STATISTIČKE METODE U GEOFIZICI
(Interna skripta)

Zoran Pasarić

Prirodoslovno - matematički fakultet
Geofizički odsjek

Zagreb, 2011.
(Zadnja promjena: 22. prosinca 2021.)

Prva verzija skripata je nastala 2011. godine samoinicijativnim trudom studentica Daniele Landeka i Iris Odak. One su pretipkale svoje bilješke s predavanja te, opet samoinicijativno, izradile sve slike koje sam, a možda i nisam, nacrtao na ploči. Pitanje je kada bi i da li bi ovaj tekst izašao da nije bilo njihovog truda.

Skripta nisu prošla postupak recenzije. Studentima su dostupna putem internih mrežnih stranica Geofizičkog odsjeka sa željom da budu korisna i olakšaju praćenje kolegija Statističke metode u geofizici. Najljepše molim da mi se dojave uočene pogreške.

Sadržaj

Poglavlje 1

UVOD

Ovim predgovorom pokušat ću ukratko objasniti ulogu (i važnost) matematičke statistike i teorije vjerojatnosti u geofizici. Budući da je statistika disciplina koja se, najgrublje rečeno, bavi varijabilnošću, recimo prvo što je to *varijabilnost*. Riječ varijabilnost je sinonim za promjenjivost, nestalnost. No kada u znanosti kažemo da je nešto varijabilno, podrazumijevamo da postoji ili se javlja mnoštvo 'pojava', pri čemu se te pojave ne daju na 'jednostavan' način razumjeti, opisati i/ili sažeti, odnosno da u rečenom mnoštvu ne možemo uočiti neku 'pravilnost'. Statistika, pak, nas uči kako (generalno, tj. bez pretpostavki specifičnih za određeno područje) opisati te kvantificirati varijabilnost; uči nas kako ispravno pristupiti varijabilnosti.

Varijabilnost je prisutna u svim područjima ljudskog djelovanja, u nekima više, u nekima manje. Za geofiziku se može reći da je nerazdvojivo, inherentno, povezana s varijabilnošću. Iako su osnovni fizikalni zakoni u geofizici (bilo da se radi o krutim tijelima bilo da se radi o fluidima) dobro poznati i (relativno) jednostavni, ti se zakoni realiziraju, odnosno ispoljavaju u golemom mnoštvu okolnosti, uz međudjelovanje i ispreplitanje, što rezultira golemom varijabilnošću. Radi se o velikom dijelu ukupne varijabilnosti koju uočavamo (ili ne uočavamo) u prirodi. Tu je varijabilnost pri provođenju geofizičkih mjerenja vrlo teško isključiti u željenoj mjeri i to vrijedi bez obzira koliko eksperiment (opažanja) bio pažljivo planiran i eventualno skup. Npr. pri dizajniranju opažanja imat će se na umu karakteristike instrumenta naspram mjerene fizikalne veličine, zatim druge veličine koje treba simultano mjeriti, trajanje mjerenja, vrijeme mjerenja, npr. dan, noć, sezona, itd, itd. Ipak, nikakvo planiranje neće elimirati činjenicu da instrument mjeri *sve* procese koji se trenutno zbivaju u okruženju, a ne samo one koje bismo htjeli izmjeriti, te da su ti procesi u samom trenutku mjerenja pod utjecajem okruženja. Dio tog utjecaja najčešće je predmet istraživanja, a ostatak je neželjena 'smetnja'. U širem smislu moglo bi

se reći da je osnovna zadaća geofizike izdvojiti 'traženu' informaciju iz golemog mnoštva drugih informacija koje u danom trenutku predstavljaju smetnju ili šum.

Osim potrebe da se varijabilnost opiše i kvantificira, stalno postoji i potreba za donošenjem odluka u prisustvu varijabilnosti. Svojstveno takvim odlukama je da ne možemo biti posve sigurni u njihovu ispravnost. Dakle varijabilnost u sustavu dovodi do neizvjesnosti odnosno nesigurnosti u odlučivanju. Često govorimo o 'procjenama' koje su onda više ili manje pouzdane, više ili manje izvjesne. Svakodnevni primjer imamo u prognozi vremena, bez obzira radi li to profesionalni meteorolog ili običan čovjek s ulice.

Može se (i formalno) pokazati da, u izvjesnom smislu, teorija vjerojatnosti predstavlja jedini ispravni teorijski okvir za razumijevanje i tretiranje neizvjesnosti (engl. *uncertainty*). Stoga se i matematička statistika, a naročito u onom dijelu koji se bavi statističkim procjenjivanjem (engl. *statistical inference*) nužno oslanja na teoriju vjerojatnosti.

Da bi se u praksi primjenilo statističko zaključivanje, nužno je imati (pretpostaviti) određeni model varijabilnosti. Standardni model varijabilnosti na kojem se temelji 'klasična' statistika (koja je predmet i ovog kolegija) počiva na, barem konceptualnoj, mogućnosti ponavljanja pokusa te na dvije pretpostavke:

1) pokusi su istovjetni,

2) pokusi su međusobno nezavisni.

Obje pretpostavke, a naročito druga, rijetko su kada ispunjene u geofizici, što značajno otežava primjenu statističke teorije. Zapravo, prvo pitanje na koje treba odgovoriti prije svake primjene glasi: "Koji je naš model varijabilnosti?". Pri tom treba upotrijebiti svo poznavanje odgovarajuće geofizičke discipline, a isto tako imati i dobro razumijevanje odgovarajućih vjerojatnosnih i statističkih pojmova. Primjena statistike bez da se zna model varijabilnosti znači da je implicitno upotrijebljen klasični model, bilo to dobro ili loše. Zadati ili odrediti razuman model varijabilnosti, delikatan je zadatak, pri čemu klasični model obično služi kao referentni model. Obično se pitamo da li su klasične pretpostavke ispunjene u dovoljnoj mjeri, te ako nisu, zašto nisu, čime ih nadomjestiti i kakve su posljedice.

Da bi student jednoga dana, u praksi, mogao odogovoriti na gornje probleme, potrebno je da u potpunosti shvati i usvoji osnovne vjerojatnosne i statističke pojmove te njihove međusobne odnose. Kada tome dodamo i usvajanje odgovarajućeg načina razmišljanja, dobili smo glavne ciljeve kolegija. Kolegij stoga počinje razmjerno detaljnim ponavljanjem gradiva koje su studenti učili na prvoj godini Studija i nastavlja pojedinim temama koje su više ili manje okrenute primjeni u geofizici. Svugdje se prednost daje 'dubini' pred 'širinom' te, koliko je moguće, naglašava geofizički kontekst.

1.1 Vjerojatnost

1.1.1 Osnovni pojmovi

Definirat ćemo najprije neke osnovne pojmove koje ćemo koristiti:

- ***slučajni pokus*** - pokus čiji ishod nije potpuno predvidiv
- ***događaj*** - skup svih mogućih ishoda nekog slučajnog pokusa, može biti jedostavan (elementaran) ili složen
- ***prostor događaja*** - Ω, skup svih mogućih elementarnih događaja vezanih uz neki slučajni pokus; najveći mogući složeni događaj.

Aksiomi vjerojatnosti (Kolmogorov, 1930.): Vjerojatnost je fukcija P koja svakom događaju A pridružuje broj $P(A)$, tako da vrijedi:

1. $\forall A \quad P(A) \geq 0$,
2. $P(\Omega) = 1$,
3. $A \bigcap B = \emptyset \quad \Longrightarrow \quad P(A \bigcup B) = P(A) + P(B)$.

Aksiomi, kao ni teorija vjerojatnosti, ne kažu koja je vjerojatnost od $P(A)$, već samo koja svojstva vjerojatnost mora imati.

Intuitivno značenje vjerojatnosti:

- *Vjerojatnost a'priori*: Ako promatramo 'pokus' koji ima konačno, konkretno n, ishoda koje označimo s $x_1, \ldots x_n$, te nemamo razloga preferirati niti jedan od njih, onda definiramo vjerojatnost $P(x_i) = \frac{1}{n}$. Radi se o tzv. principu nedovoljnog razloga koji potječe još od Jacoba Bernoullija (1645-1705).
- *Interpretacija preko čestina* (frekvencijski pristup): Vjerojatnost događaja je njegova relativna čestina u dugačkom nizu ponavljanja. Ako ponovimo pokus n puta, a događaj A se dogodi n_A puta onda je $f(A) = n_A/n$ relativna čestina od A. Tada je

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

Preciznije to možemo iskazati pomoću ***zakona velikih brojeva***:

$$\forall \varepsilon > 0, \quad \widetilde{P}\left|\frac{n_A}{n} - P(A)\right| > \varepsilon \xrightarrow[n\to\infty]{} 0.$$

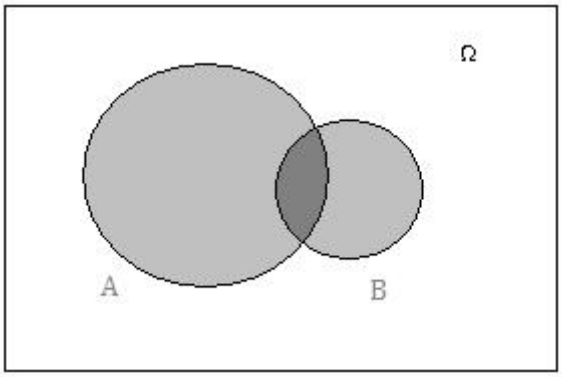
Fiksiramo ε i uzmemo dovoljno veliki n. Napravimo *puno serija* od n pokusa i svaki put izračunamo $\frac{n_A}{n}$. Među dobivenim čestinama će biti malo onih koji od $P(A)$ odstupaju više od ε. Vjerojatnost pojave takvih čestina teži k nuli kako n raste.

- *Subjektivna interpretacija* (Bayesov pristup): $P(A)$ se definira kao osobni (subjektivni) stupanj uvjerenja (*degree of belief*) u pojavljivanje događaja A. Subjektivne vjerojatnosti moraju biti usklađene (konzistentne), tj. ne mogu biti posve proizvoljne. I one moraju udovoljavati određenim pravilima (pravilu sume i pravilu produkta) iz kojih se mogu dobiti i aksiomi vjerojatnosti.

Zakoni vjerojatnosti su univerzalni i ne ovise o interpretaciji.

Svojstva vjerojatnosti: Za $A \subseteq \Omega$, pišemo $A^C := \Omega \backslash A = \{A \text{ se nije dogodio}\}$.

1. $A \cap A^C = \emptyset \Longrightarrow P(A^C) = 1 - P(A)$
2. $0 \leq P(A) \leq 1$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (*prošireno pravilo sume*)



Dokaz za 3: Sa slike se može vidjeti da vrijedi

$$A \cup B = A \cup (B \setminus A),$$

$$B = (A \cap B) \cup (B \setminus A).$$

Tada je prema 3. aksiomu vjerojatnosti

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

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

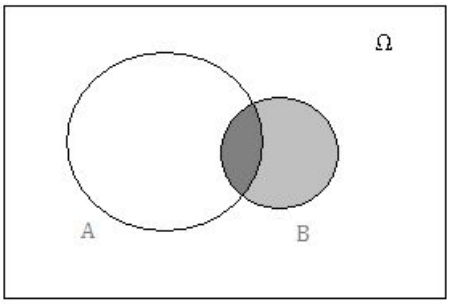
Kombinacijom dva prethodna izraza dobije se upravo pravilo sume.

1.1.2 Uvjetna vjerojatnost i Bayesov teorem

Uvjetna vjerojatnost: Zanima nas kako činjenica da znamo (ili pretpostavljamo) da se zbio događaj B utječe na vjerojatnost drugog događaja A. Uvjetna vjerojatnost A uz uvjet B je

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

B je ovdje uvjetni događaj.



$P(A|B)$ zadovoljava aksiome vjerojatnosti ako za Ω uzmeno događaj B. Također vrijedi:

1. $P(B|B) = 1$
2. $A \cap B = \emptyset \implies P(A|B) = 0$
3. $P(A \cap B) = P(A|B) \cdot P(B) = P(A|B) \cdot P(A)$ (*pravilo množenja*)

Nezavisnost dva događaja: Dva događaja A i B su nezavisni ako pojavljivanje ili nepojavljivanje jednoga od njih ne mijenja vjerojatnost pojave drugog, tj. $P(A|B) = P(A)$. Odatle je

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

Drugim riječima, pojavljivanje ili nepojavljivanje od B ne mijenja neizvjesnost u vezi pojavljivanja od A.

Zakon potpune vjerojatnosti: Neka je Ω prostor događaja, te $A_i \subseteq \Omega, i = 1, 2, \ldots, n$ skup događaja takav da je $\bigcup_{i=1}^{n} A_i = \Omega$, te $A_i \cup A_j = \emptyset$ za $i \neq j$. Tada za $D \subseteq \Omega$ vrijedi:

$$P(D) = \sum_{i=1}^{n} P(D \cap A_i) = \sum_{i=1}^{n} P(D|A_i) \cdot P(A_i).$$

Bayesov teorem: Omogućava „okretanje" uvjetnih vjerojatnosti, tj. iz poznavanja *svih* $P(D|A_i)$ može se izračunati $P(A_i|D)$ $\forall i$. Vrijedi:

$$P(D \cap A_i) = P(D|A_i) \cdot P(A_i) = P(A_i|D) \cdot P(D)$$

$$P(A_i|D) = \frac{P(D|A_i) \cdot P(A_i)}{P(D)}$$

$$P(A_i|D) = \frac{P(D|A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(D|A_j) \cdot P(A_j)},$$

pri čemu su $P(A_i)$ početne vjerojatnosti (*prior*), $P(D)$ činjenice, dokazi (*evidence*), $P(D|A_i)$ vjerodostojnosti (*likelihood*) od D ako vrijedi A_i, dok su $P(A_i|D)$ aposteriorne ili naknadne vjerojatnosti od A_i (*posterior*), tj. vjerojatnosti nakon što su opaženi podaci D.

Promotrimo poseban slučaj $n = 2$. Tada je $A_1 = A$, $A_2 = A^C$, te definiramo:
Šansa ili ***izgledi*** (*odds*) za A:

$$\sigma(A) = \frac{P(A)}{P(A^C} = \frac{P(A)}{1 - P(A)},$$

$$\sigma(A|D) = \frac{P(A|D)}{P(A^C|D)},$$

$$P(A|D) = \frac{P(D|A) \cdot P(A)}{P(D)}, \quad P(A^C|D) = \frac{P(D|A^C) \cdot P(A^C)}{P(D)} \quad \Longrightarrow$$

$$\sigma(A|D) = \underbrace{\frac{P(D|A)}{P(D|A^C)}}_{\textbf{\textit{Bayesov faktor}}} \sigma(A)$$

Bayesov faktor je kvocijent vjerodostojnosti od A i A^C uz prisustvo (podatka) D. Ako podaci D više podupiru A nego A^C, onda je $\sigma(A|D) > \sigma(A)$ i obrnuto. Želimo li iz (podatka) D procjeniti da li vrijedi A, tj. zanima nas $P(A|D)$, veličina $P(D|A)$ nije bitna sama za sebe nego u usporedbi s $P(D|A^C)$.

Odnos A i D potpuno je opisan *tablicom združenih vjerojatnosti:*

(I)	A	A^C	$\sum$
D	$P(D \cap A)$	$P(D \cap A^C)$	$P(D)$
D^C	$P(D^C \cap A)$	$P(D^C \cap A^C)$	$P(D^C)$
$\sum$	$P(A)$	$P(A^C)$	1

uz koju vežemo i *tablice uvjetnih vjerojatnosti:*

(II)	A $P(A)$	A^C $P(A^C)$
D	$P(D\|A)$	$P(D\|A^C)$
D^C	$P(D^C\|A)$	$P(D^C\|A^C)$
$\sum$	1	1

tablica vjerodostojnosti

(III)		A	A^C	$\sum$
D	$P(D)$	$P(A\|D)$	$P(A^C\|D)$	1
D^C	$P(D^C)$	$P(A\|D^C)$	$P(A^C\|D^C)$	1

tablica naknadnih vjerojatnosti

1.1.3 Perzistencija kao uvjetna vjerojatnost

Kada se za neku postaju i neki mjesec pogleda oborina po danima obično se vide grupe dana s oborinom, te dana bez oborine. To upućuje na (statističku) povezanost koja se u meteorologiji naziva *perzistencija*. Ujedno to znači da odgovarajući događaji nisu, statistički gledano, nezavisni. Označimo događaje:

0 – suho danas	1 – kišno danas
$\overline{0}$ – suho jučer	$\overline{1}$ – kišno jučer

Klimatološke vjerojatnosti:

$$p_0 = P(0) = P(\overline{0}) = p_{\overline{0}} \qquad p_1 = P(1) = P(\overline{1}) = p_{\overline{1}} \tag{1.1}$$

Tablice združenih, te uvjetnih vjerojatnosti:

$$P(\overline{0} \cap 1) = P(\overline{1} \cap 0), \tag{1.2}$$

(I)	0	1	$\sum$
$\overline{0}$	$P(\overline{0} \cap 0)$	$P(\overline{0} \cap 1)$	p_0
$\overline{1}$	$P(\overline{1} \cap 0)$	$P(\overline{1} \cap 1)$	p_1
$\sum$	p_0	p_1	1

(II)	0	1
	p_0	p_1
$\overline{0}$	$P(\overline{0}\|0)$	$P(\overline{0}\|1)$
$\overline{1}$	$P(\overline{1}\|0)$	$P(\overline{1}\|1)$
$\sum$	1	1

(III)		0	1	$\sum$
$\overline{0}$	p_0	$P(0\|\overline{0})$	$P(1\|\overline{0})$	1
$\overline{1}$	p_1	$P(0\|\overline{1})$	$P(1\|\overline{1})$	1

$P(\overline{0}|1)$ – broj početaka grupa s oborinama
$P(0|\overline{1})$ – broj krajeva grupa s oborinama

zbog (1.1) i (1.2) je $P(\overline{i}|j) = P(i|\overline{j}) =: p_{ij},\ \forall i, j = 0, 1$, pri čemu je p_{ij} tzv. *vjerojatnost prijelaza* iz (stanja) j u (stanje) i. Vrijedi:

$$p(i \cap j) = p_{ji} \cdot p_j = p_{ij} \cdot p_i, \qquad \forall i, j = 0, 1,$$

pa se tablice (II) i (III) svode na:

(II)	0	1
	p_0	p_1
$\overline{0}$	p_{00}	p_{10}
$\overline{1}$	p_{01}	p_{11}
$\sum$	1	1

(III)		0	1	$\sum$
$\overline{0}$	p_0	p_{00}	p_{01}	1
$\overline{1}$	p_1	p_{10}	p_{11}	1

,

dok je $a = \left(\begin{smallmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{smallmatrix}\right)$ tzv. *matrica prijelaza*.

