$

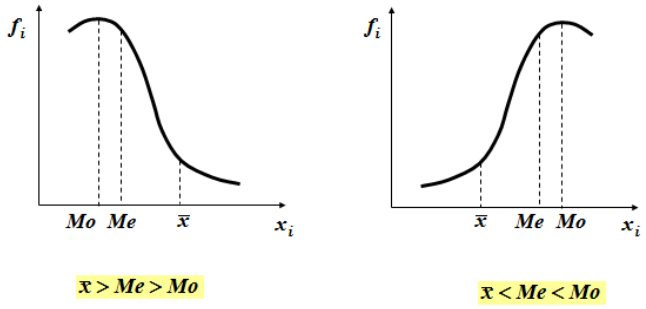


1.2.8 Medijan

Neka je $y_1, ..., y_N$ grupirani statistički niz, tj. neka su to sve vrijednosti numeričke varijable za svih N članova populacije poredane po uređaju $y_1 \leqslant \cdots \leqslant y_N$. Neka je $r = \text{Int}\left(\frac{N}{2}\right) + 1$, gdje Int označuje cijelu vrijednost broja bez decimala (primjerice $\text{Int}\, 6.9 = 6$). U slučaju da je N neparan, **medijan** Me definiramo kao vrijednost y_r središnjeg (r-tog) člana niza . U slučaju da je N paran medijan definiramo kao poluzbroj vrijednosti središnjih članova, tj. $M_e = \frac{y_{r-1}+y_r}{2}$. Medijan ima smisla računati i za ordinalne varijable.

Medijan ima svojstvo da je zbroj apsolutnih odstupanja svih vrijednosti varijable od nekog fiksnog broja minimalan upravo za medijan. U nizu od 100 brojeva s vrijednošću 10 i jednog broja 20 medijan je 10. Iz njegovog tumačenja da prva polovica članova niza ima vrijednost $\leqslant 10$ a druga polovica vrijednosti $\geqslant 10$ uočavamo manjkavosti položajnih srednjih vrijednosti. U ovom primjeru je $M_o = 10$, a $\mu = 10.099$.

Aritmetičku sredinu smijemo zamišljati kao težište poligona frekvencija, mod kao točku u bazi u kojoj je poligon najviši, a medijan kao točku u bazi u kojoj okomiti pravac dijeli poligon frekvencija na dva dijela jednakih površina.



Zadatak 1.25 *U nizu* $1, 3, 3, 2, 5, 3, 7, 7, 8, 8, 10, 11$ *plasmana reprezentacije na svjetskim prvenstvima odredite mod i medijan.*

Rješenje: $M_o = 3,\ M_e = 6.$

Zadatak 1.26 *Studenti su položili ispit sa sljedećim ocjenama: A, C, B, A, D, D, D, A, B, D. Odredite mod i medijan.*

Rješenje: *Mod je* $M_o = D$, *dočim je medijan između ocjene B i C, tj. preračunato u brojčane vrijednosti ove ordinalne varijable medijan je* $M_e = 2.5$.

Ako je distribucija numeričke varijable grupirana u (prave) razrede, onda definiramo **medijalni razred** kao prvi po redu razred $[L_1, L_2]$ čija je kumulativna frekvencija veća ili jednaka $\frac{N}{2}$. Ako je f_{med} frekvencija (nekorigirana) medijalnog razreda, $l = L_2 - L_1$ njegova veličina , $F(L_1)$ kumulativna frekvencija (zbroj svih frekvencija) do medijalnog razreda ,onda se medijan aproksimira vrijednošću

$$M_e = L_1 + \frac{\frac{N}{2} - F(L_1)}{f_{med}} l.$$

Zadatak 1.27 *U tablici su dani podaci o broju nezaposlenih osoba prijavljenih na Hrvatskom zavodu za zapošljavanje krajem* 1999. *Odredite medijan.*

Godine života	*Broj osoba*	*Kumulativni niz*	*Veličina razreda*
$15-20$	67170	67170	5
$20-25$	48482	115652	5
$25-30$	119819	235471	5
$30-40$	82263	317734	10
$40-50$	10604	328338	10
$50-(65)$	13392	341730	(15)
ukupno	341730		

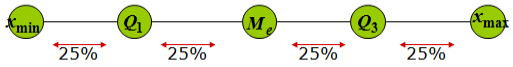
Rješenje: *Budući je* $N = 341730$, *a* $\frac{N}{2} = 170865$, *to je razred* $(25-30)$ *s kumulativnom frekvencijom* 235471 *medijalni razred. Stoga je* $L_1 = 25$, $f_{med} = 119819$, $l = 5$ *i* $F(25) = 115652$. *Slijedi da je medijan* $M_e = L_1 + \frac{\frac{N}{2}-F(25)}{f_{med}} l = 27.304$. *Zaključujemo da je dob prve polovice osoba prijavljenih na HZZ-e iznosila* 27 *ili manje godina, a druga polovica prijavljenih osoba je bila starija od* 27 *godina.*

1.2.9 Kvantili

Neka je $y_1, ..., y_N$ grupirani statistički niz tj. neka su to sve vrijednosti numeričke varijable za svih N članova populacije poredane po uređaju $y_1 \leqslant \cdots \leqslant y_N$. Označimo $r = \text{Int}\left(j\frac{N}{n}\right) + 1$. **Kvantili reda** n su vrijednosti $K_1, ..., K_{n-1}$ koje računamo po formuli

$$K_j = \begin{cases} y_r, & j\frac{N}{n} \notin \mathbb{N} \\ \frac{y_{r-1}+y_r}{2}, & j\frac{N}{n} \in \mathbb{N} \end{cases}, \quad j = 1, ..., n-1.$$

Kvantili reda n određuju n intervala $[y_1, K_1\rangle, \langle K_1, K_2\rangle, ..., \langle K_{n-1}, y_N]$ u svakom od kojih se nalazi najviše $\frac{100}{n}\%$ vrijednosti niza. Kvantil reda 2 je medijan, kvantile Q_1, Q_2, Q_3 reda 4 nazivamo **kvartilima**, kvantile reda 10 **decilima**, a reda 100 **percentilima**.



Zadatak 1.28 *Odredite kvartile u nizu* -1, -3, 0, -1, -1, 5, 0, -3, 1, 2, 3, 3.

Rješenje: *Niz poredajmo po uređaju:* $-3, -3$, -1, -1, -1, 0, 0, 1, 2, 3, 3, 5.

Iz $\frac{N}{4} = 3$, $2\frac{N}{4} = 6$, $3\frac{N}{4} = 9$ *proizlazi* $Q_1 = \frac{y_3+y_4}{2} = -1$, $Q_2 = M_e = \frac{y_6+y_7}{2} = 0$, $Q_3 = \frac{y_9+y_{10}}{2} = 2.5$.

Ako je distribucija numeričke varijable grupirana u (prave) razrede, onda j-ti **kvantilni razred reda** n definiramo kao prvi po redu razred $[L_1, L_2]$ čija je kumulativna frekvencija veća ili jednaka $j\frac{N}{n}$. Ako je f_{kvant} frekvencija (nekorigirana) j-tog kvantilnog razreda, l njegova veličina, $F(L_1)$ kumulativna frekvencija (zbroj svih frekvencija) do j-tog kvantilnog razreda, onda se j-ti kvantil aproksimira vrijednošću

$$K_j = L_1 + \frac{j\frac{N}{n} - F(L_1)}{f_{kvant}} l. \tag{1}$$

Primjer 1.29 *U donjoj tablici su zadani podaci o plaćama zaposlenika jednoga poduzeća u eurima grupirani po platnim razredima.*
Primijetimo da umjesto (kumulativnih) frekvenija f_i *smijemo promatrati (kumalativne) postotke* $p_i 100 = \frac{f_i}{N} 100$. *Stoga, množeći brojnik i nazivnik razlomka iz formule* (1) *sa* $\frac{100}{N}$ *dobivamo formulu*

$$K_j = L_1 + \frac{j\frac{N}{n}\frac{100}{N} - \frac{F(L_1)}{N}100}{\frac{f_{kvant}}{N}100} l = L_1 + \frac{j\frac{100}{n} - \frac{F(L_1)}{N}100}{p_{kvant}100} l,$$

gdje je p_{kvant} *proporcija* j*-tog kvantilnog razreda reda* n*, tj. proporcija prvog po redu razreda čiji je kumulativni postotak veći ili jednak* $j\frac{100}{n}$. *Iz* $\frac{N}{4} = 25$, $2\frac{N}{4} = 50$, $3\frac{N}{4} = 75$ *proizlazi da je* 1. *kvartilni razred* $1500.5 - 1700.5$, *medijalni razred je* $1700.5 - 1900.5$, *a* 3. *kvartilni razred je* $1900.5 - 2100.5$. *Nadalje, vrijedi* $Q_1 = 1500.5 + \frac{25-21.7}{16.5}200 = 1540.5$, $M_e = 1700.5 + \frac{50-38.2}{23.8} \cdot 200 = 1799.7$, $Q_3 = 1900.5 + \frac{75-62}{14.9}200 = 2075$. *Možemo zaključiti da do četvrtine zaposlenika ima plaću manju od* 1540.5 €, *do polovine zaposlenika ima plaću manju od* 1799.7 €, *dok do četvrtine*

zaposlenika ima plaću veću od 2075 €.

Iznos plaće	*Postotak* $p_i100\%$	*Kumulativni niz postotaka*	*Veličina razreda*
$499.5 - 700.5$	0.1	0.1	200
$700.5 - 900.5$	0.2	0.3	200
$900.5 - 1100.5$	2.6	2.9	200
$1100.5 - 1300.5$	6.5	9.4	200
$1300.5 - 1500.5$	12.3	21.7	200
$1500.5 - 1700.5$	16.5	38.2	200
$1700.5 - 1900.5$	23.8	62.0	200
$1900.5 - 2100.5$	14.9	76.9	200
$2100.5 - 2300.5$	11.1	88.0	200
$2300.5 - 2500.5$	7.0	95.0	200
$2500.5 - 3000.5$	4.2	99.2	500
$3000.5 - 4000.5$	0.8	100.00	1000

1.2.10 Nepotpune mjere disperzije

Raspon varijacije R_X numeričke varijable X jest razlika između najveće i najmanje vrijednosti varijable, ako takve postoje (kod beskonačnih populacija varijabla ne mora imati svoj minimum i maksimum) $R_X = X_{\max} - X_{\min}$.

Budući u izračun raspona ulaze samo dvije vrijednosti (koje mogu biti outlayeri) taj parametar ne uzima u obzir variranje podataka. Raspon ima smisla i za ordinalnu varijablu.

Interkvartilom I_Q numeričke ili ordinalne varijable nazivamo razliku gornjeg i donjeg kvartila tj. $I_Q = Q_3 - Q_1$. Možemo reći da je interkvartil raspon varijacije središnjih 50% članova uređenog niza. Slično i **interdecil** $I_D = D_9 - D_1$ je raspon središnjih 80% podataka, a **interpercentil** $I_P = P_{99} - P_1$ je raspon središnjih 98% podataka.

Pripadajuća relativna mjera je **koeficijent interkvartilne devijacije** $V_Q = \frac{Q_3 - Q_1}{Q_1 + Q_3}$, koji ima smisla samo za varijable s pozitivnim vrijednostima. Vrijedi $0 \leqslant V_Q < 1$. Što je V_Q bliže 0, to je varijabilnost središnjih 50% podataka manja.

Zadatak 1.30 *Odredite raspon i koeficijent interkvartilne devijacije za niz:* 1, 2, 4, 4, 6, 9, 9, 9, 10, 100.

Rješenje: $R = 100 - 1 = 99$, $Q_1 = 4$, $Q_3 = 9 \Rightarrow V_Q = \frac{9-4}{9+4} = 0.38$.

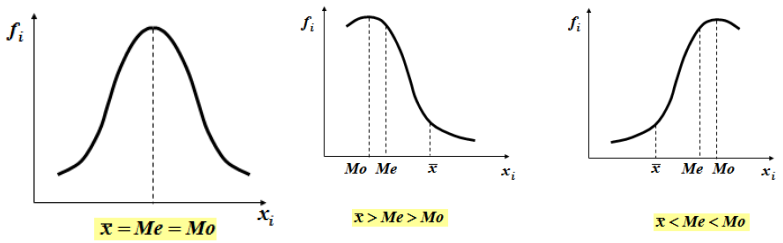
Zadatak 1.31 *Odredite raspon i koeficijent interkvartilne devijacije za niz:* 1000, 1100, 1111, 1150, 1160, 1180.

Rješenje: $R = 180$, $V_Q = \frac{60}{2260} = 0.026$.

1.2.11 Mjere asimetrije i zaobljenosti

Mjera asimetrije varijable je parametar koji daje informaciju o načinu rasporeda podataka prema aritmetičkoj sredini. Najvažnija mjere asimetrije je **koeficijent asimetrije** $\alpha_3 = \frac{\mu_3}{\sigma^3}$ (najčešće vrijednosti su iz $\langle -2, 2 \rangle$, u simetričnom rasporedu je $\alpha_3 = 0$), Pearsonova mjera, Bowleyjeva mjera.

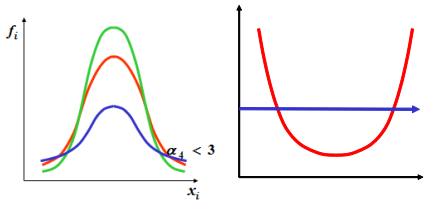
Grafovi distribucija frekvencija varijable pokazuju redom simetričnu varijablu ($\alpha_3 = 0$), pozitivno asimetričnu ($\alpha_3 > 0$) i negativno asimetričnu ($\alpha_3 < 0$) varijablu.



Mjera zaobljenosti varijable je parametar koji daje informaciju o zaobljenosti modalnog vrha na poligonu distribucije frekvencija varijable. Zaobljenost se mjeri **koeficijentom zaobljenosti** $\alpha_4 = \frac{\mu_4}{\sigma^4}$ čija vrijednost je pozitivna.

Grafovi distribucija frekvencije varijable pokazuju redom "normalno" zaobljenu varijablu ($\alpha_4 = 3$), više zaobljenu ($\alpha_3 > 3$) i manje zaobljenu ($\alpha_3 < 3$) varijablu,

te nezaobljenu $\alpha_4 = 1.8$ i U-zaobljenu $\alpha < 1.8$.



Poglavlje 2

Vjerojatnost

2.1 Događaji slučajnog pokusa

Predmet zanimanja inferencijalne statistike su **slučajni pokusi**, tj. djelatnosti mjerenja opažanja ili definirani procesi iz kojih izviru neki rezultati. Ishodi ili rezultati slučajnog pokusa nisu jednoznačno određeni i ne mogu se unaprijed predvidjeti na temelju uvjeta pokusa. Međutim, ako se takvi pokusi ponavljaju dovoljno mnogo puta, dolazi se do odgovarajućih zakonitosti. Proučavanje tih zakonitosti je predmet teorije vjerojatnosti. Za razliku od slučajnog pokusa, **deterministički pokus** je jednoznačno određen uvjetima pokusa.
Podrazumijevamo da svaki slučajni pokus ima statističke značajke: može se ponavljati proizvoljan broj puta, unaprijed je poznato što se registrira kao i svi mogući ishodi, pri čemu ishod pojedinačnog pokusa nije poznat. Ako bacamo predmet s visine h i registriramo vrijeme koje je potrebno predmetu da udari o tlo, onda je taj pokus deterministički čim se izvodi u laboratorijskim uvjetima (ovisi samo o visini h), a slučajan ako se izvodi u vanjskim uvjetima (ovisi o nizu uvjeta koji utječu na ishod).

Definicija 2.1 *Skup kojega tvore svi mogući ishodi nekog pokusa nazivamo* ***prostorom elementarnih događaja*** *toga pokusa i označujemo sa* Ω. ***Događaj*** *je neki određeni podskup prostora elementarnih događaja* Ω. *Jednočlani događaj* $\{\omega\}$ *nazivamo* ***elementarnim događajem*** *tj. to je svaki događaj koji se ne može*

rastaviti na jednostavnije događaje (često identificiramo elementarni događaj $\{\omega\}$ *sa ishodom* ω*). U protivnom kažemo da je događaj* ***složen.*** *Ako pokus ima za ishod* ω *onda kažemo da se neki događaj* A *dogodio pri tomu pokusu ako je* $\omega \in A$ *tj. ako* ω *pripada skupu* A*. Podskup* $\emptyset$ *nazivamo* ***nemogućim događajem****, a* Ω *nazivamo* ***siguranim događajem****.*

Primjer 2.2 *Prostor elementarnih događaja bacanja novčića je* $\Omega = \{P, G\}$*, gdje* P *označuje da je ishod bacanja novčića pismo, a* G *da je ishod glava. Elementarni događaji su* $\{P\}$ *i* $\{G\}$*. Jedini složeni događaji je* $\{P, G\} -$*ishod bacanja je pismo ili glava, dok* $\emptyset$ *označuje nemoguć događaj-ishod bacanja nije ni pismo ni glava. Prostor elementarnih događaja bacanja kocke dva puta za redom je* $\Omega = \{(i, j) \mid i, j = 1, ..., 6\}$*. Ukupan broj elementarnih događaja je* 36*, a ukupan broj događaja je* 2^{36} *(broj svih podskupova od* Ω*). Primjerice, skup* $\{(1,6), (2,6), (3,6), (4,6), (5,6), (6,6), (6,1), (6,2), (6,3), (6,4), (6,5)\}$ *predstavlja događaj da je u barem jednom bacanju kocke pala šestica, a događaj da je zbroj brojeva dobivenih u dva bacanja kocke jednak* 4 *je* $\{(1,3), (2,2), (3,1)\}$*. Prostor elementarnih događaja trajanja neke žarulje je* $\Omega = [0, \infty\rangle$*.*

2.1.1 Operacije s događajima

U opisivanju pokusa i događaja koristimo se jezikom teorije skupova između ostalog jer skupovne operacije najbolje ilustriraju operacije s događajima. Pretpostavimo da su $A, B \subseteq \Omega$ događaji. Kažemo da događaj A povlači događaj B ako je $A \subseteq B$ (kad god se dogodi A onda se dogodi i B). Suprotan događaj događaju A je događaj $A^c = \Omega \backslash A$ (A^c se dogodi točno onda kada se ne dogodi A). Presjek događaja A i B je događaj $A \cap B$ ($A \cap B$ se dogodi točno onda kada se dogodi i A i B). Analogno se definira presjek $A_1 \cap \cdots \cap A_n$ konačno mnogo i presjek $A_1 \cap \cdots \cap A_n \cap \cdots = \bigcap_{k=1}^{\infty} A_k$ prebrojivo mnogo događaja. Unija događaja A i B je događaj $A \cup B$ ($A \cup B$ se dogodi točno onda kada se dogodi A ili B). Analogno se definira unija $A_1 \cup \cdots \cup A_n$ konačno mnogo i unija $A_1 \cup \cdots \cup A_n \cap \cdots = \bigcup_{k=1}^{\infty} A_k$ prebrojivo mnogo događaja. Razlika događaja A i B je događaj $A \backslash B$ ($A \backslash B$ se dogodi točno onda kada se dogodi A i ne dogodi B). Događaji A i B su međusobno

isključivi ako je $A \cap B = \emptyset$.

Primjer 2.3 *Baca se novčić 3 puta. Prostor elementarnih događaja je* $\Omega = \{(P,P,P), (P,P,G), (P,G,P), (P,G,G), (G,P,P), (G,P,G), (G,G,P), (G,G,G)\}$. *Događaj da je barem dva puta pala glava je* $A = \{(P,G,G), (G,P,G), (G,G,P), (G,G,G)\}$. *Suprotan događaj događaju A je događaj da je najviše jednom pala glava tj.* $A^c = \{(P,P,P), (P,P,G), (P,G,P), (G,P,P)\}$.

Ako u istom pokusu registriramo koliko je puta palo pismo, onda je novi prostor elementarnih događaja $\Omega' = \{0,1,2,3\}$. *Događaj da je pismo palo najviše 2 puta je* $B = \{0,1,2\}$, *a događaj da je pismo palo barem jednom je* $C = \{1,2,3\}$. *Presjek događaja B i C je događaj* $B \cap C = \{1,2\}$. *Unija događaja B i C je siguran događaj* $B \cup C = \Omega'$. *Razlika događaja B i C je* $B \backslash C = \{0\}$.

2.2 Vjerojatnost događaja

Teorija vjerojatnosti daje pravila kako se polazeći od jednih (najčešće elementarnih) vjerojatnosti izračunavaju druge. Do polaznih vjerojatnosti možemo doći na različite načine, odnosno one mogu biti: subjektivne, statističke (a posteriori) i klasične matematičke (a priori i geometrijske). Intuitivno, vjerojatnošću nekog događaja smatramo broj iz intervala $[0,1]$, koji iskazuje određeni stupanj izvjesnosti da se taj događaj, vezan za slučajno pokus, dogodi. Pri tomu je vjerojatnost sigurnog događaja 1, a vjerojatnost nemogućeg događaja jednaka 0.

Subjektivna vjerojatnost je broj $p(A)$ iz intervala $[0,1]$ određen na temelju uvjerenja i prosudbe okolnosti relevantnih za nastup slučajnog događaja A. U praksi se utvrđuje kad nije moguće utvrditi vjerojatnost na klasičan način ili kad nema empirijskih podataka za statistički način. Primjerice, vjerojatnost pobjede u meču tenisača koji je 450. na ATP listi tenisača nad tenisačem koji je 451. na ATP listi, ako pri tom nisu nikad prije odmjerili snage, spada u subjektivnu prosudbu eksperata za tenis.

Primjer 2.4 *Predviđanja stručnjaka iz određenog gospodarskog instituta o inflaciji za sljedeću godinu dana tablicom:*

stupanj inflacije	*vjerojatnost*
10% i više	*0,03*
od 5% do 10%	*0,85*
manje od 5%	*0,12*

su primjer subjektivnih vjerojatnosti.

Neka neki slučajni pokus ponovimo n puta i neka se pri tome događaj A dogodi n_A puta. Tada broj $f_n(A) = \frac{n_A}{n}$ nazivamo **relativnom frekvencijom događaja** A u n **ponavljanja pokusa**. Pri malom broju ponavljanja pokusa relativna frekvencija događaja nosi u sebi slučajni karakter i može se značajno mijenjati od jedne do druge serije pokusa. Ako pri uvećanju broja ponavljanja pokusa relativna frekvencija sve više gubi slučajni karakter i sve više se grupira oko određenog broja onda taj broj nazivamo **statistička** ili **a posteriori vjerojatnost** događaja A i označujemo ga sa $p(A)$. Dakle vrijedi $p(A) = \lim_{n\to\infty} f_n(A)$. U praksi statističku vjerojatnost koristimo ako nije moguće doći do matematičke vjerojatnosti (a priori) i obično ju aproksimiramo relativnom frekvencijom $f_n(A)$ koja je to bolja što je n (broj ponavljanja pokusa) veći.

Ako su A i B međusobno isključivi događaji, onda je $n_{A\cup B} = n_A + n_B$, pa iz $f_n(A \cup B) = \frac{n_{A\cup B}}{n} = \frac{n_A}{n} + \frac{n_B}{n} = f_n(A) + f_n(B)$ zaključujemo da je

$$p(A \cup B) = p(A) + p(B). \tag{1}$$

Nadalje, iz $n_{A^c} = n - n_A$ slijedi $f_n(A^c) = \frac{n_{A^c}}{n} = \frac{n-n_A}{n} = 1 - \frac{n_A}{n} = 1 - f_n(A)$, pa zaključujemo da je

$$p(A^c) = 1 - p(A). \tag{2}$$

Očito je statistička vjerojatnost broj iz $[0, 1]$ koji iznosi 1 za siguran događaj, a 0 za nemoguć događaj. Ova prirodna svojstva statističke vjerojatnosti ćemo zahtijevati od bilo koje druge vjerojatnosti, stoga ćemo ih kasnije ugraditi u aksiome vjerojatnosti.

Primjer 2.5 *Ako je u tijeku jedne godine proizvedeno* 500 000 *komada nekog uređaja, od kojih je* 5 000 *bilo odmah neispravno, a* 1 000 *se pokvarilo tijekom prvoga tjedna, i ako pretpostavimo da se proizvodnja nastavlja u nepromijenjenim*

uvjetima, onda vjerojatnost događaja A da jedan slučajno odabrani proizvod bude odmah neispravan možemo jedino izraziti kao statističku vjerojatnost aproksimiranu relativnom frekvencijom $p(A) = f_n(A) = \frac{5000}{500000} = 0.01$. *Vjerojatnost da se slučajno odabrani proizvod ne pokvari nakon tjedan dana je* $p((A \cup B)^c) = 1 - p(A \cup B) = 1 - \left(\frac{5000}{500000} + \frac{1000}{500000}\right) = 1 - 0.01 - 0.002 = 0.988$, *gdje je B događaj da se slučajno odabrani proizvod pokvario tijekom prvoga tjedna.*

Pretpostavimo da neki pokus ima konačno mnogo ishoda $\omega_1, ..., \omega_n$ takvih da su svi elementarni događaji jednako vjerojatni, tj. da je $p(\omega_1) = \cdots = p(\omega_n)$. Tada iz $\Omega = \{\omega_1\} \cup \cdots \cup \{\omega_n\}$ i svojstva (1) slijedi $1 = p(\Omega) = p(\omega_1) + \cdots + p(\omega_n)$, što povlači $n \cdot p(\omega_i) = 1$, odnosno $p(\omega_i) = \frac{1}{n}$, $i = 1, ..., n$. Ako se događaj $A = \{\omega_{i_1}, ..., \omega_{i_m}\}$ sastoji od m ishoda (kažemo još da je m broj povoljnih elementarnih događaja za događaj A, a n broj svih mogućih ishoda), onda je vjerojatnost događaja A jednaka broju $p(A) = p(\omega_{i_1}) + \cdots + p(\omega_{i_m}) = \frac{1}{n} + \cdots + \frac{1}{n} = \frac{m}{n}$ kojega nazivamo **vjerojatnost a priori**.

Zadatak 2.6 *Iz kutije od* 50 *sijalica, od kojih je* 5 *neispravnih, se izvlači jedna. Kolika je vjerojatnost da ona bude neispravna?*

Rješenje: *Budući se svaka sijalica može izvući s jednakom vjerojatnošću, to je ukupan broj ishoda* 50, *a broj povoljnih* 5, *pa je vjerojatnost jednaka* $\frac{5}{50} = 0.1$.

Zadatak 2.7 *U šeširu se nalazi* 40 *karata (briškulice). Iz šešira izvlačimo jednu kartu. Kolika je vjerojatnost da je izvučena karta boje špadi, a kolika da nije konj ili kralj?*

Rješenje: *Svaka karta se može izvući s jednakom vjerojatnošću, a broj svih mogućih ishoda je* 40. *Broj povoljnih ishoda da se izvuče špada je* 10, *pa je vjerojatnost toga događaja jednaka* $\frac{10}{40} = 0.25$. *Ako sa A označimo događaj da je izvučen konj, a sa B da je izvučen kralj, onda je* $p(A) = \frac{4}{40} = 0.1 = p(B)$, *pa je* $p((A \cup B)^c) = 1 - p(A \cup B) = 1 - 0.1 - 0.1 = 0.8$.

Zadatak 2.8 *U siječnju 1992. je u R.H. ostvareno* 191 018 *noćenja i to: gostiju iz R.H.* 137 921, *iz republika ex. SFRJ (bez R.H.)* 29 191 *i iz ostalih stranih zemalja* 23 906. *Kolika je vjerojatnost da je slučajno odabrano noćenje ostvarila osoba koja nije iz R.H.?*

Rješenje: *Ishod ovoga slučajnoga odabira može biti bilo koje od* 191 018 *noćenja s jednakom vjerojatnošću* $\frac{1}{191018}$. *Ako je* A *događaj da je noćenje ostvarila osoba koja je iz neke republike ex. SFRJ, a* B *osoba iz neke druge strane zemlje, onda je* $p(A) = \frac{29191}{191018} = 0.15282$, $p(B) = \frac{23906}{191018} = 0.12515$, *pa je* $p(A \cup B) = p(A) + p(B) = 0.27797$. *Primijetimo da je* $p(A \cup B) = 1 - \frac{137921}{191018}$.

Primjer 2.9 *De Mere je bilježio rezultate igre koja se sastojala od bacanja 3 različite kocke. Promatrao je događaj* A_1 *da je ukupan zbroj brojeva na 3 kocke jednak 11 i događaj* A_2 *da je ukupan zbroj brojeva na kockama jednak 12. De Mere je ustanovio da se događaj* A_1 *pojavljuje češće nego* A_2, *a smatrao je da bi se ta dva događaja trebali pojavljivati podjednako često. Naime, događaj* A_1 *se sastoji od 6 mogućnosti pojavljivanja brojeva u jednom bacanju, tj. od* 6 *kombinacija:* $6,4,1$; $6,3,2$; $5,5,1$; $5,4,2$; $5,3,3$; $4,4,3$ *i događaj* A_2 *se sastoji od 6 kombinacija:* $6,5,1$; $6,4,2$; $6,3,3$; $5,5,2$; $5,4,3$; $4,4,4$. *Grešku u zaključivanju je našao Pascal koji je dokazao da ishodi koje je naveo De Mere nisu jednako vjerojatni. Naime, prostor elementarnih događaja za ovaj pokus je* $\Omega = \{(i,j,k) \mid i,j,k = 1,...,6\}$. *Svi elementarni događaji su jednako vjerojatni i vjerojatnost im je* $\frac{1}{6^3} = \frac{1}{216}$. *Događaj da su se na kockama pojavili brojevi* 6, 4 *i* 1 *se može rastaviti na sljedeće ishode:* $(1,4,6)$, $(1,6,4)$, $(4,1,6)$, $(4,6,1)$, $(6,1,4)$ *i* $(6,4,1)$, *pa je odgovarajuća vjerojatnost jednaka* $\frac{6}{216}$. *Događaj da se na kockama pojave brojevi* $4,4,4$ *se samo sastoji od ishoda* $(4,4,4)$, *pa je njegova vjerojatnost jednaka* $\frac{1}{216}$. *Događaj da se na kockama pojave brojevi* $6,3,3$ *se sastoji od ishoda:* $(6,3,3)$, $(3,6,3)$, $(3,3,6)$, *pa je odgovarajuća vjerojatnost jednaka* $\frac{3}{216}$. *Zbrajanjem odgovarajućih događaja dobijemo* $p(A_1) = \frac{27}{216}$ *i* $p(A_2) = \frac{25}{216}$.

Ako je prostor elementarnih događaja Ω neprebrojiv (skup $\mathbb{R}$), a svi elementarni događaji jednako vjerojatni, nema smisla primijeniti formulu za a priori vjerojatnost. No, ako se skup Ω može prikazati kao ograničeni skup na pravcu, ravnini ili prostoru, čija je mjera (duljina, površina ili volumen) jednaka $\mu(\Omega)$, a mjera događaja (podskupa) A je $\mu(A)$, onda je vjerojatnost događaja A jednaka broju $p(A) = \frac{\mu(A)}{\mu(\Omega)}$ kojega nazivamo **geometrijska vjerojatnost**. Primjerice, zamislimo da sa strelicom, čiji je šiljak savršeno (beskonačno) tanak, gađamo u

pluteni zid pred nama, kojemu je površina 10, u metu istaknutu na tomu zidu površine 2. Vjerojatnost pogotka mete je jednak kvocijentu površina (mjera) tih dvaju skupova (zida i mete) tj. $\frac{2}{10} = 0.2$. Vjerojatnost pogotka bilo kojega dijela ciljanog zida čija je površina jednaka 0 je ništična. Primjerice, vjerojatnost da se pogodi unaprijed određena točka na zidu ili dužina (ovi objekti su 0-dimenzionalni, odnosno 1-dimenzionalni, pa su površine 0) jednaka je 0. Iako ovi događaji nisu nemogući (razlikuju se od $\emptyset$), njihova vjerojatnost je 0, što se donekle opire našoj percepciji vjerojatnosti.