Perzistencija se može definirati i brojčano kao $C = p_{11} - p_{01} = p_{00} - p_{10}$. $C > 0$ znači da pojava (1) češće slijedi pojavu $(\overline{1})$ nego nepojavu $(\overline{0})$. Vrijedi:

a) $-1 \leq C \leq 1 \qquad$ (zbog $0 \leq p_{11}, p_{01} \leq 1$),

b) $p_1 = P(\overline{0} \cap 1) + P(\overline{1} \cap 1) = p_{01} \cdot p_0 + p_{11} \cdot p_1$. Odatle je

$$C \geq 0 \;\Leftrightarrow\; p_1 \leq p_{11}(p_0 + p_1) = p_{11}, \quad \text{tj.} \quad C \geq 0 \;\Leftrightarrow\; p_1 \leq p_{11},$$

tj. vjerojatnost pojave (1) nakon pojave $(\overline{1})$ veća je od klimatološke vjerojatnosti. Analogno $p_0 = P(\overline{0} \cap 0) + P(\overline{0} \cap 1) = p_{00} \cdot p_0 + p_{10} \cdot p_1$, odakle je

$$C \geq 0 \;\Leftrightarrow\; p_0 \leq p_{00}(p_0 + p_1) = p_{00} \quad \text{tj.} \quad C \geq 0 \;\Leftrightarrow\; p_0 \leq p_{00}.$$

c) $C = 0 \Leftrightarrow p_1 = p_{11} = p_{01}$, a također i $p_0 = p_{00} = p_{10}$. Odatle je:

$$P(\overline{0} \cap 0) = p_{00} \cdot p_0 = p_0^2,$$
$$P(\overline{1} \cap 0) = p_{10} \cdot p_1 = p_1 p_0,$$
$$P(\overline{0} \cap 1) = p_{10} \cdot p_0 = p_0 p_1,$$
$$P(\overline{1} \cap 1) = p_{11} \cdot p_1 = p_1^2,$$

iz čega izlazi da su svi događaji "danas" i "sutra" nezavisni.

d) $C = -1$ ili $C = 1 \rightarrow$ ekstremni slučajevi (nisu fizikalni).

Klime s visokom perzistencijom ($C \approx 1$) su neugodne (npr. pustinje).

1.2 Slučajne varijable i funkcije distribucije vjerojatnosti

1.2.1 Definicije i osnovna svojstva

Vezano uz ishode pokusa obično vršimo neke matematičke i logičke operacije. Praktično je, a najčešće i prirodno, ishode slučajnih pokusa opisivati brojevima.

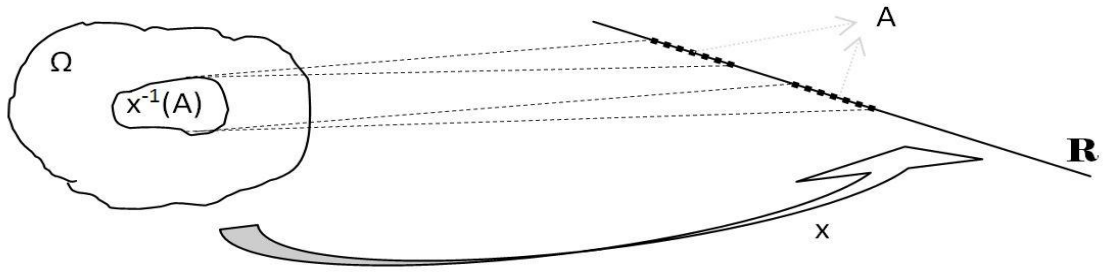
Slučajna varijabla je funkcija koja svakom elmentarnom događaju (ishodu) pridružuje realan broj. Pišemo:

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

Na Ω je zadana vjerojatnost. Tek (Ω, P) je vjerojatnosni prostor i upravo ta činjenica razlikuje slučajne varijable od običnih funkcija. Zbog toga je moguće za svaki skup vrijednosti slučajne varijable odrediti koliko je vjerojatan. Za $A \subset \mathbb{R}$ imamo

$$\text{Vj}(A) \equiv P_X(A) = P\{X^{-1}(A)\} \equiv P\{X \in A\}.$$

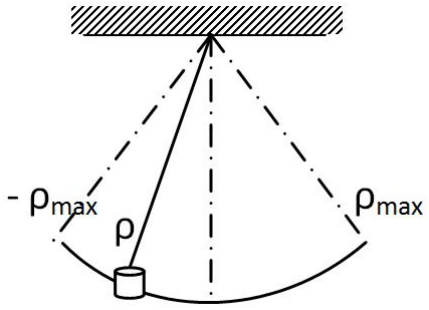
P_X je vjerojatnost na skupu svih realnih brojeva. Dakle, svaka slučajna varijabla inducira vjerojatnost na $\mathbb{R}$. Vjerojatnost *skupa vrijednosti* $A \subset \mathbb{R}$ je jednaka vjerojatnosti *događaja* $X \in A$. Stoga se često sam $\mathbb{R}$ uzima kao prostor događaja (*sample space*).



$P_X \to$ vjerojatnost na $\mathbb{R}$.

Napomena: Slučajna varijabla nema veze sa intuitivnim pojmom "slučajnosti". Naprosto, radi se o funkciji čijim vrijednostima znamo pridružiti vjerojatnosti, odnosno čestine.

Primjer: matematičko njihalo



Zanima nas položaj (otklon) njihala u vremenu, $\rho(t)$

1. Deterministički pristup – $\rho(t)$ je rješenje neke diferencijalne jednadžbe.

2. Vjerojatnosni pristup – ρ je slučajna varijabla s vrijednostima u $[-\rho_{max}, \rho_{max}]$. $P\{\rho \in [\rho_1, \rho_2]\}$ je vjerojatnost da se njihalo (tj. otklon ρ) nađe u intervalu $[\rho_1, \rho_2]$.

Ako je X slučajna varijabla krajnji (maksimalni) cilj je poznavati $P\{X \in A\}$ za $\forall A \subseteq \mathbb{R}$. Najčešće je vjerojatnosni prostor složen (kompliciran) objekt. No struktura realnih brojeva je dobro usklađena sa zakonima vjerojatnosti. Zato se vjerojatnost P_X može naročito jednostavno zadati. Umjesto da je zadajemo za $\forall A \subseteq \mathbb{R}$, dovoljno je poznavati funkciju $F_X : \mathbb{R} \to \mathbb{R}$, definiranu sa

$$F_X(x) = P\{X \leq x\} = P\{X \in <-\infty, x]\} = P_X(<-\infty, x])$$

Iz F_X možemo odrediti $P_X(A)$ za $\forall A \subseteq \mathbb{R}$. F_X je *(kumulativna) funkcija distribucije (razdiobe) vjerojatnosti* slučajne varijable X.

Svojstva:

1. $\lim\limits_{x \to -\infty} F_X(x) = \lim\limits_{x \to -\infty} P_X(<-\infty, x]) = 0,$
2. $\lim\limits_{x \to +\infty} F_X(x) = 1,$
3. Monotonost: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \quad$ (zbog $P(A) \geq 0, \quad \forall A$).

Napomena: Slučajne varijable (funkcije) označavamo velikim, štampanim slovima $X, Y, \ldots$, a vrijednosti koje te varijable poprimaju (realni brojevi) malim pisanim slovima $x, y, \ldots$. Nadalje, vjerojatnost je definirana na događajima, a ne na varijablama. Npr. ispravno je pisati $P\{X \leq x\}$, dočim $P(X)$, ili $P(x)$ nije ispravno, ili je u najmanju ruku neprecizno.

Ovisno o vrijednostima koje mogu poprimiti, razlikujemo dvije vrste slučajnih varijabli:

a) ***Diskretna slučajna varijabla*** je ona koja može poprimiti *konačno ili najviše prebrojivo beskonačno* vrijednosti. Zadaje se skupom vrijednosti koje može poprimiti i pripadnim vjerojatnostima, tj. tablicom oblika:

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix}, \qquad \sum_i p_i = 1,$$

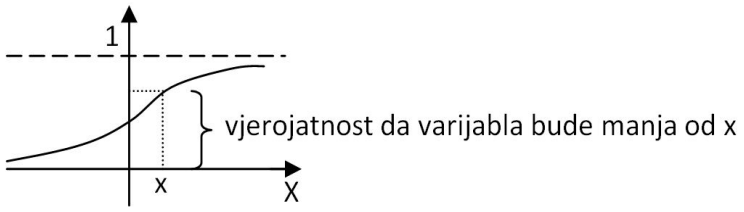
koja predstavlja *zakon razdiobe* disretne slučajne varijable. Očito je:

$$X : \Omega \to \{x_i, \quad i = 1, 2, \ldots\} \subseteq \mathbb{R}$$

$$P\{X = x_i\} = p_i, \quad \forall i, \qquad P\{X = \tilde{x}\} = 0, \quad \tilde{x} \neq x,$$

$$F_X(x) = P\{X \leq x\} = \sum_{\substack{i \\ x_i \leq x}} p_i.$$

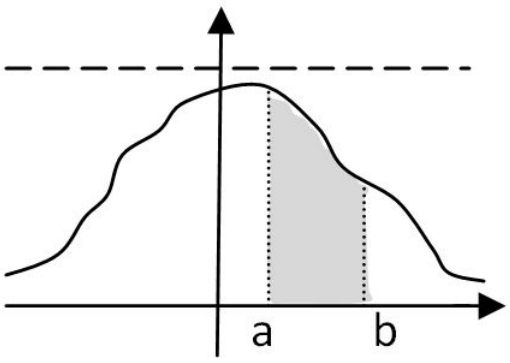
b) ***Kontinuirana slučajna varijabla*** je ona koja može poprimiti sve vrijednosti iz nekog intervala realnih brojeva ili pak cijelog $\mathbb{R}$. Funkcija distribucije tipično izgleda:



Ako je F_X derivabilna, tj. postoji funkcija $f_X : \mathbb{R} \to \mathbb{R}$ takva da je $F_X' = f_X$, tj.

$$F_X = \int_{-\infty}^{x} f_X(x')dx'$$

onda je f_X *funkcija gustoće vjerojatnosti* slučajne varijable X. Tipični oblik od f_X:



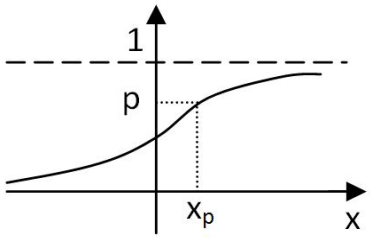
$$P\{a < X \leq b\} = F_X(b) - F_X(a) =$$
$$= \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx$$
$$= \int_{a}^{b} f_X(x)dx$$

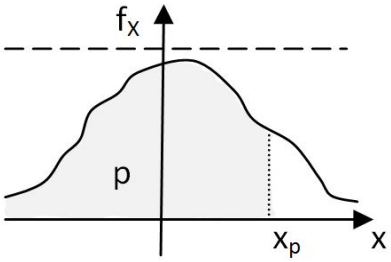
Nužni uvjeti za funkciju gustoce su: $f_X \geq 0$ i $\int_{\mathbb{R}} f_X(x)dx = 1$

Napomena: Funkcija gustoće vjerojatnosti kod kontinuiranih slučajnih varijabli odgovara zakonu razdiobe kod diskretnih slučajnih varijabli. Zbog $P\{X = x_0\} = 0$ nema smisla govoriti o pojedinačnim vrijednostima, već o (po volji) malim intervalima.

Primjer: Ako je temperatura $T = 23^o\text{C}$ to znači da imamo događaj: $\{T \in [23^o\text{C} - \Delta t/2, 23^o\text{C} + \Delta t/2]\}$, gdje je Δt preciznost termometra.

Kvantili. Neka je X kontinuirana slučajna varijabla. Za $p \in [0, 1], p$-ti kvantil je broj x_p takav da je $P\{X \leq x_p\} = p$, tj. $x_p = F_X^{-1}(p)$.





Medijan je 0.5-ti kvantil od X, tj. to je broj $m = x_{0.5}$, takav da je $P\{X \leq m\} = P\{X \geq m\} = 1/2$. U diskretnom slučaju, definicuju treba malo prilagoditi.

1.2.2 Matematičko očekivanje i momenti

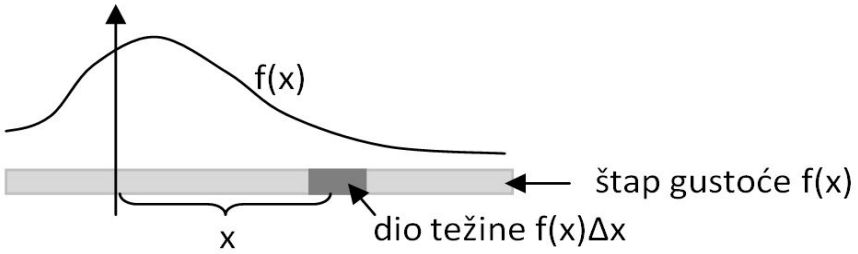
Za slučajnu varijablu X, *matematičko očekivanje* ili *srednja vrijednost* je broj:

$$\mathbb{E}(X) = \int xf(x)dx \quad \text{(kontinuirani slučaj)},$$

$$\mathbb{E}(X) = \sum_i x_i p_i \quad \text{(diskretni slučaj)}.$$

Interpretacija:

1. $\mathbb{E}(X)$ je otežani srednjak. Težine su pripadne vjerojatnosti.
2. Kod igara na sreću, matematičko očekivanje predstavlja očekivanu dobit ili (najčešće) gubitak.
3. Fizička interpretacija - težište nehomogenog štapa:



$xf(x)\Delta x =$ moment težine u odnosu na ishodište
$\overline{X} = \mathbb{E}(X) = \int xf(x)dx \quad \rightarrow$ težište (zbog $\int f(x)dx = 1$)

Svojstva očekivanja:

1. $\mathbb{E}(c) = c, \quad c$ je konstanta,

2. $\mathbb{E}(c \cdot X) = c \cdot \mathbb{E}(X)$ $\quad \to$ homogenost,

3. $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$ $\quad \to$ aditivnost.

Naravno, svojstvo homogenosti i aditivnosti zajedno čine svojstvo linearnosti.

Očekivanje funkcije slučajne varijable

Neka je $g : \mathbb{R} \to \mathbb{R}$ funkcija te $X : \Omega \to \mathbb{R}$ slučajna varijabla. Ako je

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

onda je $g(X) : \Omega \to \mathbb{R}$ opet slučajna varijabla čiji zakon razdiobe glasi:

$$g(X) \sim \begin{pmatrix} g(x_1) & g(x_2) & \dots \\ p_1 & p_2 & \dots \end{pmatrix}.$$

Posljedično,

$$\mathbb{E}(g(X)) = \sum_i g(x_i)p_i.$$

Analogno, za kontinuiranu slučajnu varijablu s gustoćom $f(x)$ vrijedi

$$\mathbb{E}(g(X)) = \int g(x)f_X(x)dx$$

.

Momenti su veličine koje daju djelomičnu informaciju o slučajnoj varijabli, odnosno o njenoj razdiobi. Neka je X slučajna varijabla s gustoćom $f(x)$. Tada definiramo *momente oko nule*:

$\mu_0 = \mathbb{E}(X^0) = 1 \qquad \to$ red 0,
$\mu_1 = \mathbb{E}(X^1) = \overline{X} \qquad \to$ red 1,
$\mu_2 = \mathbb{E}(X^2) = \int x^2 f(x)dx \quad \to$ red 2, itd.

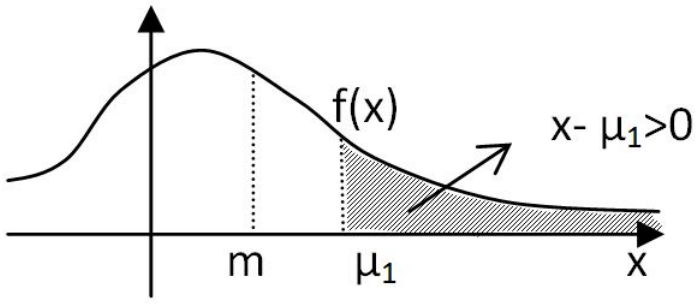
Postoje također i *momenti oko sredine* ili *centralni momenti*:

$m_0 = \mathbb{E}((X - \mu_1)^0) = 1,$
$m_1 = \mathbb{E}((X - \mu_1)^1) = \mathbb{E}(X) - \mathbb{E}(\mu_1) = 0, \quad \longrightarrow$ moment sile u odnosu na težište iščezava, pri čemu je μ_1 srednja vrijednost ili težište (mjera *položaja*).
$m_2 = \mathbb{E}((X - \mu_1)^2) =$ (koristeći svojstva očekivanja) $=$

$$= \mathbb{E}(X^2) - 2\mu_1\mathbb{E}(X) + \mathbb{E}(\mu_1^2) =$$
$$= \mu_2 - 2\mu_1^2 + \mu_1^2 = \mu_2 - \mu_1^2 =$$
$$= \mathrm{Var}(X) = \sigma^2, \quad \longrightarrow \text{varijanca (mjera } \textit{raspršenja}\text{), dočim je}$$
σ standardna devijacija.

$m_3 = \mathbb{E}((X-\mu_1)^3)$ $\longrightarrow$ treći moment oko sredine.

Ako je razdioba simetrična, tj. vrijedi $f(\mu_1 + x) = f(\mu_1 - x)$, onda je očito $m_3 = 0$. Promotrimo $m_3 = \int (x-\mu_1)^3 f(x)dx$ u slučaju kada je razdioba asimetrična:



Sveukupno je vjerojatnost $P\{X > \mu_1\} < P\{X < \mu_1\}$, ali "dugi i debeli rep" na desnoj strani će nadvladati sveukupno "deblju" lijevu stranu.

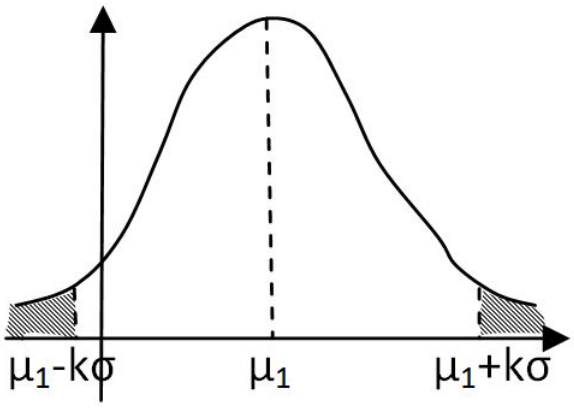
Ovdje očekujemo da vrijedi $m_3 > 0$ i u tom smislu m_3 je mjera *asimetrije*. Normiranjem se dobiva bezdimenzionalna veličina $\alpha_3 = m_3/\sigma^3$. Dakle, $\alpha_3 > 0$ sugerira da postoji rep u desno, a $\alpha_3 < 0$, rep u lijevo.

$m_4 = \mathbb{E}((X-\mu_1)^4)$ $\rightarrow$ mjera *spljoštenosti.*

Chebishevljeva nejednakost: Neka je X slučajna varijabla, te neka postoje $\mu_1 = \mathbb{E}(X)$, $\sigma^2 = \mathrm{Var}(X)$. Tada za svaki $k > 0$ vrijedi:

$$P\{|X - \mu_1| \geq k\sigma\} \leq \frac{1}{k^2}$$

Nejednakost govori: "Debljina" repova je kontrolirana varijancom.