Primjer 2.10 *Kolika je vjerojatnost da slučajno izgeneriran broj iz* $[0,1]$ *bude jednak* $\frac{1}{2}$*, a kolika da pripada segmentu* $\left[\frac{1}{4}, \frac{3}{4}\right]$*?*

Ovdje je $\Omega = [0,1]$*, događaj* $A = \left\{\frac{1}{2}\right\}$ *i* $B = \left[\frac{1}{4}, \frac{3}{4}\right]$*, pa je* $p(A) = \frac{\mu(A)}{\mu(\Omega)} = \frac{0}{1} = 0$ *i* $p(B) = \frac{\mu(B)}{\mu(\Omega)} = \frac{\frac{1}{2}}{1} = \frac{1}{2}$ *(μ označuje duljinu).*

Ovo možemo interpretirati na sljedeći način: ako "gađamo" u realne brojeve onda je vjerojatnost pogotka u unaprijed određeni segment, ma kako uzak bio, uvijek veća od 0*, dok je vjerojatnost pogotka u unaprijed određeni broj jednaka* 0*.*

2.3 Vjerojatnosni prostor

Naše intuitivno poimanje vjerojatnosti, kao i neka prirodna svojstva koja proizlaze iz takvoga poimanja, ćemo ugraditi u aksiomatski okvir vjerojatnosti. Skup aksioma kojima ćemo opisati vjerojatnosni prostor nam daje uvjete kojima vjerojatnost promatranih događaja mora udovoljiti i pomoću kojih možemo istraživati vjerojatnosti složenijih događaja. No, što će biti početna vjerojatnost, tj. vjerojatnost pojedinih elementarnih događaja, ovisiti će o svakom pojedinom slučaju i određivati će se na način opisan u prethodnom odjeljku.

Neka je Ω prostor elementarnih događaja. Na njemu treba prvo zadati familiju $\mathcal{F}$ svih mogućih događaja (to je neki podskup od partitivnog skupa $\mathcal{P}(\Omega)$) koju ćemo nazivati **familijom događaja**. Ta familija treba udovoljavati nekim razumnim zahtjevima:

- $\emptyset \in \mathcal{F}$;

- $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
- $A_i \in \mathcal{F},\ i \in \mathbb{N} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Funkcija $p : \mathcal{F} \to [0,1]$ koja svakom događaju A pridružuje broj $p(A)$ se naziva **vjerojatnost** ako je

$$p(\Omega) = 1,\ A_i \in \mathcal{F},\ i \in \mathbb{N},\ A_i \cap A_j = \emptyset \text{ za } i \neq j \Rightarrow p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i).$$

Definicija 2.11 *Uređenu trojku* $(\Omega, \mathcal{F}, p)$, *gdje je* Ω *prostor elementarnih događaja,* $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ *familija događaja i* $p : \mathcal{F} \to [0,1]$ *vjerojatnost, nazivamo* ***vjerojatnosni prostor****.*

Vrijede sljedeća svojstva vjerojatnosti:

- $p(\emptyset) = 0$ (vjerojatnost nemogućeg događaja);
- $p(A_1 \cup \cdots \cup A_n) = p(A_1) + \cdots + p(A_n)$, ako su $A_1, ..., A_n$ međusobno isključivi;
- $p(A^c) = 1 - p(A)$ (vjerojatnost suprotnog događaja);
- $A \subseteq B \Rightarrow p(A) \leq p(B)$;
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ (vjerojatnost da nastupi barem jedan od događaja A i B);
- $p(A \backslash B) = p(A) - p(A \cap B)$.

Zadatak 2.12 *Neka je* $\Omega = \{\omega_1, \omega_2, \omega_3\}$ *prostor elementarnih događaja nekoga slučuajnoga pokusa i* $\mathcal{F} = \mathcal{P}(\Omega)$ *familija događaja. Može li funkcija* $p : \mathcal{F} \to [0,1]$ *za koju vrijedi* $p(\omega_1) = 0.1$, $p(\omega_2) = 0.8$ *i* $p(\omega_3) = 0.2$ *biti vjerojatnost?*

Rješenje: *Budući su* $\{\omega_1\}$, $\{\omega_2\}$ *i* $\{\omega_3\}$ *isključivi događaji, to po aksiomima vjerojatnosti mora vrijediti* $p(\Omega) = 1$ *s jedne strane i*

$$p(\Omega) = p(\{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\}) = p(\omega_1) + p(\omega_2) + p(\omega_3) = 1.1,$$

s druge strane. Stoga p *ne može biti vjerojatnost.*

Primjer 2.13 *U nekom mjestu ima* 4111 *stanovnika. Od toga su* 3998 *hrvatski državljani, a njih* 750 *ima strano državljanstvo. Kolika je vjerojatnost da slučajno odabrani stanovnik toga mjesta ima uz hrvatsko, državljanstvo barem još jedne zemlje?*

Prostor Ω *elementarnih događaja ovog slučajnog pokusa se sastoji od svih stanovnika mjesta, tj. elementarni događaj je da je odabran jedan stanovnik od njih* 4111*. Događaj* A *je da odabrani stanovnik ima hrvatsko državljanstvo, događaj* B *je da ima strano državljanstvo, a događaj* $A \cap B$ *je da ima i hrvatsko i neko strano državljanstvo. Tada je*

$1 = p(\Omega) = p(A) + p(B) - p(A \cap B) = \frac{3998}{4111} + \frac{750}{4111} - p(A \cap B) \Rightarrow p(A \cap B) = 0.9725 + 0.1824 - 1 = 0.154\,9.$

2.3.1 Diskretni vjerojatnosni prostor

U slučaju kada je prostor elementarnih događaja Ω konačan ili prebrojiv ($\Omega = \{\omega_1, ..., \omega_n\}$ ili $\Omega = \{\omega_i \mid i \in \mathbb{N}\}$), za familiju svih događaja uzimamo partitivni skup $\mathcal{F} = \mathcal{P}(\Omega)$, pa svih mogućih događaja ima 2^n, a vjerojatnost je dovoljno zadati samo za elementarne događaje $p_i = p(\{\omega_i\})$. Takav vjerojatnosni prostor nazivamo **diskretnim**.

Tada je $p(\Omega) = \sum_i p_i = p_1 + p_2 + \cdots = 1$, a vjerojatnost proizvoljnog događaja A jednaka je zbroju vjerojatnosti svih ishoda uključenih u A (još ćemo reći povoljnih za A), tj. $p(A) = \sum_{\omega_i \in A} p_i$.

Primjer 2.14 *Iz kutije u kojoj se nalazi* 10 *crvenih,* 4 *crne i* 6 *bijelih kuglica se izvlači nasumce jedna kuglica. Ako je elementarni događaj odabir bilo koje kuglice, onda je vjerojatnost svakog elementarnog događaja* $\frac{1}{20}$*. Budući je prostor elementarnih događaja* Ω *konačan (sastoji se od* 20 *elementarnih događaja) to je odgovarajući vjerojatnosni prostor diskretan. Vjerojatnost događaja "izvučena je crna kuglica" je* $\frac{1}{20} + \frac{1}{20} + \frac{1}{20} + \frac{1}{20} = \frac{4}{20} = 0.2$*. Vjerojatnost događaja "nije izvučena crna kuglica" je* $1 - 0.2 = 0.8$ *ili* $\frac{10}{20} + \frac{6}{20} = 0.8$*.*

Primjer 2.15 *Pokus se sastoji od uzastopnog bacanja jednog novčića. Ishod pokusa je broj bacanja do prvog nastupa pisma. Pripadni vjerojatnosni prostor se sastoji od*

skupa $\Omega = \mathbb{N}$*, familije događaja* $\mathcal{P}(\mathbb{N})$ *i vjerojatnosti zadane samo na elementarnim događajima. Elementarnom događaju* $\{k\}$*(u* k*-tom bacanju novčića je prvi put palo pismo) pridružujemo vjerojatnost* $p(\{k\}) = \frac{1}{2^k}$ *(svih mogućih, jednako vjerojatnih, ishoda* $(x_1, ..., x_k)$*, gdje* $x_i \in \{P, G\}$ *označuje rezultat* i*-tog bacanja, kod bacanja novčića* k *puta ima* 2^k*, a samo je jedan povoljni ishod* $(G, ..., G, P)$*). Sada se vjerojatnost lako proširi na sve događaje. Primjerice, događaj* $A = \{3, 4, 5\}$ *da je novčić bačen od 3 do 5 puta ima vjerojatnost* $p(A) = p(\{3\}) + p(\{4\}) + p(\{5\}) = \frac{1}{2^3} + \frac{1}{2^4} + \frac{1}{2^5}$.

Događaj $B = \{2, 3, ...\}$ *da je novčić bačen barem 2 puta ima vjerojatnost* $p(B) = p(\Omega) - p(\{1\}) = 1 - \frac{1}{2} = \frac{1}{2}$.

Zadatak 2.16 *Neka je vjerojatnost (subjektivna) da će domaćin pobijediti u nogometnoj utakmici jednaka* 0.5*, vjerojatnost da će igrati neriješeno jednaka* 0.25*, vjerojatnost da neće postići zgoditak jednaka* 0.35 *i vjerojatnost da će izgubiti uz barem jedan postignuti zgoditak jednaka* 0.1*. Kolika je vjerojatnost da će rezultat biti 0:0, a kolika da će rezultat biti 1:1, 2:2 itd.?*

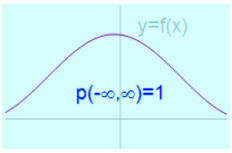
Rješenje: *U ovomu vjerojatnosnom prostoru prostor elementarnih događaja tvore svi rezultati promatrane nogometne utakmice. Vjerojatnost nije zadana za svaki elementarni događaj, već samo na nekim složenim događajima. Vjerojatnost da će domaćin izgubiti jednaka je* $1 - 0.5 - 0.25 = 0.25$*. Vjerojatnost da će domaćin izgubiti bez postignutog zgoditka je jednaka* $0.25 - 0.1 = 0.15$*. Vjerojatnost da će rezultat biti* $0 : 0$ *(neće izgubiti, a neće ni postići zgoditak) je* $0.35 - 0.15 = 0.2$*. Vjerojatnost da će rezultat biti neriješen uz postignute zgoditke je* $0.25 - 0.2 = 0.05$.

Zadatak 2.17 *U nekom društvu je* 1% *ljudi s nezavršenom osnovnom školom,* 21% *sa završenom samo osnovnom školom,* 78% *sa završenom srednjom školom ili više, te* 11% *sa završenom višom školom, fakultetom ili više. Kolika je vjerojatnost da slučajno odabrani građanin bude sa samo završenom srednjom školom?*

Rješenje: *Ako je* A*-događaj da građanin nema završenu osnovnu školu,* B*-završenu samo osnovnu,* C*-završenu samo srednju,* D*-završenu višu školu, fakultet ili više, onda je* $p(A) = 0,01$*,* $p(B) = 0,21$*,* $p(C \cup D) = 0,78$ *i* $p(D) = 0,11$*. Tada je* $1 = p(\Omega) = p(A) + p(B) + p(C) + p(D) = 0,01 + 0,21 + p(C) + 0,11 \Rightarrow p(C) = 1 - 0.01 - 0.21 - 0.11 = 0.67$.

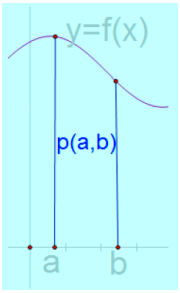
2.3.2 Nediskretni vjerojatnosni prostor

Pretpostavimo sada da su Ω i $\mathcal{F}$ neprebrojivi i da vjerojatnost događaja poprima sve vrijednosti iz $[0,1]$. Proučavat ćemo vjerojatnosni prostor $(\Omega, \mathcal{F}, p)$ takav da je $\Omega = \mathbb{R}^n$. Familija svih događaja $\mathcal{F}$ ne mora biti čitav partitivan skup (ima podskupova na $\mathbb{R}^n$ koji nisu događaji), ali mora sadržavati sve vrste intervala $\langle a,b\rangle^n$, $[a,b\rangle^n$, $\langle a,b]^n$ i mora udovoljavati standardnim zahtjevima. Proučavat ćemo vjerojatnost koja je zadana na događajima koji su produkti intervala (ishod pokusa je uređena n-torka realnih brojeva koji su u promatranom intervalu), a zatim se primjenom svojstava vjerojatnosti definicija proširuje na sve ostale događaje.
Radi jednostavnosti, ograničit ćemo se samo na slučaj $\Omega = \mathbb{R}$, tj. na vjerojatnosni prostor oblika $(\mathbb{R}, \mathcal{F}, p)$ kojemu je vjerojatnost p zadana na intervalima, i to pomoću funkcije $f : \mathbb{R} \to \mathbb{R}$, $f(x) \geqslant 0$, $x \in \mathbb{R}$, takve da je površina između osi x i krivulje $y = f(x)$ jednaka 1, tj. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.



Definicija 2.18 *Ako je vjerojatnost na nediskretnom vjerojatnosnom prostoru $(\mathbb{R}, \mathcal{F}, p)$ zadana formulom $p(\langle a,b\rangle) = \int_a^b f(x)\,dx$, koja određuje vjerojatnost intervala $\langle a,b\rangle$ (događaja da je ishod pokusa broj iz $\langle a,b\rangle$), onda funkciju f nazivamo* ***gustoćom*** *vjerojatnosti.*

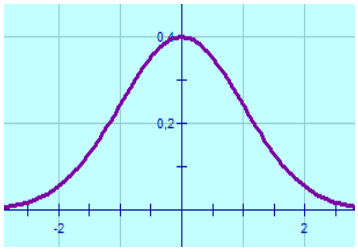
Broj $p(\langle a,b\rangle)$ odgovara površini između intervala $\langle a,b\rangle$ na osi x i krivulje $y = f(x)$. Vjerojatnost događaja $\{a\}$ (bilo kojeg jednočlanog, a onda i konačnog skupa) je uvijek nula tj. $p(\{a\}) = 0$ (površina štapića $x = a$).



Napomenimo da se analogno pomoću funkcije n-varijabli i višestrukog integrala definira vjerojatnost događaja $\langle a,b\rangle^n$ u vjerojatnosnom prostoru $(\mathbb{R}^n,\mathcal{F},p)$. U nastavku ćemo navesti neke modele nediskretnih vjerojatnosnih prostora koji se mogu prepoznati u mnogim prirodnim pojavama.

2.3.3 Normalna distribucija vjerojatnosti

Ako je gustoća vjerojatnosti $f:\mathbb{R}\to\mathbb{R}$, $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$, onda kažemo da vjerojatnost p ima **standardnu normalnu** (z) **razdiobu** ili **distribuciju**. Funkcija f je poznata kao **Gaussova** ili **normalna funkcija**. Njezin graf (Gaussova ili z-krivulja) je simetričan u odnosu na os y.



Vjerojatnost da je neki broj u intervalu $\langle a,b\rangle$ jednaka je površini lika omeđenog Gaussovom krivuljom, osi x i pravacima $x=a$ i $x=b$.

U donjoj tablici su izračunate površine iznad intervala $\langle 0,z\rangle$, tj. vjerojatnosti $p(\langle 0,z\rangle)$.

Zadatak 2.19 *Ako vjerojatnost ima standardnu normalnu razdiobu, odredite vjerojatnost da je broj z upao u interval $\langle 0,1.61\rangle$, $\langle -2,2\rangle$, $\langle -2,0.5\rangle$, $\langle 0,\infty\rangle$, $\langle 1.11,\infty\rangle$ i $\langle -\infty,-3.075\rangle$.*

Rješenje: $p(\langle 0,1.61\rangle)=0.4463$, $p(\langle -2,2\rangle)=2p(\langle 0,2\rangle)=2\cdot 0.4772=0.954\,4$,
$p(\langle -2,0.5\rangle)=p(\langle -2,0\rangle)+p(\langle 0,0.5\rangle)=p(\langle 0,2\rangle)+p(\langle 0,0.5\rangle)=0.4772+0.1915=0.668\,7$,
$p(\langle 0,\infty\rangle)=0.5$, $p(\langle 1.11,\infty\rangle)=p(\langle 0,\infty\rangle)-p(\langle 1.11,\infty\rangle)=0.5-0.3665=0.133\,5$,
$p(\langle -\infty,-3.075\rangle)=p(\langle 3.075,\infty\rangle)=p(\langle 0,\infty\rangle)-p(\langle 0,3.075\rangle)=0.5-0.49895=0.001\,05$.

Normal Curve Areas

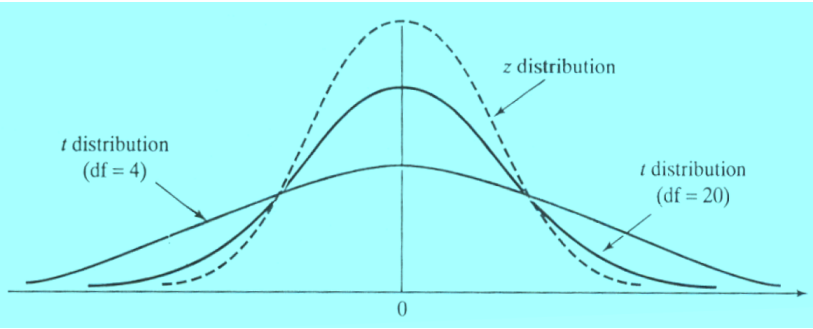
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Zadatak 2.20 *Ako vjerojatnost ima standardnu normalnu razdiobu, odredite vrijednost z takvu da je $p(\langle -z, z\rangle) = 0.9$. Odredite vrijednost $z > 0$ takvu da je $p(\langle z, \infty\rangle) = 0.005$ i takvu da je $p(\langle -\infty, z\rangle) = 0.975$.*

Rješenje: $p(\langle -z, z\rangle) = 0.9 \Rightarrow p(\langle 0, z\rangle) = 0.45 \in \langle 0.4495, 0.4505\rangle \Rightarrow z \approx 1.645$,
$p(\langle z, \infty\rangle) = 0.005 \Rightarrow p(\langle 0, z\rangle) = 0,495 \in \langle 2.57, 2.58\rangle \Rightarrow z \approx 2.575$,
$p(\langle -\infty, z\rangle) = 0.975 \Rightarrow p(\langle -\infty, 0\rangle) + p(\langle 0, z\rangle) = 0.975 \Rightarrow p(\langle 0, z\rangle) = 0.475 \Rightarrow z = 1.96$.

2.3.4 Studentova distribucija vjerojatnosti

Za svaki $\nu \in \mathbb{N}$, postoji funkcija definirana za svaki $t \in \mathbb{R}$, koju nazivamo Studentova t-funkcija s ν-stupnjeva slobode, čiji graf je simetričan u odnosu na os y i spljošteniji je nego li graf normalne funkcije. Povećavanjem stupnjeva slobode (eng. degrees of freedom-df), graf Studentove t-funkcije (t-krivulja) postaje sve sličnija z-krivulji, a za velike stupnjeve slobode su te dvije krivulje gotovo iste.



Kažemo da vjerojatnost kojoj je gustoća Studentova t-funkcija s ν stupnjeva slobode ima **Studentovu $t-$razdiobu** ili **distribuciju s v stupnjeva slobode**.

U donjoj tablici su prikazane vjerojatnosti $p(\langle t, \infty \rangle) = \alpha$, tj. površine ispod t-krivulje, između pravca $x = t$ i osi x.

Zadatak 2.21 *Ako vjerojatnost ima Studentovu razdiobu s* 12*, odnosno* 13 *stupnjeva slobode, odredite $t > 0$ takav da je $p(\langle -t, t \rangle) = 0.95$. Posebno odredite t takav da je $p(\langle -\infty, t \rangle) = 0.995$.*

Rješenje:

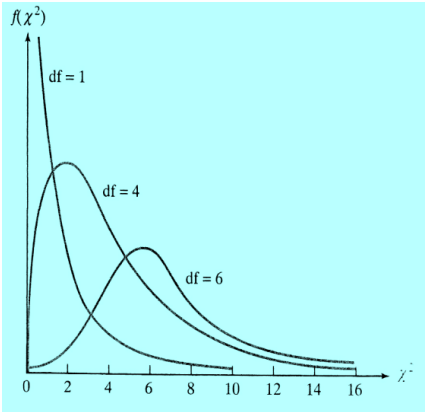
ν	$\alpha = p(\langle t, \infty \rangle)$	t
12	$(1 - 0.95)/2 = 0.025$	2.179
12	$1 - 0.995 = 0.005$	3.055
13	0.025	2.160
13	0.005	3.012

ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

$\nu = k$ (broj stupnjeva slobode)

2.3.5 Hi-kvadrat distribucija vjerojatnosti

Za svaki $\nu \in \mathbb{N}$, postoji funkcija f koju nazivamo χ^2-funkcija s ν-stupnjeva slobode, sa svojstvom $f(x) = 0$, za sve $x < 0$. Grafovi tih funkcija (χ^2-krivulje) su dolje prikazani u ovisnosti o stupnjevima slobode.



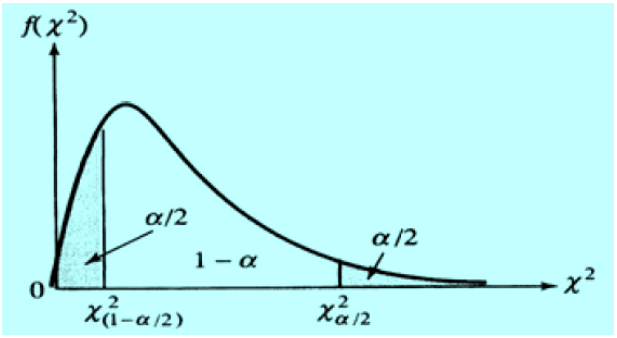
Kažemo da vjerojatnost kojoj je gustoća χ^2-funkcija s ν stupnjeva slobode ima **χ^2-razdiobu s v stupnjeva slobode**.

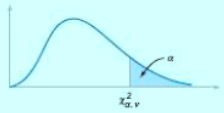
U donjoj tablici su prikazane vjerojatnosti $\alpha = p\left(\langle \chi^2, \infty \rangle\right)$, tj. površine ispod χ^2-krivulje, između pravca $x = \chi^2$ i osi x.

Zadatak 2.22 *Ako vjerojatnost ima χ^2-razdiobu s 10 stupnjeva slobode, odredite $\chi_1^2, \chi_2^2 > 0$ takav da je $p\left(\langle 0, \chi_1^2 \rangle\right) = 0.05 = p\left(\langle \chi_2^2, \infty \rangle\right)$.*

Rješenje: $p\left(\langle 0, \chi_1^2 \rangle\right) = 0.05 \Rightarrow p\left(\langle 0, \infty \rangle\right) - p\left(\langle \chi_1^2, \infty \rangle\right) = 0.05 \Rightarrow p\left(\langle \chi_1^2, \infty \rangle\right) = 1 - 0.05 = 0.95$

$\Rightarrow \chi_1^2 = 3.94,\ \chi_2^2 = 18.31.$





ν \ α	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

$\nu = k$ (broj stupnjeva slobode)

2.4 Uvjetna vjerojatnost i neovisnost događaja

Definicija 2.23 *Neka je $(\Omega, \mathcal{F}, p)$ vjerojatnosni prostor i neka su $A, B \in \mathcal{F}$. Kažemo da su događaji A i B* ***neovisni*** *ako je $p(A \cap B) = p(A)\, p(B)$.*

Neka je $(\Omega, \mathcal{F}, p)$ vjerojatnosni prostor i neka su $A, B \in \mathcal{F}$, $p(A) > 0$. ***Uvjetna vjerojatnost*** *događaja B pod uvjetom da se (prethodno) dogodio događaj A definira se kao $p(B|A) = \frac{p(A \cap B)}{p(A)}$.*

Očito, ako su događaji A i B neovisni onda je $p(B|A) = p(B)$. Nadalje, vrijedi $p(A \cap B) = p(B|A)\, p(A) = p(A|B)\, p(B)$. Općenito vrijedi:

$$p(A_1 \cap \cdots \cap A_n) = p(A_1)\, p(A_2|A_1)\, p(A_3|A_1 \cap A_2) \cdots p(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Primjer 2.24 *Vjerojatnost da iz šešira u kojem se nalaze* 3 *crne loptice numerirane sa* $1, 2$ *i* 3*, te* 4 *crvene numerirane sa* $1, 2, 3$ *i* 4*, izvučemo crnu lopticu numeriranu sa* 1 *je* $\frac{1}{7}$*. Vjerojatnost da je izvučena ista loptica, ako smo prethodno vidjeli da je njezina boja crna, je* $\frac{1}{3}$*, odnosno* $p(A|B) = \frac{p(A\cap B)}{p(B)} = \frac{\frac{1}{7}}{\frac{3}{7}} = \frac{1}{3}$ *(A je događaj da je izvučena loptica s brojem 1, a B događaj da je izvučena crna loptica).*

Primjer 2.25 *U kutiji se nalazi* 21 *bijela i* 10 *crnih kuglica. Iz kutije su izvučene dvije kuglice, bez vraćanja prve kuglice u kutiju. Kolika je vjerojatnost događaja da je druga izvučena kuglica bijela ako se zna da je prva izvučena bijela?*

Ako je A događaj da je prva izvučena kuglica bijela, a B događaj da je druga izvučena bijela, onda je $p(B|A) = \frac{20}{30}$ *(jedna bijela je izvučena, pa ih je ostalo* 30 *od čega* 20 *bijelih). Izračun preko formule je* $p(B|A) = \frac{p(A\cap B)}{p(A)} = \frac{\frac{21\cdot 20}{31\cdot 30}}{\frac{21\cdot 30}{31\cdot 30}} = \frac{2}{3}$*. Može se izračunati da je vjerojatnost događaja B, bez da znamo A jednaka* $p(B) = \frac{21\cdot 20 + 10\cdot 21}{31\cdot 30} = \frac{21}{31}$.

Zadatak 2.26 *Ako iz društva od* 5 *ljudi (*3 *muškarca i* 2 *žene) od kojih su dva punoljetna muškarca i jedna punoljetna žena na vrata pozvoni jedna osoba za koju se zna da je muško, kolika je vjerojatnost da je punoljetna?*

Rješenje: *Vjerojatnost da je osoba i muško i punoljetna je* $p(A \cap B) = \frac{2}{5}$*. Vjerojatnost da je osoba muško je* $p(A) = \frac{3}{5}$*. Tražena vjerojatnost je* $p(B|A) = \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3}$ *(kao da smo eliminirali iz razmatranja* 2 *ženske osobe, pa od* 3 *preostale muške izabiremo jednu punoljetnu osobu od moguće* 2*).*

Zadatak 2.27 *Banka raspolaže s* 3 *identična kompjutorska sustava. Sustavi rade neovisno s istom programskom podrškom. Prema proizvođaču, vjerojatnost zastoja hardwarea iznosi* 0.01*. Kolika je vjerojatnost da nastane zastoj u jednom danu sva* 3 *sustava?*

Rješenje: *Označimo li sa* A_i *događaj da je i-ti sustav u zastoju, onda očigledno tražimo* $p(A_1 \cap A_2 \cap A_3)$*, što je zbog neovisnosti događaja jednako* $p(A_1)\, p(A_2)\, p(A_3) = 0.01^3 = 0,000001$.

Zadatak 2.28 *Od* 5 *ključeva samo jedan otvara vrata. Odredite vjerojatnost događaja da su potrebna* 3 *pokušaja da se otvore vrata.*

Rješenje: *Ako je A_i događaj da će se vrata otvoriti u i-tom pokušaju, onda je traženi događaj $A_1^c \cap A_2^c \cap A_3$, pa je*

$$p\left(A_1^c \cap A_2^c \cap A_3\right) = p\left(A_1^c\right) p\left(A_2^c | A_1^c\right) p\left(A_3 | \left(A_1^c \cap A_2^c\right)\right) = \frac{4}{5}\frac{3}{4}\frac{1}{3} = \frac{1}{5}.$$

Primjer 2.29 *Izvodimo pokus bacanja dviju kocki. Ispitajte neovisnost događaja: na prvoj kocki je pao broj* 1*, a na drugoj broj* 2*. Nadalje, ako događaj A predstavlja da je na* 1*. kocki palo* $2, 3$ *ili* 4*, događaj B predstavlja da je na* 2*. kocki palo* $4, 5$ *ili* 6*, a događaj C predstavlja da je ukupan zbroj* 10*, ispitajte neovisnost ovih događaja.*

Prostor elementarnih događaja je $\Omega = \{(i, j) \mid i, j = 1, ..., 6\}$ *pa je* $p\left(\{(i, j)\}\right) = \frac{1}{36}$*. Budući da događaji da je na* 1*. kocki pao broj i, a na drugoj broj j imaju vjerojatnost* $\frac{6}{36}$*, to iz* $\frac{1}{36} = \frac{6}{36}\frac{6}{36}$ *zaključujemo da su oni neovisni. Događaji A i B imaju po 18 povoljnih ishoda, dok je* $C = \{(4, 6), (6, 4), (5, 5)\}$*. Vrijedi*
$p\left(A \cap B\right) = \frac{9}{36} = \frac{1}{4} = p\left(A\right) p\left(B\right) = \frac{18}{36}\frac{18}{36}$, $p\left(A \cap C\right) = \frac{1}{36} \neq p\left(A\right) p\left(C\right) = \frac{18}{36}\frac{3}{36}$,
$p\left(B \cap C\right) = \frac{3}{36} \neq p\left(B\right) p\left(C\right) = \frac{18}{36}\frac{3}{36}$,
$p\left(A \cap B \cap C\right) = \frac{1}{36} \neq p\left(A\right) p\left(B\right) p\left(C\right)$.

Zadatak 2.30 *Iz podataka u tablici odredite vjerojatnost da slučajno odabrani student bude ženska osoba koja studira ili medicinske ili tehničke znanosti. Nadalje, odredite vjerojatnost da je slučajno odabrani student muška osoba ako je poznato da studira medicinu, te ispitajte jesu li događaji A_1 i B_1 neovisni.*

Studij	*muški* (B_1)	*ženski* (B_2)	Σ
prirodne znanosti (A_1)	*1132*	*1775*	*2907*
tehničke znanosti (A_2)	*12883*	*5309*	*18192*
medicinske z. (A_3)	*1614*	*3098*	*4712*
biotehničke z. (A_4)	*1969*	*1422*	*3391*
društvene i humanističke z. (A_5)	*19546*	*20907*	*40453*
umjetničke akademije (A_6)	*474*	*574*	*1048*
Σ	*37618*	*33085*	*70703*

Rješenje*:*

$$p\left(B_2 \cap (A_3 \cup A_2)\right) = \frac{3098 + 5309}{70703} = 0.1189,$$

$$p\left(B_1|A_3\right) = \frac{p\left(B_1 \cap A_3\right)}{p\left(A_3\right)} = \frac{\frac{1614}{70703}}{\frac{4712}{70703}} = 0.3425.$$

Budući je $p\left(A_1 \cap B_1\right) = \frac{1132}{70703} = 0.016$,
a $p\left(A_1\right) p\left(B_1\right) = \frac{2907}{70703} \cdot \frac{37618}{70703} = 0.0219$, *to su ovi događaji ovisni.*

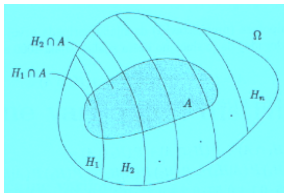
2.5 Potpuna vjerojatnost i Bayesova formula

Definicija 2.31 *Kažemo da je konačna ili prebrojiva množina nepraznih događaja* $\{H_i \mid i \in I \subseteq \mathbb{N}\}$ *vjerojatnosnog prostora* $(\Omega, \mathcal{F}, p)$ ***potpun sistem događaja na*** Ω *ako je* $\bigcup_{i \in I} H_i = \Omega$ *i* $H_i \cap H_j = \emptyset$, *za svaki* $i \neq j$.

Tada za proizvoljan događaj A vrijedi $A = (A \cap H_1) \cup (A \cap H_2) \cup \cdots$ iz čega dobivamo **formulu potpune vjerojatnosti**
$p\left(A\right) = p\left(H_1\right) p\left(A|H_1\right) + p\left(H_2\right) p\left(A|H_2\right) + \cdots$.

Dijeljenjem sa $p\left(A\right)$ dobivamo
$1 = \frac{p(H_1)p(A|H_1)}{p(A)} + \frac{p(H_2)p(A|H_2)}{p(A)} + \cdots = \frac{p(H_1 \cap A)}{p(A)} + \frac{p(H_2 \cap A)}{p(A)} + \cdots = p\left(H_1|A\right) + p\left(H_2|A\right) + \cdots$.

Time smo dobili novu, tzv. uvjetnu vjerojatnost p_A na prostoru $\left(A, \mathcal{F}_A, p_A\right)$, gdje je $\mathcal{F}_A$ familija svih događaja B_A oblika $B_A = B \cap A$, $B \in \mathcal{F}$. Uvjetna vjerojatnost p_A je definarana sa $p_A\left(B_A\right) = p\left(B|A\right)$.