Dokaz (u kontinuiranom slučaju):

Pomoćna tvrdnja: Ako je Y slučajna varijabla takva da je $Y \geq 0$, onda $\forall K > 0$

vrijedi:

$$P\{Y \geq K\} \leq \frac{\mathbb{E}(Y)}{K}.$$

Zaista

$$\mathbb{E}(Y) = \int_0^\infty y f(y)dy \geq \int_K^\infty y f(y)dy \geq K \int_K^\infty f(y)dy = K \cdot P\{Y \geq K\}.$$

Sada stavimo: $Y = (X - \mu_1)^2, \quad \mathbb{E}(Y) = \mathrm{Var}(X), \quad K = k^2\sigma^2$

$$P\{|X - \mu_1|^2 \geq k^2\sigma^2\} \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2},$$

pa dobivamo

$$P\{|X - \mu_1| \geq k\sigma\} \leq \frac{1}{k^2}.$$

1.3 Slučajni vektori i višedimenzionalne funkcije distribucije

1.3.1 Definicije i osnovna svojstva

Dvodimenzinalni slučajni vektor je par slučajnih varijabli definiranih na istom vjerojatnosnom prostoru, $(X, Y) : \Omega \to \mathbf{R}^2$. Vezuje se uz zajedničko mjerenje dviju veličina. Za svaki elementarni događaj $\omega \in \Omega$ imamo dva rezultata, $X(\omega)$ i $Y(\omega)$, ali se ne radi o dva (elementarna) događaja.

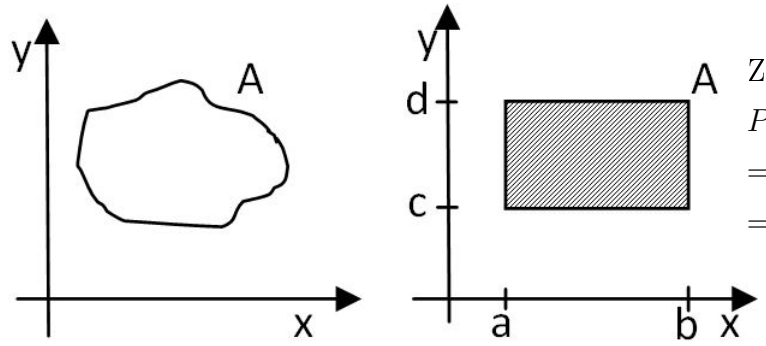
Nećemo definirati funkciju kumulativne razdiobe ($F_{X,Y}$) jer nije praktična, već definiramo zakon razdiobe, i to posebno za diskretne te posebno za kontinuirane varijable.

Neka su X i Y diskretne varijable. Neka X poprima vrijednosti $x_1, x_2, ...$, a Y vrijednosti $y_1, y_2, ...$ Slučajni vektor (X, Y) tada poprima vrijednosti u skupu $\{(x_i, y_j) : i, j = 1, 2...\}$. Označimo s $p_{ij} = P\{(X, Y) = (x_i, y_j)\} = P\{X = x_i \text{ i } Y = y_j\}$, združenu vjerojatnost da bude $X = x_i$ i istovremeno $Y = y_j$. *Zakon razdiobe* od (X, Y) je zadan tablicom združenih vjerojatnosti:

	y_1	y_2	$\dots$	$\sum$
x_1	p_{11}	p_{12}	$\dots$	$p_{1\bullet}$
x_1	p_{21}	p_{22}	$\dots$	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\sum$	$p_{\bullet 1}$	$p_{\bullet 2}$	$\dots$	1

, $\qquad \sum_{ij} p_{ij} = 1.$

Ako su X i Y kontinuirane slučajne varijable, zakon razdiobe od (X,Y) zadajemo (*združenom*) *funkcijom gustoće*. To je funkcija od dvije varijable $f:\mathbb{R}^2\to\mathbb{R}$, takva da za $\forall A\subseteq\mathbb{R}^2$ vrijedi: $P\{(X,Y)\in A\}=\iint\limits_A f(x,y)dxdy$.



Za $A=[a,b]\times[c,d]$
$P\{(X,Y)\in A\}=$
$=P\{a\le X\le b \text{ i } c\le Y\le d\}=$
$=\int_a^b\int_c^d f(x,y)dxdy$

Marginalne gustoće slučajnog vektora (X,Y) su obične gustoće vjerojatnosti slučajnih varijabli X i Y zasebno. Neka je $A\subseteq\mathbb{R}$. Tada za gustoću f_X, od X vrijedi:

$$\int_A f_X(x)dx=P\{x\in A\}=P\{x\in A \text{ i } Y\in\mathbb{R}\}=\int_A dx\underbrace{\int_{-\infty}^{\infty}dyf(x,y)}_{f_X(x)}.$$

Dakle,

$$f_X(X)=\int_{-\infty}^{\infty}f(x,y)dy.$$

Analogno je i

$$f_Y(Y)=\int_{-\infty}^{\infty}f(x,y)dx.$$

U diskretnom slučaju marginalne razdiobe su dane s:

$$P\{X=x_i\}=\sum_j p_{ij}=p_{i\bullet},$$

$$P\{Y=y_i\}=\sum_i p_{ij}=p_{\bullet j}.$$

Ako znamo združene gustoće slučajnog vektora, potpuno su određene marginalne gustoće, ali OBRAT NE VRIJEDI.

Uvjetne razdiobe vjerojatnosti

Neka je (X,Y) slučajni vektor.

Diskretni slučaj: Neka je $P\{X=x_i,Y=y_j\}=p_{ij}$ zakon razdiobe. Tada je

$$P\{X=x_i|Y=y_j\}=\frac{P\{X=x_i\cap Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\bullet j}}.$$

Odatle je $\begin{pmatrix} x_1 & x_2 & \dots \\ \frac{p_{1j}}{p_{\bullet j}} & \frac{p_{2j}}{p_{\bullet j}} & \dots \end{pmatrix}$ zakon razdiobe od X uz uvjet $Y = y_j$. Analogno vrijedi i za Y.

Kontinuirani slučaj: Neka je $f(X,Y)$ je funkcija gustoće vjerojatnosti. Za fiksno $y_0 \in \mathbb{R}$ i proizvoljni x te $h > 0$ gledamo:

$$\begin{aligned} P\{X < x \mid y_0 \le Y < y_0 + h\} &= \frac{P\{X < x \text{ i } y_0 \le Y < y_0 + h\}}{P\{y_0 \le Y < y_0 + h\}} = \\ &= \frac{\int_{y_0}^{y_0+h} dy' \int_{-\infty}^{x} dx' f(x', y')}{\int_{y_0}^{y_0+h} dy' \int_{-\infty}^{\infty} dx' f(x', y')} \to \\ &\Big(\text{budući da } \frac{1}{h}\int_{y_0}^{y_0+h} f(y')dy' \to f(y_0) \text{ kada } h \to 0\Big) \\ &\to \frac{\int_{-\infty}^{x} f(x', y')dx'}{\int_{-\infty}^{\infty} f(x', y')dx'} = \frac{\int_{-\infty}^{x} f(x', y_0)dx'}{f_Y(y_0)} = \\ &= P\{X < x \mid Y = y_0\} = F_{X|Y=y_0}(x). \end{aligned} \tag{1.3}$$

Time je dobivena *uvjetna kumulativna distribucija od X uz uvjet $Y = y$*. Deriviramo li posljednji izraz u (1.3) po x, dobit ćemo uvjetnu gustoću od X uz uvjet $\{Y = y_0\}$:

$$f_{X|Y=y_0}(x) = \frac{f(x, y_0)}{f_Y(y_0)}.$$

1.3.2 Nezavisnost i nekoreliranost

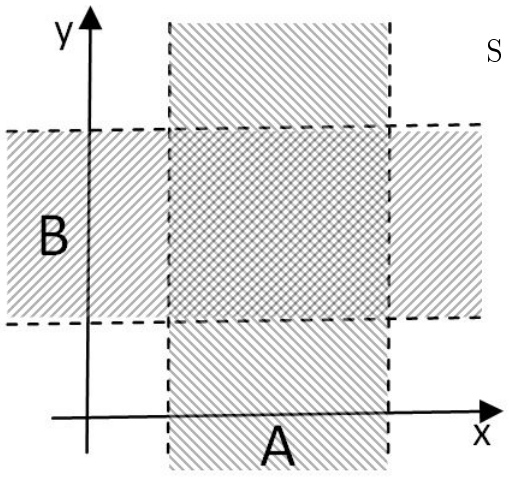
Ako je $A \subseteq \mathbb{R}$, te X slučajna varijabla, za događaj $\tilde{A} = \{X \in A\}$ kažemo da je vezan uz slučajnu varijablu X. ***Slučajne varijable X i Y su nezavisne ako su bilo koja dva događaja vezana uz njih međusobno nezavisna, tj. ako za svaki A i $B \subseteq \mathbf{R}$ vrijedi:***

$$P\{X \in A \text{ i } Y \in B\} = P\{X \in A\} \cdot P\{Y \in B\}, \quad \text{ili}$$

$$P\{\tilde{A} \cap \tilde{B}\} = P\{\tilde{A}\} \cdot P\{\tilde{B}\}.$$

Što nezavisnost znači za funkciju združene gustoće vjerojatnosti? S jedne strane je:

$$P\{X \in A \text{ i } Y \in B\} = \iint_{AxB} f(x,y)dxdy.$$



S druge strane je:

$$P\{X \in A \quad i \quad Y \in B\} =$$

$$\int_A f_X(x)dx \cdot \int_B f_Y(y)dy =$$

$$= \iint\limits_{AxB} f_X(x)f_Y(y)dxdy.$$

(Fubinijev teorem)

Posljedično, varijable su nezavisne ako za $\forall A, B$ vrijedi:

$$\iint\limits_{AxB} f(x,y)dxdy = \iint\limits_{AxB} f_X(x)f_Y(y)dxdy,$$

odakle izlazi:

$$f(x,y) = f_X(x) \cdot f_Y(y) \Leftrightarrow X \text{ i } Y \text{ su nezavisne.}$$

Smisao nezavisnosti je da bilo kakva informacija o jednoj varijabli ne utječe na vjerojatnosni prostor koji odgovara drugoj varijabli (ne mijenja neizvjesnost u vezi druge varijable). Posebno, nezavisnost u statistici nema veze s uzročno-posljedičnim odnosima. Ipak, ako postoji uzročno-posljedična veza, događaji ne mogu biti nezavisni.

Kovarijanca varijabli X ***i*** Y (mješoviti drugi moment oko sredine)

Neka je (X,Y) slučajni vektor te neka je $\mu_x = \mathbb{E}(X)$, $\mu_y = \mathbb{E}(Y)$. *Kovarijanca* varijabli X i Y je

$$\text{Cov}(X,Y) := \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \text{(u kontinuiranom slučaju)} =$$

$$= \iint (x-\mu_x)(y-\mu_y)f(x,y)dxdy.$$

Intuitivno značenje: Kovarijanca mjeri sklonost varijabli X i Y da simultano poprimaju vrijednosti veće od svojih srednjaka ili manje od njih.

Kovarijanca ovisi o jedinicama mjere. Ako je $\text{Cov}(X,Y) = 0$ kažemo da su X i Y *nekorelirane* (u protivnom su korelirane).

Momenti funkcije slučajnog vektora

Neka je g funkcija dvije varijable, $g : \mathbb{R}^2 \to \mathbb{R}$, te (X,Y) slučajni vektor s gustoćom

$f(x,y)$. Tada je $g(X,Y)$ također slučajna varijabla i vrijedi

$$\mathbf{E}(g(X,Y)) = \iint g(x,y)\cdot f(x,y)dxdy$$

tvrdnja: Ako su X i Y nezavisne onda $\forall k,l \in \mathbb{N}$ vrijedi $\mathbb{E}(X^k \cdot Y^l) = \mathbb{E}(X^k)\cdot\mathbb{E}(Y^l)$.

Dokaz (za kontinuirani slučaj): Neka je $f(x,y) = f_X(x)\cdot f_Y(y)$ funkcija gustoće slučajnog vektora (X,Y). Tada je:

$$\begin{aligned}\mathbb{E}(X^kY^l) &= \iint\limits_{\mathbb{R}^2} x^k y^l f(x,y)dxdy = \\ &= \int x^k f_X(x)dx \int y^l f_Y(y)dy = \\ &= \mathbb{E}(X^k)\cdot\mathbb{E}(Y^l).\end{aligned}$$

tvrdnja: X,Y nezavisne $\Rightarrow$ X i Y nekolerirane, tj. $\mathrm{Cov}(X,Y) = 0$.

Dokaz:

$$Cov(X,Y) = \mathbf{E}(\underbrace{(X-\mu_x)}_{X'}\underbrace{(Y-\mu_y)}_{Y'}) = \mathbf{E}(X-\mu_x)\mathbf{E}(Y-\mu_y) = 0.$$

OBRAT NE VRIJEDI!!

Malo računa s očekivanjem

Neka je zadano n međusobno nekoreliranih slučajnih varijabli $X_1, X_2, \ldots, X_n$ sa srednjacima μ_i, te varijancama σ_i^2, $i = 1,\ldots,n$. Koliki su srednjak i varijanca od $X = \sum_{i=1}^n X_i$?

a) $\mu = \mathbf{E}(X) = \sum_i \mu_i$.

b)

$$\begin{aligned}\mathrm{Var}(X) &= \mathbf{E}((X-\mu)^2) = \\ &= \mathbf{E}\Big(\Big(\sum_i (X_i-\mu_i\Big)^2\Big) = \\ &= \mathbf{E}\Big(\sum_i (X_i-\mu_i)^2 + \sum_{\substack{i,j;\\ i\neq j}} (X_i-\mu_i)(X_j-\mu_j)\Big) = \\ &= \sum_i \mathbf{E}\big((X_i-\mu_i)^2\big) + \sum_{\substack{i,j;\\ i\neq j}} \mathbf{E}\big((X_i-\mu_i)(X_j-\mu_j)\big) = \\ &= \sum_i \sigma_i^2 + \sum_{\substack{i,j;\\ i\neq j}} \underbrace{Cov(X_i,X_j)}_{0}.\end{aligned}$$

Drugim riječima, ako su varijable nekorelirane vrijedi *Pitagorin poučak*: Varijanca sume je jednaka sumi varijanci. Naravno, isto vrijedi i ako su varijable nezavisne.

c) Za varijablu $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ vrijedi:

$$\mu_{\overline{X}} = \frac{1}{n}\sum_i \mu_i; \qquad \mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2$$

Ako je $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$, onda je $\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$.

Dakle, usrednjavanjem n nekoreliranih varijabli varijanca se smanjuje za faktor n.

d) Primjena: Neka je X slučajna varijabla, $\mathbf{E}(X) = \mu$, $\mathrm{Var}(X) = \sigma^2 < \infty$. Na koji način procjenjujemo srednjak od X?

- Provedemo n nezavisnih mjerenja (opažanja), dobijemo uzorak od n vrijednosti i usrednjimo ga:
$$x_1, x_2, \ldots, x_n \to \overline{x} = \frac{1}{n}\sum_i x_i.$$
- U kakvoj je vezi dobivena vrijednost $\overline{x}$ s "točnom" vrijednošću μ?
Svaka pojedina vrijednost koju slučajna varijabla može poprimiti je jedna njezina realizacija. Svaki x_i je realizacija slučajne varijable X.
Što znači da su realizacije nezavisne? To znači da je svaki x_i realizacija slučajne varijable X_i koja je distribuirana jednako kao i X, pri čemu su varijable X_i međusobno nezavisne. Iz dijagrama
$$\begin{array}{ccccccc} x_1, & x_2, & \ldots, & x_n & \to & \frac{1}{n}\sum_i x_i = \overline{x} \\ \downarrow & \downarrow & & \downarrow & \text{je realizacija} & \downarrow \\ X_1, & X_2, & \ldots, & X_n & \to & \frac{1}{n}\sum_i X_i = \overline{X} \end{array}$$
izlazi da je i osmotrena vrijednost $\overline{x}$ samo jedna (od mogućih) realizacija slučajne varijable $\overline{X}$. Pri tom je $\mathbf{E}(\overline{X}) = \mu$, te $\mathrm{Var}(\overline{X}) = \frac{1}{n}\sigma^2 \to 0$ kada $n \to \infty$. Za veliko n velika je vjerojatnost da će $\overline{x}$ biti blizu μ (Chebishevljeva nejednakost).
- $(x_1, x_2, \ldots, x_n)$ je jedna realizacija od $(X_1, X_2, \ldots, X_n)$. Dakle jedna realizacija od $(X_1, X_2, \ldots, X_n)$ daje pouzdanu informaciju o $\mu = \mathbf{E}(X)$.

1.4 Osnovna statistička obilježja skupa podataka

1.4.1 Osnove

- Exploratory data analysis (EDA, Turkey)

  - istražiti "na prvu", pronaći činjenice bez nekih teorijskih pretpostavki i očekivanja

- Skup (numeričkih) podataka je svaka množina brojeve dobivenih mjerenjem (pokusom)
- Svaka analiza empirijskih podataka počinje ***preliminarnom statističkom obradom***

Ciljevi:

- Sažeti i rezimirati podatke,
  - relativno mali napor
  - mali broj izvedenih veličina
  - prikazati veliki dio informacija
- Uočiti neobične pojave $\rightarrow$ pogreške ili posebno zanimljivi događaji
- Steći "opći osjećaj" za dani skup podataka

Alati/metode:

- Najvažniji alat je OKO (ekstremi, trendovi, opći izgled, ...)
- Numerička obilježja
- Grafički prikazi (histogrami, dijagrami raspršenja, vremenski nizovi, ...)

1.4.2 Numerička obilježja

- Skup podataka $X = \{x_1, x_2, x_3, \ldots, x_n\}$ možemo shvatiti kao diskretnu slučajnu varijablu $\tilde{X}$ koja je potpuno određena zakonom razdiobe:

$$\tilde{X} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}$$

(svaki x_i je jednako vjerojatan)

odnosno pripadnom kumulativnom funkcijom distribucije:

$$F_{\tilde{X}}(x) = P(\tilde{X} \leq x) = \sum_{i, x_i \leq x} \frac{1}{n}$$

$F_{\tilde{X}}(x)\big(\,\cdot\,\big)$ je *empirijska f-ja distribucije* skupa podataka X.

- Numerička obilježja skupa podataka su numerička obilježja pripadne empirijske razdiobe
- Mnoge statističke metode "rade dobro" samo ako su ispunjene određene (relativno stroge) pretpostavke
- Poželjne osobine numeričkih obilježja:
  - Jakost, postojanost (engl. *robustness*)
  - Otpornost, neosjetljivost (engl. *resistance*)
- Metoda (mjera) je ***jaka*** ako ne ovisi značajno o određenim (teorijskim) pretpostavkama, već dobro radi u većini situacija. Npr. srednjak i medijan kao mjere "centra" nekog skupa podataka.
- Mjera je ***otporna*** ako možebitan manji broj "divljih" podataka, tj. podataka koji "jako" odstupaju od ostalih (engl. *outliers*), nema na nju veliki utjecaj. Primjer su opet srednjak i medijan.

Kvantili (fraktili):

- Pomoćne veličine za konstrukciju drugih parametara
- Ekvivalentni su percentilima
- *Definicija:* za $p \in [0, 1]$, *kvantil* q_p je broj koji dijeli skup podataka na dva dijela tako da je p-ti udio podataka manjih od q_p, te $(1-p)$-ti udio većih od q_p.
- Kvantil ima iste jedinice kao i podaci na koje se odnosi
- Određuje se sortiranjem pa brojanjem
- Za skup podataka $\{x_1, x_2, \ldots, x_n\}$ s $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ označimo sortirani skup

- **Medijan**, $q_{0.5}$ je najčešće korišteni kvantil. Označava središte skupa podataka u smislu da postoji jednako podataka koji su manji i onih koji su veći od njega

$$q_{0.5} = \begin{cases} x_{([n+1]/2)} & \text{za n neparan} \\ 1/2\Big(x_{(n/2)} + x_{(n/2+1)}\Big) & \text{za n paran} \end{cases}$$

$$x_{(1)} \quad x_{(2)} \quad x_{(3)} \quad \underbrace{x_{(4)}}_{\uparrow} \quad x_{(5)} \quad x_{(6)} \quad x_{(7)}$$

$$x_{(1)} \quad x_{(2)} \quad x_{(3)} \quad \underbrace{\quad}_{\uparrow} \quad x_{(4)} \quad x_{(5)} \quad x_{(6)}$$

Najčešći kvantili	Broj podskupova n	Oznaka
Medijan	2	$q_{0.5}$
Tercili	3	$q_{0.33}$ $q_{0.66}$
Kvartili	4	$q_{0.25}$ $q_{(0.5)}$ $q_{0.75}$
Kvintili	5	$q_{0.2}$ $q_{0.4}$ $q_{0.6}$ $q_{0.8}$
Decili	10	$q_{0.1}$ $q_{0.2}$ $\cdots$ $q_{0.9}$
Percentili	100	$q_{0.01}$ $q_{0.02}$ $\cdots$ $q_{0.99}$

- Opća formula: $q_p = x_{(k)}$, gdje je $k = p \cdot (n+1)$ zaokruženo na cijeli broj.
- Npr. $n = 1000$, $p = 0.1$; $\quad 1001 \cdot 0.1 = 100.1 \;\Rightarrow\; q_{0.1} = x_{(100)}$

Podjela numeričkih obilježja:

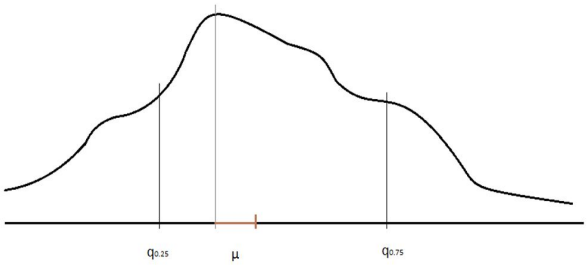
Parametri lokacije određuju "središnju tendenciju", tj. "generalnu amplitudu"

- medijan, $\quad q_{0.5}$, $\quad$ jaka i otporna mjera
- srednjak, $\quad \overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\quad$ ni jaka ni otporna
- otežani srednjak, $\quad \overline{x}_w = \dfrac{\sum_i w_i x_i}{w_i}$
- trimean $= \dfrac{q_{0.25} + 2q_{0.5} + q_{0.75}}{4}$

Parametri (mjere) raspršenja određuju "tipično odstupanje" od "srednjeg položaja"

- amplituda, $\quad A = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$, $\quad$ ni jaka ni otporna

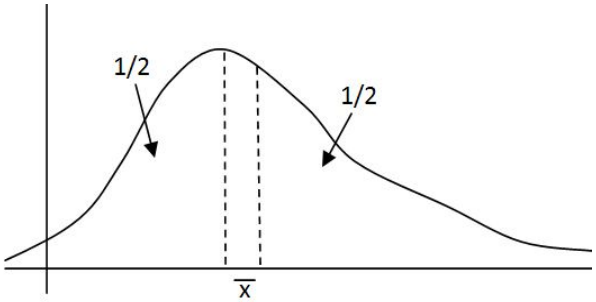
- interkvartilni raspon $IQR = q_{0.75} - q_{0.25}$, jaka i otporna



- srednje apsolutno odstupanje $\delta = \frac{1}{n}\sum_i |x_i - \overline{x}|$
- standardna devijacija $s = \sqrt{\frac{1}{n}\sum_i (x_i - \overline{x})^2}$
  - najčešće korištena, nije niti jaka niti otporna
  - s^2 je *varijanca* skupa podataka
- koeficijent varijacije, $C_v = \frac{s}{\overline{x}} \cdot 100$, izražen u postotcima

Parametri simetrije mjere asimetričnost skupa podataka u odnosu na "srednji položaj"

- koeficijent asimetrije, $C_s \equiv \gamma = \frac{\frac{1}{n}\sum_i (x_i - \overline{x})^3}{s^3}$
  - nije niti jak niti otporan
- Yule-Kendall-ov index $\gamma_{YK} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{IQR}$
  - jaka i otporna mjera
  - razlika udaljenosti između medijana i kvartila normirana interkvartilnim rasponom



("Debeli" rep u desno čini da je i $q_{0.75}$ pomaknut u desno.

Parametar spljoštenosti

- $g = \dfrac{\frac{1}{n}\sum_i (x_i - \overline{x})^4}{s^4}$
  - Za normalnu razdiobu $g = 3$ (dogovorno se uzima za referentnu)
  - Eksces (odstupanje), $E = g - 3$

1.4.3 Grafički prikazi

Stabljika i lišće (stem-and-leaf)

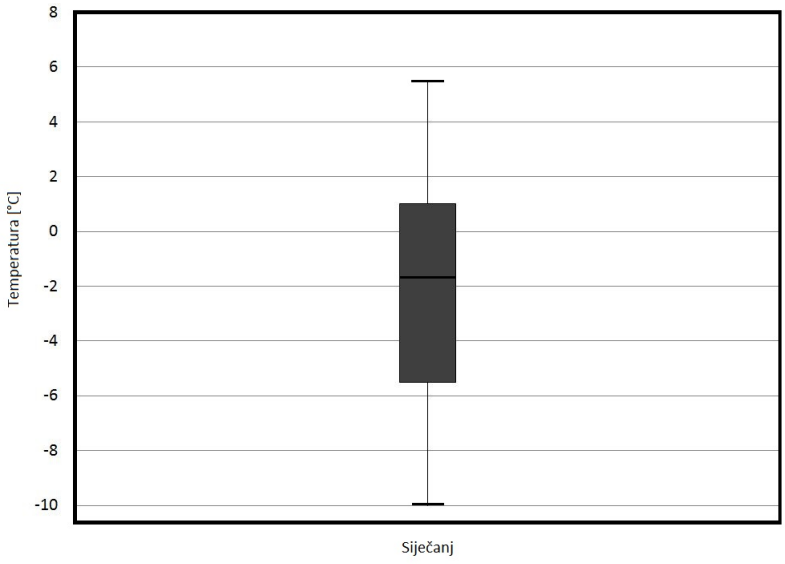
- Prikaz pogodan za stjecanje općeg dojma
- Primjenjiv na nevelike skupove (do cca 200) podataka
- Tipično, podaci se grupiraju uzevši u obzir sve *osim* najmanje značajne znamenke (NZZ)
- NZZ (listovi) se ispisuju lijevo od vertikalne crte, redom uz odgovarajuću grupu (stabljiku)
- *Primjer*: min. temperatura na Griču, siječanj, 1960.

-9	1 9
-8	
-7	9 6
-6	1 5 3
-5	1 3
-4	4 3
-3	3
-2	1
-1	4 2 5
-0	6 9 1 9 9
0	3 7 1
1	2 9 9 5
2	
3	7 0
4	
5	5

- Sve vrijednosti su jednako vjerojatne ali vidimo koje su češće
- Puno informacija (položaj, oblik, raspršenje ...)
- Ukoliko su (pojedine) grupe prevelike, mogu se dalje podijeliti
- Ako je prikaz previše "rijedak" mogu se grupe povećati
- Izdvojenice se zapišu iznad ili ispod grafa
- Stabljike se mogu naknadno sortirati
- Prikaz olakšava sortiranje cijelog skupa, određivanje kvantila, ...

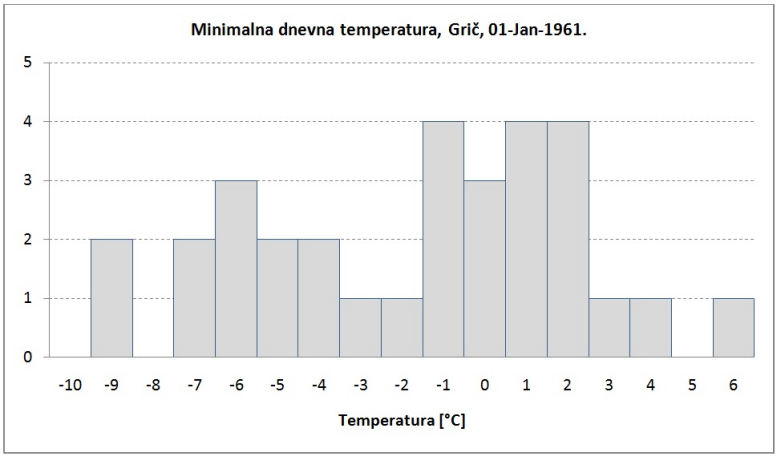
Dijagram s pravokutnikom (box-plot)

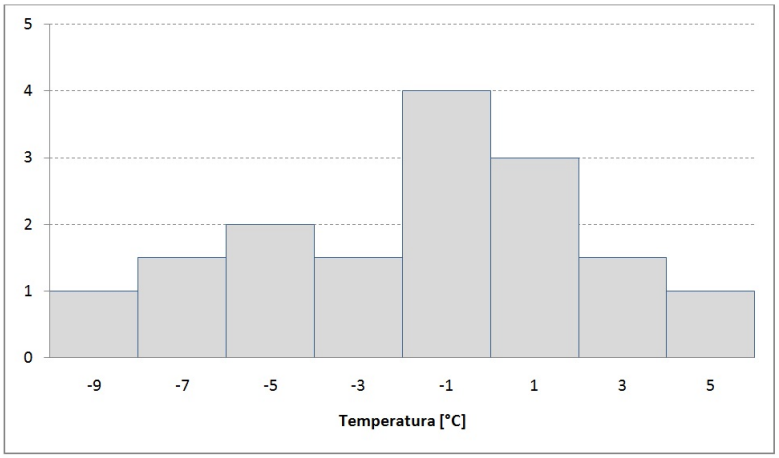
- Box-plot ili box-and-whisker plot (pravokutnik i vlas)
- Daje izvstan sažetak empirijske razdiobe putem pet kvantila i to:
  - min ($X_{(1)}$)
  - donji kvartil ($q_{0.25}$)
  - medijan kvartil ($q_{0.5}$)
  - gornji kvartil ($q_{0.75}$)
  - max ($X_{(n)}$)



Histogram

- 2D prikaz koji daje *razdiobu čestina* nekog skupa podataka po *razredima* (klasama)
  - Čitav *raspon* se podijeli u klase
  - Broj podataka u klasi je *čestina* ili *frekvencija* te klase
  - Za svaki razred crta se pravokutnik kojemu je baza određena *širinom klase*, Δx_r, a *površina* proporcionalna čestini za tu klasu
- Kako odrediti širine razreda Δx_r?
  - Kompromis! Ako je Δx_r previše veliko histogram je izglađen te ne daje puno informacije; ako je Δx_r suviše malo histogram je jako nepravilan te se javljaju i prazni razredi.
  - Charlierovo pravilo: br. klasa $= N_r = 20, \quad \Delta x_r = \dfrac{A}{N_r} = \dfrac{x_{\max} - x_{\min}}{N_r}$.
  - Brooksovo pravilo: $N_r = 5 \log_{10} n$.
  - U praksi Δx_r je *fizikalno smislen* i (obično) prirodan broj.





Ogiva

- Grafički prikaz kumulativne razdiobe čestina
  - Ako skup podataka $\{x_1, \ldots, x_n\}$ interpretiramo kao diskretnu slučajnu varijablu $\tilde{X}$, onda njena kumulativna razdioba glasi
$$F_{\tilde{X}}(x) = P\{\tilde{X} \leq x\} = \sum_{i, x_i \leq x} \frac{1}{n}.$$
- 2D graf takav da:
  - Na osi x su sami podaci (npr. temperatura).
  - Na osi y su pripadne *normirane, kumulativne čestine*:
    $p(x)$ = procjena od $F_{\tilde{X}}(x) = P(\tilde{X} \leq x)$;
    $p(x)$ = relativna frekvencija koja daje procjenu vjerojatnosti da proizvoljni (ili čak budući) podatak neće premašiti vrijednost x;
  - Alternativno, svaki x je neki kvantil, a $p(x)$ kaže koji je to kvantil. Kvantile zato možemo lako i precizno odrediti.
  - Podatke prvo sortiramo: $\{x_1, x_2, \ldots, x_n\} \to \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, a potom definiramo $p(x_{(i)}) = \dfrac{i}{n+1}$. Postoje i druge formule. (Smatramo da su podaci, od najmanjeg do najvećeg, svi jednako vjerojatni. Formula $p(x_{(i)}) = \frac{i}{n}$ se ne koristi jer bi sugerirala da je $x_{(1)}$ najmanja, a $x_{(n)}$ najveća moguća vrijednost.)

Poglavlje 2

PRIDJELJIVANJE TEORIJSKE RAZDIOBE EMPIRIJSKOJ RAZDIOBI ČESTINA

Teorijske rezdiobe (distribucije) vjerojatnosti u užem smislu:

- jedan manji broj posebnih funkcija razdiobe
- često nastaju kao posljedica određenih pretpostavki o načinu generiranja nekog skupa podataka
- sume distribucije i odnosi među njima su dobro proučeni
- konkretne vrijednosti svake razdiobe ovise o malom broju veličina, tzv. parametrima dane razdiobe.

2.1 Od empirijske razdiobe k teorijskoj

- Empirijska razdioba je razdioba vjerojatnosti u skupu podataka konačne duljine, pri čemu sve vrijednosti smatramo jednako vjerojatnima. Potpuno je predstavljena ogivom, ali se bolje "vidi" histogramom.
- Prednost histograma: sažimanje i preglednost.
- Mane histograma: nepravilnost koja je posljedica ograničenosti skupa podataka (prazne i stršeće klase); izostanak ekstremnih događaja ne znači da su oni nemogući.

- Korist od pridruživanja teorijske razdiobe:

  → daljnje sažimanje: cijeli skup podataka (uzorak), pa i sama statistička populacija iz koje je uzorak uzet, opisani su malim brojem parametara.

  → olakšana je usporedba (dviju postaja)

  → izglađivanje i interpolacija

  → prirodno je smatrati da teorijska distribucija daje "bolju" ("točniju") informaciju od histograma.

  → ekstrapolacija: mogućnost procjene vjerojatnosti i za događaje koji su izvan granica postojećih podataka

- Postupak pridruživanja:

  1. odabrati teorijsku razdiobu prema:
     - svojstvima (klimatološkog) elementa
     - analognim svojstvima teorijskih razdioba
  2. iz uzorka procjeniti parametre određene teorijske distribucije

Digresija: Kratki uvod u statističko zaključivanje

Skup podataka i uzorak fizicki su jedno te isto. Razlika među njima je samo konceptualna, ali bitna i netrivijalna. Ako skup podataka vidimo kao dio populacije onda takav skup podataka zovemo *uzorak*. Pri tom je *populacija* skup svih mogućih vrijednosti koje su, ili bi, mogle biti izmjerene. Bez obzira da li smo u mogućnosti ponoviti uzorkovanje, smatramo da je naš uzorak samo jedan od puno mogućih uzoraka koji su, ili bi, mogli biti izvučeni iz populacije. Pomoću statističkog zaključivanja iz uzorka želimo procijeniti određena svojstava populacije. To se radi tako da se naš konkretni, stvarno izmjereni ili osmotreni uzork usporedi sa svim mogućim, premda neosmotrenim, uzorcima. Pri tom se koriste razne statistike, odnosno veličine izračunate iz uzorka. *Statistika* je dakle svaka funkcija uzorka (npr. srednja vrijednost uzorka). Obzirom da njena vrijednost ovisi o konkretnom uzorku, statistika je slučajna varijabla. Uočimo da pojam statistika ima dva značenja: U širem smislu to je cijela disciplina, a u užem smislu je to svaka funkcija uzorka.

2.2 Metoda momenata

Parametri se procjenjuju direktno iz skupa podataka (uzorka) računanjem odgovarajućih numeričkih mjera (statistika).

Npr. Za $\{x_1, \ldots, x_n\}$, iz pretpostavljeno normalne raspodjele $N(\mu, \sigma^2)$ uzima se

$$\hat{\mu} = \frac{1}{n}\sum_i x_i = \overline{x}, \qquad \text{procjena od } \mu,$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2 = s^2, \qquad \text{procjena od } \sigma^2.$$

2.3 Metoda maksimalne vjerodostojnosti (ML)

Odabiru se parametri koji maksimiziraju funkciju vjerodostojnosti (engl. *likelyhood function*; *likelyhood* = mogućnost, vjerojatnost, izvjesnost).

- Pretpostavimo da je populacija opisana funkcijom gustoće $f(x;\Theta)$, te da je $\{x_1, x_2, \ldots, x_n\}$ slučajni uzorak iz te populacije. Kada se kaže da je uzorak slučajan obično se smatra da su vrijednosti x_i međusobno nezavisne, odnosno da pojava neke od njih ne mijenja vjerojatnost pojave ostalih. Preciznije, zamišljamo da svakom mjerenju x_i, odgovara jedna slučajna varijabla X_i, odnosno da je x_i realizacija slučajne varijable X_i. Pri tom sve varijable $X_j, j = 1, \ldots, n$ imaju jednu te istu funkciju gustoće $f(x;\Theta)$ i međusobno su nezavisne.

- Za kontinuirane varijable je $P\{X_i = x_i\} = 0$, no mjerenja imaju točnost Δx, te izmjerena vrijednost x_i znači da se zbio događaj $\{x_i \in A_i\}$, $A_i = \left[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right]$.

- Zbog nezavisnosti je

$$P\{X_1 \in A_1 \text{ i } X_2 \in A_2 \ldots \text{ i } X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\} \approx$$

$$\approx \prod_{i=1}^{n} f(x_i;\Theta)\Delta x = \prod_{i=1}^{n} f(x_i,\Theta)\cdot(\Delta x)^n = \underbrace{L(\Theta)}_{\text{funkcija vjerodostojnosti}} (\Delta x)^n$$

- Θ se odredi iz zahtjeva $L(\Theta) \to \max$. Dakle, određujemo Θ tako da osmotreni događaj $A = \{X_1 \in A_1 \text{ i } X_2 \in A_2 \ldots \text{ i } X_n \in A_n\}$ ima najveću vjerojatnost, odnosno da Θ bude što vjerodostojniji u svjetlu opaženih podataka.

- Umjesto $L(\Theta) \to \max$ obično je lakše riješiti $\ln(L(\Theta)) \to \max$, što je ekvivalentno sa $\sum_{i=1}^{n} \ln(f(x_i;\Theta)) \to \max$.