Primjer 2.32 *Neka je* $\left(\Omega = \{1,2,3,4,5,6\}, \mathcal{F} = \mathcal{P}\left(\Omega\right), p\right)$ *vjerojatnosni prostor slučajnog pokusa bacanja kocke. Označimo sa* A *događaj: "na kocki je pao paran broj". Događaji* $H_1 = \{1,2\}$, $H_2 = \{3,4\}$, $H_3 = \{5,6\}$ *tvore potpuni sistem događaja na* Ω. *Događaj* $A \cap H_1$ *predstavlja događaj koji se sastoji od ishoda "na kocki*

je pao broj 2 *" i tumačimo ga kao događaj: "na kocki je pao paran broj i pao je broj* 1 *ili* 2 *". S druge strane događaj* $H_1|A$ *se također sastoji od istoga ishoda a tumačimo kao događaj: "na kocki je pao broj* 1 *ili* 2 *uz uvjet da znamo da je pao paran broj". Skupovno su događaji* $A \cap H_1$ *i* $H_1|A$ *jednaki, no razliku uočavamo tek promatrajući ih kao događaje vjerojatnosnih prostora* $(\Omega, \mathcal{F}, p)$ *i* $(A, \mathcal{F}_A = \{A \cap X \mid X \subseteq \Omega\}, p_A)$ *redom. Vjerojatnost prvog događaja je* $p(A \cap H_1) = \frac{1}{6}$*, a (uvjetna) vjerojatnost drugog događaja je* $p_A(H_{1A}) = \frac{1}{3}$ *što je jednako* $p(H_1|A) = \frac{p(A \cap H_1)}{p(A)} = \frac{\frac{1}{6}}{\frac{1}{2}}$.

Ako je $H_1, H_2, ...$ potpuni sistem događaja na Ω, onda, za svaki i, vrijedi sljedeća formula koju zovemo **Bayesovom formulom**:

$p(H_i|A) = \frac{p(H_i \cap A)}{p(A)} = \frac{p(H_i)p(A|H_i)}{\sum_j p(H_j)p(A|H_j)}$.

Primjer 2.33 *Sijalica može pripadati trima raznim serijama* S_1, S_2 *i* S_3*, pri čemu su* $p_1 = 0.25$*,* $p_2 = 0.5$ *i* $p_3 = 0.25$ *vjerojatnost da sijalica pripada seriji* S_1*,* S_2 *i* S_3*, redom. Vjerojatnost da će sijalica iz* S_1 *sijati barem* 1000 *sati je* 0.1*, iz* S_2 *je* 0.2*, a iz* S_3 *je* 0.4*. Odredite vjerojatnost događaja da će slučajno izabrana sijalica sijati barem* 1000 *sati. Nadalje, ako je poznato da je sijalica sijala barem 1000 sati odredite vjerojatnost da je ona iz treće serije.*

Ako je A *događaj da je slučajno odabrana sijalica sijala barem 1000 sati, a* H_i *događaj da sijalica pripada seriji* S_i*,* $i = 1, 2, 3$*, onda vrijedi* $A = (A \cap H_1) \cup (A \cap H_2) \cup (A \cap H_3)$ *i* H_1, H_2 *i* H_3 *čine sistem potpunih događaja. Stoga vrijedi*

$p(A) = p(A \cap H_1) + p(A \cap H_2) + p(A \cap H_3) \Rightarrow$

$p(A) = p(H_1)\, p(A|H_1) + p(H_2)\, p(A|H_2) + p(H_3)\, p(A|H_3) =$

$0.25 \cdot 0.1 + 0.5 \cdot 0.2 + 0.25 \cdot 0.4 = 0.225.$

Napokon, vjerojatnost da slučajno odabrana sijalica iz skupa onih koji će sijati barem 1000 *sati bude iz serije* S_3 *je* $p(H_3|A) = \frac{p(H_3)p(A|H_3)}{p(A)} = \frac{0.25 \cdot 0.4}{0.225} = 0.\dot{4}$.

Zadatak 2.34 *Na ispit je izašlo* 70% *studenata koji polažu prvi put. Na prethodnom roku od* 100 *studenata koji su izašli prvi put prošlo ih je* 50*, a od* 50 *studenata koji su i prije izlazili na ispit prošlo je njih* 20*. Odredite vjerojatnost da će slučajno odabrani student na novom roku proći ispit, te vjerojatnost da je slučajno odabrani iz skupine studenata koji su već polagali taj ispit, ako znamo da je taj student položio ispit .*

Rješenje: *Ako je A događaj da je student položio ispit, H_1 događaj da student prvi put polaže i H_2 događaj da ne polaže prvi put, onda su zadane sljedeće (a priori i a posteriori) vjerojatnosti $p(H_1) = 0.7$, $p(H_2) = 0.3$, $p(A|H_1) = 0.5$ i $p(A|H_2) = 0.4$. Slijedi*

$p(A) = p(A \cap H_1) + p(A \cap H_2) = p(H_1)\,p(A|H_1) + p(H_2)\,p(A|H_2) = 0.47$ *i*
$p(H_2|A) = \frac{p(H_2)p(A|H_2)}{p(A)} = \frac{0.3 \cdot 0.4}{0.47} = 0.225.$

Istaknimo da uvjetne vjerojatnosti $p(A|H_1)$ i $p(A|H_2)$ tumačimo kao vjerojatnosti da slučajno odabrani student iz skupa studenata koji prvi put polažu, odnosno ne polažu prvi put, položi ispit. S druge strane vjerojatnosti $p(A \cap H_1)$ i $p(A \cap H_2)$ tumačimo kao vjerojatnost da slučajno odabrani student iz skupa studenata koji su izašli na ispit prvi puta polaže, odnosno ne polaže prvi put, i položi taj ispit.

Poglavlje 3

Slučajna varijabla

3.1 Diskretna slučajna varijabla

Pri praćenju nekoga pokusa zanimamo se za neko numeričko obilježje X toga pokusa. Kod numeričkih varijabli populaciju su tvorili svi pokusi i svakom pokusu varijabla je pridruživala njegovo numeričko obilježje. Kod konačne populacije svakom numeričkom obilježju se pridruživala njezina relativna frekvencija i takvu funkciju smo nazivali distribucijom relativne frekvencije numeričke varijable. No, kod beskonačnih populacija pojam relativne frekvencije nekoga numeričkoga obilježja nema smisla i umjesto njega koristimo se pojmom vjerojatnosti događaja kojemu pridružujemo to numeričko obilježje. No, to znači da moramo krenuti od vjerojatnosnoga prostora $(\Omega, \mathcal{F}, p)$ i svakom ishodu (elementarnom događaju) pokusa $\omega \in \Omega$ pridružiti neki broj (obilježje).

Definicija 3.1 *Neka je $(\Omega, \mathcal{F}, p)$ vjerojatnosni prostor. Neka funkcija $X : \Omega \to \mathbb{R}$ koja svakom ishodu ω pridružuje realni broj $x = X(\omega)$ poprima konačno ili najviše prebrojivo vrijednosti $x_1, x_2, \ldots$. Ako je $X^{-1}(x_i) \in \mathcal{F}$, za svaki i, onda X nazivamo* ***diskretnom slučajnom varijablom****.*

Ako je $(\Omega, \mathcal{F}, p)$ diskretni vjerojatnosni prostor (Ω je prebrojiv i $\mathcal{F} = \mathcal{P}(\Omega)$) onda je svaka funkcija $X : \Omega \to \mathbb{R}$ diskretna slučajna varijabla.

Koristit ćemo oznake $(X = x_0)$ za $\{\omega \in \Omega \mid X(\omega) = x_0\}$ i analogno $(X < x_0) = \{\omega \in \Omega \mid X(\omega) < x_0\}$ i slično.

Neka diskretna slučajna varijabla poprima vrijednosti $x_1, x_2, ...$ i neka je $p_i = p(X = x_i)$, $i = 1, 2, ...$, vjerojatnost događaja koji se sastoji od onih ishoda kojima slučajna varijabla pridruži broj x_i (događaj da slučajna varijabla ima vrijednost x_i). Skup uređenih parova $\{(x_1, p_1), (x_2, p_2), ...\}$ nazivamo **distribucijom slučajne varijable** X.

Definicija 3.2 *Ako postoji suma* $x_1p_1+x_2p_2+\cdots$ *onda ju nazivamo* ***očekivanjem slučajne varijable*** X *i taj broj označujemo sa* $E[X]$.

Očekivanje $E[X]$ odgovara aritmetičkoj sredini μ numeričke varijable zadane na konačnoj populaciji s vrijednostima $x_1, x_2, ..., x_N$ i odgovarajućim relativnim frekvencijama $p_1, ..., p_N$, pa se ovi brojevi često i jednako označuju sa μ.

Ako je μ očekivanje slučajne varijable X, onda očekivanje slučajne varijable $(X - \mu)^2$ nazivamo **varijancom slučajne varijable** X i označujemo ju sa $Var[X] = \sigma^2$, a **standardnom devijacijom** $D[X] = \sigma$ nazivamo broj $\sqrt{\sigma^2}$.

Vrijedi $\sigma^2 = E\left[(X - \mu)^2\right] = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots$.

Primjedba 3.3 *Svaku (statističku) numeričku varijablu* $X : S \to \mathbb{R}$ *zadanu na konačnoj populaciji* S *od* N *elemenata koji poprimaju* k *različitih vrijednosti* $x_1, ..., x_k$*, možemo promatrati kao diskretnu slučajnu varijablu na sljedeći način: Definiramo prostor elementarnih događaja* Ω *elementi kojega su ishodi* $\omega_i =$ *"element populacije ima obilježje* x_i*", i* $= 1, ..., k$*. Za svaki taj ishod uzimamo vjerojatnost* p_i *koja odgovara relativnoj frekvenciji obilježja* x_i*. Na takav način je definiran vjerojatnosni prostor* $(\Omega, \mathcal{F}, p)$*, a slučajna varijabla* $X : \Omega \to \mathbb{R}$ *pridružuje ishodu* ω_i *broj* x_i*. Očekivanje i varijanca te slučajne varijable odgovaraju aritmetičkoj sredini i varijanci početne statističke varijable. Nadalje, svaku diskretnu (statističku) varijablu zadanu na beskonačnom skupu možemo promatrati kao diskretnu slučajnu varijablu pri čemu su vjerojatnosti* p_i *elementarnih događaja* $\omega_i =$ *"element populacije ima obilježje* x_i*" definirane kao limesi* $\lim\limits_{N\to\infty} \frac{f_i(N)}{N}$ *odgovarajućih relativnih frekvencija, kad opseg populacije* N *teži u beskonačnost.*

Primjer 3.4 *Temperature zraka (zaokružene) izmjerene u travnju 2011. u Splitu*

u 12 sati su dane u tablici.

temperatura	18	19	20	21	22	23	24	25
f_i	10	4	8	2	3	1	1	1

Tablica predstavlja distribuciju numeričke varijable $X : \{1., 2., ..., 30.\} \to \mathbb{R}$ *koja svakom danu u mjesecu travnju pridružuje izmjerenu temperaturu. Aritmetička sredina (prosječna temperatura za travanj) je* $\mu = \frac{10}{30} \cdot 18 + \frac{4}{30} \cdot 19 + \cdots + \frac{1}{30} \cdot 25 = 19.867$. *Definirajmo prostor elementarnih događaja* $\Omega = \{\omega_i \mid i = 0, ..., 40\}$ *kojeg tvore elementarni događaji* $\omega_i =$ *"temperatura u Splitu izmjerena u 12 sati u nekom danu u travnju je* i*-stupnjeva Celzijusa". Za vjerojatnost elementarnog događaja* ω_i *ćemo staviti relativnu frekvenciju* $p_i = \frac{f_i}{30}$ *temperature* i^0C *izmjerene u travnju 2011. tj.* $p_i = 0$, *za* $i < 18$, $p_{18} = \frac{10}{30}$, $p_{19} = \frac{4}{30}, ..., p_{25} = \frac{1}{30}$. *Sada je diskretna vjerojatnost* p *definirana na svakom događaju iz* $\mathcal{F} = \mathcal{P}(\Omega)$. *Na takav način smo definirali diskretni vjerojatnosni prostor* $(\Omega, \mathcal{F}, p)$, *pa statističku varijablu* X *možemo promatrati kao slučajnu varijablu* $X : \Omega \to \mathbb{N}$ *definiranu sa* $X(\omega_i) = i$. *Očekivanje te slučajne varijable je* $E[X] = 19.867$.

U prethodnom primjeru statistička varijabla opisuje samo postojeće stanje dočim slučajna varijabla daje model kojim se anticipiraju događaji u budućnosti. Primijetimo da je taj model utemeljen na empirijskim (aposteriori) vjerojatnostima. Slučajna varijabla koja modelira određeni slučajni pokus iz realnog života je vjerodostojnija ako su vjerojatnosti dobivene kao relativne frekvencije obilježja proizašlih iz što je moguće više izvedenih pokusa. Slučajna varijabla iz prethodnog primjera predstavlja model za određivanje zaokružene temeperature u 12 sati bilo kojega dana u travnju u Splitu. No, taj model bi bio bolji da smo za odgovarajuće vjerojatnosti mogli staviti relativne frekvencije pojedinih zaokruženih temperatura izmjerenih u posljednjih 100 godina u travnju. Slučajna varijabla može i modelirati neke pojave koristeći teorijske (apriori) vjerojatnosti koje se odnose na neke pokuse. Naime, kod izvođenja nekog pokusa deskriptivna statistika pomoću statističke varijable može samo detektirati rezultate toga pokusa iz kojih nije razvidna neka zakonitost jer se odnose na samo konkretno izvedene pokuse. S druge strane, inferencijalna statistika pomoću slučajne varijable nudi teorijski očekivane rezultate koji se odnose na svaki pokus takve vrste.

Primjer 3.5 *Neka se pokus sastoji u bacanju novčića 3 puta zaredom i neka slučajna varijabla bilježi koliko puta je palo pismo. Ova slučajna varijabla X je definirana na vjerojatnosnom prostoru $(\Omega, \mathcal{P}(X), p)$ gdje je* $\Omega = \{(i,j,k) \mid i,j,k = 0,1\}$. *Prostor elementarnih događaja se sastoji od ishoda $\omega_1 = (0,0,0)$ (niti jednom nije palo pismo), $\omega_2 = (0,0,1)$ (samo jednom, u trećem bacanju je palo pismo), $\omega_3 = (0,1,0)$, $\omega_4 = (0,1,1)$, $\omega_5 = (1,0,0)$, $\omega_6 = (1,0,1)$, $\omega_7 = (1,1,0)$ $\omega_8 = (1,1,1)$. Slučajna varijabla poprima ove vrijednosti $X(\omega_1) = 0$, $X(\omega_2) = X(\omega_3) = X(\omega_5) = 1$, $X(\omega_4) = X(\omega_6) = X(\omega_7) = 2$ i $X(\omega_8) = 3$, a pripadne vjerojatnosti su*

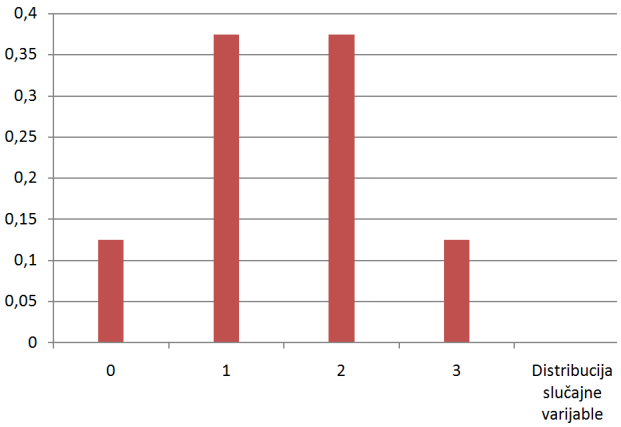
$$p(X=0) = \tfrac{1}{8} = p(X=3),\ p(X=1) = \tfrac{3}{8} = p(X=2),$$

što kratko zapisujemo kao $X = \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$.

Primjerice, događaj da slučajna varijabla poprima vrijednost veću od 1 pišemo $(X > 1) = (X = 2) \cup (X = 3)$ i on ima vjerojatnost $p(X > 1) = \frac{3}{8} + \frac{1}{8} = 0.5$.

Očekivanje ove slučajne varijable je $E[X] = \frac{1}{8}\cdot 0 + \frac{3}{8}\cdot 1 + \frac{3}{8}\cdot 2 + \frac{1}{8}\cdot 3 = \frac{12}{8} = \frac{3}{2} = 1.5$.

Varijanca je $\sigma^2 = \frac{1}{8}\cdot(0-1.5)^2 + \frac{3}{8}\cdot(1-1.5)^2 + \frac{3}{8}\cdot(2-1.5)^2 + \frac{1}{8}\cdot(3-1.5)^2 = 0.75$.



U bilo kojem konkretnom pokusu bacanja tri novčića odjednom, ponovljenom konačno mnogo puta, statistička varijabla bilježi broj pojavljivanja pisma svakog izvedenog pokusa, a relativne frekvencije obilježja $0, 1, 2$ i 3 ove statističke varijable se ne moraju podudarati s teorijskim vjerojatnostima $p(X=0)$, $p(X=1)$, $p(X=2)$ i $p(X=3)$ događaja iz gore definiranog vjerojatnosnog prostora. No, za dovoljno veliki broj ponavljanja one će biti približno jednake.

3.1.1 Bernoullijev pokus i binomna razdioba

Pretpostavimo da u nekom pokusu vjerojatnost nekog događaja A iznosi p, odnosno vjerojatnost da ne nastupi taj događaj, tj. vjerojatnost od A^c, je $1-p$. Nadalje, pretpostavimo da se vjerojatnost toga događaja ne mijenja pri ponavljanju pokusa. Takav događaj i pokus nazivamo **Bernoullijevim**.

Bernoullijev pokus je npr. bacanje novčića, bacanje kocke, izvlačenje kuglice iz neke kutije tako da kuglicu vraćamo u kutiju nakon izvlačenja, a odgovarajući Bernoullijevi događaji su: "palo je pismo", "pala je šestica", "izvučena je točno određena kuglica".

Definicija 3.6 *Ako Bernoullijev pokus ponavljamo n puta, a slučajna varijabla X_n svakoj seriji od n pokusa (svakoj uređenoj n-torki ishoda jednog pokusa) pridružuje ukupan broj ishoda događaja A, onda varijablu X_n nazivamo* ***binomnom slučajnom varijablom*** *i kažemo da ima binomnu $B\{n,p\}$ razdiobu.*

Teorem 3.7 *Distribucija binomne slučajne varijable $B\{n,p\}$ (**binomna distribucija**) je zadana sa $p(X_n=k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0,...,n$.*

Svaku diskretnu slučajnu varijablu koja poprima vrijednosti iz skupa $\{0,1,...,n\}$, a za koju postoje p i n takvi da joj je distribucija $B\{n,p\}$ nazivamo binomnom slučajnom varijablom.

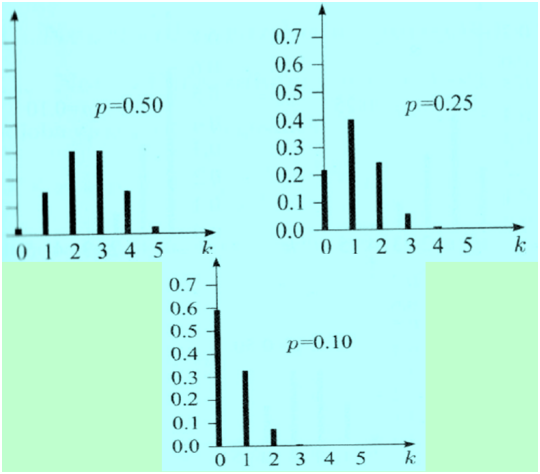
Broj $p(k)$ označuje vjerojatnost da se u n ponavljanja pokusa događaj A dogodi točno k puta.

Primjerice, vjerojatnost da u 5 bacanja kocke 2 puta padne broj 6 je $p(X=2) = \frac{5!}{2!3!}\cdot\left(\frac{1}{6}\right)^2\cdot\left(\frac{5}{6}\right)^3 = \frac{625}{3888} \approx 0.16$.

Teorem 3.8 *Očekivanje binomne slučajne varijable s $B\{n,p\}$ razdiobom je $E[X_n] = np$, a standardna devijacija je $D[X_n] = \sigma = \sqrt{np(1-p)}$.*

Za jako velike n ($np > 4$ i $n(1-p) > 4$) binomna distribucija se može dovoljno dobro opisati s normalnom distribucijom (odgovarajućom formulom za normalnu distribuciju).

Na slici su prikazane binomne distribucije $B\{5, 0.5\}$, $B\{5, 0.25\}$ i $B\{5, 0.1\}$.



Primjer 3.9 *Iz kutije koja sadrži* 4 *bijele i* 6 *crnih kuglica izvlačimo* 5 *puta za redom po jednu kuglicu i vraćamo ju u kutiju. Kolika je vjerojatnost da smo točno* 3 *puta, odnosno najviše* 3 *puta, izvukli bijelu kuglicu? Koje je očekivanje slučajne varijable koja registrira broj povoljnih ishoda u* 5 *izvlačenja?*

Ova slučajna varijabla je binomna s distribucijom $B\left\{5, \frac{4}{10}\right\}$. *Tražene vjerojatnosti su* $p(X = 3) = \frac{5!}{3!\cdot 2!} \cdot \left(\frac{4}{10}\right)^3 \cdot \left(\frac{6}{10}\right)^2 = \frac{144}{625} = 0,2304,$

$p(X \leq 3) = 1 - p(x > 3) = 1 - p(X = 4) - p(X = 5) = 1 - \frac{5!}{4!\cdot 1!} \cdot \left(\frac{4}{10}\right)^4 \cdot \left(\frac{6}{10}\right)^1 - \frac{5!}{5!\cdot 0!} \cdot \left(\frac{4}{10}\right)^5 \cdot \left(\frac{6}{10}\right)^0 = \frac{2853}{3125} \approx 0.913.$

Očekivanje je $E[X] = np = 2.$

Primjer 3.10 *Prodavač u dogovoru s proizvođačem daje jednogodišnje jamstvo na neki uređaj. Prema podacima iz prijašnjeg razdoblja,* 15 % *kupaca prijavljuje kvar u jamstvenom roku. Ako je jednoga dana prodano* 8 *uređaja i ako varijabla* X *bilježi broj prijavljenih kvarova uređaja u jamstvenom roku, kako glasi distribucija slučajne varijable* X, *te kolika je njezina očekivana vrijednost i devijacija.*

Slučajna varijabla je binomna s distribucijom $B\{8, 0.15\}$. *Distribucija je dana formulom* $p(k) = \frac{8!}{k!(8-k)!} 0.15^k \cdot 0.85^{8-k}$, $k = 0, ..., 8$, *a vrijednosti su dane u tablici:*

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6,7,8*
$p(X = x_i)$	*0.272*	*0.384*	*0.237*	*0.083*	*0.018*	*0.002*	*0*

Očekivanje je $\mu = np = 8 \cdot 0.15 = 1.2$, *a standardna devijacija je* $\sigma = \sqrt{np(1-p)} = \sqrt{1.02} \approx 1.00995.$

Zadatak 3.11 *Ako se bacaju dvije kocke istovremeno* 30 *puta, odredite očekivani broj dobitka broja* 3 *i* 4 *u istom bacanju.*

Rješenje: *Vjerojatnost da padnu* 3 *i* 4 *u istom bacanju je* $p = \frac{2}{36} = \frac{1}{18} = 0.0\dot{5}$. *Slučajna varijabla koja registrira broj dobitka para* 3 *i* 4 *u* 30 *bacanja je binomna s distribucijom* $B\left\{30, 0.0\dot{5}\right\}$. *Distribucija je zadana formulom*

$$p\left(X = k\right) = \frac{30!}{k!\left(30-k\right)!} 0.0\dot{5}^k \left(1 - 0.0\dot{5}\right)^{30-k}.$$

Očekivanje je $\mu = E\left[X\right] = 0p\left(0\right) + 1p\left(1\right) + \cdots + 30p\left(30\right) = np = 30 \cdot 0.0\dot{5} = 1.\dot{6}$.

3.1.2 Poissonova razdioba

Definicija 3.12 *Kažemo da diskretna slučajna varijabla* X *koja poprima vrijednosti u* $\mathbb{N}_0$ *ima* ***Poissonovu distribuciju*** *ako postoji* $\lambda > 0$ *takav da je njezina distribucija zadana sa* $p\left(X = k\right) = \frac{e^{-\lambda}\lambda^k}{k!}$, $k = 0, 1, ...$

Teorem 3.13 *Očekivanje varijable* X *s Poissonovom distribucijom je* $\mu = E\left[X\right] = \lambda$, *a standardna devijacija je* $\sigma = D\left[X\right] = \sqrt{\lambda}$.

Poissonova razdioba je granični slučaj niza binomnih varijabli X_n s parametrima n, p_n uz granični prijelaz $n \to \infty$, ali tako da $n \cdot p_n$ ostaje konstanta.

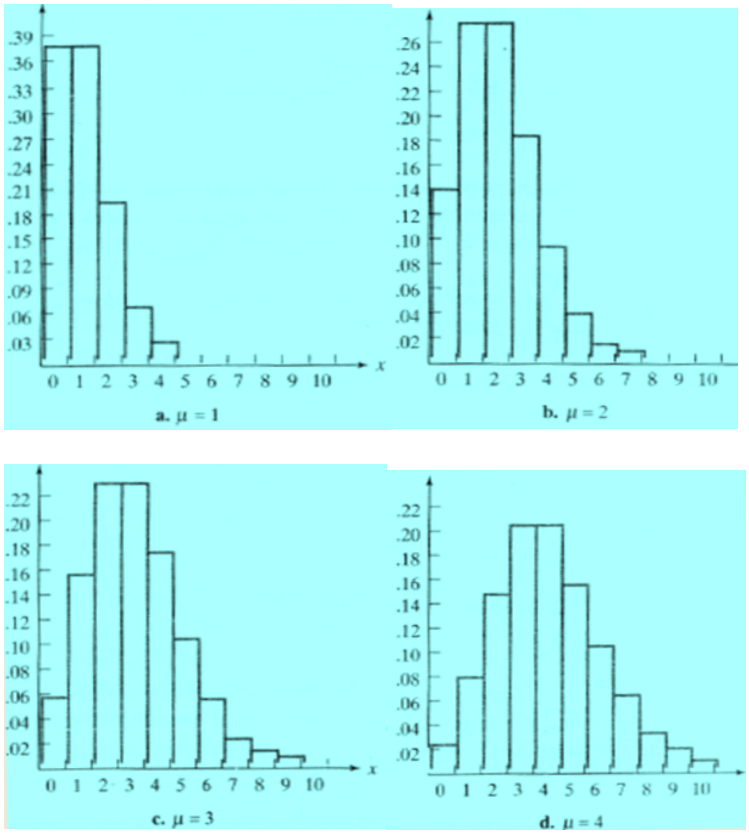
Za jako male vjerojatnosti ($p \leq 0.08$) i veliki broj pokušaja ($n \geqslant 1500p$), binomna distribucija se može dovoljno dobro opisati pomoću Poissonove i njezina analiza je tada lakša (primjerice broj dobitaka brojeva 3 i 4 u istom bacanju dvije kocke, ako se broj bacanja n stalno ponavlja mnogo puta, se aproksimira slučajnom varajablom s Poissonovom distribucijom uz $\lambda = n\frac{2}{36}$). Za dovoljno veliki λ Poissonova distribucija se približno može opisati normalnom.

Poissonova distribucija je prikladna za opis pokusa koji se sastoji u mjerenju broja povoljnih ishoda u određenoj (jednakoj) vremenskoj jedinici, jediničnoj površini, udaljenosti, volumenu i sl., a vjerojatnost nastanka toga događaja je jednaka za svaku jedinicu vremena, površine, udaljenosti, volumena itd., ishodi pokusa su neovisni, a očekivana vrijednost broja povoljnih ishoda događaja po jedinici je jednaka λ. Primjerice, takvi pokusi su oni koji bilježe: broj ljudi oboljelih od

gripe u nekom fiksnom periodu, broj autobusa koji dođu na stajalište u nekom fiksnom vremenu, broj meteora vidljivih kroz neko fiksno vrijeme, broj gledatelja određene utakmice, broj umrlih stanica u dijelu organizma u nekom fiksnom vremenu, broj čestica nastalih u nekom fizikalnom eksperimentu...

Kod navedenih primjera parametar λ Poissonove razdiobe se u svakom pojedinom slučaju određuje eksperimentalno u ovisnosti o promatranom uzorku (izjednačimo očekivanje $\mu = \lambda$ s aritmetičkom sredinom uzorka).

Na slici su prikazane Poissonave distribucije za parametre $\mu = \lambda = 1, 2, 3, 4$.



Primjer 3.14 *Ako je srednja vrijednost dolazaka autobusa na istu stanicu u određenom vremenskom intervalu u nekoliko tjedana promatranja jednaka* 3*, odredite vjerojatnost da u tom razdoblju ne dođe niti jedan autobus, da ih dođe* 3*, te da ih dođe* 4*.*

Koristimo empirijski podatak o srednjoj vrijednosti $\overline{x}$ za očekivanje ($\mu = E[X] = \lambda$) slučajne varijable s Poissonovom distribucijom. Naime, iskustveno polazimo od pretpostavke da slučajna varijabla koja bilježi broj dolazaka autobusa na isto mjesto u fiksnom periodu ima Poissonovu razdiobu. Stoga je vjerojatnost zadana formulom $p(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}$. Tražene vjerojatnosti su $p(X=0) = \frac{e^{-3}3^0}{0!} = e^{-3} \approx 0.049$, $p(X=3) = \frac{e^{-3}3^3}{3!} \approx 0.224$, $p(X=4) = \frac{e^{-3}3^4}{4!} \approx 0.168$.

3.1.3 Hipergeometrijska razdioba

Bernoulijev pokus najlakše opisujemo kao izvlačenje, n puta uzastopce, jednog elementa iz skupa od N elemenata od kojih M elemenata ima svojstvo A , a preostalih $N-M$ elemenata nema svojstvo A. Nakon što smo izvukli neki element, vraćamo ga natrag u skup, i postupak ponavljamo. No, ako u n izvlačenja element ne vraćamo natrag u skup, već ga ostavljamo sa strane, vjerojatnost da se izvuče element sa svojstvom A se mijenja u svakom sljedećem pokušaju. Samo u prvom pokušaju je $p = \frac{M}{N}$. (Ovaj pokus smijemo zamišljati kao da odjednom izvlačimo n elemenata iz skupa od N elemenata). Varijablu koja bilježi koliko puta je nastupio događaj A (tj. izvučen element koji ima svojstvo A) nazivamo **hipergeometrijskom slučajnom varijablom**.

Distribucija hipergeometrijske slučajne varijable (**hipergeometrijska distribucija** s parametrima n, N i M) je zadana sa

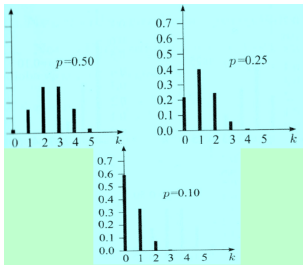
$p(X=k) = \frac{\frac{M!(N-M)!}{(M-k)!k!(N-M-n+k)!(n-k)!}}{\frac{N!}{(N-n)!n!}} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$, $k = 0, ..., n$, $k \leq M$, $n-k \leq N-M$.

Svaku diskretnu slučajnu varijablu koja poprima vrijednosti iz skupa $\{0, 1, ..., n\}$, a za koju postoje n, N i M takvi da joj je distribucija hipergeometrijska s parametrima n, N i M nazivamo hipergeometrijskom slučajnom varijablom.

Teorem 3.15 *Očekivanje hipergeometrijske slučajne varijable s parametrima n, N i M je $\mu = E[X] = n\frac{M}{N}$, a standardna devijacija je*
$\sigma = D(X) = \sqrt{n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}}$.

Kad je $n < 0.005N$ hipergeometrijska distribucija se aproksimira binomnom.

Na slikama su prikazane hipergeometrijske distribucije za $n = 5$, $N = 100$ i $M = 50, 25$ i 10.



Zadatak 3.16 *Proizvođač isporučuje kupcu pošiljku od* 15 *komada nekog proizvoda, među kojima se nalaze* 4 *neispravna. Zbog nemogućnosti kompletne provjere kupac može provjeriti samo* 4 *slučajno odabrana komada iz pošiljke. Kolika je vjerojatnost da će kupac prihvatiti pošiljku ako u njoj može tolerirati najviše jedan neispravni komad? Koje je očekivanje?*

Rješenje: *Budući da kupac provjerava* 4 *različita slučajno odabrana komada i nakon provjere komad ne vraća natrag, već uzima neki drugi, to je slučajna varijabla, koja bilježi broj neispravnih u uzorku od* 4*, hipergeometrijska s parametrima* $n = 4$, $N = 15$ *i* $M = 4$. *Tražena vjerojatnost je* $p(X \leq 1) = p(X = 0) + p(X = 1) = \frac{\binom{4}{0}\binom{11}{4}}{\binom{15}{4}} + \frac{\binom{4}{1}\binom{11}{3}}{\binom{15}{4}} = 0.593$. $E[X] = 4\frac{4}{15} = \frac{16}{15}$.

3.1.4 Geometrijska razdioba

Ako se Bernoulijev pokus ponavlja sve dok se ne dogodi Bernoulijev događaj A koji u svakom pokusu ima istu vjerojatnost p, onda se slučajna varijabla X koja bilježi koliko je puta pokus izveden dok nije nastupio događaj A naziva **geometrijska slučajna varijabla.** Ona poprima vrijednosti iz $\mathbb{N}$, a njezina distribucija je $p(X = k) = p(1-p)^{k-1}$ i nazivamo ju **geometrijskom distribucijom** s parametrom p. Svaku varijablu s ovakvom distribucijom nazivamo geometrijskom slučajnom varijablom.

Teorem 3.17 *Očekivanje geometrijske slučajne varijable s parametrom* p *je* $E[X] = 1 + \frac{1-p}{p}$.

Varijabla koja bilježi koliko je bacanja dviju kocki potrebno dok ne padnu istovremeno dvije šestice je geometrijska slučajna varijabla s $p = \frac{1}{36} = 0.02\dot{7}$.

Primjerice $p(X = 1) = 0.02\dot{7}$, $p(X = 2) = 0.02\dot{7} \cdot \left(1 - 0.02\dot{7}\right) \approx 0.0262$,

$p(X = 3) = 0.02\dot{7} \cdot \left(1 - 0.02\dot{7}\right)^2 \approx 0.255$, $p(X = 4) = 0.02\dot{7} \cdot \left(1 - 0.02\dot{7}\right)^3 \approx 0.0248$,

$p(X = 10) = 0.02\dot{7} \cdot \left(1 - 0.02\dot{7}\right)^9 \approx 0.0211$.

Očekivanje je $\mu = 1 + \frac{\frac{35}{36}}{\frac{1}{36}} = 36$.

3.1.5 Pascalova razdioba

Ako se Bernoulijev pokus ponavlja sve dok se ne dogodi Bernoulijev događaj A (koji u svakom pokusu ima istu vjerojatnost p) točno n puta, onda se slučajna varijabla X koja bilježi koliko je puta pokus izveden dok nije nastupio događaj A točno n puta naziva **Pascalova slučajna varijabla.** Ona poprima vrijednosti $n, n+1, n+2, ...$, a njezina distribucija je $p(X = k) = \frac{(k-1)!}{(n-1)!(k-n)!} p^n (1-p)^{k-n} = \binom{k-1}{n-1} p^n (1-p)^{k-n}$ i nazivamo ju **Pascalovom distribucijom** s parametrom p i n. Svaku varijablu s ovakvom distribucijom nazivamo Pascalovom slučajnom varijablom.

Teorem 3.18 *Očekivanje Pascalove slučajne varijable s parametrima p i n je* $E[X] = n\left(1 + \frac{1-p}{p}\right)$.

Primjer 3.19 *Izračunajte očekivani broj bacanja kocke dok se šestica ne pojavi* 3 *puta, te vjerojatnost da je bilo potrebno najviše* 4 *bacanja za to.*

Poznato je: $n = 3$, $p = \frac{1}{6} = 0.1\dot{6}$. *Stoga je* $E[X] = 3\left(1 + \frac{\frac{5}{6}}{\frac{1}{6}}\right) = 18$, $p(X \leq 4) = p(X = 3) + p(X = 4) \approx 0.167^3 + 3 \cdot 0.167^3 \cdot (1 - 0.167) = 0.016296$.

3.1.6 Jednolika distribucija

Definicija 3.20 *Diskretnu slučajnu varijablu X koja poprima konačno mnogo vrijednosti $x_1, ..., x_n$ i za koju je $p(x_1) = \cdots = p(x_n) = \frac{1}{n}$ nazivamo* ***jednoliko (uniformno) distribuiranom varijablom****.*

Primjerice varijabla koja bilježi broj karte koja je izvučena od 40 karata je jednoliko distribuirana slučajna varijabla jer svi brojevi imaju istu vjerojatnost da budu izvučeni i ona iznosi $\frac{4}{40} = 0.1$.

3.2 Slučajna i kontinuirana varijabla

Definicija 3.21 *Neka je $(\Omega, \mathcal{F}, p)$ vjerojatnosni prostor. Funkciju $X : \Omega \to \mathbb{R}$ nazivamo **slučajnom varijablom** ako je $X^{-1}(\langle a, b\rangle) \in \mathcal{F}$ za svaki $a, b \in \mathbb{R}$.*

Kao i kod diskretne slučajne varijable koristimo oznake $p(a < X < b) = p(X^{-1}\langle a, b\rangle), p(X < a) = p(X^{-1}\langle -\infty, a\rangle), ...$

Očito je i diskretna slučajna varijabla X slučajna varijabla, no sada taj poseban tip slučajne varijable možemo poopćiti dopuštajući da X poprimi neprebrojivo mnogo vrijednosti, ali da postoji konačan ili prebrojiv skup $D \subseteq \mathbb{R}$ takav da je $p(X \in D) = 1$.

Definicija 3.22 ***Funkcijom distribucije*** *od X nazivamo funkciju $F_X : \mathbb{R} \to [0, 1]$ definiranu izrazom $F_X(a) = p(X < a)$.*

Očito za diskretnu slučajnu varijablu X, funkcija distribucije $F_X : \mathbb{R} \to [0, 1]$ je definirana sa $F(x) = \sum_{\substack{d \in D \\ d < x}} p(X = d)$, a očekivanje je $\mu = E[X] = \sum_{d \in D} d \cdot p(X = d)$, gdje je $D \subseteq \mathbb{R}$ takav da je $p(X \in D) = 1$.

3.2.1 Kontinuirana slučajna varijabla

Definicija 3.23 *Za slučajnu varijablu X kažemo da je **neprekidna (kontinuirana)** ako postoji nenegativna funkcija $f : \mathbb{R} \to \mathbb{R}$ takva da je $p(a < X < b) = \int_a^b f(x)\,dx$, za svaki $a, b \in \mathbb{R}$, $a < b$. Funkciju f nazivamo **gustoćom** slučajne varijable a njezin graf **krivuljom distribucije**.*

Primijetimo da neprekidna slučajna varijabla može poprimiti bilo koju vrijednost iz barem jednog intervala $\langle a, b\rangle$ (pa se ponekad u literaturi tako i definira),

a isto tako da je diskretna ona koja može poprimiti samo konačno ili prebrojivo mnogo vrijednosti. U nastavku ćemo pod slučajnom varijablom podrazumijevati diskretnu ili kontinuiranu slučajnu varijablu.

Funkcija distribucije neprekidne slučajne varijable X gustoće f je dana sa $F_X(x_0) = \int_{-\infty}^{x_0} f(x)\,dx$.

Vrijedi $1 = p(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\,dx$ i $p(X = x_0) = 0 = \int_{x_0}^{x_0} f(x)\,dx$.

3.2.2 Očekivanje

Definicija 3.24 ***Očekivanjem*** *neprekidne slučajne varijable X gustoće f nazivamo broj (ako postoji) $\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$, a odgovarajućom* ***standardnom devijacijom*** *nazivamo broj $\sigma = D[X] = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx}$, odnosno* ***varijancom*** *broj $\sigma^2 = Var[X]$.*

Sljedeće tvrdnje vrijede za svaku slučajnu varijablu.

Teorem 3.25 *Ako su X i Y slučajne varijable s očekivanjima $E[X]$ i $E[Y]$, onda varijable $c = const$, cX i $X + Y$ imaju sljedeća očekivanja $E[c] = c$, $E[cX] = cE[X]$ i $E[X + Y] = E[X] + E[Y]$. Nadalje, vrijedi $D[X] = E[X^2] - (E[X])^2$.*

Teorem 3.26 (Čebišev) *Neka je X slučajna varijabla s očekivanjem μ i standardnom devijacijom σ, te $k \in \mathbb{R}$, $k \geqslant 1$. Tada je $p(X \in \langle \mu - k\sigma, \mu + k\sigma \rangle) > 1 - \frac{1}{k^2}$.*

Primjerice, vjerojatnost da vrijednost slučajne varijable X bude u intervalu $\langle \mu - 2\sigma, \mu + 2\sigma \rangle$ je uvijek barem 0.75, a vjerojatnost da bude u intervalu $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ je $0.\dot{8}...$

Ako je X slučajna varijabla s očekivanjem μ i devijacijom σ, slučajnu varijablu $Z = \frac{X-\mu}{\sigma}$ ($Z : \Omega \to \mathbb{R}$, $Z(\omega) = \frac{X(\omega)-\mu}{\sigma}$) nazivamo pripadnom standardiziranom varijablom. Standardizirana varijabla ima očekivanje 0 i devijaciju 1.

Primjer 3.27 *Ako slučajna varijabla trajnosti auto guma nekog proizvođača ima očekivanje* 40000 *km i devijaciju* 4000, *kolika je najmanja vjerojatnost da će gume trajati između* 34000 *i* 46000 *km?*

Nađimo najprije standardizirane vrijednosti $z_1 = \frac{34000-40000}{4000} = -1.5$ *i* $z_2 = \frac{46000-40000}{4000} = 1.5$. *Tražena vjerojatnost je*

$$p\left(z \in \langle -1.5, 1.5\rangle\right) = p\left(z \in \langle 0 - 1.5 \cdot 1, 0 + 1.5 \cdot 1\rangle\right)$$

i ona je, po Čebiševom teoremu, veća od $1 - \frac{1}{1.5^2} = 0.\dot{5}$.

3.3 Modeli kontinuiranih slučajnih varijabli

U prirodi i u statističkim primjenama najčešći primjeri kontinuiranih slučajnih varijabli su one čija je gustoća f:

- studentova t-funkcija s ν stupnjeva slobode. Za takvu varijablu kažemo da je t-**distribuirana**, njezino očekivanje je 0, a devijacija je $\sigma = \sqrt{\frac{\nu}{\nu-2}}$;
- χ^2-funkcija s ν stupnjeva slobode. Za takvu varijablu kažemo da je χ^2-**distribuirana**, njezino očekivanje je $\mu = \nu$, a devijacija je $\sigma = \sqrt{2\nu}$;
- Fisherova F-funkcija sa stupnjevima slobode ν_1 i ν_2 (površine ispod ove krivulje, tj. vjerojatnosti su zadane tabelarno). Za takvu varijablu kažemo da je F-**distribuirana**. Njezino očekivanje je $\mu = \frac{\nu_2}{\nu_2-2}$, $\nu_2 \geqslant 3$;
- eksponencijalna funkcija $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$, s parametrom $\lambda \geqslant 0$. Za takvu varijablu kažemo da je **eksponencijalno distribuirana**, njezino očekivanje je $\mu = \frac{1}{\lambda}$, a devijacija je $\sigma = \frac{1}{\lambda}$;
- konstantna funkcija $f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \in \mathbb{R}\setminus[a,b] \end{cases}$ s parametrima a i b, $a < b$. Za takvu varijablu kažemo da je **neprekidno uniformno distribuirana**, a njezino očekivanje je $\mu = \frac{a+b}{2}$, a devijacija je $\sigma = \frac{b-a}{2\sqrt{3}}$.

Primjer 3.28 *Neki stroj neprekidno puni butelje čija zapremina može biti bilo koja vrijednost između* 0.7 *i* 0.755 *dcl. Ako pretpostavimo da su te vrijednosti neprekidno uniformno distribuirane, onda je očekivana vrijednost slučajne varijable koja bilježi zapreminu* $\mu = \frac{a+b}{2} = 0.725$ *dcl. Vjerojatnost da je zapremina slučajno odabrane boce veća od* 0.75 *je*

$$p\left(X > 0.75\right) = \int_{0.75}^{0.755} \frac{1}{0.755-0.7} dx = \frac{1}{0.055} x|_{0.75}^{0.755} = \frac{0.755-0.75}{0.055} = \frac{0.005}{0.055} \approx 0.09.$$

Glavna krakteristika ovoga modela slučajne varijable koja opisuje navedeni proizvodni proces jest da je vjerojatnost da napunjena boca ima zapremninu iz nekog intervala širine d *uvijek ista, za svaki interval širine* d*, tj.*

$$p\left(t < X < t + d\right) = p\left(t' < X < t' + d\right), \text{ za svaki } t \in [0.75, 0.7], \ d \in \mathbb{R}.$$

Primjer 3.29 *Vrijeme posluživanja neke stranke na jednom šalteru banke u prosjeku iznosi* 10 *minuta. Ako je utrošak vremena po stranci na tom šalteru eksponencijalno distribuirana slučajna varijabla, onda (stavljajući da je* $\lambda = \frac{1}{\mu} = \frac{1}{10}$*) je vjerojatnost da usluživanje slučajno prispjele stranke bude do 6 minuta jednaka*

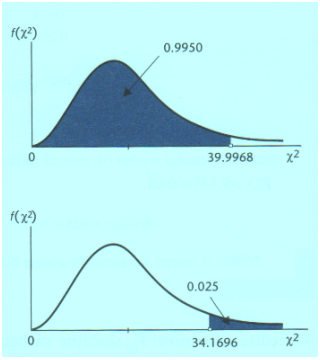
$$p\left(X < 6\right) = \int_{0}^{6} \frac{1}{10} e^{-\frac{1}{10}x} dx = -e^{-\frac{1}{10}x}|_{0}^{6} = -e^{-0.6} + 1 \approx 0.451.$$

Glavna krakteristika ovoga modela slučajne varijable koja opisuje navedenu situaciju jest da je vjerojatnost da vrijeme opsluživanja stranke bude iz nekog intervala $[a, a + d]$ *uvijek veća od vjerojatnosti da vrijeme opsluživanja stranke bude iz nekog intervala* $[b + d, b + d]$*, za svaki* $a < b$*,* $d \in \mathbb{R}$*.*

Primjer 3.30 *Slučajna varijabla koja je* χ^2*-distribuirana s* 20 *stupnjeva slobode ima očekivanje* $\mu = 20$*, a devijaciju jednaku* $\sigma = 2\sqrt{10}$*. Vjerojatnost da je vrijednost slučajne varijable manja od 39.9968 je* $p\left(X < 39.9968\right) = 1 - p\left(X \geqslant 39.9968\right) = 1 - 0.005 = 0.995$*, a vjerojatnost da je između 34.1696 i 39.9968 je*

$p\left(X \in \langle 34.1696, 39.9968 \rangle\right) = p\left(X > 34.1696\right) - p\left(X \geqslant 39.9968\right) = 0.025 - 0.005 =$

0.02.



3.3.1 Normalno distribuirana slučajna varijabla

Ako je gustoća slučajne varijable X funkcija $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$, onda za nju kažemo da je **normalno distribuirana** (ili samo **normalna**) i pišemo$X \sim N(\mu, \sigma)$ (time naglašavamo činjenicu, da gustoća normalno distribuirane varijable ovisi o parametrima μ i σ)

Očekivanje normalno distribuirane varijable $X \sim N(\mu, \sigma)$ je $E[X] = \mu$, a standardna devijacija je $D[X] = \sigma$.

Standardizirana varijabla $Z = \frac{X-\mu}{\sigma}$ od normalno distribuirane varijable X je normalno distribuirana varijabla $Z \sim N(0, 1)$, čija je gustoća $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

Normalnu distribuciju imaju sljedeće varijable:

- visina i težina ljudi;
- inteligencija i razne fizičke i mentalne karakteristike ljudi i drugih živih bića.

Parametri Gaussove razdiobe se u svakom pojedinom slučaju određuju eksperimentalno u ovisnosti o populaciji koju promatramo (pronađu se aritmetička sredina i standardna devijacija uzorka).

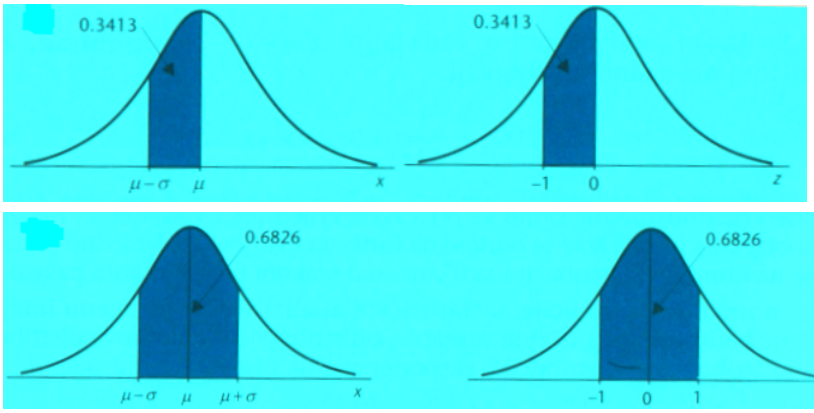
Zadatak 3.31 *Čaj se pakira u vrećice nominalne mase* 50 *g. Masa vrećica je normalno distribuirana slučajna varijabla s očekivanjem jednakim nominalnoj masi i devijacijom od* 2 *g. Ako se slučajno odabere vrećica, kolika je vjerojatnost da je njezina masa manja od* 51 *g, te između* 48 *i* 49 *g?*

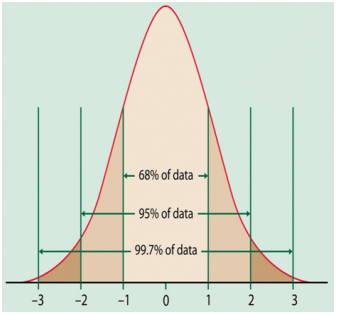
Rješenje: $p(X < 51) = p\left(Z < \frac{51-50}{2}\right) = p(Z < 0.5) = p(Z < 0) +$
$p(0 \leq Z < 0.5) = 0.5 + 0.1915 = 0.6915$

$p(48 < X < 49) = p(-1 < Z < -0.5) = p(-1 < Z < 0) - p(-0.5 < Z < 0) =$
$p(0 < Z < 1) - p(0 < Z < 0.5) = 0.3413 - 0.1915 = 0.149\,8.$

Zadatak 3.32 *Ako je slučajna varijabla* $X \sim N(\mu, \sigma)$ *normalno distribuirana, odredite vjerojatnost da* X *poprimi vrijednost iz intervala* $\langle \mu - \sigma, \mu + \sigma \rangle$ *te iz intervala*
$\langle \mu - 2\sigma, \mu + 2\sigma \rangle$, $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$.

Rješenje: $p(\mu - \sigma < X < \mu + \sigma) = p(-1 < Z < 1) = 2p(0 < Z < 1) =$
$2 \cdot 0.3413 = 0.6826,$
$p(\mu - 2\sigma < X < \mu + 2\sigma) = p(-2 < Z < 2) = 2p(0 < Z < 2) = 2 \cdot 0.4772 =$
$0.9544,$
$p(\mu - 3\sigma < X < \mu + 3\sigma) = p(-3 < Z < 3) = 2p(0 < Z < 3) = 2 \cdot 0.4987 =$
$0.9974.$





Primjedba 3.33 *Često puta se neprecizno i za diskretnu slučajnu varijablu X kaže da je distribuirana razdiobom neke kontinuirane slučajne varijable čija je gustoća f, primjerice, da je normalno distribuirana. No, to zapravo podrazumijeva da se skup točaka $\{(x_1, p_1), (x_2, p_2), ...\}$ distribucije te slučajne varijable X nalazi na grafu funkcije gustoće f i da ove dvije varijable imaju jednaka (ili približno jednaka) očekivanja i varijance. Napose, svaku statističku numeričku varijablu zadanu na konačnom skupu, ili općenitije svaku diskretnu statističku numeričku varijablu, možemo konvertirati u diskretnu slučajnu varijablu kao što je opisano u Primjedbi 3.3, pa ako za nju možemo reći da je distribuirana razdiobom neke kontinuirane slučajne varijable, onda to isto kažemo i za početnu statističku varijablu. Zbog toga možemo tolerirati izjave: duljina životnog vijeka (broj napunjenih godina života) Hrvata je normalno distribuirana, uspjeh studenata na ispitu iz nekog kolegija je normalno distribuiran...*

3.4 Primjene slučajnih varijabli

Osnovna upotreba slučajne varijable u inferencijalnoj statistici je opisivanje nekog slučajnog pokusa iz realnog života. Pri tome slučajnim pokusom u statističkoj analizi smatramo svaku djelatnost iz koje izvire neki broj kao rezultat. Slučajni pokus može biti nasumični odabir nekog elementa populacije, čiji elementi se neprestano mijenjaju, pri čemu je njegovo numeričko obilježje ishod toga pokusa. Primjerice, mjerenje težine ili visine na populaciji svih Hajdukovih pretplatnika ili istraživanje o starosti hrvatskih državljana. Za fiksnu populaciju nam je dostatna i deskriptivna statistika, odnosno statsitička varijabla koja se odnosi samo na zadanu i određenu fiksnu populaciju. Primjerice broj gledatelja na domaćim utakmicama Jugoplastike u sezoni 1990/1991 je zadan i određen, na fiksnoj i konačnoj populaciji svih domaćih utakmica te sezone. Općenito, slučajni pokus može biti promatranje nekih procesa koji se mogu proizvoljno mnogo puta izvesti. Primjerice, broj ubačenih koševa s crte slobodnih bacanja određenoga igrača u seriji od 30 bacanja, broj automobila koji se skupe na prvom crvenom svjetlu na semaforu na raskrižju ulica Hrvatske mornarice i Domovinskog rata u Splitu iza 12 sati. Slučа-

jna varijabla modelira gore opisane pokuse i odnosi se na bilo koji izvedeni pokus određene vrste.

Pri izboru slučajne varijable za opis pojedinog procesa možemo reći da će ona biti diskretna ako može poprimiti konačno mnogo ili najviše prebrojivo mnogo (ne nužno cjelobrojnih) vrijednosti. Izabrati ćemo konkretan model diskretne slučajne varijable ako su u tom slučajnom pokusu ispunjene temeljne značajke toga modela do čega dolazimo iskustvenim saznanjima ili teorijskim razmatranjem. Primjerice, budući da teoretski svako slobodno bacanje određenog igrača ima jednaku vjerojatnost za koš, to pokus sa slobodnim bacanjima opisujemo s binomnom slučajnom varijablom s binomnom $B\{30, p\}$ distribucijom pri čemu za p možemo uzeti postotak šuta iz slobodnih bacanja promatranog igrača. Spomenuti pokus koji ispituje broj automobila koji čekaju da se upali zeleno svjetlo na točno određenom semaforu iskustveno vjerojatno najbolje modelira slučajna varijabla s Poissonovom distribucijom, pri čemu za λ uzimamo srednju vrijednost automobila u prethodnom promatranom razdoblju. Ako iz fiksne populacije svih automobila s hrvatskim tablicama slučajnim odabirom iz registra odaberemo 100 automobila među kojima brojimo one koji imaju zadarske tablice onda je ovaj pokus opisan hipergeometrijskom varijablom gdje je N broj registriranih automobila u R.H., M je broj automobila sa zadarskim registarskim oznakama i $n = 100$.

Ne tako matematički strogo, možemo reći da će slučajna varijabla koja opisuje određeni pokus biti kontinuirana ako ona može poprimiti bilo koju realnu vrijednost (neprebrojivo mnogo vrijednosti), odnosno ne možemo se ograničiti na prebrojivo mnogo vrijednosti koje ta slučajna varijabla može poprimiti. Primjerice, istraživanje visine svih Dalmatinaca moramo opisati kontinuiranom slučajnom varijablom (vjerojatno normalno distribuiranom). Naime, iako u ovom trenutku ima samo konačno mnogo brojeva koji predstavljaju visine svih trenutno živećih Dalmatinaca (pa čak i onih koji su umrli) ne možemo biti sigurni da će visina nekih Dalmatinaca u budućnosti biti na tom popisu visina, odnosno ta visina u cm može biti bilo koji realni broj (broj iz intervala $\langle 0, 300\rangle$). Nadalje, točna temperatura zraka izmjerena na određenom mjestu u određeno vrijeme, rezultat trčanja na 100 metara učenika drugih razreda šibenskih srednjih škola su primjeri pokusa koji se

moraju opisati kontinuiranom slučajnom varijablom.

Poglavlje 4

Dvodimenzionalna slučajna varijabla. Korelacija

4.1 Dvodimenzionalna slučajna varijabla

Česte su situacije u kojima ishodu ω određenoga pokusa pridružujemo više realnih brojeva, tj. uređenu n-torku realnih brojeva $(X_1(\omega), ..., X_n(\omega))$. Primjerice, ako svim studentima ω Sveučilišta u Splitu registriramo prosječnu ocjenu $X_1(\omega)$, duljinu studiranja $X_2(\omega)$ i broj komisijskih ispita $X_3(\omega)$, onda $X = (X_1, X_2, X_3)$ možemo promatrati kao varijablu koja svakom slučajno odabranom studentu Sveučilišta u Splitu koji ima prosječnu ocjenu x_1, koji studira x_2 godina i koji je x_3 puta polagao pred povjerenstvom, pridružuje uređenu trojku (x_1, x_2, x_3). Mi ćemo se ograničiti na varijable $Z = (X, Y)$ koje svakom ishodu ω pridružuju uređeni par $Z(\omega) = (X(\omega), Y(\omega))$. Primjerice događaju ω da slučajno odabrani automobil troši $x = X(\omega)$ litara na 100 km pri brzini od $y = Y(\omega)$ km/h, slučajna varijabla $Z = (X, Y)$ pridružuje uređeni par (x, y).

Definicija 4.1 *Neka je $(\Omega, \mathcal{F}, p)$ vjerojatnosni prostor. Funkciju $Z = (X, Y) : \Omega \to \mathbb{R}^2$ nazivamo **dvodimenzionalnom slučajnom varijablom** ako je $Z^{-1}(\langle a, b\rangle \times \langle c, d\rangle) \in \mathcal{F}$ za svaki $a, b, c, d \in \mathbb{R}$, $a < b$, $c < d$. Ako dvodimenzionalna slučajna varijabla $Z = (X, Y) : \Omega \to \mathbb{R}^2$ poprima najviše konačno ili prebrojivo vrijednosti (x_1, y_1), $(x_1, y_2), ...,$ (x_2, y_1), $(x_2, y_2), ..., (x_i, y_j), ...$ tada ju nazivamo*

diskretnom dvodimenzionalnom varijablom.

Ako je $(\Omega, \mathcal{F}, p)$ diskretni vjerojatnosni prostor (Ω je prebrojiv i $\mathcal{F} = \mathcal{P}(\Omega)$), onda je svaka funkcija $Z = (X, Y) : \Omega \to \mathbb{R}^2$ diskretna slučajna varijabla.

Definicija 4.2 *Neka diskretna dvodimenzionalna slučajna varijabla $Z = (X, Y)$ poprima vrijednosti (x_i, y_j) i neka je $p_{ij} = p(X = x_i, Y = y_j)$ vjerojatnost događaja koji se sastoji od onih ishoda kojima slučajna varijabla pridruži uređeni par (x_i, y_j) (događaj da slučajna varijabla ima vrijednost (x_i, y_j)). Skup svih uređenih parova $((x_i, y_j), p_{ij})$ nazivamo* ***distribucijom slučajne varijable*** *$Z = (X, Y)$.*

Distribuciju diskretne slučajne varijable $Z = (X, Y)$ prikazujemo *tablicom kontigencije*:

$X \backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Primijetimo da je $\sum_{i,j} p_{ij} = p_{11} + p_{12} + \cdots + p_{21} + \cdots + p_{ij} + \cdots = 1$.

Primjer 4.3 *Kod istovremenog bacanja novčića i kocke uređeni par (x, y) pridružujemo događaju da se novčiću pojavilo $x \in \{0, 1\}$ (0-pismo, 1-glava), a na kocki $y \in \{1, ..., 6\}$. Dvodimenzionalna varijabla (X, Y) može poprimiti 12 različitih vrijednosti, a vjerojatnost događaja da varijabla poprimi bilo koju od tih vrijednosti je $\frac{1}{12}$. Distribucija te varijable je prikazana tablicom*

$X \backslash Y$	1	2	3	4	5	6
0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

.

Primjer 4.4 *Dvije trake proizvode određeni artikl. U jedinici vremena, kapacitet proizvodnje prve trake je 4, a druge 3 artikla. Neka* (X, Y) *predstavlja broj proizvedenih artikala prve i druge trake, uz pretpostavku da je proizvodnja slučajna. Neka je distribucija te varijable dana tablicom:*

$X\backslash Y$	0	1	2	3
0	*0.01*	*0.02*	*0.02*	*0.02*
1	*0.02*	*0.04*	*0.04*	*0.04*
2	*0.04*	*0.06*	*0.07*	*0.05*
3	*0.06*	*0.06*	*0.07*	*0.08*
4	*0.08*	*0.07*	*0.07*	*0.08*

.

Odredite vjerojatnost da prva traka proizvede 2 artikla i vjerojatnost da prva traka proizvede više od druge.

Događaju da je prva traka proizvela 2 artikla odgovaraju ishodi $(2,0)$, $(2,1)$, $(2,2)$ *i* $(2,3)$ *pa je* $p(X=2) = p_{20} + p_{21} + p_{22} + p_{23} = 0.22$ *i* $p(X > Y) = p_{10} + p_{20} + p_{21} + p_{30} + \cdots + p_{42} + p_{43} = 0.61$.

4.1.1 Marginalne distribucije

Neka je (X, Y) diskretna slučajna varijabla kojoj je distribucija određena vjerojatnostima $p_{ij} = p(X = x_i, Y = y_j)$, $i = 1, ..., j = 1, ...$

Događaj $(X = x_i)$ se definira kao skup koji se sastoji od svih ishoda u kojima varijabla X poprima vrijednost x_i, tj. od ishoda (x_i, y_1), (x_i, y_2), ...

Vjerojatnost tog događaja je $p_{i\cdot} = p(X = x_i) = p_{i1} + p_{i2} + p_{i3} + \cdots$

Vrijednosti x_i i vjerojatnosti $p_{i\cdot}$ određuju distribuciju slučajne varijable X (a time i slučajnu varijablu X) koju nazivamo **marginalnom distribucijom slučajne varijable** X.

Na sličan način događaj $(Y = y_j)$ se definira kao skup koji se sastoji od svih ishoda u kojima varijabla Y poprima vrijednost y_j, tj. od ishoda (x_1, y_j), (x_2, y_j), ...

Vjerojatnost tog događaja je $p_{\cdot j} = p(Y = y_j) = p_{1j} + p_{2j} + p_{3j} + \cdots$

Vrijednosti y_j i vjerojatnosti $p_{\cdot j}$ određuju distribuciju slučajne varijable Y (a time i slučajnu varijablu Y) koju nazivamo **marginalnom distribucijom slučа-**

jne varijable Y.

$X\backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	$\sum$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	$p_{1\cdot}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	$p_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot j}$	$\cdots$	1

U Primjeru 4.4 možemo odrediti slučajnu varijablu X koja predstavlja proizvodnju na prvoj traci i može poprimiti vrijednosti $0, 1, 2, 3, 4$, te slučajnu varijablu Y koja predstavlja proizvodnju na drugoj traci, a koja može poprimiti vrijednosti $0, 1, 2, 3$. Marginalne distribucije ovih varijabli su dane tablicama:

x_i	0	1	2	3	4
$p_{i\cdot}$	0.07	0.14	0.22	0.27	0.30

Npr. $p_{0\cdot} = p_{00} + p_{01} + p_{02} + p_{03} = 0.1 + 0.2 + 0.2 + 0.2 = 0.7$....

y_j	0	1	2	3
$p_{\cdot j}$	0.21	0.25	0.27	0.27

Npr. $p_{\cdot 3} = p_{03} + p_{13} + p_{23} + p_{33} + p_{43} = 0.02 + 0.04 + 0.05 + 0.08 + 0.08 = 0.27$.

4.1.2 Uvjetne distribucije

Neka je (X, Y) diskretna slučajna varijabla kojoj je distribucija određena vjerojatnostima $p_{ij} = p\left(X = x_i, Y = y_j\right)$, $i = 1, ..., j = 1, ...$ Ako unaprijed znamo da je varijabla Y poprimila vrijednost y_j, vjerojatnost da varijabla X poprimi vrijednost x_i je uvjetna vjerojatnost $p\left(x_i|y_j\right) = p\left(X = x_i|y_j\right) = \frac{p_{ij}}{p_{\cdot j}}$. Očito vrijedi $p\left(x_1|y_j\right) + p\left(x_2|y_j\right) + \cdots = 1$. Vrijednosti x_i i vjerojatnosti $p\left(x_i|y_j\right)$ određuju **uvjetnu slučajnu varijablu** $X|\left(Y = y_j\right)$ (varijabla koja bilježi vrijednosti varijable X ako unaprijed znamo da je varijabla Y poprimila vrijednost y_j). Njezinu distribuciju nazivamo **uvjetnom distribucijom slučajne varijable** $X|\left(Y = y_j\right)$.