Primjer (metoda maksimalne vjerodostojnosti za normalnu razdiobu):

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

Funkcija vjerodostojnosti glasi:

$$L(x_1,\ldots,x_n;\mu,\sigma)=\frac{1}{\sigma^n(2\pi)^{n/2}}\prod_{i=1}^{n}e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}. \qquad /\ln$$

$$\ln(L(x;\mu,\sigma))=-n(\ln\sigma+\ln\sqrt{2\pi})-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}=$$

$$=-n\ln\sigma-n\ln\sqrt{2\pi}-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2 \quad \to \quad \max.$$

Nužan uvjet ekstrema je:

$$\frac{\partial}{\partial\mu}=0 \qquad \text{i} \qquad \frac{\partial}{\partial\sigma}=0.$$

Odatle je:

$$\frac{2}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \quad \Rightarrow \quad \mu=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}.$$

$$-\frac{n}{\sigma}-\frac{1}{2}\frac{(-2)}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0 \quad \Rightarrow$$

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}ns^2=0 \quad \Rightarrow \quad \sigma^2=s^2.$$

Poglavlje 3

NEKE DISKRETNE RAZDIOBE

3.1 Binomna razdioba

Binomna razdioba je vezana uz nezavisno ponavljanje istovrsnih jednostavnih pokusa. Jednostavni slučajni pokus je pokus sa dva ishoda (1 - uspjeh, 0 - neuspjeh).

$$X \sim \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix} \qquad \rightarrow \text{Bernoullijeva slučajna varijabla}$$

Pokus ponavljamo n puta, nezavisno. Označimo s X_i slučajnu varijablu koja odgovara i-tom pokusu. Dakle, $X_1, \ldots, X_n$ je n nezavisnih slučajnih varijabli distribuiranih kao i X.

- *Binomna* slučajna varijabla je broj uspjeha koji se ostvare u n nezavisnih ponavljanja, tj.
$$Y = X_1 + X_2 + \cdots + X_n$$
- Zakon razdiobe od Y:
  - mogući ishodi: $\{0, 1, 2, \ldots, n\}$. Odredimo $P\{Y = k\}$
  - Skupu $\{l_1, l_2, \ldots, l_k\} \subseteq \{1, 2, \ldots, n\}$ pridružujemo značenje:
    Uspjeh se dogodio u l_1-om, l_2-om, $\ldots, l_k$-tom pokusu. Zbog *nezavisnosti*, vjerojatnost takvog događaja je
    $$P\{X_{l_1} = 1, \ \ldots, \ X_{l_k} = 1, \ \text{ostali} = 0\} = p^k \cdot q^{n-k}.$$
    Takvih događaja ima $\dfrac{n(n-1)\ldots(n-k+1)}{k!} = \binom{n}{k}$. Zaključujemo da je
    $$P\{Y = k\} = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n,$$

što predstavlja vjerojatnost da se u n pokusa uspjeh dogodi k puta. Parametri ove razdiobe su dakle p i n, te pišemo:

$$Y \sim B(n, p).$$

Vjerojatnost za uspjeh u jednom pokusu je p, a za neuspjeh $q = 1 - p$.

- Binomni poučak:

$$1^n = (p+q)^n = \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} = 1 \quad \Rightarrow \quad \Big(\sum_{k=1}^{n} P\{Y = k\} = 1 \Big).$$

- U praksi su dva bitna uvjeta:

  1. Vjerojatnost je približno ista za sve pokušaje,
  2. Pokušaji su mogu smatrati međusobno nezavisnima.

- ***Momenti:***

$$Y = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$$

  - Srednja vrijednost:

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(X_1 + X_2 + \cdots + X_n) = \\ &= \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mathbb{E}(X) = n(1 \cdot p + 0 \cdot q) = np. \end{aligned}$$

  - Varijanca:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) = (\text{nekoreliranost}) = \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) = n\mathrm{Var}(X) = \\ &= n\mathbb{E}\big((X - \mathbb{E}(X))^2\big) = n\mathbb{E}\big((X-p)^2\big) = \\ &= n\big[(1-p)^2 p + (0-p)^2 q\big] = np(1-p)(1-p+p) = npq. \end{aligned}$$

  - Koeficijent simetrije, $\mathbf{C_s} = \dfrac{\mathbb{E}\big((Y - \mathbb{E}(Y))^3\big)}{\sigma^3}$:

$$\mathbb{E}\big((Y-\mathbb{E}(Y))^3\big) = \mathbb{E}\big[\big(\sum_i X_i\big) - np\big]^3 =$$
$$= \mathbb{E}\big[(X_1-p)+(X_2-p)+\cdots+(X_n-p)\big]^3 =$$
$$= \mathbb{E}\big[\big(\sum_i (X_i-p)^3\big) + \sum_{i,j;i\neq j}\big((X_i-p)^2(X_j-p)\big)\big] = (\text{nezavisnost}) =$$
$$= \sum_i \mathbb{E}\big((X_i-p)^3\big) + \sum_{i,j;i\neq j}\mathbb{E}\big((X_i-p)^2\big)\underbrace{\mathbb{E}(x_j-p)}_{0} =$$
$$= n\mathbb{E}\big((X-p)^3\big) = n\big[(1-p)^3p + (0-p)^3q\big] = npq(1-2p) \Rightarrow$$

$$\mathbf{C_s} = \frac{1-2p}{\sqrt{npq}.}$$

Posebni slučajevi:

$p = 1/2 \quad \Rightarrow \mathbf{C_s} = 0$, tj. varijabla je simetrična,
$p \longrightarrow 0 \quad \Rightarrow \mathbf{C_s} \longrightarrow +\infty$
$p \longrightarrow 1 \quad \Rightarrow \mathbf{C_s} \longrightarrow -\infty$

- Rekurzivna formula:

$$P\{Y=0\} = q^n$$
$$P\{Y=k\} = \frac{p}{q}\frac{n-k+1}{k}P\{Y=k-1\}$$

- Procjena parametara metodom momenata:
  Skupu podataka (uzorku) $\{y_1, y_2, \ldots, y_N\}$ želimo prilagoditi $Y \sim B(n,p)$. Metoda momenata (a isto i ML metoda) daje $\mathbb{E}(Y) = np = m_y = \frac{1}{N}\sum_{i=1}^N y_i \Rightarrow \hat{p} = \frac{m_y}{n}$, što je procjena vjerojatnosti uspjeha. Broj ponavljanja (parametar n) je obično određen prirodom problema.

- *Zakon velikih brojeva*: Sjetimo se Chebischevljeve nejednakosti: $\forall X$, t.d. je $\mathrm{Var}(X) < \infty$ i $\forall \epsilon > 0$ vrijedi:
$$P\{|X - \mathbf{E}(X)| \geq \epsilon\} \leq \frac{\mathrm{Var}(X)}{\epsilon^2}.$$
  Primjenimo nejednakost na $Y \sim B(n,p)$:
$$P\{|Y-np| \geq \epsilon\} \leq \frac{npq}{\epsilon^2},$$
$$P\{|\frac{Y}{n} - p| \geq \underbrace{\frac{\epsilon}{n}}_{\epsilon'}\} \qquad \text{uz} \qquad \epsilon^2 = \epsilon'^2 n^2,$$
$$P\{|\frac{Y}{n} - p| \geq \epsilon'\} \leq \frac{p(1-p)}{\epsilon'^2}\frac{1}{n},$$

pri čemu je $\frac{Y}{n}$ je relativna čestina uspjeha u n ponavljanja. Vjerojatnost jednog uspjeha je p. Ako pokus ponovimo puno puta ($n \gg$) mala je šansa da će se relativne čestine jako razlikovati od p.

- Na primjer:

$$\text{Za } \epsilon' = 0.01 \text{ i } p = 0.5, \text{ je } \frac{p(1-p)}{\epsilon^2} = 2500.$$

Odatle,

$$n = 10000 \qquad \Rightarrow \qquad \frac{2500}{n} = \frac{1}{4} = 25\%,$$

$$n = 100000 \qquad \Rightarrow \qquad \frac{2500}{n} = \frac{1}{40} = 2.5\%,$$

3.2 Poissonova razdioba

- Granični slučaj binomne razdiobe kada $p \to 0$, a $n \to \infty$ tako da je np, tj. očekivani broj uspjeha konstantan.

- Primjenjuje se kad imamo dva ishoda od kojih se jedan pojavljuje rijetko i njega smatramo uspjehom.

- Praktična korist: za veliko n, nepraktično je računati binomne koeficijente, a još važnije je da imamo jedan parametar manje.

- Izvod: Neka je $n \cdot p = \lambda$ odakle je $p = \frac{\lambda}{n}$. Krenimo od $Y \sim B(n,p)$ s idejom da pustimo $n \to \infty$.

$$\begin{aligned} P\{Y = k\} &= \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \\ &= \frac{\lambda^k \left(1 - \frac{\lambda}{n}\right)^n}{k!} \frac{n!}{(n-k)!} \frac{1}{\left(1 - \frac{\lambda}{n}\right)^k n^k} = \\ &= \frac{\lambda^k \left(1 - \frac{\lambda}{n}\right)^n}{k} \frac{n(n-1)\dots(n-k+1)}{(n-\lambda)^k} \xrightarrow[n\to\infty]{} \frac{\lambda^k e^{-\lambda}}{k!}. \end{aligned}$$

Ako slučajna varijabla X poprima vrijednosti $\{0, 1, 2, \dots\}$ s vjerojatnostima

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

onda je X *Poissonova* varijabla s parametrom λ. Pišemo:

$$X \sim P(\lambda).$$

- ***Momenti:*** Neka je $Y \sim B(n,p), \quad np = \lambda, \quad X \sim P(\lambda)$.

$$\mu_Y = \mathbb{E}(Y) = np \quad \to \quad \mathbb{E}(X) = \lambda, \quad \text{kada } n \to \infty,$$

$$\sigma_Y^2 = \mathrm{Var}(Y) = np(1-p) \quad \to \quad \mathrm{Var}(X) = \lambda,$$
$$C_{sY} = \frac{1-2p}{\sqrt{npq}} \quad \to \quad \frac{1}{\sqrt{\lambda}} = C_{sX}.$$

- Rekurzivne relacije:

$$P\{X=0\} = e^{-\lambda},$$
$$P\{X=k\} = \frac{\lambda}{k} P\{X=k-1\}.$$

- Procjena parametara metodom momenata:
  Skupu podataka (uzorku) $\{x_1, x_2, \ldots, x_N\}$ želimo prilagoditi $X \sim P(\lambda)$. Metoda momenata daje $\mathbb{E}(X) = \lambda = m_x = \frac{1}{N}\sum_{i=1}^{N} x_i \Rightarrow \hat{\lambda} = m_x$.

3.3 Negativna binomna razdioba

- I dalje gledamo događaj s dva ishoda; gledamo broj uspjeha u nekom vremenskom intervalu. Za skup podataka (uzorak) $\{x_1, \ldots, x_n\}$ računamo:

$$m_X = \overline{X} = \frac{1}{n}\sum_i x_i, \quad s_x^2 = \frac{1}{n}\sum_i (x_i - m_x)^2,$$

  te rezoniramo: $m_X > s_x^2 \longrightarrow$ binomna razdioba $\quad (\mu_X = np, \sigma_x^2 = npq)$,
  $m_X = s_x^2 \longrightarrow$ poissonova razdioba $\quad (\mu = \sigma^2 = \lambda)$,
  $m_X < s_x^2 \longrightarrow (q > 1, p < 1, n < 0, p+q = 1) \longrightarrow$ negativna binomna razdioba.

- Neka su $k \in \mathbb{N}$ i $p > 0$ dva parametra. Neka je X slučajna varijabla koja poprima vrijednosti u skupu $\{0, 1, 2, \ldots\}$ s vjerojatnostima

$$\begin{aligned} P\{X=x\} &= \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)} \cdot \frac{p^x}{(1+p)^{k+x}} \\ &= \frac{(x+k-1)!}{x!(k-1)!} \cdot \frac{p^x}{(1+p)^{k+x}} = \\ &= \binom{x+k-1}{k-1} \frac{p^x}{(1+p)^{k+x}}, \end{aligned}$$

pri čemu je Γ tzv. gama funkcija definirana sa:

$$\Gamma(\alpha+1) = \int_0^\infty x^\alpha e^{-x} dx = \alpha!.$$

Tada kažemo da X ima negativnu binomnu razdiobu i pišemo $X \sim NB(k,p)$.

- Za $m \in \mathbb{R}$ i $n \in \mathbb{N}$ definiramo poopćeni binomni koeficijent:

$$\binom{m}{n} = \frac{m(m-1)\dots(m-n+1)}{n!}.$$

Tada za $x, k \in \mathbb{N}$ vrijedi:

$$\binom{x+k-1}{k-1} = (-1)^x \binom{-k}{x},$$

odakle imamo:

$$\begin{aligned} P\{X=x\} &= (-1)^x \binom{-k}{x} p^x (1+p)^{-k-x} = \\ &= \binom{-k}{x} (-p)^x (1-(-p))^{-k-x}. \end{aligned}$$

- Općenito vrijedi tzv. poopćeni binomni teorem: Za $|x| > |y]$ i $m \in \mathbb{R}$ vrijedi:

$$(x+y)^m = \sum_{k=0}^{\infty} \binom{m}{k} x^{m-k} y^k.$$

Odatle je:

$$\sum_{x=0}^{\infty} P\{X=x\} = (-p+(1-(-p)))^{-k} = 1^{-k} = 1,$$

što pokazuje da se zaista radi o zakonu razdiobe diskretne slučajne varijable.

- Formalna veza s binomnom razdiobom:

$$n \longleftrightarrow -k$$

$$p \longleftrightarrow -p$$

$$q \longleftrightarrow 1+p$$

- Momenti (u analogiji s binomnom razdiobom):

$$\begin{aligned} \mathbb{E}(X) = k \cdot p \qquad &= (-k)\dot{(}-p) \qquad \cdots \qquad (np), \\ \mathrm{Var}(X) = k \cdot p \cdot (1+p) \qquad &= (-k)\dot{(}-p)\dot{(}1-(-p)) \qquad \cdots \qquad (npq). \end{aligned}$$

- Procjena parametara metodom momenata:

$$m_x = kp, \qquad (m_x < s_x^2)$$
$$s_x^2 = kp(1+p),$$

Odatle je

$$\hat{p} = \frac{s_x^2}{m_x} - 1 > 0,$$
$$\hat{k} = \frac{m_x^2}{s_x^2 - m_x} > 0.$$

U ovom slučaju bolja je metoda maksimalne vjerodostojnosti, no ona zahtijeva relativno složen numerički postupak.

- Diskusija:
  - Binomna i Poissonova varijabla se temelje na *nezavisnom ponavljanju jednostavnog* slučajnog pokusa $\longrightarrow m_x \geq s_x^2$
  - $m_x < s_x^2$ je signal da je nešto od gornjeg narušeno
  - Nezavisnost povlači približno ravnomjeran vremenski raspored uspjeha, pa su situacije kad je to narušeno (tj. kada rijetki događaj ima tendenciju da se javlja u grupama u određenom dijelu godine) potencijalno povoljne za primjenu NBR.
  - NBR se može dobiti i kao kompozitna Poissonova razdioba, tj. $X \sim NB(k,p)$ se može izgraditi pomoću Poissonove varijable $\tilde{X} - P(\lambda)$, pri čemu λ nije fiksan, već je i sam slučajna varijabla s Γ razdiobom, $\lambda \sim \Gamma(k,p)$. ($\Gamma(k,p), \mu = kp, \sigma^2 = kp^2$). Na ovaj način uvažavamo (modeliramo) činjenicu da se vjerojatnost uspjeha, tj. čestine mijenjaju.

- Alternativni opis i interpretacija NBR:

$$P\{X = x\} = \binom{x+k-1}{k-1} \underbrace{\left(\frac{p}{1+p}\right)^x}_{1-p'} \underbrace{\left(\frac{1}{1+p}\right)^k}_{p' \in (0,1)} = \binom{x+k-1}{k-1} p'^k (1-p')^x =$$

= vjerojatnost da se (unutar Bernulijeve sheme s vjerojatnošću uspjeha p') fiksan broj od k uspjeha dogodi točno nakon $k+x$ pokušaja. Drugim riječima, X je broj neuspjeha koji će se dogoditi prije nego se dogodi točno k uspjeha.

Poglavlje 4

NEKE TEORIJSKE KONTINUIRANE FUNKCIJE DISTRIBUCIJE

4.1 Normalna ili Gaussova razdioba

- Normalna razdioba je ponajvažnija razdioba, kako u teoriji, tako i u praksi (vidi centralni granični teorem, naprijed)

- Funkcija gustoće vjerojatnosti normalne varijable X ovisi o dva parametra, μ i σ i glasi

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

  Pišemo $X \sim \mathrm{N}(\mu,\sigma)$.

- Srednja vrijednosti i varijanca su:

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

$$\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \sigma^2,$$

  dok je koeficijent simetrije $C_s = 0$, tj. normalna razdioba je simetrična.

- Kumulativna funkcija razdiobe (distribucija)

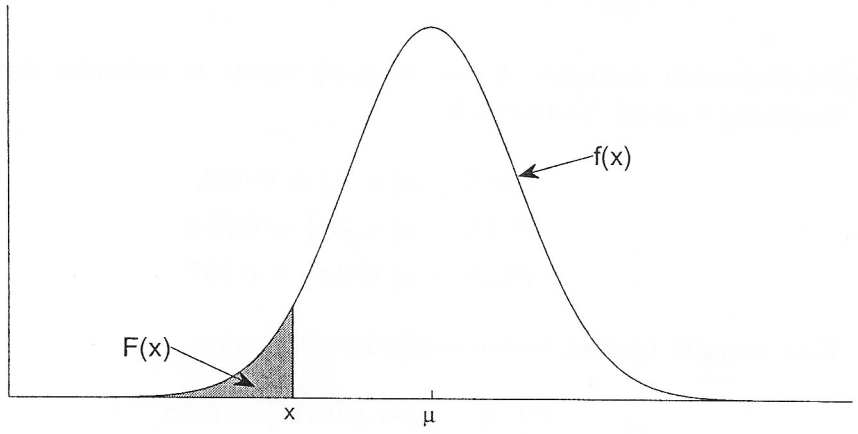
$$F(x;\ \mu,\sigma) = \int_{-\infty}^{x} f(x';\mu,\sigma)dx'$$

nije elementarna, pa joj se vrijednosti dobivaju iz odgovarajućih tablica ili pak korištenjem odgovarajućeg softvera (npr. funkcija normcdf u Matlabu). Tablice sadrže kumulativne vrijednosti *standardne normalne razdiobe* $\mathrm{N}(0,1)$, tj. razdiobe koja ima srednjak $\mu = 0$ te standardnu devijaciju $\sigma = 1$. U slučaju proizvoljnih μ i σ treba koristiti linearnu transformaciju:

$$X \sim \mathrm{N}(\mu, \sigma) \qquad \longrightarrow \qquad Y = \frac{X - \mu}{\sigma} \sim \mathrm{N}(0,1).$$

Tabelirana je funkcija distribucije od Y, tj.

$$\Phi(y) \equiv F_Y(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left\{-\frac{1}{2}t^2\right\} dt.$$



Slika 4.1: Funkcija gustoće vjerojatnosti, $f(x)$, i kumulativna razdioba vjerojatnosti, $F(x)$, normalne razdiobe.