Ako unaprijed znamo da je varijabla X poprimila vrijednost x_i, vjerojatnost da varijabla Y poprimi vrijednost y_j je uvjetna vjerojatnost $p(y_j|x_i) = p(Y = y_j|x_i) = \frac{p_{ij}}{p_{i\cdot}}$. Očito vrijedi $p(y_1|x_i) + p(y_2|x_i) + \cdots = 1$. Vrijednosti y_j i vjerojatnosti $p(y_j|x_i)$ određuju **uvjetnu slučajnu varijablu** $Y|(X = x_i)$ (varijabla koja bilježi vrijednosti varijable Y ako unaprijed znamo da je varijabla X poprimila vrijednost x_i). Njezinu distribuciju nazivamo **uvjetnom distribucijom slučajne varijable** $Y|(X = x_i)$.

U Primjeru 4.4 uvjetna distribucija varijable $X|(Y = 1)$, odnosno varijable koja bilježi broj proizvedenih artikla na prvoj traci ako je u istoj jedinici vremena na drugoj traci proizveden samo jedan artikal, je dana u tablici

x_i	0	1	2	3	4
$p(i\|1) = \frac{p_{i1}}{p_{\cdot 1}}$	$\frac{0,02}{0,25} = 0,08$	$\frac{0,04}{0,25} = 0,16$	$\frac{0.06}{0.25} = 0,24$	$\frac{0.06}{0.25} = 0,24$	$\frac{0.07}{0.25} = 0,28$

.

Uvjetna distribucija varijable $Y|(X = 0)$ je:

y_j	0	1	2	3
$p(j\|0) = \frac{p_{0j}}{p_{0\cdot}}$	$\frac{0,01}{0,07} = 0,142$	$\frac{0,02}{0,07} = 0,285$	$\frac{0,02}{0,07} = 0,285$	$\frac{0.02}{0.07} = 0,285$

.

Primjer 4.5 *Distribucija varijable (X, Y) koja bilježi ishode istovremenog bacanja novčića i kocke je dana u tablici kontigencije zajedno s marginalnim distribucijama:*

$X\backslash Y$	1	2	3	4	5	6	$p_{i\cdot}$
0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
$p_{\cdot j}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

.

Uvjetna distribucija varijable $X|(Y = j)$, $j = 1, ..., 6$, je

x_i	0	1
$p(i\|j) = \frac{p_{ij}}{p_{\cdot j}}$	$\frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}$	$\frac{1}{2}$

, odnosno uvjetna distribucija je identična distribuciji varijable X. Analogno se vidi da je uvjetna distribucija varijable $Y|(X = i)$ identična distribuciji varijable Y.

4.1.3 Neovisnost slučajnih varijabli

Definicija 4.6 *Neka je (X,Y) diskretna slučajna varijabla kojoj je distribucija određena vjerojatnostima $p_{ij} = p(X = x_i, Y = y_j)$, $i = 1, ..., j = 1, ...$ Kažemo da su slučajne varijable X (s distribucijom $(x_i, p_{i\cdot})$) i Y (s distribucijom $(y_j, p_{\cdot j})$) **neovisne** ako je $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ za sve i, j.*

Varijable X (bacanje novčića) i Y (bacanja kocke), iz Primjera 4.5, su neovisne. Zaista, $\frac{1}{12} = p_{ij} = p_{i\cdot}p_{\cdot j} = \frac{1}{2} \cdot \frac{1}{6}$, za svaki i, j. Varijable X (proizvodnja artikala na prvoj traci) i Y (proizvodnja artikala na drugoj traci), iz Primjera 4.4, nisu neovisne. Zaista, $p_{00} = 0.01 \neq p_{0\cdot} \cdot p_{\cdot 0} = 0.07 \cdot 0.21 = 0.014\,7$.

Teorem 4.7 *Neka su slučajne varijable X i Y neovisne. Tada je $E[XY] = E[X]E[Y]$.*

4.2 Kovarijanca i koeficijent korelacije

Neka je (X,Y) dvodimenzionalna slučajna varijabla i neka je $\mu_X = E[X]$, $\mu_Y = E[Y]$, $\sigma_X = D[X]$, $\sigma_Y = D[Y]$. **Kovarijanca varijabli** X i Y je broj

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y.$$

Za diskretne varijable vrijedi $cov(X,Y) = \sum_i \sum_j p_{ij}x_iy_j - \mu_X\mu_Y$.

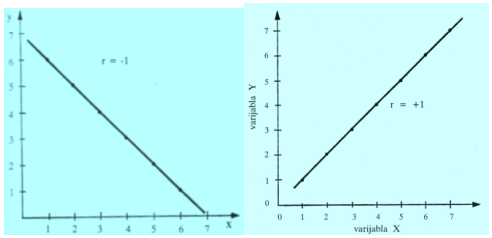
Kovarijanca mjeri stupanj linearne povezanosti varijabli X i Y (ona je i najčešća u prirodi). Ako su varijable neovisne onda je njihova kovarijanca jednaka 0. No, ako im je kovarijanca jednaka 0, varijable ne moraju biti neovisne (mogu biti nelinearno povezane).

Kao mjera stupnja linearne povezanosti još se koristi **koeficijent korelacije** $r(X,Y) = \rho(X,Y) = \rho = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$.

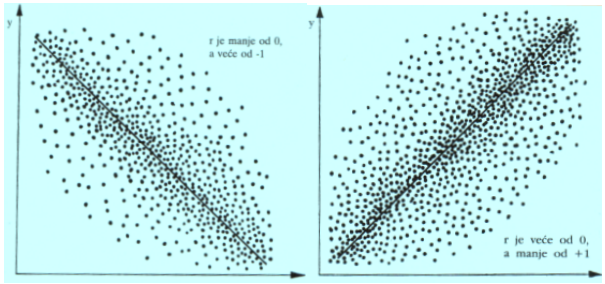
Koeficijent korelacije je broj sa svojstvom $-1 \leq \rho \leq 1$. Prikažemo li u pravokutnom koordinatnom sustavu točke $(x,y) = (X(\omega), Y(\omega))$, dobivamo tzv. **dijagram rasipanja**. Koeficijent ρ je bliži 1 ili -1 što taj dijagram uspješnije možemo aproksimirati pravcem.

Ako je $\rho(X,Y) > 0$ kažemo da su varijable pozitivno korelirane (pravac je rastući-rast varijable Y odgovara rastu varijable X), odnosno ako je $\rho(X,Y) < 0$ kažemo da su negativno korelirane (pravac je padajući-pa rastu varijable X odgovara pad varijable Y).

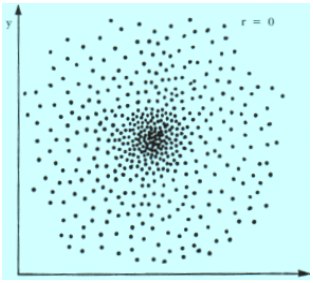
Vrijedi $|\rho(X,Y)| = 1$ ako i samo ako je $Y = aX + b$ (varijable su u linearnoj funkcionalnoj ovisnosti).



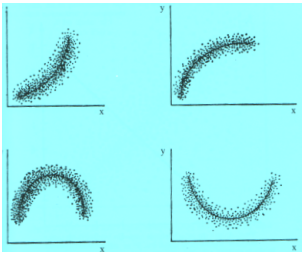
Ako je $0.7 \leq |\rho| < 1$, onda su varijable vrlo visoko linearno povezane. Primjerice, visina i težina ljudi ($\rho \geqslant 0.7$) ili stupanj utreniranosti i puls u prvoj minuti oporavka nakon vježbe ($\rho \leq -0.7$)



Ako je $0.4 \leq |\rho| < 0.7$ tada su varijable značajno korelirane. Ako je $0.2 \leq |\rho| < 0.4$ tada su varijable slabo korelirane. Ako je $0 < |\rho| < 0.2$ tada su varijable neznatno korelirane, odnosno ako je $\rho(X,Y) = 0$ kažemo da su varijable **nekorelirane**. Varijable su nekorelirane ako i samo ako je $cov(X,Y) = 0$.



Ako su varijable neovisne, onda su nekorelirane. Obrat ne vrijedi. Naime, povezanost među varijablama može biti vrlo visoka i nelinearna (dijagram rasipanja se može aproksimirati nekom drugom krivuljom umjesto pravca). Primjerice, koeficijent korelacije između broja ponavljanja i količine upamćenog gradiva je malen iako je povezanost očita. Naime, s prvim ponavljanjima, količina naučenog gradiva naglo raste, a kasnije porast blago stagnira. Nadalje, istraživanjem povezanosti između inteziteta rasvjete i radnog učinka u nekom preciznom poslu dolazimo u početku do velikog porasta učinka s porastom rasvjete, a kod jakih inteziteta rasvjete njezina promjena nema efekta na učinak, dok kod prevelikog porasta inteziteta rasvjete dolazi do zaslijepljenosti radnika i do padajućeg učinka.



Koeficijent korelacije je jedan od najčešće upotrebljavanih ali i zloupotrebljavanih statističkih podataka.
Činjenicu da su dvije varijable visoko korelirane treba oprezno interpretirati. Česta pogreška je neuvažavanje da su obje varijable uzročno povezane s trećom varijablom. Primjerice, broj električnih aparata i broj djece u domaćinstvima je visoko koreliran, ali ne zato što broj djece u obitelji djeluje na broj aparata, već je

to posljedica standarda koji djeluje i na jedno i na drugo. U podrobnijoj analizi bismo trebali isključiti utjecaj ekonomskog standarda i promatrati domaćinstva s približno jednakim ekonomskim mogućnostima. Korelacija duljine stopala i sposobnosti pisanja djece od 1. do 8. razreda je velika, a to je odraz starenja. U podrobnijoj analizi bismo trebali promatrati učenike iste dobi, te bismo zaključili nekoreliranost. Slično se može dobiti besmislica o broju kino dvorana i mačaka lutalica u gradovima...

Teorem 4.8 *Neka je* (X,Y) *dvodimenzionalna slučajna varijabla. Tada je* $E[X+Y]=E[X]+E[Y]$, $Var[X+Y]=Var[X]+Var[Y]-2cov(X,Y)$.

Ako su X i Y neovisne slučajne varijable, onda je $Var[X+Y]=Var[X]+Var[Y]$.

Zadatak 4.9 *Zadani su podaci o prodaji motora u nekom salonu za 60 radnih dana. U tablici su prikazane frekvencije dana kada je broj prodavača bio* x_i*, a broj prodanih motora* y_j

$x_i \backslash y_i$	0	1
1	0	20
2	20	0
3	0	20

.

Ispitajte neovisnost i koreliranost slučajnih varijabla X *i* Y*.*

Rješenje: *Tablica distribucije varijable* (X,Y) *zajedno s marginalnim distribucijama varijabli* X *i* Y *je:*

$x_i \backslash y_i$	0	1	$p_{i\cdot}$
1	0	$\frac{1}{3}$	$\frac{1}{3}$
2	$\frac{1}{3}$	0	$\frac{1}{3}$
3	0	$\frac{1}{3}$	$\frac{1}{3}$
$p_{\cdot j}$	$\frac{1}{3}$	$\frac{2}{3}$	1

.

Slučajne varijable X *i* Y *nisu neovisne jer je* $p_{00}=0\neq\frac{1}{9}=p_{0\cdot}\cdot p_{\cdot 0}$..., *a varijable nisu ni korelirane jer je* $cov(X,Y)=0$*. Zaista,*

$$E[XY]=\sum_{i=1}^{3}\sum_{j=0}^{1} i\cdot j\cdot p_{ij}=1\cdot 0\cdot 0+1\cdot 1\cdot\tfrac{1}{3}+2\cdot 0\cdot\tfrac{1}{3}+2\cdot 1\cdot 0+3\cdot 0\cdot 0+3\cdot 1\cdot\tfrac{1}{3}=\tfrac{4}{3},$$

$E[X] = \sum_{i=1}^{3} i \cdot p_{i\cdot} = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} = 2,$
$E[Y] = \sum_{j=0}^{1} j \cdot p_{\cdot j} = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3},$
$cov(X,Y) = E[XY] - E[X]E[Y] = \frac{4}{3} - 2 \cdot \frac{2}{3} = 0.$

4.3 Kontinuirana dvodimenzionalna slučajna varijabla

Za slučajnu varijablu (X,Y) kažemo da je **neprekidna** ili **kontinuirana** ako postoji nenegativna funkcija $f : \mathbb{R}^2 \to \mathbb{R}$ takva da je $p(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x,y)\,dxdy$, za svaki $a, b, c, d \in \mathbb{R}$, $a < b$, $c < d$. Funkciju f nazivamo **gustoćom** slučajne varijable (X,Y).

Primjerice, funkcija

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]}$$

je gustoća normalno distribuirane dvodimenzionalne slučajne varijable (X,Y).

Funkcije $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$, $x \in \mathbb{R}$ i $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$, $y \in \mathbb{R}$ su **marginalne funkcije gustoće** slučajnih varijabli X i Y redom.

Funkcije $f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$, $x \in \mathbb{R}$, i $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ su **uvjetne funkcije gustoće vjerojatnosti** za slučajnu varijablu X uz uvjet $Y = y$ i slučajnu varijablu Y uz uvjet $X = x$, redom.

Kontinuirane slučajne varijable X i Y su neovisne ako je $f(x,y) = f_X(x) \cdot f_Y(y)$, za sve $x, y \in \mathbb{R}$.

Za kontinuiranu slučajnu varijablu (X,Y) se definira jednako kovarijanca i koeficijent korelacije i vrijede analogne tvrdnje i svojstva kao i za diskretnu varijablu.

Poglavlje 5

Intervali povjerenja

5.1 Metoda uzoraka

Neka je S populacija i $X : S \to O = \{o_1, ..., o_k\} \subseteq \mathbb{R}$ numerička varijabla koja svakom članu s_i populacije (statističkog skupa) $S = \{s_1, s_2, ..., s_N\}$ pridružuje neko numeričko obilježje $X(s_i)$. Primijetimo da X možemo promatrati kao slučajnu varijablu $X : \Omega \to O$, pri čemu je prostor elementarnih događaja $\Omega = \{\omega_1, ..., \omega_k\}$ gdje je ω_i događaj da neki član populacije ima obilježje $o_i \in O$, tj. $X(\omega_i) = o_i$, a vjerojatnost događaja je $p(\omega_i)$ jednaka relativnoj frekvenciji $p_i = \frac{f_i}{N}$ obilježja o_i. Ako je populacija beskonačna, onda je prikladnije govoriti o slučajnoj varijabli, budući da pojam relativne frekvencije obilježja u tom slučaju nema smisla već je zamjenjen pojmom vjerojatnosti da varijabla poprimi to obilježje. Početak razmatranja može biti i slučajna varijabla X (neovisno o statističkoj varijabli) koja opisuje neki slučajni pokus za koju ćemo onda reći da ima beskonačnu populaciju koja se sastoji od svih mogućih pokusa. Tada nam je, u statističkom istraživanju, za izračun parametara te varijable, kao što su očekivanje slučajne varijable μ (isto što i aritmetička sredina numeričke varijable), varijanca σ^2 (parametri koji brojčano iskazuju osobitost populacije), ili za izračun proporcije p (relativne frekvencije ili vjerojatnosti) pojedinog obilježja, potrebno imati sve vrijednosti varijable X, odnosno treba biti dostupno numeričko obilježje svakog člana populacije, tj. treba biti dostupna vjerojatnost (relativna frekvencija) svakog obilježja.

Ako je varijabla X nenumerička, tj. ako se svakom članu populacije pridružuje neko nenumeričko obilježje, onda se pripadna slučajna varijabla dobiva kodiranjem. Primjerice, događaju ω_i da neki član populacije ima obilježje o_i slučajna varijabla X pridruži broj i. Tako nenumeričku varijablu koja bilježi ishod svakog mogućeg bacanja novčića možemo promatrati kao slučajnu varijablu koja svakom događaju "palo je pismo" pridruži broj 0, a događaju "pala je glava" pridruži 1, ili varijablu koja svakom ispitaniku iz populacije pridružuje odgovor na postavljeno pitanje "DA" ili "NE" možemo promatrati kao slučajnu varijablu koja događaju "ispitanik je odgovorio DA" pridružuje broj 1, odnosno "ispitanik je odgovorio NE" pridruži 0. U tom slučaju vjerojatnost da slučajna varijabla poprimi vrijednost 1 (0) je jednaka proporciji obilježja "DA" ("NE"). Također, i ovom slučaju, u statističkom istraživanju, za izračun proporcije (relativne frekvencije) ili vjerojatnosti pojedinog obilježja potrebno je imati sve vrijednosti varijable X, odnosno treba biti dostupno numeričko obilježje svakog člana populacije, tj. treba biti dostupna vjerojatnost (relativna frekvencija) svakog obilježja.

Često puta nije moguće prikupiti sve podatke, odnosno obilježja svakog elementa populacije radi raznih razloga: zbog prevelikog ili beskonačnog opsega populacije, složenosti istraživanja, previsokih financijskih troškova takvog istraživanja, ako se istraživanjem uništavaju elementi populacije (ispitivanje biometrijskih karakteristika ljudi, testiranje tehničkih proizvoda, kemijska analiza prehrambenih konzervi, stavovi ljudi neke regije o najvećim političkim strankama, njihovim liderima, programima i do sada ostvarenim, a obećanim ciljevima...).
Tada se moramo zadovoljiti dijelom podataka, odnosno vrijednostima varijable na **uzorku** (podskupu S_0 populacije S). U tom slučaju želimo donijeti što kvalitetniji zaključak o cijeloj populaciji temeljem podataka na uzorku. Naravno, svaki takav zaključak, osim ako uzorak nije jednak čitavoj populaciji, sadrži grešku, pa zaključke možemo donijeti s nekom razinom **pouzdanosti** (vjerojatnosti da je zaključak o osobitosti cijele populacije točan). S ciljem dobivanja što reprezentativnijih, vjerodostojnijih zaključaka o cijeloj populaciji moramo se pobrinuti da prikupljeni podaci budu na **reprezentativnom** uzorku. Isto tako da bismo mogli definirati sve parametre teorijske slučajne varijable možemo izvesti samo ograničen

broj pokusa koje tada smatramo uzorkom na kojemu možemo izračunati relativne frekvencije pojedinih ishoda kojima ćemo procijeniti njihove vjerojatnosti (broj posjetitelja negog kafića subotom, broj točnih odgovora sudionika nekog kviza...).

Najreprezentativniji uzorak je **slučajni uzorak** koji se formira na način da svaki element populacije ima jednaku vjerojatnost da bude izabran u uzorak. Najbolji način sastavljanja slučajnog uzorka, a da pri tomu eliminiramo nesvjesno psihološko-praktično preferiranje pojedinih elemenata, jest da se elementi numeriraju i onda nasumce, računalnim programom, izabiru brojevi. Naravno, postoji mogućnost i da slučajni uzorak ne bude reprezentativan, tj. da bude **pristran**, na što uostalom upućuje i činjenica da svaki zaključak temeljen na uzorku ima određenu razinu (ne)pouzdanosti.

Jednako reprezentativan je i **sistemski uzorak** u kojem se odabere nasumce prvi član uzorka iz numerirane populacije, a nakon njega u uzorak ulazi svaki n-ti član populacije.

Stratificirani uzorak je u mnogim slučajevima reprezentativniji nego li slučajni uzorak. Kod njegovog formiranja se populacija prvo podijeli, prema nekim karakteristikama, u slojeve (stratume), a potom se iz svakog sloja uzima slučajni uzorak tako da njegova veličina u odnosu na veličinu cijelog uzorka bude proporcionalna veličini sloja u odnosu na veličinu cijele populacije.

Klaster uzorak je lošija varijanta slučajnog uzorka, a koristi se u velikim ekonomskim, političkim ili tržišnim istraživanjima. Formira se na način da se cijela populacija podijeli u više manjih blokova (primjerice grad se podijeli na više kvartova ili blokova), pa se nasumce odabere jedan od tih blokova koji onda predstavlja klaster uzorak. Ovaj uzorak je praktičan jer su anketari koncentrirani na jednom području.

Kvotni uzorak se formira na način da organizator istraživanja, poznajući strukturu stanovništva obzirom na predmet istraživanja, unaprijed odredi broj ljudi iz svakog pojedinog stratuma, a anketar sam odabire te ljude dok ne ispuni kvotu. Ovakav uzorak često nije reprezentativan, jer anketar sam, hodajući gradom ili jednom ulicom, po svojim afinitetima i atrakcijama, odabire ispitanike.

Prigodni uzorak je uzorak koji nam je, u datim oklolnostima, jedini dostupan

i može biti ekstremno pristran.

5.2 Procjenitelj parametra

Na svakom uzorku parametar $\hat{\theta}$ koji je izračunat pomoću vrijednosti obilježja članova uzorka nazivamo **procjeniteljem** toga istoga parametra θ izračunatog na cijeloj populaciji. Procjenitelj $\hat{\theta}$ možemo promatrati kao varijablu koja svakom uzorku $\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}$ veličine $n < N \leq \infty$ (podskupu bilo konačne populacije $S = \{s_1, s_2, ..., s_N\}$, bilo beskonačne) pridružuje parametar $\hat{\theta}\left(s_{i_1}, s_{i_2}, ..., s_{i_n}\right)$. Označimo li sa $\mathcal{P}_n$ skup svih uzoraka veličine n, tj. skup svih n-članih podskupova od S, onda je procjenitelj $\hat{\theta}$ varijabla $\hat{\theta} : \mathcal{P}_n \to \mathbb{R}$. Svaki procjenitelj možemo tretirati kao slučajnu varijablu (na gore opisan način) koja događaju da n-uzorak ima parametar $\hat{\theta}$ pridruži upravo broj $\hat{\theta}$. **Sampling distribucija** je distribucija te slučajne varijable. Primjerice, $\overline{x}$ je procjenitelj aritmetičke sredine koji svakom uzorku $\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}$ veličine n pridružuje aritmetičku sredinu toga uzorka $\overline{x}\left(\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}\right) = \frac{s_{i_1}+s_{i_2}+\cdots+s_{i_n}}{n}$.
Ako je početno razmatranje slučajna varijabla koja opisuje neki slučajni pokus (koji se može izvesti neograničeno mnogo puta) onda uzorkom iz $\mathcal{P}_n$ smatramo niz od uzastopnih n pokusa, a aritmetička sredina $\overline{x}$ ishoda tih n pokusa je procjenitelj očekivanja te slučajne varijable. Prisjetimo se da smo za ovakva razmatranja dogovorno govorili da imaju beskonačnu populaciju.

Ako je populacija konačna ili prebrojiva (najčešći slučaj kod društvenih istraživanja), onda uzoraka fiksne veličine ima konačno ili prebrojivo, pa je procjenitelj $\hat{\theta}$ diskretna slučajna varijabla. U slučaju kada je populacija neprebrojiva ili kad je početno razmatranje slučajna varijabla koja opisuje neki slučajni pokus (koji se može izvesti neograničeno mnogo puta), onda i uzoraka ima neprebrojivo, a procjenitelj $\hat{\theta}$ je neprekidna slučajna varijabla. To možemo prepoznati kod ispitivanja nekih kontinuiranih procesa, primjerice mjerenje temperature zraka u jednom danu, gdje temperatura poprima vrijednost u svakom dijeliću vremena, a mi raspolažemo s podacima na uzorku koji se sastoji od n mjerenja. Strogo govoreći, elementi uzorka su ovdje mali intervali vremena.

Primjer 5.1 *Neka populacija* $S = \{A, B, C, D\}$ *ima sljedeća obilježja*

s	A	B	C	D
$X(s)$	1	5	3	7

. Skup uzoraka veličine 2 je $\mathcal{P}_2 = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}\}$. *Vrijednosti procjenitelja aritmetičke sredine* $\overline{x} : \mathcal{P}_2 \to \mathbb{R}$ *su*

uzorak u	A, B	A, C	A, D	B, C	B, D	C, D
$\overline{x}(u)$	3	2	4	4	6	5

.

Sampling distribucija diskretne slučajne varijable $\overline{x}$ *je zadana vrijednostima* $(\overline{x_i}, p_i)$*, gdje je* $\overline{x_i}$ *vrijednost aritmetičke sredine, a* p_i *vjerojatnost (relativna frekvencija) da 2-uzorak ima aritmetičku sredinu jednaku* $\overline{x_i}$*.*

$\overline{x_i}$	3	2	4	6	5
$p(\overline{x_i})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Očekivanje slučajne varijable $\overline{x}$ *jednako je* $E[\overline{x}] = 3\frac{1}{6} + 2\frac{1}{6} + 4\frac{2}{6} + 6\frac{1}{6} + 5\frac{1}{6} = 4$*, što je jednako aritmetičkoj sredini cijele populacije* $\mu = \frac{1+5+3+7}{4} = 4$*.*

Definicija 5.2 *Kažemo da je procjenitelj* $\hat{\theta}$ *parametra* θ ***nepristran*** *ako je očekivanje slučajne varijable* $\hat{\theta}$ *jednako parametru* θ *cijele populacije, tj.* $E\left[\hat{\theta}\right] = \theta$*.*

Teorem 5.3 *Varijabla* $\overline{x}$ *koja svakom* n*-uzorku pridružuje njegovu aritmetičku sredinu je nepristran procjenitelj aritmetičke sredine (ili očekivanja)* μ *cijele populacije, tj. vrijedi* $E[\overline{x}] = \mu$*.*

Ako na konačnoj populaciji $S = \{s_1, s_2, ..., s_N\}$ M elemenata ima isto obilježje o, tada je proporcija (relativna frekvencija) toga obilježja jednaka $p = \frac{M}{N}$. U slučaju beskonačne populacije označimo sa p vjerojatnost događaja da element ima obilježje o, a sa $\hat{p}$ procjenitelja proporcije koji svakom uzorku $\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}$ veličine n pridružuje proporciju obilježja o, tj. $\hat{p}(\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}) = \frac{m}{n}$, gdje je m broj elemenata u uzorku koji imaju obilježje o.

Teorem 5.4 *Varijabla* $\hat{p} : \mathcal{P}_n \to \mathbb{R}$ *je nepristrani procjenitelj proporcije (vjerojatnosti)* p *nekog obilježja na cijeloj populaciji, tj* $E[\hat{p}] = p$*.*

Primjer 5.5 *Populacija se sastoji od* 5 *glasača koji odgovaraju na referendumsko pitanje sa DA (=*1*) ili NE (=*0*). Rezultati glasovanja su*

glasač	*A*	*B*	*C*	*D*	*E*
odgovor	*DA*	*NE*	*NE*	*DA*	*NE*

. *Proporcija odgovora DA je* $p = \frac{M}{N} = \frac{2}{5} = 0.4$. *Svih uzoraka veličine 2 ima 10, a proporcije uzoraka su*

Glasači u uzorku	*odgovori*	m_i	$\hat{p}_i = \frac{m_i}{2}$
A,B	$1, 0$	1	0.5
A,C	$1, 0$	1	0.5
A,D	$1, 1$	2	1
A,E	$1, 0$	1	0.5
B,C	$0, 0$	0	0
B,D	$0, 1$	1	0.5
B,E	$0, 0$	0	0
C,D	$0, 1$	1	0.5
C,E	$0, 0$	0	0
D,E	$1, 0$	1	0.5

.

Sampling distribucija varijable $\hat{p}$ *je*

$\hat{p}_i$	0	0.5	1
$p(\hat{p}_i)$	$\frac{3}{10}$	$\frac{6}{10}$	$\frac{1}{10}$

.

Očekivanje od $\hat{p}$ *je* $E[\hat{p}] = 0.3 \cdot 0 + 0.6 \cdot 0.5 + 0.1 \cdot 1 = 0.4$ *što je jednako proporciji* p *odgovora DA na cijeloj populaciji.*

Označimo sa $\hat{\sigma}^2$ procjenitelja varijance σ^2 koji uzorku $\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}$ veličine n uzetom iz konačne populacije veličine N pridružuje broj

$$\hat{\sigma}^2\left(\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}\right) = s^2 \frac{n}{n-1} \frac{N-1}{N},$$

gdje je s^2 varijanca uzorka. Ako je populacija beskonačna, onda procjenitelja varijance definiramo sa

$$\hat{\sigma}^2\left(\{s_{i_1}, s_{i_2}, ..., s_{i_n}\}\right) = s^2 \frac{n}{n-1}.$$

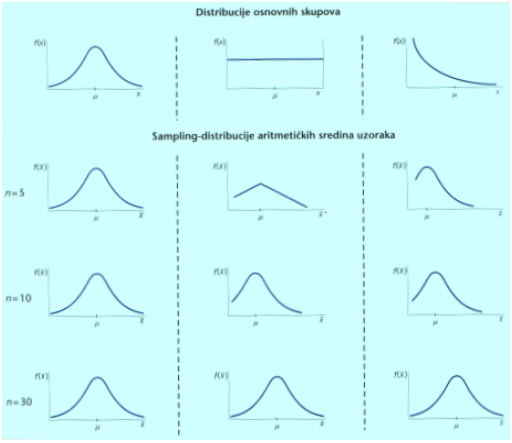
Pokaže se da u tom slučaju vrijedi $\hat{\sigma}^2 = \frac{\left(s_{i_1} - \overline{x}\right)^2 + \cdots + (s_{i_n} - \overline{x})^2}{n-1}$, gdje je $\overline{x}$ aritmetička sredina uzorka.

Teorem 5.6 *Varijabla* $\hat{\sigma}^2 : \mathcal{P}_n \to \mathbb{R}$ *je nepristrani procjenitelj varijance* σ^2 *cijele populacije, tj* $E\left[\hat{\sigma}^2\right] = \sigma^2$.

U praksi se najčešće zanemaruje faktor $\frac{N-1}{N}$ za dovoljno velike populacije jer je $\lim\limits_{N\to\infty} \frac{N-1}{N} = 1$.

5.2.1 Sampling distribucije procjenitelja

Procjenjuje li se parametar samo brojem, nije moguće donijeti sud o preciznosti procjene, niti o razini pouzdanosti s kojom možemo upotrijebiti tu procjenu. Zato su nam potrebne informacije o sampling distribucijama procjenitelja.



Ako slučajni uzorak potječe iz normalno distribuirane populacije $N(\mu, \sigma)$, onda je sampling distribucija aritmetičkih sredina $\overline{x}$ također normalno distribuirana i to s očekivanjem $\mu_{\overline{x}} = \mu$ i standardnom devijacijom $\sigma_{\overline{x}}$ (još se kaže standardna greška sredine).

Ako je slučajni uzorak izabran iz proizvoljno distribuirane populacije s parametrima μ, $\sigma > 0$, onda je, u slučaju da je uzorak **dovoljno velik**, tj. ako je uzorak čija je veličina $n > 30$, sampling distribucija aritmetičkih sredina približno normalno distribuirana s očekivanjem $\mu_{\overline{x}} = \mu$ i standardnom devijacijom $\sigma_{\overline{x}}$. Ovo je posljedica Centralnog graničnog teorema koji tvrdi da sampling distribucije teže ka normalnoj distribuciji $N(\mu, \sigma_{\overline{x}})$ kad veličina uzorka n teži u beskonačno.

Budući se u statističkim istraživanjima najviše bavimo konačnim populacijama, pa su i odgovarajuće varijable procjenitelja diskretne, prisjetimo se da, po Primjedbi 3.33, takve smatramo normalno distribuiranima ako se skup točaka $\{(x_1, p_1), (x_2, p_2), ...\}$ distribucije te diskretne slučajne varijable nalazi na grafu Gaussove krivulje s parametrima μ i σ. Možemo to shvatiti kao da su vrijednosti konačne populacije jedan pogodan uzorak uzet iz skupa vrijednosti normalno distribuirane slučajne varijable.

Standardna devijacija procjenitelja $\overline{x}$ je jednaka $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ ako je populacija konačna i veličine N, odnosno $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ ako je populacija beskonačna. U praksi se faktor $\sqrt{\frac{N-n}{N-1}}$ izostavlja i za konačne populacije ako je $\frac{n}{N} < 0.05$ (populacije čija je veličina puno veća od veličine uzorka).

Primjer 5.7 *U Primjeru 5.1 aritmetička sredina (očekivanje) varijable X na cijeloj populaciji jednaka je $\mu = 4$. Standardna devijacija varijable X je* $\sigma = \sqrt{\frac{(1-4)^2+(5-4)^2+(3-4)^2+(7-4)^2}{4}} = \sqrt{5}$.

Očekivanje varijable $\overline{x}$ je $E[\overline{x}] = 4$, a standardna devijacija je $\sigma_{\overline{x}} = \sqrt{\frac{(3-4)^2+(2-4)^2+2(4-4)^2+(6-4)^2+(5-4)^2}{6}} = \sqrt{\frac{5}{3}}$ *što je jednako* $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{5}}{\sqrt{2}}\sqrt{\frac{4-2}{4-1}}$.