- Za $X \sim \mathrm{N}(\mu, \sigma)$, te $a < b \in \mathbb{R}$ je

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

a nakon transformacije $a \to a' = \frac{a-\mu}{\sigma}$, te $b \to b' = \frac{b-\mu}{\sigma}$ imamo

$$P\{a < X < b\} = F_X(b) - F_X(a) = \Phi(b') - \Phi(a').$$

- Iz tablica se najčešće očitava tzv. *funkcija kvantila*, tj. inverzna funkcija od Φ:

$$z(\alpha) \equiv z_\alpha = \Phi^{-1}(\alpha).$$

Iz definicije vrijedi:

$$P\{Y \leq z_\alpha\} = \alpha, \qquad P\{Y \geq z_{1-\alpha}\} = \alpha,$$

$$P\{X - \mu \leq z_\alpha \sigma\} = \alpha, \qquad P\{X - \mu \geq z_{1-\alpha}\sigma\} = \alpha,$$

a odatle i

$$P\{|X - \mu| \leq z_{1-\alpha/2}\sigma\} = 1 - \alpha, \qquad P\{|X - \mu| \geq z_{1-\alpha/2}\sigma\} = \alpha.$$

- Odatle dolazi nekoliko standardnih brojeva koji se vezuju uz normalnu varijablu $X \sim \mathrm{N}(\mu, \sigma)$ ($\approx$ znači 'približno'):

$$P\{|X - \mu| < \sigma\} \approx 0.682, \qquad \text{tj. } z_{0.841} \approx 1,$$

$$P\{|X - \mu| < 2\sigma\} \approx 0.954, \qquad \text{tj. } z_{0.977} \approx 2,$$

$$P\{|X - \mu| < 3\sigma\} \approx 0.997. \qquad \text{tj. } z_{0.999} \approx 3,$$

Kod mnogih testova trebamo sljedeće činjenice:

$$P\{|X - \mu| < 1.960\sigma\} \approx 0.95, \qquad \text{tj. } z_{0.975} \approx 1.960,$$

$$P\{|X - \mu| < 2.576\sigma\} \approx 0.99, \qquad \text{tj. } z_{0.995} \approx 2.576,$$

$$P\{X - \mu < 1.645\sigma\} \approx 0.95, \qquad \text{tj. } z_{0.95} \approx 1.645,$$

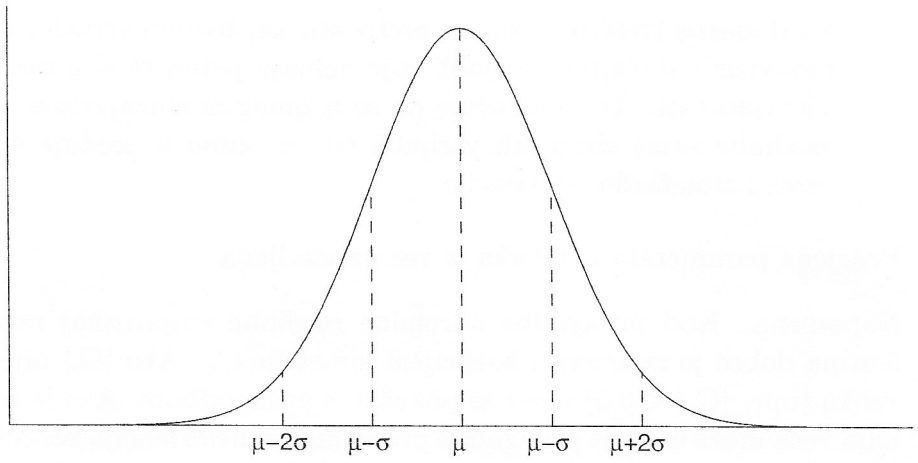
$$P\{X - \mu < 2.326\sigma\} \approx 0.99. \qquad \text{tj. } z_{0.99} \approx 2.326,$$

- Važnost normalne razdiobe potječe iz sljedećeg teorema:

  - **Centralni granični teorem:** Ako je $X_i, i = 1, 2, \ldots$ niz međusobno nezavisnih, jednako distribuiranih slučajnih varijabli, tako da je $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, onda za svako $a < b$ vrijedi:

$$\lim_{n\to\infty} P\left\{a \leqslant \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \leqslant b\right\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{\frac{1}{2}t^2} dt.$$

  Drugim riječima, niz slučajnih varijabli $Y_n = \sum_{i=1}^n X_i$ nakon odgovarajućeg centriranja i normiranja konvergira, kada $n \longrightarrow \infty$, k standardnoj normalnoj slučajnoj varijabli. Kažemo da je normalna razdioba *granična* razdioba sume nezavisnih, jednako distribuiranih slučajnih varijabli.



Slika 4.2: Funkcija gustoće vjerojatnosti normalne razdiobe.

- Uočimo da teorem vrijedi neovisno o tome kakav je (jedan te isti) zakon razdiobe varijabli X_i, u čemu i leži njegova snaga i praktična upotrebljivost. Ipak, treba imati na umu da je rezultat asimptotski, tj. varijabla $Y_n = \sum_{i=1}^{n} X_i$ ima *približno* normalnu razdiobu $\mathrm{N}(n\mu, \sqrt{n}\sigma)$ samo ako je n *dovoljno velik*. Teorem ništa ne govori o tome koliki stvarno n treba biti za dobiti 'dobru' aproksimaciju. To ovisi o zakonu razdiobe od X_i.
- Drugim riječima, suma velikog broja slučajnih varijabli (koje ne mogu poprimiti jako velike vrijednosti previše često) je približno normalno raspodijeljena. U praksi, srednje vrijednosti velikih uzoraka bit će normalno raspodijeljene.
- Uz dodatne (relativno slabe) pretpostavke, teorem vrijedi i za niz nezavisnih slučajnih varijabli koje nemaju jednu te istu razdiobu vjerojatnosti. To omogućuje da se u mnogim slučajevima zakon razdiobe sume slučajnih varijabli odredi samo iz srednje vrijednosti i standardne devijacije.
- Prije nego što teorem primjenimo na binomnu slučajnu varijablu $Y_n \sim B(n,p)$, pogledajmo što se događa kada $n \to \infty$? Tada imamo

$$\binom{n}{k} p^k q^{n-k} = \frac{1}{k!} \left(\frac{p}{q}\right)^k n(n-1)\cdots(n-k+1)\, q^n \quad \to \quad 0,$$

kada $n \to \infty$. Dakle, za svako (fiksno) k, $P(Y_n = k)$ teži k nuli, što daje degeneriranu razdiobu. No ne idu svi $P(Y_n = k)$ u nulu jednako brzo.

Budući je $Y_n = \sum_{i=1}^{n} X_i$, pri čemu su X_i Bernoullijeve slučajne varijable, to

centralni granični teorem daje:

$$\frac{Y_n - np}{\sqrt{(npq)}} \to N(0,1), \qquad \text{kada } n \to \infty,$$

odnosno $Y_n \dot{\sim} N(np, \sqrt{npq})$, za dovoljno veliko n (znak $\dot{\sim}$ znači "ima približno razdiobu"). Ovaj je opis vrlo precizan te omogućava da se precizno "izmjere" repovi od Y_n za razumno velike n.

– Iz prethodnog je

$$P\left\{\left|\frac{Y_n}{n} - p\right| \geq z_{1-\alpha/2}\sqrt{\frac{pq}{n}}\right\} = \alpha.$$

Ako stavimo $z_{1-\alpha/2}\sqrt{pq/n} = \epsilon$, onda je $\alpha = 2(1 - \Phi(\epsilon\sqrt{pq/n})$. Za $\epsilon = 0.01$ imamo:

n	α
100	0.8415
1000	0.5271
10000	0.0455
100000	$2.536 \cdot 10^{-10}$

- Procjena parametara iz uzorka je već napravljena.

- Napomena: Kod prilagodbe normalne razdiobe empirijskoj razdiobi čestina dobro je izračunati koeficijent simetrije C_s. Ako $|C_s|$ nije preveliko (npr. $|C_s| < 0.5$) može se pokušati s prilagodbom. Ako je asimetrija veća može se prije prilagodbe primjeniti transformacija koordinata $y = x^\lambda$ gdje je $\lambda = \frac{1}{2}$, $\frac{1}{3}$ ili $y = \ln x$, naravno, ako je transformacija definirana.

4.2 Eksponencijalna razdioba

- Funkcija gustoće *eksponencijalne razdiobe* ovisi o parametru β i glasi:

$$f(x;\beta) = \frac{1}{\beta}\exp\left\{-\frac{x}{\beta}\right\}, \quad x \geqslant 0,$$

Kumulativna funkcija distribucije se može eksplicitno izračunati:

$$F(x;\beta) = 1 - \exp\left\{-\frac{x}{\beta}\right\}.$$

- Srednja vrijednost i varijanca su

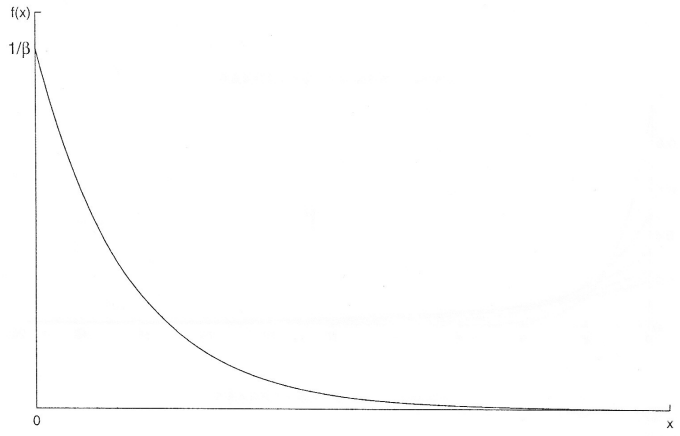
$$\mu = \beta, \quad \sigma^2 = \beta^2,$$

odakle izlazi da je koeficijent varijacije $C_V = 1$.

- U praksi se koristi za opisivanje količine oborine u nekom kraćem razdoblju, npr. za satne ili dnevne količine.

- Prilagodba empirijskoj razdiobi: Procjena parametara metodom momenata, a jednako i metodom maksimalne vjerodostojnosti vodi na

$$\hat{\beta} = m_x = \frac{1}{n}\sum_{i=1}^{n} x_i.$$



Slika 4.3: Funkcije gustoće eksponencijalne razdiobe

4.3 Gama razdioba

- Funkcija gustoće *gama razdiobe* ovisi o dva parametra, α i β i glasi:

$$f(x;\ \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} \exp\left\{-\frac{x}{\beta}\right\}, \qquad x \geqslant 0,$$

pri čemu je $\Gamma(\cdot)$ tzv. gama funkcija (koju smo već prije definirali). Ako slučajna varijabla X ima gama razdiobu s parametrima α i β, onda pišemo $X \sim \Gamma(\alpha, \beta)$.

- α je parametar oblika (jer je u igri funkcija $z^{\alpha-1}\exp\{-z\}$), a β je parametar skale (jer je u igri kvocjent $\frac{x}{\beta}$).

- Srednja vrijednost i varijanca su

$$\mu = \alpha\beta, \quad \sigma^2 = \alpha\beta^2,$$

- Tjeme razdiobe je $x_T = (\alpha - 1)\beta$ dok je koeficijent simetrije $C_s = 2\alpha^{-\frac{1}{2}}$. Gama razdioba je, dakle, uvijek pozitivno asimetrična (rep u desno). Kada $\alpha \to \infty$ asimetrija opada, a gama razdioba teži k normalnoj razdiobi (premda razmjerno sporo).

- U statističkim tablicama nalazi se nekompletna gama funkcija

$$F(x;\ \alpha) = \frac{1}{\Gamma(x)} \int_0^x x^{\alpha-1} e^{-z} dz,$$

  što je (kumulativna) funkcija distribucije slučajne varijable $X \sim \Gamma(\alpha, 1)$. U slučaju da je $\beta \neq 1$ treba primijeniti transformaciju $x \longrightarrow \frac{x}{\beta}$.

- Neka je $\alpha \in \mathbb{N}$, te neka su $X_i, i = 1, \ldots, \alpha$ međusobno nezavisne slučajne varijable pri čemu svaka ima eksponencijalnu razdiobu s parametrom β. Tada je $\sum_{i=1}^{\alpha} X_i \sim \Gamma(\alpha, \beta)$.

- Gama razdioba se koristi za opis kontinuiranih, pozitivnih veličina. Tako se u klimatologiji često koristi za opis količina oborine u nekom duljem razdoblju (npr. mjesečne količine).

- Prilagodba empirijskoj razdiobi:

  - Kod metode momenata izračunamo srednjak (m_x) i varijancu (s_x^2) uzorka, te riješimo sustav jednadžbi:

$$m_x = \hat{\alpha}\hat{\beta}$$
$$s_x^2 = \hat{\alpha}\hat{\beta}^2$$

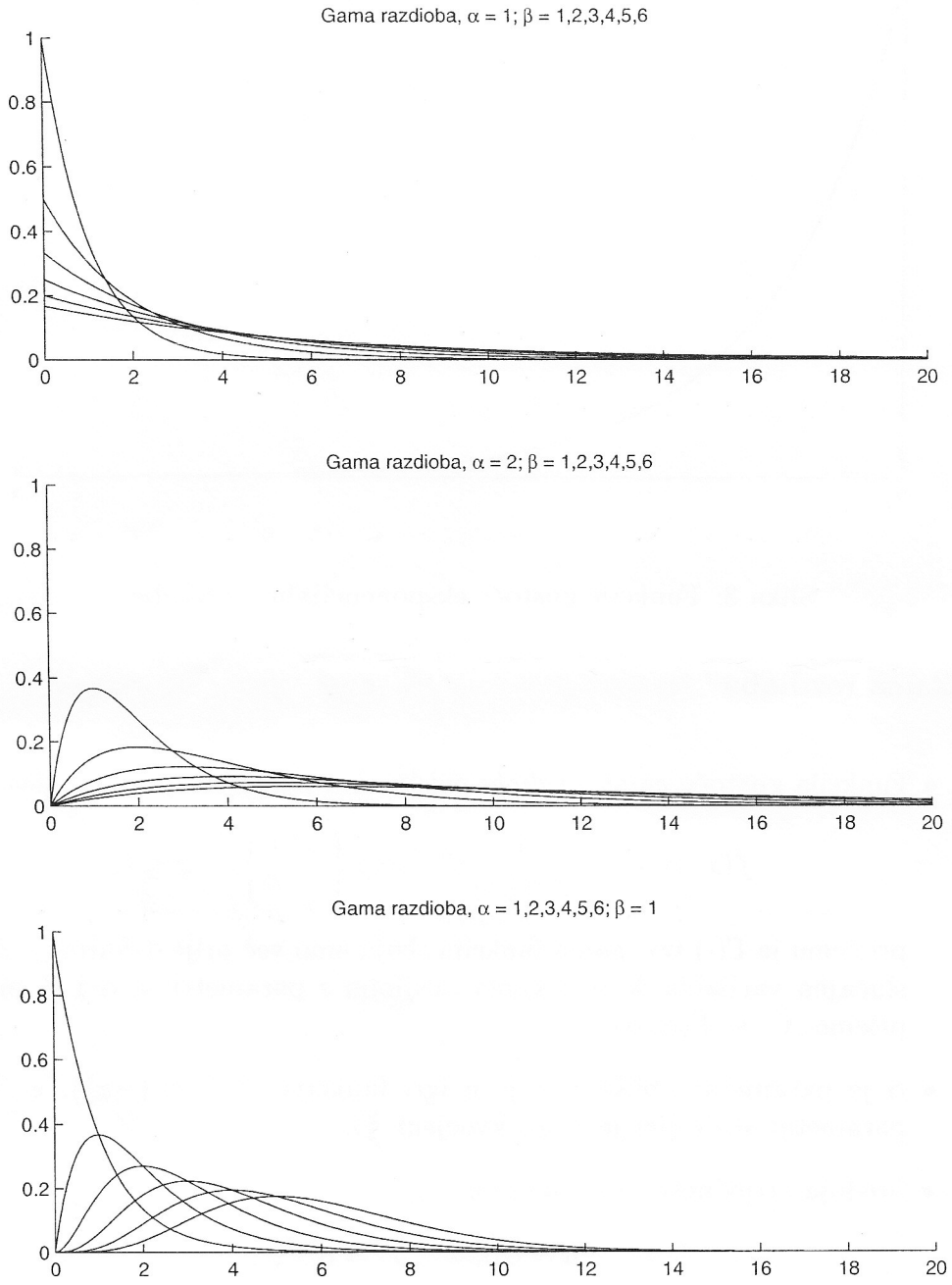
    odakle se dobije

$$\hat{\beta} = \frac{s_x^2}{m_x},$$
$$\hat{\alpha} = \frac{s_x^2}{m_x^2} = \frac{1}{C_v^2}.$$

    Na ovaj način dobivaju se relativno dobre procjene ako je parametar oblika α dovoljno velik.

  - Metoda maksimalne vjerodostojnosti zahtijeva složeni numerički postupak, no može se upotrijebiti i sljedeća aproksimacija. Prvo se izračuna

$$A = \ln(m_x) - \frac{1}{n} \sum_{i=1}^{n} \ln x_i,$$



Slika 4.4: Funkcije gustoće gama razdiobe za razne vrijednosti parametrara

dakle logaritam srednjaka uzorka minus srednjak logaritama (pri čemu uzorak ne smije sadržavati nule) i potom

$$\hat{\beta} = \frac{1 + \sqrt{1 + 4A/3}}{4A}, \qquad \hat{\alpha} = \frac{m_x}{\hat{\beta}}.$$

Dodatni problem nastaje u praksi ako se pojave vrijednosti jednake nuli (npr. mjesec bez oborine).