Sampling distribucija proporcija $\hat{p}$ uzoraka veličine n uzetih iz konačnog skupa veličine N je hipergeometrijska s očekivanjem jednakim stvarnoj proporciji p i standardnom devijacijom jednakoj $\sigma_{\hat{p}} = \sqrt{\frac{p\cdot(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$. Ako je populacija beskonačna, onda je ona binomna s očekivanjem p i standardnom devijacijom jednakoj $\sigma_{\hat{p}} = \sqrt{\frac{p\cdot(1-p)}{n}}$. U praksi se faktor $\sqrt{\frac{N-n}{N-1}}$ izostavlja i za konačne populacije ako je $\frac{n}{N} < 0.05$. Ako je uzorak **dovoljno velik**, tj. ako je $np \geqslant 5$ i $n(1-p) \geqslant 5$, onda je u oba slučaja sampling distribucija približno normalna $N(p, \sigma_{\hat{p}})$, $\sigma_{\hat{p}} = \sqrt{\frac{p\cdot(1-p)}{n}}$.

Sampling distribucija varijanci $\hat{\sigma}^2$, ako slučajni uzorak veličine n potječe iz normalno distribuirane populacije, ima oblik χ^2-distribucije. Preciznije, varijabla $\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$ ima χ^2-distribuciju s $n-1$ stupnjeva slobode.

5.3 Intervali povjerenja

Kvalitetnija procjena parametra θ populacije od točkaste procjene (procjenitelj $\hat{\theta}$) je **intervalna procjena** koja se sastoji u određivanju intervala $\langle a, b\rangle$ za kojeg možemo s vjerojatnošću $1-\gamma$ tvrditi da sadrži parametar θ, tj. $p\left(\theta \in \langle a, b\rangle\right) = 1-\gamma$, tako da granice a i b ovise o vrijednostima na uzorku. Vjerojatnost $1-\gamma$ nazivamo **razinom pouzdanosti** procjene koji se obično izražava postotkom $(1-\gamma)\,100\%$ i uzima 90%, 95%, 99%. Širinu intervala nazivamo **preciznošću** procjene, a sam interval **intervalom povjerenja**.

Razina pouzdanosti i preciznost procjene su obrnuto proporcionalni. Uži interval, odnosno veća preciznost se može postići uz smanjenje pouzdanosti, odnosno, veća pouzdanost rezultira smanjenjem preciznosti, odnosno širim intervalom.

5.3.1 Procjena aritmetičke sredine

Ako je uzorak, uzet iz proizvoljno distribuirane populacije, nepoznatog očekivanja μ, dovoljno velik ($n > 30$), onda je sampling distribucija aritmetičkih sredina $\overline{x}$ normalna $N\left(\mu, \sigma_{\overline{x}}\right)$. Tada se u intervalu $\left\langle \mu - z_{\gamma/2}\sigma_{\overline{x}}, \mu + z_{\gamma/2}\sigma_{\overline{x}} \right\rangle$ nalazi $(1-\gamma)\,100\%$ aritmetičkih sredina uzoraka, odnosno vjerojatnost da aritmetička sredina $\overline{x}$ nekog uzorka bude u ovom intervalu je $1-\gamma$. Broj $z_{\gamma/2}$ je vrijednost standardizirane normalne varijable Z koja ima svojstvo $p\left(Z > z_{\gamma/2}\right) = \gamma/2$. No, to znači da je vjerojatnost da aritmetička sredina μ populacije bude u intervalu $\left\langle \overline{x} - z_{\gamma/2}\sigma_{\overline{x}}, \overline{x} + z_{\gamma/2}\sigma_{\overline{x}} \right\rangle$ također $1-\gamma$.

Interval

$$\left\langle \overline{x} - z_{\gamma/2}\sigma_{\overline{x}}, \overline{x} + z_{\gamma/2}\sigma_{\overline{x}} \right\rangle$$

nazivamo **intervalom povjerenja aritmetičke sredine za velike uzorke s razinom pouzdanosti** $1-\gamma$.

Interval interpretiramo na način da se s vjerojatnošću $1-\gamma$ očekuje da nepoznata aritmetička sredina populacije bude veća od donje, a manja od gornje granice intervala. Budući $\sigma_{\overline{x}}$ u slučaju i konačne i beskonačne populacije ovisi o standardnoj devijaciji σ cijele populacije koja je najčešće nepoznata, to ju u formuli

smijemo zamijeniti s njezinim procjeniteljem

$$\hat{\sigma} = \sqrt{s^2 \frac{n}{n-1} \frac{N-1}{N}}$$

izračunatim na uzorku, odnosno s

$$\hat{\sigma} = \sqrt{s^2 \frac{n}{n-1}},$$

ako je N nepoznat i dovoljno velik $\frac{N-1}{N} \approx 1$.
Ovaj interval možemo primjenjivati i na male uzorke ($n \leq 30$) ako je populacija iz koje je uzorak uzet normalna i ako joj je poznata standardna devijacija σ.

Primjer 5.8 *Ako su vrijednosti populacije* 6 *s frekvencijom* 50, 10 *s frekvencijom* 20 *i* 30 *s frekvencijom* 10, *onda je aritmetička sredina populacije* $\mu = 10$, *a standardna devijacija je* $\sigma = \sqrt{60}$. *Standardna devijacija aritmetičkih sredina* $\overline{x}$ *uzoraka veličine 31 je* $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{60}}{\sqrt{31}}\sqrt{\frac{49}{79}} \approx 1.095$. *Ako iz populacije odaberemo uzorak veličine* 31 *koji se sastoji od elemenata populacije koji imaju obilježje* 6, *onda je aritmetička sredina uzorka* $\overline{x} = 6$, *a odgovarajući interval povjerenja s razinom pouzdanosti* 90% *je* $\left\langle \overline{x} - z_{\gamma/2}\sigma_{\overline{x}}, \overline{x} + z_{\gamma/2}\sigma_{\overline{x}} \right\rangle$. *Budući je* $1 - \gamma = 0.9$, *to je* $\gamma/2 = 0.05$ *i* $z_{0.05} \approx 1.645$, *(*$0.05 = p\,(Z > z_{0.05}) = 0.5 - p\,(Z < z_{0.05}) = 0.5 - 0.4495$, *pa je po tablici vrijednosti normalne distribucije* $z_{0.05} \in \langle 1.64, 1.65 \rangle$*). Stoga je interval jednak* $\langle 6 - 1.645 \cdot 1.095, 6 + 1.645 \cdot 1.095 \rangle = \langle 4.1987, 7.8013 \rangle$. *Očito je da* $\mu = 10$ *ne pripada ovom intervalu. Za ovakav uzorak niti povećanje razine pouzdanosti na* 99% *ne bi dalo zadovoljavajući interval. Naime, u tom slučaju je* $\gamma/2 = 0.005$ *i* $z_{0.005} = 2.58$, *pa interval očekivanja* $\langle 6 - 2.58 \cdot 1.095, 6 + 2.58 \cdot 1.095 \rangle = \langle 3.1749, 8.825\,1 \rangle$ *također ne sadrži* μ. *No, uzmemo li primjerice uzorak koji se sastoji od* 29 *elemenata s obilježjem* 6 *i* 2 *elementa s vrijednošću* 30, *onda je aritmetička sredina uzorka* $\overline{x} = 7.54$, *a pripadni interval očekivanja s* 99 % *razine pouzdanosti je* $\langle 7.54 - 2.58 \cdot 1.095, 7.54 + 2.58 \cdot 1.095 \rangle = \langle 4.714\,9, 10.365 \rangle$ *i daje reprezentativnu informaciju o aritmetičkoj sredini. Prvi uzorak spada među* 1% *uzoraka za koje interval povjerenja razine pouzdanosti 99% neće sadržavati* μ.

Zadatak 5.9 *U slučajnom uzorku od* 64 *naloga izdana na terminalu neke banke zabilježeni su podaci o vremenu potrebnom za obradu tih naloga. Prosječno vrijeme toga uzorka je* $\overline{x} = 9.70906$ *minuta, a standardna devijacija uzorka je* $s =$

3.04569. *Odredite granice u kojima se može očekivati prosječno trajanje obrade naloga komitenata te banke? Razina pouzdanosti procjene neka je 95%.*

Rješenje: *Budući je* $n = 64 > 30$ *u pitanju je veliki uzorak. Budući je* N *nepoznat, a možemo pretpostaviti da je dovoljno velik tj. da uzorak čini manje od* 5% *svih naloga, to uzimamo* $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ *(izostavljamo faktor* $\sqrt{\frac{N-n}{N-1}}$*). No budući je standardna devijacija* σ *cijele populacije nepoznata, smijemo umjesto* σ *uzeti procjenitelja* $\hat{\sigma} = s\sqrt{\frac{n}{n-1}} = 3.04569 \cdot \sqrt{\frac{64}{63}} = 3.06977$ *(izostavljamo faktor* $\sqrt{\frac{N-1}{N}}$*). Vrijedi* $1 - \gamma = 0.95 \Rightarrow \gamma/2 = 0.025 \Rightarrow z_{0.025} = 1.96$*, stoga su granice traženog intervala* $9.70906 \pm 1.96 \cdot \frac{3.06977}{\sqrt{64}}$*, pa je interval* $\langle 8.957, 10.461 \rangle$ *.*

Ako je slučajni uzorak mali ($n \leq 30$) i ako je **izabran iz populacije čija je distribucija normalna**, ali s nepoznatim parametrima μ i σ, onda interval $\left\langle \overline{x} - t_{\gamma/2}\sigma_{\overline{x}}, \overline{x} + t_{\gamma/2}\sigma_{\overline{x}} \right\rangle$ sadrži μ s vjerojatnošću $1 - \gamma$. Broj $t_{\gamma/2}$ je vrijednost t-distribuirane varijable s $n - 1$ stupnjeva slobode koja ima svojstvo $p\left(t > t_{\gamma/2}\right) = \gamma/2$. Budući da je σ nepoznat, to u formuli za $\sigma_{\overline{x}}$ umjesto σ uvrštavamo procjenitelja $\hat{\sigma}$ izračunatog na uzorku.

Interval

$$\left\langle \overline{x} - t_{\gamma/2}\sigma_{\overline{x}}, \overline{x} + t_{\gamma/2}\sigma_{\overline{x}} \right\rangle$$

nazivamo **intervalom povjerenja aritmetičke sredine za male uzorke s razinom pouzdanosti** $1 - \gamma$.

Zadatak 5.10 *Iz evidencije od* 8967 *telefonskih razgovora ispituje se prosječno trajanje razgovara temeljem slučajnog uzorku o trajanju 10 razgovara:*
2,1,1,2,3,4,2,1,1,3. Ako pretpostavimo da je trajanje razgovora normalno distribuirano, odredite granice za koje se s pouzdanošću 95% može očekivati da obuhvaćaju prosječno trajanje razgovora.

Rješenje: *Zadano je* $n = 10$*, stupnjevi slobode*$= 9$*,* $1 - \gamma = 0.95$*,* $\gamma/2 = 0.025$*, pa je* $t_{0.025}(9) = 2.262$*.*

Slijedi $\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{2+1+\cdots+3}{10} = 2$ *min,*
$\hat{\sigma} = \sqrt{\frac{(x_1-2)^2+\cdots+(x_n-2)^2}{n-1}} = \sqrt{\frac{x_1^2+\cdots+x_n^2-n\overline{x}^2}{n-1}} = \frac{\sqrt{10}}{3}$,
$\sigma_{\overline{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{1}{3}$ *(budući je* $\frac{n}{N} < 0.05$ *smijemo izostaviti faktor* $\sqrt{\frac{N-n}{N-1}}$*).*
Traženi interval je $\left\langle 2 - 2.262 \cdot \frac{1}{3}, 2 + 2.262 \cdot \frac{1}{3} \right\rangle = \langle 1.246, 2.754 \rangle$ *.*

5.3.2 Procjena proporcije

Procjenitelj $\hat{p}$ proporcije p nekog obilježja populacije je proporcija uzorka $\hat{p} = \frac{m}{n}$, gdje je m broj članova uzorka s određenim obilježjem, a n veličina uzorka. Ako slučajni uzorak potječe iz beskonačnog skupa, onda je sampling distribucija proporcija $\hat{p}$ binomna distribucija s očekivanjem $E[\hat{p}] = p$ i standandardnom devijacijom

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot (1-p)}{n}}.$$

Ako je populacija konačna i veličine N onda je sampling distribucija proporcija $\hat{p}$ hipergeometrijska s očekivanjem p i standardnom devijacijom

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot (1-p)}{n} \left(\frac{N-n}{N-1}\right)}.$$

Budući da obje distribucije teže normalnoj distribuciji, u oba slučaja se kao praktično pravilo uzima da je sampling distribucija proporcija $\hat{p}$ aproksimativno normalno distribuirana $N(p, \sigma_{\hat{p}})$ ako je $np \geqslant 5$ i $n(1-p) \geqslant 5$. U tom slučaju, koristeći svojstva normalne distribucije, interval

$$\left\langle \hat{p} - z_{\gamma/2}\sigma_{\hat{p}}, \hat{p} + z_{\gamma/2}\sigma_{\hat{p}} \right\rangle$$

sadrži pravu proporciju p s vjerojatnošću $1-\gamma$. Taj interval nazivamo **intervalom povjerenja proporcije s razinom pouzdanosti** $1-\gamma$.

Budući da je u formuli za izračun granica intervala povjerenja proporcije potrebna proporcija p, a koja nije dostupna i koju procjenjujemo tim istim intervalom, smijemo izraz $\frac{p\cdot(1-p)}{n}$ zamijeniti izrazom $\frac{\hat{p}\cdot(1-\hat{p})}{n-1}$. Također, u slučaju konačne populacije i $\frac{n}{N} < 0.05$ faktor $\frac{N-n}{N-1}$ izostavljamo.
Nadalje, budući je interval povjerenja proporcije dobiven aproksimacijom binomne i hipergeometrijske razdiobe s normalnom distribucijom i aproksimacijom standardne devijacije $\sigma_{\hat{p}}$ pomoću gornjeg izraza, to je potrebna dodatna provjera je li postupak tj. aproksimacija zadovoljavajući. Rezultate možemo prihvatiti kao relevantne ako je gornja granica intervala strogo manja od 1.

Primjer 5.11 *Od* 6432 *osiguranika neke osiguravajuće kuće u uzorku od* 400 *osiguranika njih* 320 *nije sudjelovalo u prometnoj nezgodi u prethodnoj godini. Odredite*

interval povjerenja proporcije osiguranika koji su bili sudionici prometne nezgode u prošloj godini s razinom pouzdanosti $95\,\%$.

Zadano je $N = 6432$, $n = 400$, $m = 80$, $\hat{p} = \frac{m}{n} = 0.2$. *Nadalje* $1 - \gamma = 0.95 \Rightarrow \gamma/2 = 0.025 \Rightarrow z_{0.025} = 1.96$, $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n-1}\left(\frac{N-n}{N-1}\right)} = 0.01939$ *(budući je* $\frac{n}{N} = 0.062 > 0.05$ *faktor* $\frac{N-n}{N-1}$ *ne zanemarujemo). Interval pouzdanosti je* $\langle 0.2 - 1.96 \cdot 0.01939, 0.2 + 1.96 \cdot 0.01939\rangle = \langle 0.162, 0.238\rangle$, *pa s* $95\,\%$ *sigurnosti možemo reći da je između* $16.2\,\%$ *i* $23.8\,\%$ *osiguranika imalo prometnu nezgodu u prošloj godini.*

Možemo prihvatiti ove granice jer je gornja granica manja od 1 *i* $np \approx n\hat{p} = 80$ *i* $n\,(1-p) \approx n\,(1-\hat{p}) = 320$ *što je veće od* 5, *pa je implementacija ove procedure prihvatljiva.*

Primjer 5.12 *Na uzorku od* 576 *dobitnika lutrije u SAD-u u zadnjih* 10 *godina u iznosu preko* 100 000 *\$ samo je njih* 63 *dalo otkaz na poslu. Procijenite proporciju ljudi koji su dobitnici iznosa većega od* 100 000 *\$ na lutriji u SAD-u u zadnjih* 10 *godina, a koji su nakon dobitka dali otkaz i to s razinom pouzdanosti* $95\,\%$.

Zadano je $n = 576$, $m = 63$. *Nadalje, možemo izračunati* $\hat{p} = \frac{63}{576} = 0.1$, $\gamma/2 = 0.025 \Rightarrow z_{0.025} = 1.96$, $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n-1}} = 0.013$ *(ovdje moramo zanemariti faktor* $\frac{N-n}{N-1}$ *jer je* N *nepoznat, odnosno pretpostavljamo da je dovoljno velik). Tražena proporcija* p *se nalazi u intervalu*
$\langle 0.11 - 1.96 \cdot 0.013, 0.11 + 1.96 \cdot 0.013\rangle = \langle 0.084, 0.135\rangle$ *s* $95\,\%$ *vjerojatnosti, tj. sigurni smo s vjerojatnošću* $95\,\%$ *da je između* 8.4% *i* $13.5\,\%$ *ljudi napustilo posao nakon ovakvog dobitka.*

5.3.3 Procjena varijance

Ako je uzorak veličine n **uzet iz normalno distribuirane populacije** varijance σ^2, a $\hat{\sigma}^2$ je njezin nepristrani procjenitelj, onda su vrijednosti $\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$ distribuirane po χ^2-distribuciji s $n-1$ stupnjeva slobode. Tada, s vjerojatnošću $1-\gamma$ smijemo tvrditi da se $\frac{(n-1)\hat{\sigma}^2}{\sigma^2}$ nalazi u intervalu $\left\langle \chi^2_{(1-\gamma/2);n-1}, \chi^2_{\gamma/2;n-1} \right\rangle$, gdje su brojevi $\chi^2_{(1-\gamma/2);n-1}$ i $\chi^2_{\gamma/2;n-1}$ vrijednosti χ^2-distribuirane varijable s $n-1$ stupnjeva slobode koje imaju svojstvo $p\left(\chi^2 > \chi^2_{1-\gamma/2;n-1}\right) = 1 - \gamma/2$, odnosno

$p\left(\chi^2 > \chi^2_{\gamma/2;n-1}\right) = \gamma/2$. Tada se, s istom vjerojatnošću, vrijednost $\frac{\sigma^2}{(n-1)\hat{\sigma}^2}$ nalazi u intervalu $\left\langle \frac{1}{\chi^2_{\gamma/2;n-1}}, \frac{1}{\chi^2_{(1-\gamma/2);n-1}} \right\rangle$. Napokon, s vjerojatnošću $1-\gamma$ varijanca populacije σ^2 pripada intervalu $\left\langle \frac{(n-1)\hat{\sigma}^2}{\chi^2_{\gamma/2;n-1}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2_{(1-\gamma/2);n-1}} \right\rangle$. Ovaj interval nazivamo **intervalom povjerenja varijance s razinom pouzdanosti** $1-\gamma$. Interval povjerenja standardne devijacije σ je

$$\left\langle \frac{\hat{\sigma}\sqrt{(n-1)}}{\sqrt{\chi^2_{\gamma/2;n-1}}}, \frac{\hat{\sigma}\sqrt{(n-1)}}{\sqrt{\chi^2_{(1-\gamma/2);n-1}}} \right\rangle.$$

Zadatak 5.13 *Procijenite disperziju trajnosti žarulja određenog tipa s razinom pouzdanosti* 95% *ako je pomoću uzorka od* 21 *žarulje izračunata veličina* $\sum_{i=1}^{21} (x_i - \overline{x})^2 = 208080$ *sati.*

Rješenje: *Vrijedi* $\hat{\sigma}^2 = \frac{\sum_{i=1}^{21}(x_i-\overline{x})^2}{n-1} = \frac{208080}{20} = 10404 \Rightarrow \hat{\sigma} = 102$ *(faktor* $\frac{N-1}{N}$ *je izostavljen, tj. uzimamo da je* N *jako velik). Nadalje,* $1-\gamma = 0.95 \Rightarrow \gamma/2 = 0.025 \Rightarrow 1-\gamma/2 = 0.975$, *a budući su stupnjevi slobode jednaki* 20 *slijedi* $\chi^2_{0.025;20} = 34.1696$, $\chi^2_{0.975;20} = 9.59083$. *Traženi interval je* $\left\langle \frac{102\sqrt{20}}{\sqrt{34.1696}}, \frac{102\sqrt{20}}{\sqrt{9.59083}} \right\rangle = \langle 78.036, 147.294 \rangle$, *pa kažemo da se na razini pouzdanosti od* 95% *očekuje da je prosječno odstupanje trajnosti žarulje od njihove prosječne trajnosti između* 78 *i* 147 *sati, uz opravdanu pretpostavku da je distribucija trajnosti žarulja približno normalna.*

5.3.4 Procjena razlike sredina pomoću neovisnih uzoraka

Neka vrijednosti $x_{11}, x_{21}, ..., x_{n_1 1}$ tvore uzorak veličine n_1 iz populacije S_1 s aritmetičkom sredinom μ_1 i standardnom devijacijom σ_1, te neka vrijednosti $x_{12}, x_{22}, ..., x_{n_2 2}$ tvore uzorak veličine n_2 iz populacije S_2 s aritmetičkom sredinom μ_2 i standardnom devijacijom σ_2. Procjenitelj razlike $D = \mu_1 - \mu_2$ aritmetičkih sredina populacija je razlika $\hat{D} = \overline{x_1} - \overline{x_2}$ između aritmetičkih sredina $\overline{x_1}$ i $\overline{x_2}$ uzoraka. Ovaj procjenitelj je nepristran, tj. $E\left[\overline{x_1} - \overline{x_2}\right] = \mu_1 - \mu_2 = D$. Ako su uzorci, uzeti iz populacija, veliki i međusobno neovisni, onda je pripadna sampling

distribucija približno normalna, a standardna devijacija joj je

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Ako su uzorci uzeti iz populacija veliki i neovisni, onda interval

$$\left\langle \hat{D} - z_{\gamma/2}\sigma_{\overline{x_1}-\overline{x_2}}, \hat{D} + z_{\gamma/2}\sigma_{\overline{x_1}-\overline{x_2}} \right\rangle$$

nazivamo **intervalom povjerenja za razliku aritmetičkih sredina dviju populacija za velike i neovisne uzorke s razinom pouzdanosti** $1-\gamma$, i on sadrži razliku $D = \mu_1 - \mu_2$ s vjerojatnošću $1-\gamma$. U formuli za granice ovog intervala smijemo zamijeniti (najčešće nepoznate) varijance populacija σ_1^2 i σ_2^2 s njihovim procjeniteljima $\hat{\sigma}_1^2$ i $\hat{\sigma}_2^2$.
Ako se razlika sredina procjenjuje pomoću malih uzoraka izabranih iz **normalno distribuiranih populacija** različitih sredina i **jednakih, poznatih, varijanci** $\sigma_1 = \sigma_2 = \sigma$, onda je odgovarajući interval povjerenja

$$\left\langle \hat{D} - t_{\gamma/2}\sigma_{\overline{x_1}-\overline{x_2}}, \hat{D} + t_{\gamma/2}\sigma_{\overline{x_1}-\overline{x_2}} \right\rangle,$$

gdje je $\sigma_{\overline{x_1}-\overline{x_2}} = \sigma\sqrt{\frac{n_1+n_2}{n_1 n_2}}$, a broj $t_{\gamma/2}$ je odgovarajuća vrijednost t-distribuirane varijable s $n_1 + n_2 - 2$ stupnjeva slobode.
Ako je zajednička varijanca σ nepoznata, onda se koristi izraz

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{(n_1-1)\,\hat{\sigma}_1^2 + (n_2-1)\,\hat{\sigma}_2^2}{n_1+n_2-2}\left(\frac{n_1+n_2}{n_1 n_2}\right)}.$$

Ako se prethodno utvrdi da su varijance populacija različite onda se koristi izraz $\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$, a broj stupnjeva slobode kojeg koristimo za izračun vrijednosti $t_{\gamma/2}$ je jednak

$$s.s. = \text{Int}\left[\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2 \cdot \left(\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2-1}\right)\right],$$

gdje Int označuje najveće cijelo decimalnog broja (broju se odbacuju decimale).

Zadatak 5.14 *Aritmetička sredina brzine otpremanja stranke službenice A u uzorku od* 40 *stranki je* 503 *sekunde s prosječnim odstupanjem* 145*. Aritmetička sredina brzine otpremanja stranke službenice B u uzorku od* 52 *stranke je* 407 *sekundi s prosječnim odstupanjem* 132*. Procijenite razliku prosječnih vremena potrebnih za otpremanje stranke kod službenice A i službenice B s razinom pouzdanosti* 94 %*.*

Rješenje: *Vrijedi* $\overline{x_1} = 503$, $s_1 = 145$, $n_1 = 40$, $\overline{x_2} = 407$, $s_2 = 132$, $n_2 = 52$, $\hat{D} = \overline{x_1} - \overline{x_2} = 96$, $1 - \gamma = 0.94 \Rightarrow \gamma/2 = 0.03 \Rightarrow z_{0.03} = 1.88$, $\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{s_1^2\left(\frac{n_1}{n_1-1}\right)}{n_1} + \frac{s_2^2\left(\frac{n_2}{n_2-1}\right)}{n_2}} = 29.67743$. *Stoga je traženi interval*

$$\langle 96 - 1.88 \cdot 29.67743, 96 + 1.88 \cdot 29.67743\rangle = \langle 40.206, 151.793\rangle .$$

Možemo reći da je razlika prosječnih vremena potrebnih za otpremanje stranke kod službenice A i službenice B s razinom pouzdanosti 94 % *između* 40 *i* 151 *sekundu.*

5.3.5 Procjena razlike sredina pomoću ovisnih (uparenih) uzoraka

Procjenjivanje razlike aritmetičkih sredina dviju populacija zavisnim uzorcima obično se provodi pomoću razlike vrijednosti parova, pogotovo ako se populacije sastoje od istih članova s dva različita numerička obilježja. Primjerice, možemo promatrati populaciju ljudi koji su počeli primjenjivati dijetu, a istoj osobi bilježimo težinu prije i poslije dijete, ili promatrati uspjehe učenika istog razreda kod dva različita nastavnika, ili proizvodnost populacije radnika prije i poslije stručnog usavršavanja. Kod ovakvih primjera zapravo radimo s istom populacijom $\{s_1, ..., s_N\}$ na kojoj operiraju varijable X_1 i X_2. No, onda se može formirati jedinstvena varijabla razlika $d = X_1 - X_2$ koja djeluje na istoj populaciji. Ta varijabla ima očekivanje $\mu_d = E[d] = E[X_1 - X_2] = \mu_1 - \mu_2$ i standardnu devijaciju σ_d. Uzorak $\{d_{i_1}, ..., d_{i_n}\}$ veličine n, kojeg uzimamo iz ovakve populacije razlika, je zapravo par međusobno ovisnih uzoraka $\{x_{i_1}, ..., x_{i_n}\}$ i $\{y_{i_1}, ..., y_{i_n}\}$ jednake veličine vrijednosti kojih promatramo u parovima, odnosno za i-ti član promatramo par (x_i, y_i) čija je razlika $d_i = x_i - y_i$. Varijablu koja svakom takvom uzorku $\{d_1, ..., d_n\}$ razlike parova pridruži aritmetičku sredinu $\overline{d} = \frac{d_1+\cdots+d_n}{n}$ je nepristrani procjenitelj

aritmetičke sredine μ_d, pa onda i razlike aritmetičkih sredina prve i druge populacije $\mu_1 - \mu_2 = \mu_d$, tj. vrijedi $E\left[\overline{d}\right] = \mu_1 - \mu_2$. Za velike uzorke sampling distribucija varijable aritmetičkih sredina razlika uzoraka $\overline{d}$ je približno normalna $N\left(\mu_d, \sigma_{\overline{d}}\right)$, gdje je $\sigma_{\overline{d}} = \frac{\sigma_d}{\sqrt{n}}$. Tada je vjerojatnost da razlika aritmetičkih sredina $\mu_d = \mu_1 - \mu_2$ populacija (isto što aritmetička sredina razlika) bude u intervalu $\left\langle \overline{d} - z_{\gamma/2}\sigma_{\overline{d}}, \overline{x} + z_{\gamma/2}\sigma_{\overline{d}} \right\rangle$ jednaka $1 - \gamma$. Interval

$$\left\langle \overline{d} - z_{\gamma/2}\sigma_{\overline{d}}, \overline{d} + z_{\gamma/2}\sigma_{\overline{d}} \right\rangle$$

nazivamo **intervalom povjerenja razlike aritmetičke sredine za velike uzorke na temelju ovisnih (uparenih) uzoraka s razinom pouzdanosti** $1 - \gamma$. U izrazu za granice intervala u $\sigma_{\overline{d}}$ smijemo umjesto (obično nepoznate) standardne devijacije varijable razlika σ_d uvrštavati njezin procjenitelj $\hat{\sigma}_d = s_d\sqrt{\frac{n}{n-1}}$, gdje je s_d standardna devijacija uzorka razlika $\{d_1, ..., d_n\}$, tj.

$$s_d = \sqrt{\frac{\left(d_1 - \overline{d}\right)^2 + \cdots + \left(d_n - \overline{d}\right)^2}{n}}.$$

Ako je uzorak mali, a **populacije normalno distribuirane**, onda u izrazima za granice intervala umjesto $z_{\gamma/2}$ treba staviti vrijednost $t_{\gamma/2}$ s $n - 1$ stupnjeva slobode.

Zadatak 5.15 *Zadana je tablica s težinama* 5 *proizvoljno odabranih ljudi koji su se podvrgli određenoj dijeti i to neposredno prije početka dijete i mjesec dana kasnije. Napravite interval povjerenja s* 95% *razinom pouzdanosti za razliku između prosječne težine svih ljudi koji su prihvatili program prije početka dijete i mjesec dana nakon (isto što i prosječna razlika u težinama prije i poslije). Koje pretpostavke trebaju biti ispunjene da bismo mogli prihvatiti te rezultate?*

Osoba	*Težina prije*	*Težina poslije*
A	150	143
B	195	190
C	188	185
D	197	191
E	204	200

Rješenje: *Formirajmo tablicu razlika*

Osoba	*Razlike d*
A	7
B	5
C	3
D	6
E	4

. *Vrijedi:* $n = 5$, $\overline{d} = \frac{7+\cdots+4}{5} = 5$, $\hat{\sigma}_d = s_d\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\left(d_1-\overline{d}\right)^2+\cdots+\left(d_n-\overline{d}\right)^2}{n-1}} = \sqrt{\frac{(7-5)^2+\cdots+(4-5)^2}{4}} = 1.58$, $1-\gamma = 0.95 \Rightarrow \gamma/2 = 0.025$, *stupnjevi slobode*$= 4 \Rightarrow t_{0.025} = 2.776$. *Traženi interval je* $\left\langle 5 - 2.776\frac{1.58}{\sqrt{5}}, 5 + 2.776\frac{1.58}{\sqrt{5}} \right\rangle = \langle 3.04, 6.96 \rangle$, *koji možemo prihvatiti jedino uz (opravdanu) pretpostavku da su težine svih ljudi u programu prije dijete i poslije normalno distribuirane.*

5.3.6 Procjena razlike proporcija

Neka su n_1 i n_2 veličine uzoraka uzetih iz populacija kojima su p_1 i p_2 proporcije nekog obilježja redom. Ako u prvom uzorku ima m_1 elemenata s promatranim obilježjem, a u drugom m_2, onda su $\hat{p}_1 = \frac{m_1}{n_1}$ i $\hat{p}_2 = \frac{m_2}{n_2}$ proporcije uzoraka s promatranim obilježjem redom. Razlika $\hat{p}_1 - \hat{p}_2$ je nepristrani procjenitelj prave razlike populacijskih proporcija $p_1 - p_2$. Nadalje, ako su **uzorci dovoljno veliki** (kao kod procjene proporcije) sampling distribucija razlika proporcija $\hat{p}_1 - \hat{p}_2$ je približno normalna $N\left(p_1 - p_2, \sigma_{\hat{p}_1-\hat{p}_2}\right)$, gdje je

$$\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\frac{\left(n_1\hat{p}_1 + n_2\hat{p}_2\right)\left(n_1\left(1-\hat{p}_1\right) + n_2\left(1-\hat{p}_2\right)\right)}{\left(n_1+n_2\right)n_1n_2}}.$$

Tada je vjerojatnost da razlika populacijskih proprcija $p_1 - p_2$ bude u intervalu

$$\left\langle \hat{p}_1 - \hat{p}_2 - z_{\gamma/2}\sigma_{\hat{p}_1-\hat{p}_2}, \hat{p}_1 - \hat{p}_2 + z_{\gamma/2}\sigma_{\hat{p}_1-\hat{p}_2} \right\rangle$$

jednaka $1-\gamma$.
Interval $\left\langle \hat{p}_1 - \hat{p}_2 - z_{\gamma/2}\sigma_{\hat{p}_1-\hat{p}_2}, \hat{p}_1 - \hat{p}_2 + z_{\gamma/2}\sigma_{\hat{p}_1-\hat{p}_2} \right\rangle$ nazivamo **intervalom povjerenja razlike populacijskih proporcija s razinom pouzdanosti** $1-\gamma$.

Primjer 5.16 *Od* 100 *splitskih domaćinstava odabranih u uzorak,* 50 *je na bar jednoj televiziji pratilo svečanost povodom* 100 *godišnjice Hajduka, dok je od* 200

zagrebačkih domaćinstava odabranih u uzorak njih 75 *pratilo na TV-u isti događaj. Procijenite* 95% *intervalom razliku proporcija.*

Vrijedi: $n_1 = 100$, $m_1 = 50$, $\hat{p}_1 = \frac{m_1}{n_1} = 0.5$, $n_2 = 200$, $m_2 = 75$, $\hat{p}_2 = \frac{75}{200} = 0.375$, $\hat{p}_1 - \hat{p}_2 = 0.125$, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{(n_1\hat{p}_1+n_2\hat{p}_2)(n_1(1-\hat{p}_1)+n_2(1-\hat{p}_2))}{(n_1+n_2)n_1n_2}} = 0.06038$, $z_{\gamma/2} = z_{0.025}$. *Traženi interval je* $\langle 0.125 - 1.96 \cdot 0.06038, 0.125 + 1.96 \cdot 0.06038\rangle = \langle 0.0066, 0.243\rangle$. *Možemo zaključiti, s* 95% *sigurnošću, da je razlika gledanosti TV programa vezanog uz proslavu* 100 *godišnjice Hajduka između splitskih i zagrebačkih domaćinstava između* 0.6% *i* 24%.

5.4 Određivanje veličine uzorka za procjenu parametra

Veličina slučajnog uzorka uzetog iz populacije za procjenu parametra θ ovisi, između ostalog, o željenoj preciznosti procjene i razini pouzdanosti procjene. Naravno, da veći uzorak implicira i veću preciznost i veću pouzdanost, no često puta i povećava troškove istraživanja. Stoga je uputno na početku istraživanja, prije uzimanja uzorka, iskazati željenu preciznost i razinu pouzdanosti, te temeljem toga izračunati potrebnu veličinu uzorka da bi se ostvarili postavljeni ciljevi. Ako je interval povjerenja do kojeg želimo doći simetričan $\left\langle \hat{\theta} - d, \hat{\theta} + d \right\rangle$, onda brojem $2d$ iskazujemo preciznost procjene, tj. najveću dopuštenu grešku d (između $\hat{\theta}$ i θ) u apsolutnom iznosu ili u relativnom iznosu $\frac{d}{\hat{\theta}}$ (u jedinicama procijenjenog parametra $\hat{\theta}$). Veličina uzorka n se izračunava iz formula za granice intervala povjerenja i to za zadani d i 1-γ. Ako želimo načiniti interval povjerenja $\langle \overline{x} - d, \overline{x} + d\rangle$ za procjenu sredine μ populacije s razinom pouzdanosti $1 - \gamma$, onda, za beskonačne populacije, iz $d = z_{\gamma/2}\frac{\sigma}{\sqrt{n}}$ slijedi $n = \left(\frac{z_{\gamma/2}\sigma}{d}\right)^2$.

U postupku određivanja veličine uzorka n, treba nam, osim preciznosti d i pouzdanosti $1 - \gamma$, i standardna devijacija populacije σ koja je redovito nepoznata. No, σ se kao planska veličina prosuđuje pomoću pilot istraživanja ili se uzima kao iskustveno pravilo da je raspon cijele populacije približno 6σ (Čebišev teorem). Ako se pogreška izražava relativno (koliki postatak vrijednosti sredine toleriramo

za grešku) onda je $n = \left(\frac{z_{\gamma/2}\frac{\sigma}{\mu}}{d}\right)^2$. Ako je populacija konačna, onda se n nađe iz jednadžbe $d = z_{\gamma/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, i to samo u slučaju kada je $\frac{n_0}{N} > 0.05$, gdje je $n_0 = \left(\frac{z_{\gamma/2}\frac{\sigma}{\mu}}{d}\right)^2$. U protivnom je $n = n_0$.

Zadatak 5.17 *Restoran s dostavom želi ispitati srednje vrijeme dostave koje traje od narudžbe do uručivanja na adresu. Ako menadžment restorana želi procjenu s toleriranom greškom od 5 minuta i razinom pouzdanosti od 95% koliku veličinu uzorka moraju promatrati ako se zna da je približno najbrža dostava bila 10 minuta, a najsporija 100.*

Rješenje: *Vrijedi:* $d = 5$, $z_{\gamma/2} = z_{0.025} = 1.96$, $6\sigma \approx 100 - 10 = 90 \Rightarrow \sigma \approx \frac{90}{6} = 15$. *Stoga je* $n = \left(\frac{z_{\gamma/2}\sigma}{d}\right)^2 = \left(\frac{15\cdot 1.96}{5}\right)^2 = 34.574$. *Potrebno je, dakle, uzeti u uzorak (barem) 35 dostava da bi zaključak bio tražene preciznosti i pouzdanosti.*

Ako želimo načiniti interval povjerenja $\langle \hat{p} - d, \hat{p} + d\rangle$ za procjenu proporcije p nekog obilježja populacije s razinom pouzdanosti, onda, za beskonačne populacije, iz $d = z_{\gamma/2}\sqrt{\frac{p(1-p)}{n}}$ slijedi

$$n = \left(\frac{z_{\gamma/2}}{d}\right)^2 p\,(1-p)\,.$$

U postupku određivanja veličine uzorka n, treba nam, osim preciznosti d i pouzdanosti $1-\gamma$, i proporcija p koju procjenjujemo. Za nju uzimamo plansku, približnu veličinu. No, ako nema pouzdane osnove za plansku veličinu, onda se uzima najnepovoljniji slučaj, tj. kada je $p\,(1-p)$ maksimalan, a to će biti za $p = 0.5$. Ako je populacija konačna, onda se n nađe iz jednadžbe

$$d = z_{\gamma/2}\sqrt{\frac{p\,(1-p)}{n}\left(\frac{N-n}{N-1}\right)},$$

i to samo u slučaju kada je $\frac{n_0}{N} > 0.05$, gdje je $n_0 = \left(\frac{z_{\gamma/2}}{d}\right)^2 p\,(1-p)$. U protivnom je $n = n_0$.

Na sličan način se određuju veličine n_1 i n_2 uzoraka koje treba uzeti iz dviju populacija za procjenu razlike sredina $\mu_1 - \mu_2$ i razlike proporcija $p_1 - p_2$. U izračunu se uzima da je $n_1 = n_2$.

Primjer 5.18 *Koliko dobitnika lutrije u iznosu od preko* $100\,000$ *\$ u zadnjih* 10 *godina treba uključiti u istraživanje o postotku onih dobitnika koji su napustili posao nakon takvog dobitka, ako želimo preciznost od* 0.02 *i razinu pouzdanosti* $90\,\%$*.*

Vrijedi: $2d = 0.02 \Rightarrow d = 0.01$, $1 - \gamma = 0.9 \Rightarrow z_{\gamma/2} = z_{0.05} = 1.645$*. Za* p *možemo uzeti* 0.11 *proporciju iz Primjera 5.12 kojega možemo tretirati kao pilot uzorak od* 576 *članova (u slučaju da nemamo ovaj podatak od prije morali bismo staviti* $p = 0.5$*). Sada je* $n = \left(\frac{z_{\gamma/2}}{d}\right)^2 p\,(1-p) = \left(\frac{1.645}{0.01}\right)^2 0.11 \cdot 0.89 = 2649.2$*. Zaključujemo da je potrebno promatrati uzorak od barem* 2650 *dobitnika.*

Poglavlje 6

Testiranje hipoteza

6.1 Testiranje hipoteza o parametru

Statistička hipoteza je tvrdnja o veličini parametra θ ili o obliku distribucije populacije čija se vjerodostojnost ispituje pomoću slučajnog uzorka. Postupak kojim se donosi odluka o prihvaćanju ili neprihvaćanju hipoteze temeljem podataka iz uzorka se naziva **testiranjem statističkih hipoteza**. Statistički testovi se dijele na parametarske i neparametarske. Testiranje polazi od formiranja **nulte hipoteze** H_0 i **alternativne hipoteze** H_1 koja je komplementarna nultoj hipotezi. Moguće odluke su prihvaćanje nulte hipoteze i odbacivanje nulte hipoteze (što je ekvivalentno prihvaćanju alternativne). Budući se odluka donosi na temelju podataka iz uzorka, u postupku testiranja su moguće dvije vrste pogreške.

	H_0 je istinita	H_0 je lažna
H_0 je prihvaćena (tj. nije odbačena)	ispravno	pogreška tipa II.
H_0 je odbačena	pogreška tipa I.	ispravno

Pogreška tipa I. se iskazuje vjerojatnošću α odbacivanja istinite nulte hipoteze. Još se naziva **razinom značajnosti ili signifikantnosti** (ili razinom rizika). Pogreška tipa II. se iskazuje vjerojatnošću β, a vjerojatnost $1-\beta$ se naziva **snagom statističkog testa** i označava vjerojatnost odbacivanja neistinite H_0 hipoteze.

U slučaju prihvaćanja hipoteze H_0 kažemo da je opravdano prihvatiti kao vjerojatno istinitu tvrdnju iz hipoteze (ili još bolje da je nije opravdano odbaciti) s razinom signifikantnosti α, odnosno u slučaju odbacivanja kažemo da je na danoj razini signifikantnosti α opravdano odbaciti tvrdnju iz hipoteze kao vjerojatno neistinitu.

U slučaju parametarskog testa, u kojem se testira vrijednost parametra θ pomoću vrijednosti procjenitelja parametra $\hat{\theta}$, razlikujemo 3 vrste testa. U dvosmjernom testu pretpostavljamo da je vrijednost parametra jednaka unaprijed fiksiranoj vrijednosti θ_0, tj. hipoteza H_0 glasi $\theta = \theta_0$. U jednosmjernim testovima pretpostavljamo da je $\theta \geqslant \theta_0$, odnosno $\theta \leqslant \theta_0$.

6.1.1 Z i t test

Ako se za zadanu razinu pouzdanosti $1 - \alpha$ može pomoću vrijednosti $\hat{\theta}$ načiniti interval povjerenja $\left\langle \hat{\theta} - \sigma_{\hat{\theta}} z_{\alpha/2}, \hat{\theta} + \sigma_{\hat{\theta}} z_{\alpha/2} \right\rangle$ ili $\left\langle \hat{\theta} - \sigma_{\hat{\theta}} t_{\alpha/2}, \hat{\theta} + \sigma_{\hat{\theta}} t_{\alpha/2} \right\rangle$, onda ako je $\theta = \theta_0$ slijedi da je vjerojatnost da vrijednost $\hat{\theta}$ na uzorku pripada intervalu $\left\langle \theta_0 - \sigma_{\hat{\theta}} z_{\alpha/2}, \theta_0 + \sigma_{\hat{\theta}} z_{\alpha/2} \right\rangle$, odnosno $\left\langle \theta_0 - \sigma_{\hat{\theta}} t_{\alpha/2}, \theta_0 + \sigma_{\hat{\theta}} t_{\alpha/2} \right\rangle$, jednaka $1-\alpha$. Stoga, ako je $\hat{\theta} \in \left\langle \theta_0 - \sigma_{\hat{\theta}} z_{\alpha/2}, \theta_0 + \sigma_{\hat{\theta}} z_{\alpha/2} \right\rangle$, odnosno $\hat{\theta} \in \left\langle \theta_0 - \sigma_{\hat{\theta}} t_{\alpha/2}, \theta_0 + \sigma_{\hat{\theta}} t_{\alpha/2} \right\rangle$, onda prihvaćamo hipotezu $\theta = \theta_0$. Takvu odluku donosimo ako je $-z_{\alpha/2} < z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} < z_{\alpha/2}$, odnosno $t_{\alpha/2} < t = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} < t_{\alpha/2}$, gdje z, odnosno t, nazivamo testnom veličinom. *Tada kažemo da se na razini signifikantnosti α parametar θ statistički značajno ne razlikuje od vrijednosti θ_0 ili da ne možemo odbaciti hipotezu $\theta = \theta_0$.*
Ako je $z \leqslant -z_{\alpha/2}$ ili $z \geqslant z_{\alpha/2}$, odnosno $t \leqslant -t_{\alpha/2}$ ili $t \geqslant t_{\alpha/2}$, onda odbacujemo hipotezu $\theta = \theta_0$, i kažemo da *se na razini signifikantnosti α parametar θ statistički značajno razlikuje od vrijednosti θ_0.* Vjerojatnost α iskazuje rizik da ipak bude $\theta = \theta_0$ istinito ali je procjenitelj $\hat{\theta}$ među onih $\alpha 100\%$ procjenitelja koji ne upadaju u odgovarajući interval povjerenja oko θ_0.
Vjerojatnost da vrijednost $\hat{\theta}$ na uzorku pripada intervalu $\langle -\infty, \theta_0 + \sigma_{\hat{\theta}} z_\alpha \rangle$, odnosno $\langle -\infty, \theta_0 + \sigma_{\hat{\theta}} t_\alpha \rangle$, jednaka je $1 - \alpha$. Stoga, ako je $\hat{\theta} \in \langle -\infty, \theta_0 + \sigma_{\hat{\theta}} z_\alpha \rangle$, odnosno $\hat{\theta} \in \langle -\infty, \theta_0 + \sigma_{\hat{\theta}} t_\alpha \rangle$, onda prihvaćamo hipotezu $\theta \leq \theta_0$. Takvu odluku donosimo ako je $z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} < z_\alpha$, odnosno $t = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} < t_\alpha$. *Tada kažemo da je na razini signifikantnosti α parametar θ statistički značajno manji ili se ne razlikuje od vri-*

jednosti θ_0 ili da ne možemo odbaciti hipotezu $\theta \leq \theta_0$.
Ako je $z \geqslant z_\alpha$, odnosno $t \geqslant t_\alpha$, onda odbacujemo hipotezu $\theta \leq \theta_0$, i kažemo da *je na razini signifikantnosti α parametar θ statistički značajno veći od vrijednosti θ_0.* Vjerojatnost α iskazuje rizik da ipak bude $\theta \leq \theta_0$ istinito ali je procjenitelj $\hat{\theta}$ među onih $\alpha 100\%$ procjenitelja koji ne upadaju u odgovarajući interval $\langle -\infty, \theta_0 + \sigma_{\hat{\theta}} z_\alpha \rangle$, odnosno $\langle -\infty, \theta_0 + \sigma_{\hat{\theta}} t_\alpha \rangle$.
Vjerojatnost da vrijednost $\hat{\theta}$ na uzorku pripada intervalu $\langle \theta_0 - \sigma_{\hat{\theta}} z_\alpha, \infty \rangle$, odnosno $\langle \theta_0 - \sigma_{\hat{\theta}} t_\alpha, \infty \rangle$, jednaka je $1 - \alpha$. Stoga, ako je $\hat{\theta} \in \langle \theta_0 - \sigma_{\hat{\theta}} z_\alpha, \infty \rangle$, odnosno $\hat{\theta} \in \langle \theta_0 - \sigma_{\hat{\theta}} t_\alpha, \infty \rangle$, onda prihvaćamo hipotezu $\theta \geqslant \theta_0$. Takvu odluku donosimo ako je $z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > -z_\alpha$, odnosno $t = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} > -t_\alpha$. *Tada kažemo da je na razini signifikantnosti α parametar θ statistički značajno veći ili se ne razlikuje od vrijednosti θ_0 ili da ne možemo odbaciti hipotezu $\theta \geq \theta_0$.*
Ako je $z \leqslant -z_\alpha$, odnosno $t \leqslant -t_\alpha$, onda odbacujemo hipotezu $\theta \geqslant \theta_0$, i kažemo da *je razini signifikantnosti α parametar θ statistički značajno manji od vrijednosti θ_0.* Vjerojatnost α iskazuje rizik da ipak bude $\theta \geqslant \theta_0$ istinito ali je procjenitelj $\hat{\theta}$ među onih $\alpha 100\%$ procjenitelja koji ne upadaju u odgovarajući interval $\langle \theta_0 - \sigma_{\hat{\theta}} z_\alpha, \infty \rangle$, odnosno $\langle \theta_0 - \sigma_{\hat{\theta}} t_\alpha, \infty \rangle$.

Ako se testiranje na razini značajnosti α izvodi pomoću normalne sampling distribucije, onda govorimo o z-testu i postupamo kao u sljedećoj tablici.

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hip.	Područje odbacivanja H_0
$H_0 ... \theta = \theta_0$	$H_1 ... \theta \neq \theta_0$	$-z_{\alpha/2} < z < z_{\alpha/2}$	$z \leq -z_{\alpha/2}$ ili $z \geqslant z_{\alpha/2}$
$H_0 ... \theta \leq \theta_0$	$H_1 ... \theta > \theta_0$	$z < z_\alpha$	$z \geqslant z_\alpha$
$H_0 ... \theta \geqslant \theta_0$	$H_1 ... \theta < \theta_0$	$z > -z_\alpha$	$z \leq z_\alpha$

Ako se testiranje na razini značajnosti α izvodi pomoću Studentove sampling distribucije (s odgovarajućim stupnjevima slobode), onda govorimo o t-testu i postupamo kao u sljedećoj tablici.

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hip.	Područje odbacivanja H_0
$H_0...\theta = \theta_0$	$H_1...\theta \neq \theta_0$	$-t_{\alpha/2} < t < t_{\alpha/2}$	$t \leq -t_{\alpha/2}$ ili $t \geqslant t_{\alpha/2}$
$H_0...\theta \leq \theta_0$	$H_1...\theta > \theta_0$	$t < t_\alpha$	$t \geqslant t_\alpha$
$H_0...\theta \geqslant \theta_0$	$H_1...\theta < \theta_0$	$t > -t_\alpha$	$t \leq t_\alpha$

Primjer 6.1 *Radi povećanja prometa lanac trgovina razmišlja o uvođenju mogućnosti plaćanja karticom ako prosječni mjesečni promet bude veći od* 300 000 €. *Uvedeno je pokusno plaćanje u* 15 *trgovina (manje od* 5 % *ukupnog broja trgovina). Ako je prosječni mjesečni promet u uzorku bio* 317 543 €, *a prosječno odstupanje* 4768, *kakvu odluku treba donijeti s razinom signifikantnosti* 5 %? *Pretpostavlja se da je mjesečni promet po prodavaonicama normalno distribuiran.*

Treba testirati aritmetičku sredinu μ *u odnosu na* $\mu_0 = 300\,000$. $H_0...\mu \leq 300\,000$, $H_1...\mu > 300\,000$, *pri čemu je zadano* $\overline{x} = 317\,543$, $s = 4768$, $n = 15$. *Budući je uzorak mali, to provodimo t-test:* $t = \frac{\overline{x}-\mu_0}{\sigma_{\overline{x}}} = \frac{\overline{x}-\mu_0}{\frac{s\sqrt{\frac{n}{n-1}}}{\sqrt{n}}} = 13.767$, *stupnjevi slobode su* 14, *a* $t_\alpha = t_{0.05} = 1.761$ *povlači* $t > t_\alpha$. *Rezultat testa upućuje da je, na razini signifikantnosti* 5 %, *opravdano zaključiti da je prosječan mjesečni promet nakon uvođenja kartičnog plaćanja statistički značajno veći od* 300 000 €.

Zadatak 6.2 *Direktor farme pilića razmišlja o uvođenju novog prehrambenog sredstva za piliće koje bi se jedino isplatilo ako bi tim sredstvom hranjeni pilići bili barem* 500 *g teži. U kontrolnom uzorku od* 400 *pilića hranjenih standardnom hranom prosječna težina jednaka je* 2350 *g, a prosječno odstupanje je* 200 *g. U eksperimentalnom uzorku od* 361 *pilića prosječna težina je* 3040 *g, a prosječno odstupanje je* 220 *g. Odluku treba donijeti s* 1 % *značajnosti.*

Rješenje: *Treba testirati razliku aritmetičkih sredina* $\mu_2 - \mu_1$ *težina populacije pilića hranjenih standardno i hranjenih novim sredstvom, tj.* $H_0...\mu_1 - \mu_2 \geqslant 500$, $H_1...\mu_2 - \mu_1 < 500$, *gdje je zadano* $\overline{x_1} = 2350$, $s_1 = 200$, $n_1 = 400$, $\overline{x_2} = 3040$, $s_2 = 220$, $n_2 = 361$. *Budući su uzorci veliki provodimo z-test:* $z = \frac{\overline{x_2}-\overline{x_1}-500}{\sigma_{\overline{x}_1-\overline{x}_2}} = \frac{\overline{x_2}-\overline{x_1}-500}{\sqrt{\frac{s_1^2\frac{n_1}{n_1-1}}{n_1}+\frac{s_2^2\frac{n_2}{n_2-1}}{n_2}}} = \frac{690}{\sqrt{100.25+134.44}} = 2.940\,0$, *a* $z_\alpha = z_{0.01} = 2.33$ *implicira* $z > -z_\alpha$.

Rezultat testa upućuje da je, na razini signifikantnosti 1%*, opravdano zaključiti da je razlika prosječnih težina pilića hranjenih novim sredstvom i starim statistički značajno veća ili jednaka* $500g$*.*

Zadatak 6.3 *U uzorku od* 400 *birača prve županije njih* 54 *% se izjasnilo za stranku HAHA, a od* 625 *birača druge županije njih* 48% *se izjasnilo za tu istu stranku. Na razini značajnosti* 0.08 *testirajte hipotezu da ne postoji razlika u raspoloženju birača prema stranki HAHA.*

Rješenje: *Treba testirati razliku proporcija* $p_1 - p_2$*.* $H_0 ... p_1 - p_2 = 0$*,* $H_1 ... p_1 - p_2 \neq 0$*, pri čemu je poznato* $\hat{p}_1 = 0.54$*,* $n_1 = 400$*,* $\hat{p}_2 = 0.48$*,* $n_2 = 0.48$*. Iz* $z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{0.54 - 0.48}{\sqrt{\frac{(n_1\hat{p}_1 + n_2\hat{p}_2)(n_1(1-\hat{p}_1) + n_2(1-\hat{p}_2))}{(n_1+n_2)n_1n_2}}} = 1.87$ *i* $z_{\alpha/2} = z_{0.04} = 1.75$ *proizlazi* $1.75 < z$*, pa rezultat testa upućuje da se na danoj razini signifikantnosti odbaci hipoteza* H_0*, tj. postoji statistički značajna razlika u podršci birača stranci HAHA u dvije promatrane županije.*

Zadatak 6.4 *Na referendum je izašlo* 1 000 000 *birača. Do 22.00 sata prebrojeno je* 900 000 *glasačih listića od kojih je* 55% *zaokružilo odgovor NE na referendumsko pitanje. Testirajte hipotezu da će nakon prebrojavanja svih listića konačna odluka građana na referendumsko pitanje biti negativna i to s* 99% *sigurnosti. Može li ova hipoteza biti prihvaćena s* 100% *sigurnosti?*

Rješenje: *Potrebno je testirati proporciju* p *birača koji su se negativno izrazili na referendumsko pitanje, tj.* $H_0 ... p \leq 0.5$*,* $H_1 ... p > 0.5$*. Zadano je* $\alpha = 0.01$*,* $N = 1\,000\,000$*,* $n = 900\,000$ *i* $\hat{p} = 0.55$*. Iz* $z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{\hat{p} - 0.5}{\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}\left(\frac{N-n}{N-1}\right)}} = 301.51$ *i* $z_{0.01} = 2.33$ *slijedi* $z > z_\alpha$ *što nas upućuje na prihvaćanje alternativne hipoteze* $p > 0.05$ *s vjerojatnošću* 0.01 *da je nulta hipoteza* $p \leq 0.5$ *istinita. Ipak, teorijski ne možemo biti u potpunosti sigurni u ishod referenduma temeljem prebrojanih listića. Naime, ako bi svih preostalih* 100 000 *listića bilo s pozitivnim odgovorom onda bi proporcija negativnih odgovora bila* $p = \frac{m}{N} = \frac{m}{N}\frac{n}{n} = \frac{m}{n}\frac{n}{N} = \hat{p}\frac{n}{N}$ *gdje je* n *veličina uzorka i* m *broj glasačkih listića u uzorku s odgovorom NE. Tada bi slijedilo* $p = 0.55 \cdot 0.9 = 0.495$*, odnosno konačni odgovor na referendumsko pitanje bi bio pozitivan. Da bismo temeljem uzorka veličine* n *s proporcijom* $\hat{p} > 0.5$ *(ili* $\hat{p} < 0.5$*) bili sigurni u ishod referenduma treba biti* $\hat{p}\frac{n}{N} > 0.5$ *(ili* $\hat{p}\frac{n}{N} < 0.5$*).*

6.1.2 Snaga testa

Vjerojatnost β označuje vjerojatnost pogreške tipa II. tj. prihvaćanja lažne nulte hipoteze. Ako je prava vrijednost parametra $\theta = \theta_1$, onda se za svaku vrstu testa može izračunati β i snaga testa $1-\beta$ (vjerojatnost odbacivanja lažne nulte hipoteze) u ovisnosti o θ_1. Što je prava vrijednost θ_1 bliža pretpostavljenoj, to je veća vjerojatnost β i bliža je $1-\alpha$.

Ako se hipoteza $H_0 ... \theta = \theta_0$ testira z-testom na razini α i ako je $c_1 = \theta_0 - z_{\alpha/2}\sigma_{\hat{\theta}}$ i $c_2 = \theta_0 + z_{\alpha/2}\sigma_{\hat{\theta}}$, onda je vjerojatnost prihvaćanja hipoteze, unatoč tome što je pravi parametar jednak $\theta = \theta_1$, jednaka površini što ju odsijeca interval $\langle c_1, c_2 \rangle$ ispod normalne krivulje $N(\theta_1, \sigma_{\hat{\theta}})$:

$\theta_1 > c_2$	$\beta = p\left(\frac{c_1-\theta_1}{\sigma_{\hat{\theta}}} < z < 0\right) - p\left(\frac{c_2-\theta_1}{\sigma_{\hat{\theta}}} < z < 0\right)$
$\theta_1 < c_1$	$\beta = p\left(0 < z < \frac{c_2-\theta_1}{\sigma_{\hat{\theta}}}\right) - p\left(0 < z < \frac{c_1-\theta_1}{\sigma_{\hat{\theta}}}\right)$
$c_1 < \theta_1 < c_2$	$\beta = p\left(\frac{c_1-\theta_1}{\sigma_{\hat{\theta}}} < z < 0\right) + p\left(0 < z < \frac{c_2-\theta_1}{\sigma_{\hat{\theta}}}\right)$

Analogno vrijedi i za t-test.

Ako se hipoteza $H_0 ... \theta \leq \theta_0$ testira z-testom na razini α i ako je $c_1 = \theta_0 - z_{\alpha}\sigma_{\hat{\theta}}$ i $c_2 = \theta_0 + z_{\alpha}\sigma_{\hat{\theta}}$, onda je vjerojatnost prihvaćanja hipoteze, unatoč tome što je pravi parametar jednak $\theta = \theta_1 > \theta_0$, jednaka površini što ju odsijeca interval $\langle -\infty, c_2 \rangle$ ispod normalne krivulje $N(\theta_1, \sigma_{\hat{\theta}})$:

$\theta_1 > c_2$	$\beta = 0.5 - p\left(\frac{c_2-\theta_1}{\sigma_{\hat{\theta}}} < z < 0\right)$
$\theta_1 < c_2$	$\beta = 0.5 + p\left(0 < z < \frac{c_2-\theta_1}{\sigma_{\hat{\theta}}}\right)$

Vjerojatnost prihvaćanja hipoteze $H_0 ... \theta \geqslant \theta_0$, unatoč tome što je pravi parametar jednak $\theta = \theta_1 < \theta_0$, jednaka je površini što ju odsijeca interval $\langle c_1, \infty \rangle$ ispod normalne krivulje $N(\theta_1, \sigma_{\hat{\theta}})$:

$\theta_1 < c_1$	$\beta = 0.5 - p\left(0 < z < \frac{c_1-\theta_1}{\sigma_{\hat{\theta}}}\right)$
$\theta_1 > c_1$	$\beta = 0.5 + p\left(\frac{c_1-\theta_1}{\sigma_{\hat{\theta}}} < z < 0\right)$

Analogno vrijedi i za t-test.

6.1.3 Testiranje hipoteza o varijancama pomoću F i Hi kvadrat-distribucije

Ako je uzorak veličine n uzet iz normalno distribuirane populacije varijance σ_0^2, a $\hat{\sigma}^2$ njezin nepristrani procjenitelj, onda su vrijednosti $\frac{(n-1)\hat{\sigma}^2}{\sigma_0^2}$ distribuirane po χ^2-distribuciji s $n-1$ stupnjeva slobode. Tada, s vjerojatnošću $1-\alpha$ smijemo tvrditi da se testna veličina $\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2}$ nalazi u intervalu $\left\langle \chi^2_{(1-\alpha/2);n-1}, \chi^2_{\alpha/2;n-1} \right\rangle$, gdje su brojevi $\chi^2_{(1-\alpha/2);n-1}$ i $\chi^2_{\alpha/2;n-1}$ vrijednosti χ^2-distribuirane varijable s $n-1$ stupnjeva slobode koje imaju svojstvo $p\left(\chi^2 > \chi^2_{1-\alpha/2;n-1}\right) = 1-\alpha/2$ odnosno $p\left(\chi^2 > \chi^2_{\alpha/2;n-1}\right) = \alpha/2$.
Moguće odluke uz testiranje o pretpostavljenoj varijanci na razini značajnosti α su dane u tablici

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja H_0	Područje odbacivanja H_0
$H_0..\sigma^2 = \sigma_0$	$H_1..\sigma^2 \neq \sigma_0^2$	$\chi^2_{(1-\alpha/2);n-1} < \chi^2$ ili $\chi^2 < \chi^2_{\alpha/2;n-1}$	$\chi^2 \leq \chi^2_{(1-\alpha/2);n-1}$ ili $\chi^2 \geqslant \chi^2_{\alpha/2;n-1}$
$H_0..\sigma^2 \leq \sigma_0^2$	$H_1..\sigma^2 > \sigma_0^2$	$\chi^2 < \chi^2_{\alpha/2;n-1}$	$\chi^2 \geqslant \chi^2_{\alpha/2;n-1}$
$H_0..\sigma^2 \geqslant \sigma_0^2$	$H_1..\sigma^2 < \sigma_0^2$	$\chi^2 > \chi^2_{(1-\alpha);n-1}$	$\chi^2 \leq \chi^2_{(1-\alpha);n-1}$

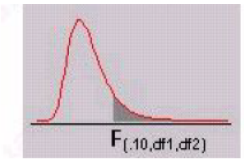
Primjer 6.5 *Broker koji trguje dionicama prosuđuje stabilnost, tj. varijabilnost tečaja određene dionice. On pretpostavlja da je prosječno odstupanje tečaja od prosjeka dionica u desetogodišnjem razdoblju* 38 *centi. Može li se prihvatiti ta hipoteza na razini signifikantnosti od* 5%*, ako je iz baze podataka o kotacijama burze izabran uzorak od* 10 *dnevnih cijena dionica čija je aritmetička sredina* $\overline{x} = 450$*, a procjenitelj varijance je* $\hat{\sigma}^2 = \sigma^2 \frac{n}{n-1} = 36$ *(faktor* $\frac{N-1}{N}$ *je zanemaren), uz pretpostavku da su cijene normalno distribuirane?*

Testiramo hipotezu $H_0...\sigma^2 = 38$, $H_1...\sigma^2 \neq 38$ *uz poznate podatke:* $n = 10$, $\alpha = 0.05$. *Vrijednosti pročitane iz tablice su* $\chi^2_{1-\alpha/2;n-1} = \chi^2_{0.975;9} = 2.70039$, $\chi^2_{\alpha/2;n-1} = \chi^2_{0.025;9} = 19.0228$, *a testna veličina je* $\chi^2 = \frac{(n-1)\hat{\sigma}^2}{\sigma_0^2} = \frac{9 \cdot 36}{38} = 9.52632$. *Budući se testna veličina nalazi unutar granica prihvaćanja nulte hipoteze, na razni značajnosti od* 5% *se prihvaća* H_0.