4.4 Beta razdioba

- Funkcija gustoće *beta razdiobe* ovisi o dva parametra, α i β i glasi:

$$f(x;\ \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leqslant x \leqslant 1, \quad \alpha, \beta > 0.$$

- Srednja vrijednost i varijanca su

$$\mu = \frac{\alpha}{\alpha + \beta},$$

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

- Prilagodba empirijskoj razdiobi se obično vrši metodom momenata, pri čemu treba riješiti sustav jednadžbi:

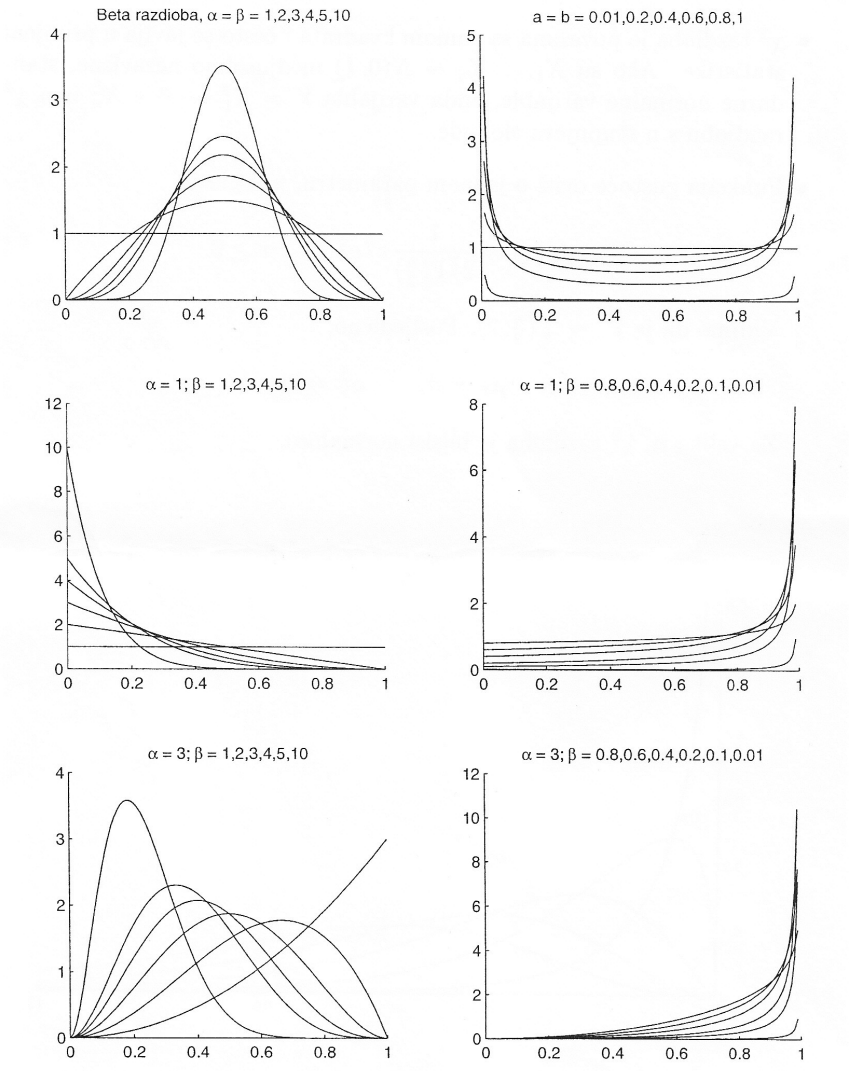
$$m_x = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}$$

$$s_x^2 = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)}.$$

Uputa:

$$m_x = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = 1 - \frac{\hat{\beta}}{\hat{\alpha} + \hat{\beta}}.$$

- Beta razdioba, ovisno o vrijednostima parametara, može poprimiti različite oblike. Uz eventualnu linearnu transformaciju, koristi se za opis kontinuiranih, ograničenih veličina (npr. naoblaka).



Slika 4.5: Funkcije gustoće beta razdiobe za razne vrijednosti parametara

4.5 χ^2 razdioba

- χ^2 razdioba je povezana sa sumom kvadrata i često se javlja u primjeni statistike. Ako su $X_1, \ldots, X_n \sim \mathrm{N}(0,1)$ međusobno nezavisne, standardne normalne varijable, onda varijabla $Y = X_1^2 + \cdots + X_n^2$ ima χ^2 razdiobu s n stupnjeva slobode.

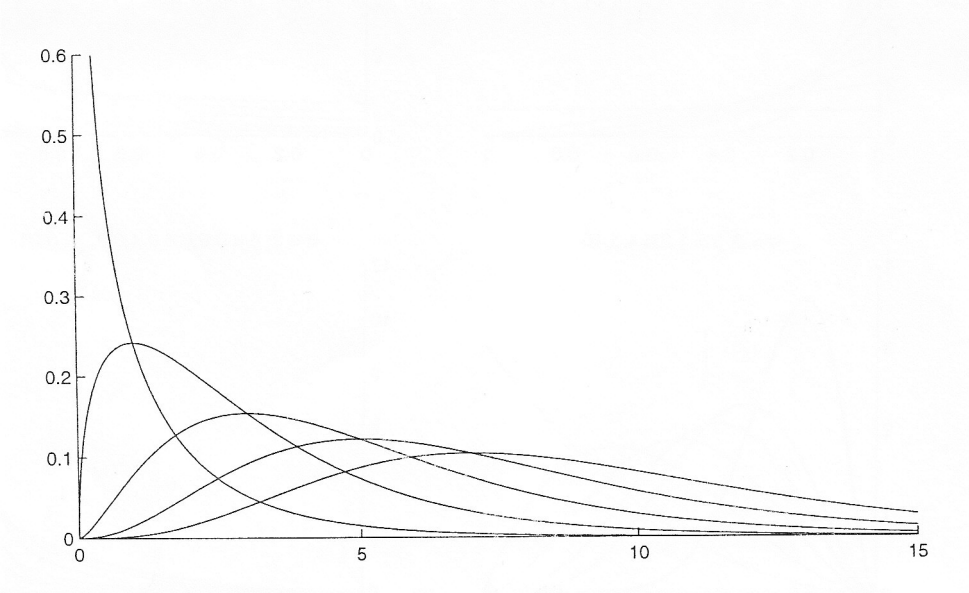
- Funkcija gustoće ovisi o jednom parametru, n, i glasi:

$$f_Y(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x \geqslant 0.$$

  Vidimo da je $Y \sim \Gamma(\frac{n}{2}, 2)$. Posljedično,

$$\mu_Y = n, \sigma_Y^2 = 2n.$$

  Za veliko n, χ^2 razdioba je bliska normalnoj.



Slika 4.6: Funkcije gustoće χ^2 razdiobe za razne vrijednosti parametara

4.6 Razdioba ekstremnih vrijednosti

Cilj je procjeniti ekstremne vrijednosti (max, min) neke veličine u određenom razdoblju (najčešće jedna godina), tj. ispitati statistička svojstva ekstrema.

Neka je zadano n nezavisnih, jednako distribuiranih slučajnih varijabli $X_1, X_2, ..., X_n$. Zanima nas slučajna varijabla

$$Y_n = \max\{X_1, ..., X_n\},$$

tj. njena funkcija distribucije.

Mogući pristupi:

- Neka je F_X funkcija distribucije od X_i, tada je

$$F_{Y_n}(y) = P\{Y_n \leqslant y\} = P\{X_1 \leqslant y, X_2 \leqslant y, ..., X_n \leqslant y\} =$$

$$= \text{ (nezavisnost) } = \prod_{i=1}^{n} P\{X_i \leqslant y\} = F_X(y)^n.$$

- Možemo pokušati direktno procjeniti gustoću vjerojatnosti od Y (pomoću histograma). Međutim i to je prilično teško budući su nizovi, tj. uzorci ekstremnih vrijednosti u pravilu kratki, a sami ekstremi rijetki događaji.

- Treća mogućnost je pokušati pronaći graničnu razdiobu za Y_n (ako takva postoji).

  Promatramo:

$$F_{Y_n}(y) = F_X(y)^n \quad \to \quad \begin{cases} 0, & \text{ako je } F_X(y) < 1 \\ 1, & \text{ako je } F_X(y) = 1 \end{cases} \text{, kada } n \to \infty.$$

  Pretpostavimo da granična distribucija postoji. To znači da postoje brojevi a_n i b_n takvi da

$$Y_n{}' = a_n \cdot Y_n + b_n \longrightarrow Y,$$

  pri čemu varijabla Y nije degenerirana. Tada je $Y_n = (Y_n{}' - b_n)/a_n$ i vrijedi

$$F_{Y_n}(y) = P\{Y_n \leqslant y\} = P\{Y_n{}' \leqslant a_n y + b_n\} = F_{Y_n'}(a_n y + b_n) \approx F_Y(a_n y + b_n),$$

  za dovoljno veliko n. Ako označimo $F_Y \equiv G$ vidimo da varijabla Y_n približno ima razdiobu $G(a_n y + b_n)$.

  Odredimo svojstva od G. Za $n, N \in \mathbb{N}$ gledamo slučajne varijable (nezavisne i jednako distribuirane):

$$\begin{array}{cc} X_1, \ldots X_n & \max\{X_1, ...X_n\} \\ X_{n+1}, \ldots X_{2n} & \max\{X_{n+1}, ...X_{2n}\} \\ \vdots & \vdots \\ X_{(N-1)n+1}, \ldots X_{Nn} & \max\{X_{(N-1)n+1}, ...X_{Nn}\} \end{array}$$

Kada $n \to \infty$, $\max\{X_1, \ldots, X_{Nn}\}$ možemo gledati na dva načina:

1. kumulativna funkcija distribucije svakog retka je $G(X) \Rightarrow$ max od N takvih ima CDF jednaku $G(X)^N$
2. max svih X_i zajedno mora imati CDF oblika $G(a_N X + b_N)$.

Zaključujemo da G mora zadovoljavati

$$[G(X)]^N = G(a_N X + b_N), \qquad \forall N \in \mathbb{N},$$

što je tzv. ***postulat stabilnosti*** za razdiobu ekstremnih vrijednosti (Frechet, 1927). Razdioba ekstremnih vrijednosti, G je stabilna u smislu da je maksimum od N varijabli s razdiobom G opet varijabla s razdiobom G do na linearnu transformaciju.

Pokazuje se da postulat stabilnosti zadovoljavaju samo 3 familije razdioba ovisno o tome da li je $a_N = 1$ ili $a_N \neq 1$.
Izvod za $a_N = 1$:

$$[G(x)]^{NM} = [G(x + b_N)]^M = G(x + b_N + b_M),$$

$$[G(x)]^{NM} = G(x + b_{NM}), \qquad \forall x,$$

$$\Rightarrow b_{NM} = b_N + b_M$$
$$\Rightarrow b_N = -\alpha \ln N, \quad \alpha = \text{ konst.}$$

Odatle imamo redom:

$$[G(x)]^N = G(x - \alpha \ln N) \qquad / \ln / \cdot(-1) / \ln$$

$$\ln N + \underbrace{\ln(-\ln G(x))}_{h(x)} = \ln(-\ln G(x - \alpha \ln N))$$

$$\ln N + h(x) = h(x - \alpha \ln N)$$

$$h(x) - h(x \underbrace{-\alpha \ln N}_{z}) = -\ln N, \qquad \forall N,$$

odnosno

$$h(x) + h(x + z) = \frac{z}{\alpha} \qquad \forall z.$$

Ako stavimo $x = 0$, te nakon toga preimenujemo z u x dobivamo:

$$h(0) - h(x) = \frac{x}{\alpha},$$

odnosno

$$h(x) = h(0) - \frac{x}{\alpha}$$

iz čega proizlazi da je $h(x)$ linearna funkcija. Budući je $h(x) = \ln(-\ln(G(x)))$, to je $h(x)$ neopadajuća i posljedično je $\alpha > 0$. Dobili smo:

$$G(x) = \exp\left[-\exp\left(-\frac{x-\alpha h(0)}{\alpha}\right)\right],$$

što uz oznaku $\alpha h(0) \equiv x_0$ daje

$$G(x) = \exp\left[-\exp\left(-\frac{x-x_0}{\alpha}\right)\right].$$

Dobivena razdioba je tzv. *Gumbellova razdioba* ili asimptotska razdioba ekstremnih vrijednosti tipa 1. Ta razdioba je određena s 2 parametra.

Svojstva Gumbelove razdiobe: $F_Y(x) = \exp\left[-\exp(-\frac{x-x_0}{\alpha})\right], \quad x \in \mathbb{R}, \quad \alpha > 0$

- Gustoća: $F'_Y(x) = f_Y(x) = \frac{1}{\alpha}\exp\left\{-\left[\exp(-\frac{x-x_0}{\alpha}) + \frac{x-x_0}{\alpha}\right]\right\}$.
- Mod ili tjeme:

$$F''_Y(x) = f'_Y(x) =$$
$$= -\frac{1}{\alpha}\exp\left\{-\left[\exp(-\frac{x-x_0}{\alpha}) + \frac{x-x_0}{\alpha}\right]\right\}\left(\frac{1}{\alpha} - \frac{1}{\alpha}\exp\left[-\frac{x-x_0}{\alpha}\right]\right) = 0$$
$$\Rightarrow \frac{1}{\alpha} - \frac{1}{\alpha}\exp\left[-\frac{x-x_0}{\alpha}\right] = 0$$
$$\Rightarrow \exp\left[-\frac{x-x_0}{\alpha}\right] = 1$$

Dakle, parametar x_0 je mod ili tjeme Gumbeove razdiobe.

- momenti:
  - $\mu_Y = x_0 + \alpha\gamma$, pri čemu je $\gamma \approx 0.577$ tzv. Eulerova konstanta,
  - $\sigma_Y^2 = \frac{\pi^2}{6} \cdot \alpha^2$,
  - $C_s \approx 1.14$ *(ne ovisi o parametrima)*

Za standardnu varijablu $W = \frac{X-X_0}{\alpha}$ postoje tablice za $F(w)$ i $f(w)$.

Napomena. Neka je F_X (jedna te ista) razdioba varijabli $X_i, i = 1, 2, \ldots$. Tip razdiobe ekstrema od X_i ovisi o ponašanju razdiobe F_X na rubovima. Ako f_X eksponencijalno opada u području ekstrema (npr. normalna razdioba) onda se dobiva Gambellova razdioba.

Analiza ekstremnih vrijednosti obično ima za cilj procjeniti kvantile za velike kumulativne vjerojatnosti. Te kvantile obično nije moguće procjeniti direktvno iz relativno

kratkih nizova. Procjena putem razdiobe ekstrema je razumna i objektivna. Drugi način za iskazivanje kvantila je ***povratni period***.

Intuitivno, ako vršimo nezavisna mjerenja (opažanja) onda će $q_{0.99}$ biti premašen *u prosjeku* jednom u 100 mjerenja. Općenito, za premašiti $(1-p)$-ti kvantil u prosjeku je treba izvršiti $1/p$ mjerenja. Ako su mjerenja povezana s vremenom, onda se njihov broj može iskazati potrebnim vremenom, tj. povratnim periodom T kojemu pripada određena povratna vrijednost X, odnosno traženi kvantil:

$$T = T(X) = \frac{1}{1 - F(X)}.$$

Formula vrijedi ako imamo jedno opažanje godišnje, npr. godišnji maksimum i T se tada izražava u godinama. Kada bi imali N_g opažanja godišnje, vrijedilo bi:

$$T = T(X) = \frac{1}{N_g}\frac{1}{1 - F(X)}.$$

Npr. za godišnji ($N_g = 1$) maksimum temperature, ako je $t = 43^oC$, te $F(t) = 0.99$. Onda je $T = 100$ godina, tj očekujemo da $t = 43^oC$ bude u prosjeku premašena jednom u 100 godina. Pri tom $T(43^oC) = 100$ ne znači da će se temperatura od 43^oC zaista i premašiti u 100 godina. Vjerojatnost za događaj $A = \{T > 43^oC\}$ u bilo kojoj godini iznosi $p = 0.01$, tj. radi se o Bernulijevoj slučajnoj varijabli:

$$X \sim \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}.$$

Broj godina koje treba čekati do *prve* pojave događaja (uz pretpostavku da su događaji međusobno nezavisni) je i sam slučajna varijabla Y s geometrijskom razdiobom.

$$P\{Y = y\} = (1-p)^{y-1}p,$$

te vrijedi:

$$\mathbb{E}(Y) = \frac{1}{p}.$$

Drugim riječima, ako pri osmatranju ili mjerenju neke slučajne veličine želimo premašiti $(1-p)$-ti kvantil, u prosjeku je potrebno izvršiti $1/p$ nezavisnih osmatranja. Ako su osmatranja vezana uz vrijeme, onda se njihov broj može iskazati i potrebnim vremenom, tj. povratnim periodom. *Povratni period* je, dakle, srednje ili očekivano vrijeme potrebno da odgovarajuća *povratna vrijednost* (kvantil) bude premašena, odnosno prosječno vrijeme potrebno da nastupi rečeni ekstremni događaj.

Opća razdioba ekstrema (Jenkinson, 1969): Postoje samo tri tipa asimptotskih razdioba ekstrema (ovisno o ponašanju na repovima) i sve se mogu opisati sa

$$F_Y(x) = e^{(-e^{-y(x)})},$$

pri čemu je

$$x = x(y) = x_0 + \alpha \frac{1 - e^{-ky}}{k}.$$

Za $k \neq 0$, vrijedi

$$F_Y(x) = e^{-(1-\frac{k}{\alpha}(x-x_0))^{1/k}}.$$

Svojstva u ovisnosti o novom parametru k:

$k \to 0$ $\qquad x = x_0 + \alpha y$ $\qquad$ Gumbell (Tip I),

$k < 0 \quad x \to x_0 + \frac{\alpha}{k} \quad$ kad $\quad y \to -\infty \Rightarrow x \in [x_0 + \frac{\alpha}{k}, +\infty >$ $\qquad$ Frechet (Tip II),

$k > 0 \quad x \to x_0 + \frac{\alpha}{k} \quad$ kad $\quad y \to +\infty \Rightarrow x \in < -\infty, x_0 + \frac{\alpha}{k}]$ $\qquad$ Weibul (Tip III).

Prilagodba empirijskoj razini se vrši metodom maksimlne vjerodostojnosti (ML) što zahtijeva numerički postupak.

4.7 Bivarijantna normalna razdioba

Ova razdioba predstavlja poopćenje jednodimenzionalne normalne razdiobe. $X \sim N(\mu, \sigma)$. Za slučajni vektor kažemo da ima bivarijantnu normalnu razdiobu i pišemo $(X, Y) \sim$ BND$(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ ako mu funkcija gustoće glasi:

$$f(x, y;\ \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \frac{1}{\sigma_x \sigma_y 2\pi \sqrt{1-\rho^2}} \cdot \exp\left\{ \frac{-1}{2\left(1-\rho^2\right)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \frac{x-\mu_x}{\sigma_x} \frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}.$$

Poseban slučaj: $\mu_x = \mu_y = 0$ (centriranje), $\sigma_x = \sigma_y = 1$ (normiranje), vodi na gustoću:

$$f(x, y;\ \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2\left(1-\rho^2\right)} \left[x^2 - 2\rho xy + y^2 \right] \right\},$$

koja se, slično kao i kod standardne normalne razdiobe, $N(0, 1)$, dobije transformacijom varijabli:

$$(X, Y) \longrightarrow \left(\frac{X-\mu_x}{\sigma_x}, \frac{Y-\mu_y}{\sigma_y} \right)$$

.