U mnogim situacijama važno nam je samo znati jesu li varijance različitih populacija jednake. Ta informacija sama po sebi je značajna jer iskazuje odnos stupnjeva disperzije dviju populacija. No, često nam služi i da bismo primijenili odgovarajuće tehnike testiranja koje možemo uporabiti samo uz pretpostavku o jednakosti populacijskih varijanci. Neka je $\hat{\sigma}_1^2$ nepristrani procjenitelj varijance σ_1^2 normalno distribuirane populacije izračunat na osnovu uzorka veličine n_1. Neka je $\hat{\sigma}_2^2$ nepristrani procjenitelj varijance σ_2^2 normalno distribuirane populacije izračunat na osnovu uzorka veličine n_2. Ako su uzorci međusobno neovisni i uzeti iz **normalno distribuiranih populacija**, onda su brojevi (omjeri)

$$\frac{\frac{\hat{\sigma}_1^2}{\sigma_1^2}}{\frac{\hat{\sigma}_2^2}{\sigma_2^2}},$$

za bilo koja dva takva uzorka, distribuirani po F-distribuciji s $[n_1 - 1, n_2 - 1]$ stupnjeva slobode. Testiranje hipoteza $\frac{\sigma_1^2}{\sigma_2^2} = 1$, $\frac{\sigma_1^2}{\sigma_2^2} \geqslant 1$, $\frac{\sigma_1^2}{\sigma_2^2} \leq 1$, odnosno njihovih negacija, se provodi uporabom F-testa, tj. usporedbom testne F-**vrijednosti** $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ i tabelarne vrijednosti F_α ili $F_{\alpha/2}$ za F-distribuciju s $[n_1 - 1, n_2 - 1]$ stupnjeva slobode, gdje je α razina signifikantnosti.



$\alpha = 0{,}01$

df_1

df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

$\alpha = 0{,}05$

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

df_1 broj stupnjeva slobode brojnika, df_2 broj stupnjeva slobode nazivnika.

Moguće odluke uz testiranje o pretpostavljenom odnosu dviju varijanci na razini značajnosti α su dane u tablici

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hip.	Područje odbacivanja H_0
$H_0 ... \frac{\sigma_1^2}{\sigma_2^2} = 1$	$H_1 ... \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$F < F_{\alpha/2}$ ili $F > \frac{1}{F_{\alpha/2}}$	$F > F_{\alpha/2}$ ili $F < \frac{1}{F_{\alpha/2}}$
$H_0 ... \frac{\sigma_1^2}{\sigma_2^2} \leq 1$	$H_1 ... \frac{\sigma_1^2}{\sigma_2^2} > 1$	$F < F_{\alpha}$	$F > F_{\alpha}$
$H_0 ... \frac{\sigma_1^2}{\sigma_2^2} \geqslant 1$	$H_1 ... \frac{\sigma_1^2}{\sigma_2^2} < 1$	$F > \frac{1}{F_{\alpha}}$	$F < \frac{1}{F_{\alpha}}$

Primjer 6.6 *Broker prosuđuje rizik trgovanja dionicama dviju firmi temeljem promatranja varijanci dnevnih zaključnih cijena. Prosječno odstupanje od prosjeka*

dnevnih cijena prve vrste dionica u uzorku od 21 *dana je* 20 *centi, a kod druge vrste dionica prosječno odstupanje od prosjeka dnevnih cijena u uzorku od* 26 *dana je* 15 *centi. Može li se prihvatiti pretpostavka da je poslovanje dionicama obiju firmi jednako rizično, uz razinu signifikantnosti* $10\,\%$.

Testiramo sljedeću hipotezu $H_0 ... \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_1 ... \frac{\sigma_1^2}{\sigma_2^2} \neq 1$. *Za implementaciju* F*-testa moramo pretpostaviti da su zaključne dnevne cijene dionica normalno distribuirane. Zadani podaci su* $n_1 = 21$, $s_1 = 20$, *što povlači* $\hat{\sigma}_1^2 = s_1^2 \frac{n_1}{n_1 - 1} = 420$, *te* $n_2 = 26$, $s_2 = 15$, *iz čega slijedi* $\hat{\sigma}_2^2 = s_2^2 \frac{n_2}{n_2 - 1} = 234$. *Tabelarna vrijednost je* $F_{\alpha/2,[n_1-1,n_{2-1}]} = F_{0.05,[20,25]} = 2.01$, *testna vrijednost je* $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{420}{234} = 1.79 < 2.01$, *pa na danoj razini značajnosti prihvaćamo da je poslovanje dionicama jednako rizično.*

Zadatak 6.7 *Zadani su podaci o proizvodnji nekog proizvoda u komadima*

I.smjena: $112, 78, 69, 97, 109, 123, 95, 111, 92, 95, 116, 92, 91, 96, 98$;

II.smjena: $46, 78, 102, 100, 92, 43, 103, 55, 111, 108, 112, 94, 93, 117, 117, 116$.

Može li se prihvatiti pretpostavka da je stupanj varijabilnosti proizvodnje II.smjene veći od stupnja varijabilnosti I.smjene, uz razinu značajnosti $\alpha = 5\%$*? Može li se prihvatiti pretpostavka da se prosječne proizvodnje u obje smjene ne razlikuju?*

Rješenje: Za oba testa potrebno je pretpostaviti da su komadi proizvoda normalno distribuirani u obje smjene. Testiramo hipotezu $H_0 ... \frac{\sigma_2^2}{\sigma_1^2} \geqslant 1 \quad H_1 ... \frac{\sigma_2^2}{\sigma_1^2} < 1$. *Vrijedi* $\overline{x_1} = \frac{112 + \cdots + 98}{15} = 98.267$, $\overline{x_2} = \frac{46 + \cdots + 116}{16} = 92.938$, $n_1 = 15$, $n_2 = 16$, $\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n_1} x_{i1}^2 - n_1 \overline{x_1}^2}{n_1 - 1} = 201.352$, $\hat{\sigma}_2^2 = 613.396$ *i* $F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{613.396}{201.352} = 3.046$. *Budući je* $F > \frac{1}{F_{\alpha,[n_2-1,n_1-1]}} = \frac{1}{F_{0.05,[15,14]}} = \frac{1}{2.463}$, *to prihvaćamo pretpostavku* H_0.

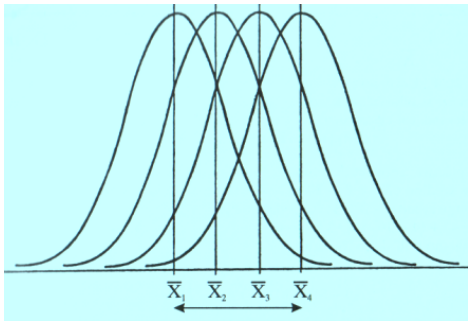
Nadalje, testiramo hipotezu $H_0 ... \mu_1 - \mu_2 = 0 \quad H_1 ... \mu_1 - \mu_2 \neq 0$. *Direktnim izračunom dobivamo* $\overline{x_1} - \overline{x_2} = 5.329$, $\sigma_{\overline{x_1} - \overline{x_2}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = 3.8795$, $t = \frac{\overline{x_1} - \overline{x_2}}{\sigma_{\overline{x_1} - \overline{x_2}}} = 0.728$. *Broj stupnjeva slobode kojeg koristimo za izračun vrijednosti* $t_{\alpha/2}$ *je jednak* $s.s. = \operatorname{Int}\left[\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2 \left(\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2 - 1}\right)\right] = \operatorname{Int}[24.2] = 24$, *pa je* $t_{\alpha/2} = 2.06$. *Budući je* $-t_{\alpha/2} < t < t_{\alpha/2}$ *to prihvaćamo pretpostavku da se prosječne proizvodnje u obje smjene statistički ne razlikuju na razini značajnosti od 0.05.*

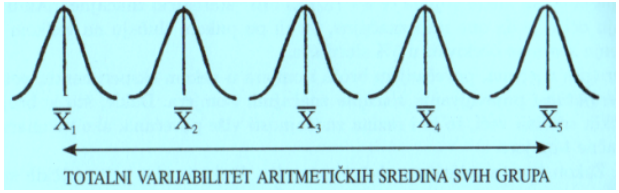
6.1.4 Testiranje hipoteza o jednakosti sredina K populacija

Analiza varijance (**ANOVA**) se sastoji od skupa postupaka kojima se raščlanjuje varijanca prema izvorima varijabilnosti njezinih vrijednosti. Upotrebljava se u mnogim područjima statistike (analiza nacrta statističkih pokusa, testiranje hipoteze o parametru u regresijskim modelima...), a u prvom redu za testiranje hipoteze o jednakosti aritmetičkih sredina $\mu_1, ..., \mu_K$ od K populacija temeljem neovisnih uzoraka:

$H_0 ... \mu_1 = \mu_2 = \cdots = \mu_K = \mu$; $\quad H_1 ...$, $\mu_i \neq \mu_j$, za neke $i, j = 1, ..., K$.

Provođenje testiranja sredina u parovima, "jedan po jedan", bi u konačnici povećalo razinu značajnosti, osnosno vjerojatnost moguće greške. Osnovna zamisao analize varijance se sastoji u tome da se usporedi varijabilitet među aritmetičkim sredinama uzoraka s varijabilitetima unutar uzoraka, pa ako je on statistički značajno veći, onda odbacujemo H_0.





Neka su $n_1, ..., n_K$ veličine uzoraka i neka je $n = n_1 + \cdots + n_K$. Označimo sa x_{ij} i-tu vrijednost u j-tom uzorku. Neka vrijednosti $x_{1j}, ..., x_{n_j j}$ tvore j-ti uzorak. Označimo sa

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{K} \sum_{i=1}^{n_j} x_{ij}$$

zajedničku aritmetičku sredinu svih vrijednosti u svim uzorcima, a sa

$$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$

aritmetičku sredinu j-tog uzorka.
Odstupanje x_{ij} od zajedničke sredine se može prikazati kao zbroj odstupanja

$$x_{ij} - \overline{x} = (\overline{x}_j - \overline{x}) + (x_{ij} - \overline{x}_j)\,.$$

Zbroj kvadrata odstupanja vrijednosti u svim uzorcima od zajedničke sredine označujemo sa SST (Sum squares total) i on iznosi

$$SST = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2 = \sum_{j=1}^{K} n_j (\overline{x}_j - \overline{x})^2 + \sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2\,.$$

Izraz

$$\sum_{j=1}^{K} n_j (\overline{x}_j - \overline{x})^2$$

označujemo sa SSB (Sum squares between) i on predstavlja dio ukupne varijabilnosti koji izvire iz međusobne varijacije sredina uzoraka. Izraz $MSB = SSB/(K-1)$ označuje sredinu takvih kvadratnih odstupanja (Mean squares).
Izraz

$$\sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

označujemo sa SSW (Sum squares within) i on predstavlja dio ukupne varijabilnosti koji izvire iz unutarnje varijabilnosti svake pojedine grupe.
Izraz $MSW = SSW/(n-K)$ označuje sredinu kvadratnih odstupanja.
Ako su sve populacije iz kojih su uzeti uzorci **normalno distribuirane s međusobno jednakom varijancom**, onda su vrijednosti $F = \frac{MSB}{MSW}$ distribuirane po F-distribuciji sa $(K-1)$ i $(n-K)$ stupnjeva slobode. U slučaju da je $F \leqslant F_{\alpha,[(K-1),(n-K)]}$ prihvaćamo nultu hipotezu da sredine svih K populacija nisu statistički značajno različite, s razinom značajnosti α.

Primjer 6.8 *Proizvođač igračaka želio je ustanoviti ima li boja igračke utjecaj na njezinu atraktivnost pa je na* 4 *uzorka od po* 10*-ero djece mjerio minute koliko se pojedino dijete zadržalo u igri s tom igračkom:*

crveni	1	2	5	7	6	1	2	2	4	4
žuti	2	3	6	3	2	8	7	5	6	8
zeleni	2	4	2	1	2	3	4	1	3	2
modri	5	3	1	2	1	3	4	2	3	1

.

Mogu li se razlike među bojama igračaka smatrati statistički značajnim na razini 5*%?*

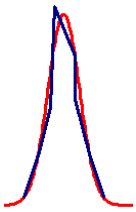
Zadani su podaci $K = 4$, $n_1 = \cdots = n_4 = 10$, $n = 40$, *iz kojih se izračuna:*
$\overline{x}_1 = 3.4$, $\overline{x}_2 = 5$, $\overline{x}_3 = 2.4$, $\overline{x}_4 = 2.5$, $\overline{x} = 3.325$,
$SSB = \sum_{j=1}^{K} n_j \left(\overline{x}_j - \overline{x}\right)^2 =$
$10[(3.4 - 3.325)^2 + (5 - 3.325)^2 + (2.4 - 3.325)^2 + (2.5 - 3.325)^2] = 43.475$,
$MSB = \frac{SSB}{K-1} = \frac{43.475}{3} = 14.487$,
$SSW = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{x}_j\right)^2 = (1 - 3.4)^2 + \cdots + (4 - 3.4)^2 + \cdots + (5 - 2.5)^2 + \cdots +$
$(1 - 2.5)^2 = 171.36$,
$MSW = \frac{SSW}{n-K} = \frac{171.36}{36} = 3.258$,
$F = \frac{14.487}{3.258} = 4.45 > F_{0.05,[3,36]} \approx 2.872$.
Zaključujemo da su razlike statistički značajne. Analizom po parovima se može pokazati da se statistički razlikuju sredine između žute i zelene, te žute i modre boje.

6.2 Neparametarski testovi

6.2.1 Hi kvadrat test

χ^2-test se primjenjuje u različitim statističkim postupcima. Najčešće se provodi radi testiranja oblika distribucije populacije. Postupak se sastoji u prikupljanju empirijske distribucije uzorka, potom se odabire model teorijske distribucije s kojom se uspoređuje empirijska distribucija. Parametre teorijske distribucije se

računa- procjenjuju pomoću uzorka. Ako testiranje pokaže da se distribucija populacije statistički ne razlikuje (približno je jednaka) pretpostavljenoj teorijskoj distribuciji, onda možemo računati očekivane vjerojatnosti (frekvencije) svih populacijskih vrijednosti.



Neka u uzorku uzetom iz populacije ima n elemenata (slučajna varijabla X poprimi n vrijednosti), a od toga točno k-različitih vrijednosti (k različitih intervala kojima pripadaju te vrijednosti) pri čemu njih f_1 ima vrijednosti x_1 (ili pripada intervalu x_1),..., f_k ima vrijednosti x_k (ili pripada intervalu x_k). Tada je $f_1 + \cdots + f_k = n$. Postavljamo sljedeću hipotezu:
H_0...distribucija populacije je točno određenog oblika; H_1...distribucija nije pretpostavljenog oblika.
Testiranje se provodi pomoću test veličine $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$, gdje je e_i očekivana apsolutna frekvencija prema pretpostavljenoj distribuciji, tj. $e_i = np_i$, gdje je $p_i = p(X = x_i)$ za diskretnu varijablu X, odnosno $p(X \in x_i)$ ako je kontinuirana. Odluka se donosi usporedbom test veličine χ^2 s teorijskom vrijednošću χ^2_α s $(k - g - 1)$ stupnjeva slobode, gdje je α razina signifikantnosti, a g je broj procijenjenih parametara pretpostavljene distribucije pomoću uzorka. H_0 se prihvaća ako je $\chi^2 \leqslant \chi^2_\alpha$.

Primjena χ^2 testa za testiranje oblika distribucije je valjana ako su ispunjeni sljedeći uvjeti:

1. Uzorak je dovoljno velik, tj. $n \geqslant 30$;

2. Očekivane frekvencije nisu suviše male, tj. sve e_i moraju biti $\geqslant 2$, a 50% očekivanih frekvencija mora biti $\geqslant 5$.

Ako se pojave očekivane frekvencije manje od propisanog, onda se spajanjem susjednih grupa one povećaju ali se i mijenja k, a time se smanjuju stupnjevi slobode.

Napomenimo da pitanje je li neki uzorak s parametrima $\overline{x}$ i s točno određeno distribuiran (primjerice normalno) znači da je taj uzorak uzet iz populacije čija distribucija ima upravo parametre koji se podudaraju s $\overline{x}$ i s (ili $\hat{\sigma}$) (u smislu Primjedbe 3.33). To pitanje se terminološki razlikuje od pitanja je li uzorak uzet iz populacije čija distribucija je određena parametrima μ i σ koji se općenito ne moraju podudarati s parametrima uzorka $\overline{x}$ i s ($\hat{\sigma}$).

Primjer 6.9 *Promatra se broj prometnih nezgoda po danima*

Broj nezgoda	0	1	2	3	$\geqslant 4$
Broj dana	44	37	15	3	1

.

Može li se prihvatiti pretpostavka da je distribucija nezgoda po danima raspoređena po Poissonovoj distribuciji s parametrom $\lambda = 0.9$*? Testira se na razini* 1% *značajnosti.*

Testiramo H_0*...distribucija nezgoda po danima se ravna po Poissonovoj distribuciji,* H_1*...distribucija se ne ravna po Poissonovoj distribuciji.*

Zadano je $n = 100$, $\alpha = 0.01$, *a pretpostavljena distribucija je* $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, *pa su apsolutne frekvencije* $e_i = np(x_i) = 100\frac{e^{-0.9}0.9^{x_i}}{x_i!}$.

x_i *(broj nezgoda)*	f_i *(broj dana)*	$p_i = p(x_i)$	$e_i = 100p_i$	$\frac{(f_i-e_i)^2}{e_i}$
0	44	0.4066	40.66	0.2744
1	37	0.3659	36.59	0.0046
2	15	0.1647	16.47	0.1312
3	3	0.0494	$4.94 + 1.34$	0.8278
≥ 4	1	0.0134	(1.34)	$**$
Σ	100	1.00	100	$\chi^2 = 1.237$

Budući je očekivana frekvencija posljednje grupe $1.34 < 2$, *to je ta frekvencija pridodana prethodnoj očekivanoj frekvenciji* 4.94. *No, onda je broj različitih vrijednosti u uzorku* $k = 4$, *a ne više 5. Test veličina je* $\chi^2 = 1.237$. *Stupnjevi slobode*

su $(k-g-1)=(4-0-1)=2$, *a teorijska vrijednost je jednaka* $\chi^2_{0.01}=11.344$. *Stoga na danoj razini značajnosti prihvaćamo hipotezu.*

Primjer 6.10 *Možemo li, mjereći visinu* 135 *dvadesetogodišnjaka, o kojima su podaci u tablici, donijeti zaključak o odstupanju distribucije visine od normalne distribucije?*

Interval x_i *(cm)*	*Frekvencija* f_i
$153.5-156.5$	0
$156.5-159.5$	1
$159.5-162.5$	2
$162.5-165.5$	9
$165.5-168.5$	15
$168.5-171.5$	25
$171.5-174.5$	28

Interval x_i *(cm)*	*Frekvencija* f_i
$174.5-177.5$	20
$177.5-180.5$	16
$180.5-183.5$	13
$183.5-186.5$	5
$186.5-189.5$	1
$189.5-192.5$	0
Σ	135

Budući je $\overline{x}=173.47$ *(vagana aritmetička sredina) i* $s=5.37$, *to ćemo testirati hipotezu je li populacija visina dvadesetogodišnjaka iz koje je uzet uzorak veličine* 135 *distribuirana po normalnoj distribuciji* $N(173.47, 5.37)$.
Da bismo lakše izračunali p_i, *tj. površinu iznad intervala* $x_i=(x_{i1},x_{i2})$ *ispod Gaussove krivulje, a time i teorijsku frekvenciju* $e_i=135p_i$, *promatrajmo standardiziranu normalnu varijablu* $z=\frac{x-\mu}{\sigma}$, *odnosno preračunajmo granice* (x_{i1},x_{i2}) *svakog pojedinog intervala u* z *vrijednosti* $\left(z_{i1}=\frac{x_{i1}-\mu}{\sigma}, z_{i2}=\frac{x_{i2}-\mu}{\sigma}\right)$ *odstupanja od aritmetičke sredine u jedinicama standardne devijacije.*
Primjerice: $z_{11}=\frac{153.5-173.47}{5.37}=-3.72$,
$z_{12}=\frac{156.5-173.47}{5.37}=-3.16$,
$p\left(Z\in\langle -3.72,-3.16\rangle\right)=p\left(\langle 0,3.72\rangle\right)-p\left(\langle 0,3.16\rangle\right)=0.0003$,
$e_i=135\cdot 0.0003\approx 0.04$.

Dobivene podatke prikažimo u donjoj tablici.

z vrijednosti (z_{i1}, z_{i2}) *intervala* x_i	*tablična vjerojatnost (površina)* p_i	*očekivana frekvencija* e_i
-3.72 do -3.16	0.0003	0.04
-3.16 do -2.6	0.0039	0.53
-2.6 do -2.04	0.016	2.16
-2.04 do -1.48	0.0487	6.57
-1.48 do -0.93	0.1068	14.42
-0.93 do -0.37	0.1795	24.23
-0.37 do 0.19	0.2197	29.66
0.19 do 0.75	0.1980	26.73
0.75 do 1.31	0.1315	17.75
1.31 do 1.87	0.0644	8.69
1.87 do 2.43	0.0232	3.13
2.43 do 2.98	0.0061	0.82
2.98 do 3.54	0.0014	0.19

Grupiranjem četiri prve te četiri zadnje grupe, te zbrajanjem odgovarajućih podataka postižemo mogućnost primjene χ^2 *test. Relevantne podatke prikažimo tabelarno.*

f_i	e_i	$f_i - e_i$	$(f_i - e_i)^2$	$\frac{(f_i-e_i)^2}{e_i}$
$12 = 1 + 2 + 9$	$0.04 + 0.53 + 2.16 + 6.57 = 9.3$	2.7	7.29	0.784
15	14.42	0.58	0.336	0.023
25	24.23	0.77	0.593	0.024
28	29.66	-1.66	2.756	0.093
20	26.73	-6.73	45.293	1.694
16	17.75	-1.75	3.063	0.173
19	$8.69 + 3.13 + 0.82 + 0.19 = 12.83$	-6.17	38.069	2.967
$\Sigma = 135$	$\Sigma = 135$			5.758

Test veličina je $\chi^2 = 5.758$. *Budući da smo oba parametra koja određuju normalnu distribuciju procijenili pomoću uzorka, to je* $g = 2$, *a nakon spajanja razreda s*

premalom frekvencijom je $k = 7$*, pa su stupnjevi slobode* $(k - g - 1) = 4$*. Budući je* $\chi^2 < \chi^2_{0.05} = 9.488$*, to prihvaćamo hipotezu da je populacija visina dvadestogodišnjaka iz koje je uzet uzorak normalno distribuirana.*

Zadatak 6.11 *Ako je od* 200 *studenata njih* 40 *palo na ispitu kod nekog profesora,* 110 *dobilo ocjenu manju od* 5*, a njih* 50 *dobilo ocjenu* 5*, možemo li reći da ovaj rezultat odstupa od "normalne" raspoređenosti uspjeha po kojoj je 50% prosječnih i po 25% loših i izvrsnih.*

Rješenje: *Zadane podatke prikažimo tabelarno*

f_i	e_i	$f_i - e_i$	$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
40	50	-10	100	2
110	100	10	100	1
50	50	0	0	0

.

Vrijedi $\chi^2 = 3$*,* $k = 3$*,* $n = 200$*. Stupnjevi slobode su* $(k - g - 1) = 2$*, pa je za* $\alpha = 0.05$ $\chi^2_{0.05} = 5.991$*. Budući je* $\chi^2 < \chi^2_{0.05}$ *to zaključujemo da ova distribucija ne odstupa statistički značajno od "normalne".*

Zadatak 6.12 *Ispituje se učestalost zastoja strojeva na jednoj proizvodnoj liniji po jednoj smjeni. Analizom* 400 *smjena registrirano je:*

Broj zastoja	0	1	2	3	4	5	6
Broj smjena	35	115	130	75	30	10	5

.

Može li se prihvatiti pretpostavka da se učestalost zastoja ravna po binomnoj distribuciji?

Rješenje: *Pretpostavljena binomna distribucija u ovom slučaju je oblika* $p(x) = \frac{6!}{(6-x)!x!} p^x (1-p)^{6-x}$*. Budući* p *nije poznat moramo ga procijeniti. Očekivana vrijednost binomne varijable je* $6p$*, pa ćemo ju izjednačiti s aritmetičkom sredinom uzorka, tj.* $\overline{x} = 2$*. Izlazi da je* $\hat{p} = \frac{2}{6}$*. Sada lako izračunamo očekivane vjerojatnosti* $p(x_i) = \frac{6!}{(6-x_i)!x_i!} \left(0.\dot{3}\right)^{x_i} \left(1 - 0.\dot{3}\right)^{6-x_i}$ *i očekivane frekvencije* $e_i = np(x_i) = 400p(x_i)$*.*

x_i	f_i	$p(x_i)$	e_i	$\frac{(f_i-e_i)^2}{e_i}$
0	35	0.0878	35.12	0.00041
1	115	0.2634	105.36	0.88202
2	130	0.3292	131.68	0.02143
3	75	0.2195	87.80	1.86606
4	30	0.0823	32.92	0.259
5	10	0.0165	$6.60 + 0.52 = 7.12$	9.72112
6	5	0.0013	(0.52)	**
Σ	400	1.00	400	11.75004

Očekivana frekvencija zadnje grupe je < 2*, pa je dodana prethodnoj očekivanoj frekvenciji* 6.6*. Stupnjevi slobode su* $(k-g-1) = (6-1-1) = 4$*. Teorijska* χ^2 *vrijednost za* $\alpha = 0.05$ *je* $\chi^2_{0.05} = 9.4877$*, a budući je manja od empirijske* $\chi^2 = 11.75004$*, hipotezu ne prihvaćamo.*

χ^2-test rabimo i za testiranje hipoteze o jednakosti proporcija triju ili više populacija:

$H_0 ... p_1 = \cdots = p_k = p$; $\quad H_1 .. p_i \neq p_j$, za neke $i, j = 1, ..., k$.

Neka u k uzoraka veličine $n_1, ..., n_k$, redom, uzetih iz k različitih populacija, $m_1, ..., m_k$ elemenata u tim populacijama ima traženo obilježje. Ispitujemo hipotezu da je proporcija traženog obilježja u svim populacijama međusobno jednaka i iznosi p.

Ako proporcija p nije zadana uzima se $\hat{p} = \frac{m_1+\cdots+m_k}{n_1+\cdots+n_k}$. Test veličina je $\chi^2 = \sum_{i=1}^{k} \frac{(f_i-e_i)^2}{e_i}$, pri čemu je $f_i = m_i$, a $e_i = n_i p$ ili $e_i = n_i \hat{p}$. Teorijsku vrijednost χ^2_α uzimamo na razini signifikantnosti α i sa $k-1$ stupnjeva slobode. Ako je $\chi^2 \leq \chi^2_\alpha$, onda prihvaćamo hipotezu da se proporcije statistički značajno ne razlikuju.

Primjer 6.13 *Iz* 4 *najveća dalmatinska grada anketirani su konzumenti kave u uzorcima od* $100, 200, 150, 250$*. Od toga je na pitanje o povremenom kupovanju nove marke kave potvrdno odgovorilo* $20, 35, 37, 43$*, redom. Može li se pretpostaviti*

da je proporcija kupaca nove marke kave jednaka u sva 4 *grada. Testira se na razini* 5% *značajnosti.*

Vrijedi $\hat{p} = \frac{m_1+\cdots+m_4}{n_1+\cdots+n_4} = \frac{20+35+37+43}{100+200+150+250} = 0.1928.$

Broj potrošača u uzorku n_i	*Broj kupaca određene kave u uzorku* $m_i = f_i$	*Očekivani broj kupaca* $e_i = n_i\hat{p}$	$\frac{(f_i-e_i)^2}{e_i}$
100	20	19.286	0.02643
200	35	38.572	0.33079
150	37	28.929	2.25176
250	43	48.215	0.56406
$\Sigma = 700$	$\Sigma = 135$	$\Sigma = 135$	$\chi^2 = 3.17304$

Testna veličina je $\chi^2 = 3.173$. *Stupnjevi slobode su* $k-1 = 3$, *pa je* $\chi^2_{0.05} = 7.81473$. *Budući je* $\chi^2 < \chi^2_\alpha$ *to možemo prihvatiti da se proporcija kupaca kave određene marke u sva* 3 *grada statistički značajno ne razlikuje, s razinom signifikantnosti* 5%.

Neka diskretna dvodimenzionalna slučajna varijabla $Z = (X, Y)$ poprima vrijednosti (x_i, y_j) i neka je $p_{ij} = p(X = x_i, Y = y_j)$ vjerojatnost događaja koji se sastoji od onih ishoda kojima slučajna varijabla pridruži uređeni par (x_i, y_j) (događaj da slučajna varijabla ima vrijednost (x_i, y_j)). Neka je distribucija diskretne slučajne varijable (X, Y) prikazana *tablicom kontigencije*:

$X\backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

U uzorku od n elemenata pojavile su se vrijednosti (x_i, y_j), $i = 1, ..., r$; $j = 1, ..., c$, s frekvencijama f_{ij}, što znači da od n elemenata u uzorku njih f_{ij} ima vrijednost

varijable (obilježje) X jednaku x_i i vrijednost varijable Y jednaku y_j, pri čemu varijable ne moraju biti numeričke, tj. obilježja x_i i y_j mogu biti nenumerička.

$X\backslash Y$	y_1	$\cdots$	y_c	$\sum$
x_1	f_{11}	$\cdots$	f_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	f_{r1}	$\cdots$	f_{rc}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot c}$	n

Želimo testirati hipotezu o neovisnosti događaja $X = x_i$ i $Y = y_j$, tj. obilježja x_i i y_j, odnosno o jednakostima $p_{ij} = p_{i\cdot}p_{\cdot j}$, za svaki $i = 1, ..., r$; $j = 1, ..., c$, gdje su $p_{i\cdot}$ i $p_{\cdot j}$ marginalne vjerojatnosti.

$H_0 ... p_{ij} = p_{i\cdot}p_{\cdot j}$, $\forall\ i = 1, ..., r$; $j = 1, ..., c$

$H_1 ... \exists i \in \{1, ..., r\}\ \exists j \in \{1, ..., c\}\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$

Test veličina je $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$, gdje je $e_{ij} = n\hat{p}_{ij} = n\frac{n_{i\cdot}n_{\cdot j}}{n^2} = \frac{n_{i\cdot}n_{\cdot j}}{n}$ očekivana frekvencija. Ako je $\chi^2 \leqslant \chi^2_\alpha$, gdje je α razina signifikantnosti testa, a χ^2_α vrijednost χ^2 distribucije sa $(r-1)(c-1)$ stupnjeva slobode, onda se prihvaća hipoteza o međusobnoj neovisnosti svih obilježja x_i i y_j.

Primjer 6.14 *Na uzorku od 900 potrošača istražuje se ovisnost između visine mjesečne plaće i sklonosti kupovine određenog proizvoda s razinom signifikantnosti od 5%. Dobiveni su sljedeći rezultati:*

	sklonost potrošnji			
plaća	*stalni kupac*	*povremeno*	*ne kupuje*	*ukupno*
< 1000 €	70	17	21	108
$1000 - 1500$	165	56	28	249
$1500 - 2500$	195	85	26	306
> 2500	170	42	25	237
ukupno	600	200	100	900

.

Očekivane frekvencije računamo prema formuli $e_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$, $i = 1, ..., 4$; $j = 1, 2, 3$, *pa je* $e_{11} = \frac{n_{1\cdot}n_{\cdot 1}}{900} = \frac{108\cdot 600}{900} = 72$, $e_{12} = \frac{n_{1\cdot}n_{\cdot 2}}{900} = \frac{108\cdot 200}{900} = 24, ..., e_{43} = \frac{n_{4\cdot}n_{\cdot 3}}{900} =$

$\frac{237\cdot 100}{900} = 26.33$. *Testna veličina je* $\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = \frac{(70-72)^2}{72} + \frac{(17-24)^2}{24} + \cdots + \frac{(25-26.33)^2}{26.33} = 18.633$. *Stupnjevi slobode su:* $(4-1)(3-1) = 6$, *pa teorijska vrijednost* χ^2 *distribucije s 6 stupnjeva slobode za* $\alpha = 0.05$ *iznosi* $\chi^2_{0.05} = 12.5916$. *Budući je* $\chi^2 > \chi^2_{0.05}$ *to, na danoj razini značajnosti, ne prihvaćamo pretpostavku da je sklonost kupovini nekog proizvoda neovisna o plaći.*

Bibliografija

[1] I. Pavlić, *Statistička teorija i primjena*, Tehnička knjiga, Zagreb, 1977.

[2] B. Petz, *Osnove statističke metode za nematematičare*, Naklada Slap, Jastrebarsko, 2007.

[3] T. Sincich, *Business Statistics by Example*, Prentice-Hall, New Jersey, 1996.

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

[5] I. Šošić. *Primijenjena statistika*, Školska knjiga, Zagreb, 2006.

[6] B. Vrdoljak, *Vjerojatnost i statistika*, Građevinsko-arhitektonski fakultet Sveučilišta u Splitu, Split, 2007.