Marginalne gustoće:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy =$$
$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left\{ \frac{-1}{2\left(1-\rho^2\right)} \left[x^2 - 2\rho xy + y^2\right] \right\} =$$

$$= \left[\begin{array}{ll} y' = \frac{y-\rho x}{\sqrt{1-\rho^2}} & y'^2 = \frac{y^2-2\rho xy+\rho^2x^2}{1-\rho^2} \\ dy' = \frac{dy}{\sqrt{1-\rho^2}} & \frac{y^2-2\rho xy}{1-\rho^2} = y'^2 - \frac{\rho^2}{1-\rho^2}x^2 \end{array} \right] =$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(x^2 + y'^2\right) \right] dy' =$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y'^2}}_{1} \Rightarrow$$

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

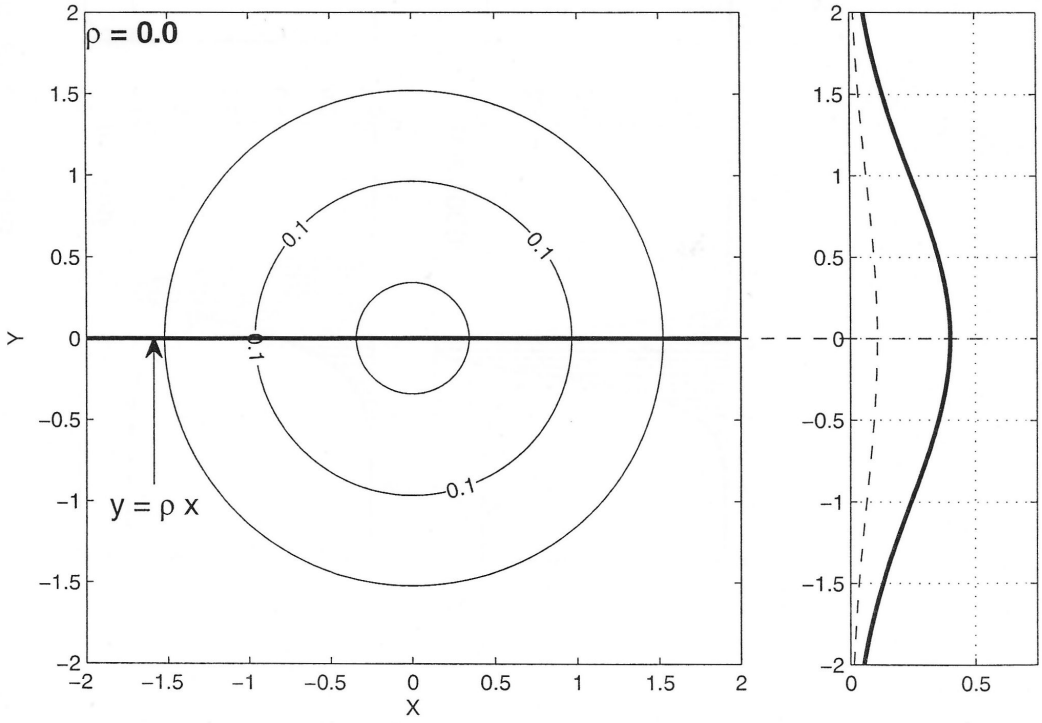
Zaključujemo da je $X \sim \mathrm{N}(0,1)$, te $Y \sim \mathrm{N}(0,1)$. Općenito je $X \sim \mathrm{N}(\mu_x, \sigma_x)$, te $Y \sim \mathrm{N}(\mu_y, \sigma_y)$. Odatle izlazi i značenje prva četri parametra: $\mu_x = \mathrm{E}(X), \sigma_x^2 = \mathrm{Var}(X)$, $\mu_y = \mathrm{E}(Y), \sigma_x^2 = \mathrm{Var}(Y)$.

Uvjetne gustoće:

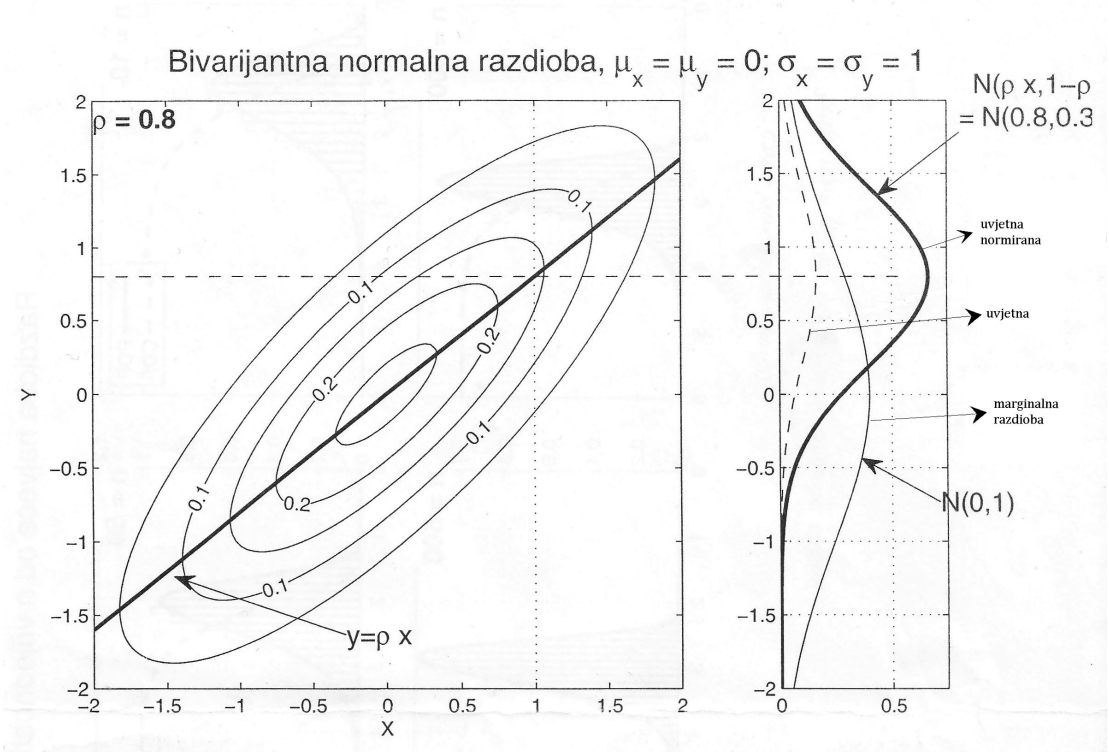
$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} =$$
$$= \frac{\sqrt{2\pi}}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(x^2-2\rho xy+y^2\right)+\frac{1}{2}x^2} =$$
$$= \frac{1}{\sqrt{2\pi\left(1-\rho^2\right)}} e^{-\frac{1}{2(1-\rho^2)}(y-\rho x)^2} \Rightarrow$$

$$f_{Y|X=x}(y) \sim \mathrm{N}\left(\rho x, \sqrt{1-\rho^2}\right).$$

Uvjetna razdioba od Y uz uvjet $X = x$ je $\mathrm{N}\left(\rho x, \sqrt{1-\rho^2}\right)$. Ako je $\rho \neq 0$, onda poznavanje x-a smanjuje neizvjesnost oko y-na. Što je ρ veći elipse su uže, tj. manja je neizvjesnost.



Slika 4.7



Slika 4.8

Pretpostavimo da je $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$. Tada je

$$\begin{aligned}
\operatorname{Cov}(X,Y) &= \mathbb{E}\left((X-\mu_x)(Y-\mu_y)\right) = \mathbb{E}(X\cdot Y) = \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int\int xy e^{-\frac{1}{2(1-\rho^2)}\left(x^2-2\rho xy+y^2\right)} dx\, dy = \left[y' = \ldots, dy' = \ldots\right] = \\
&= \frac{1}{2\pi} \int\int x\left(y'\sqrt{1-\rho^2}+\rho x\right) e^{-\frac{1}{2}\left(x^2+y'^2\right)} dx\, dy' = \\
&= \frac{1}{2\pi} \int\int \rho x^2 e^{-\frac{1}{2}\left(x^2+y'^2\right)} dx\, dy' + \underbrace{\frac{1}{2\pi} \int\int xy' e^{-\frac{1}{2}\left(x^2+y'^2\right)} dx\, dy'}_{0} = \\
&= \rho \underbrace{\frac{1}{\sqrt{2\pi}} \int x^2 e^{-\frac{1}{2}x^2} dx}_{\sigma_x^2=1} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}y'^2} dy'}_{1} = \rho.
\end{aligned}$$

Da smo imali $\sigma_x \neq 1$, $\sigma_y \neq 1$ dobili bi $\operatorname{Cov}(X,Y) = \sigma_x\sigma_y\rho$, odnosno

$$\rho = \frac{\operatorname{Cov}(X,Y)}{\sigma_x\sigma_y} \qquad = \text{koeficijent korelacije.}$$

Za $\rho = 0$ imamo:

$$f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$$

$$f(x,y) = f_X(x)\cdot f_Y(y) \Rightarrow X \text{ i } Y \text{ nezavisne.}$$

Ako je $(X,Y) \sim \text{BND}$***, onda vrijedi:*** $\rho = 0 \iff X$ ***i*** Y ***su nezavisne.***

U tom slučaju je

$$f_{Y|X=x} = \frac{f(x,y)}{f_X(x)} = \{\rho = 0\} = \frac{f_X(x)\cdot f_Y(y)}{f_X(x)} = f_Y(y),$$

tj. uvjetna gustoća je jednaka marginalnoj (informacije o varijabli X ništa ne govore o variabli Y).

Poglavlje 5

PROVJERA STATISTIČKIH PRETPOSTAVKI

5.1 Testovi

Statistička pretpostavka (hipoteza) je svaka tvrdnja u vezi nepoznate funkcije distribucije neke slučajne varijable (ili pripadnog osnovnog skupa tj. populacije).

Statistički test je postupak kojim na temelju uzorka (skup podataka dobiven mjerenjem ili opažanjem) ispitujemo valjanost neke statističke pretpostavke. razlikujemo dvije vrste testova:

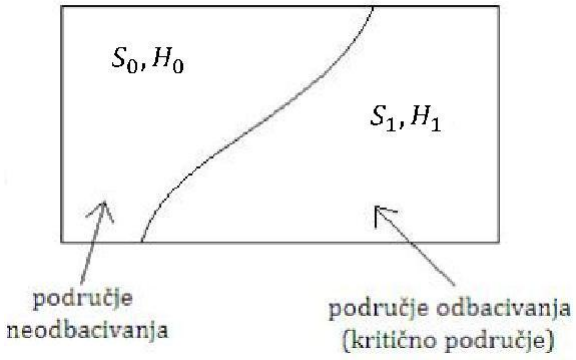
- Parametarski testovi ('vanjske' pretpostavke uključuju konkretnu teorijsku razdiobu).
- Neparametarski testovi ('vanjske' pretpostavke NE uključuju konkretnu teorijsku razdiobu).

Hipoteze:

- H_0, hipoteza koju želimo testirati (obično odbaciti), nul hipoteza.
- H_1, alternativna hipoteza (često "nije H_0").

Test je pravilo koje svakom uzorku pridružuje jednu od dvije moguće odluke:

- H_0 se odbacuje, tj prihvaća se H_1
- H_0 se ne odbacuje (ALI se ne može reći niti da se prihvaća).



Skup svih mogućih uzoraka je, dakle, podijeljen na dva dijela, S_0 i S_1, koji se konstruiraju tako da S_0 bude što vjerojatnije ako je H_0 točno, te da istovremeno S_1 bude što vjerojatnije ako je H_1 točno. Ako se uzorak kojim raspolažemo nađe u S_0, prihvatit ćemo H_0, a ako se nađe u S_1, odbacit ćemo ga. S tim u vezi, uz svaki test se vežu dva broja:

$\alpha = P\{(X_i)_{i=1}^n \in S_1 | H_0 \text{ točno}\}$ = pogreška I vrste (odbacivanje *ispravnog* H_0),

$\beta = P\{(X_i)_{i=1}^n \in S_0 | H_1 \text{ točno}\}$ = pogreška II vrste (prihvaćanje *neispravnog* H_0).

Moć testa se definira kao $= 1 - \beta$. To je vjerojatnost odbacivanja neispravne nul hipoteze, tj. vjerojatnost da će test otkriti da istraživani fenomen 'zaista' postoji. Naravno, što je β manji to je test bolji.

Struktura te provedba testa:

1. Formulirati hipoteze H_0 i H_1.
2. Odabrati statistiku, tj. funkciju uzorka: $U = U(X_1, X_2, \ldots, X_n)$ koja je pogodna za problem koji rješavamo.
3. Odrediti razdiobu vjerojatnosti $F(u|H_0)$, tj. nul distribuciju, odnosno distribuciju od U (egzaktnu ili približnu) koja vrijedi ako je H_0 ispravno. $\longrightarrow$ Ključni korak koji ograničava izbor H_0.
4. **a)** Odabrati α (nivo značajnosti, $\alpha = 0.05, \alpha = 0.01$), te odrediti da li se kritično područje nalazi na lijevom, desnom ili na oba kraja nul distribucije. Pitamo se koje vrijednosti od U podržavaju H_0, a koje H_1.

   b) Odrediti granice za kritično područje, tj:

   $P\left\{|U| \geqslant u_{1-\frac{\alpha}{2}}\right\} = \alpha \longrightarrow$ dvostrano

   $P\{U > u_{1-\alpha}\} = \alpha \longrightarrow$ jednostrano (u desno)

   $P\{U < u_{\alpha}\} = \alpha \longrightarrow$ jednostrano (u lijevo)
5. Izračunati u iz uzorka i odlučiti (npr. dvostrani test)

   $|u| > u_1 - \frac{\alpha}{2} \longrightarrow$ odbacujemo H_0 i prihvaćamo H_1

$|u| < u_{\frac{\alpha}{2}} \longrightarrow$ na temelju postojećeg uzorka ne možemo odbaciti H_0 (na nivou značajnosti α)

Napomena: Po samoj konstrukciji, testom ne možemo pokazati ispravnost nul-hipoteze. Testom ustanovljavamo (oslanjajući se na određene 'vanjske' pretpostavke) u kojoj je mjeri uzorak u skladu s nul-hipotezom. Ako uzorak 'dovoljno' odstupa od nul-hipoteze onda je odbacujemo (uz rizik da napravimo pogrešku prve vrste). Sve što u protivnom možemo reći jest da uzorak ne daje dovoljno dokaza (engl. *evidence*) protiv nul-hipoteze.

Primjer: Da li su ljeta zadnjih 30 *godina toplija nego prije?*

Za period 1862–1980. izračunamo μ_0 *i varijancu* σ_0 *(za srednju temperaturu za srpanj). Pretpostavimo da se varijanca nije mijenjala u cijelom razdoblju te da se srednja mjesečna temperatura vlada po normalnoj razdiobi. Imamo uzorak od* 30 *mjesečnih srednjaka za srpanj* $(x_1, x_2, \ldots, x_n)$ *za period 1981–2010. Dakle,*

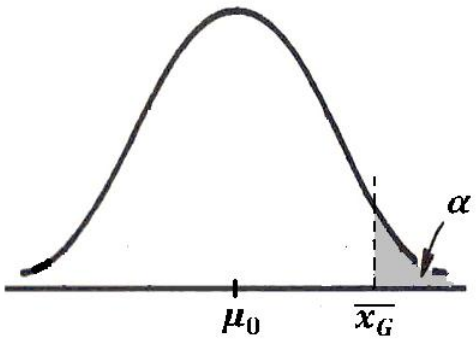
Za 1862–1980 je $X \sim \mathrm{N}(\mu_0, \sigma_0)$,

za 1981–2010 je $X \sim \mathrm{N}(\mu, \sigma_0)$.

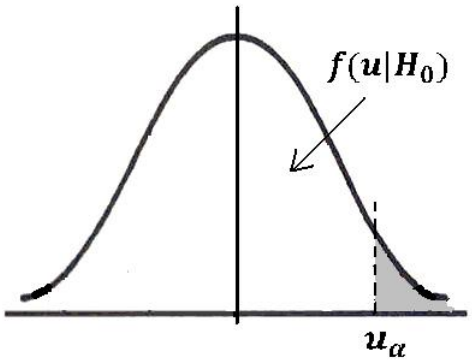
Provedimo testiranje.

1. $H_0 : \mu = \mu_0, \quad H_1 : \mu > \mu_0.$
2. $\overline{X} = \frac{1}{30} \sum_{i=1}^{30} X_i.$
3. *Nul distribucija je* $X \sim \mathrm{N}(\mu_0, \frac{\sigma_0}{\sqrt{n}})$ *(iz pretpostavke da se raspodjela ne mijenja te da su podaci iz utorka uzeti nezavisno).*
4. a) $\alpha = 0.05$ *a kritično područje je na desnom kraju (zbog* H_1*).*

   b) *Želimo granični* $\overline{x}_G$*, tako da ako vrijedi* H_0*, krivo odbacivanje napravimo s (malom) vjerojatnošću* α*.*



5. $\overline{x} > \overline{x}_G \Rightarrow$ *odbacujemo* H_0 *i prihvaćamo* H_1
   $\overline{x} < \overline{x}_G \Rightarrow$ *ne možemo odbaciti* H_0



U praksi se $\overline{X}$ *transformira tj. koristi se statistika*

$$U = \frac{\overline{X} - \mu_0}{\frac{\sigma_0}{n}} \sim \mathrm{N}(0,1) \text{ ako je } H_0 \text{ točno.}$$

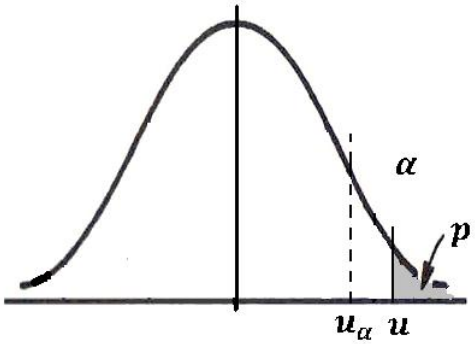
$$P\{U > u_{1-\alpha}\} = \alpha$$

.

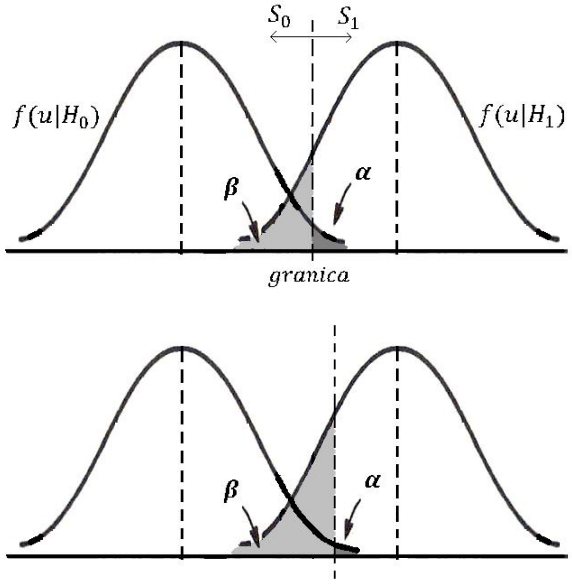
Napomena: U praksi se često, umjesto da se iskaže odluka, samo navede tzv. p vrijednost.

$$p = P\{U > u\}$$

$u \longrightarrow$ *vrijednost dobivena iz uzorka*



5.2 O pogrešci prve i druge vrste

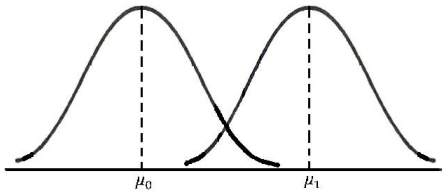


Moguća je samo jedna pogreška (I ili II vrste) ovisno o tome da li je točno H_0 ili H_1. U praksi se poznaje samo $f(u|H_0)$ (i to uz neke apriorne, 'vanjske' pretpostavke, koje testom ne provjeravamo). Zaključuje se samo na temelju $f(u|H_0)$ i α(!!), dok β ovisi o α i o $f(u|H_1)$. Dakle, za izračunati moć testa ($1-\beta =$ vjerojatnost da će test prepoznati ispravan H_1) potrebno je precizno zadati H_1.

Ako α pada, onda β raste. Istovremeno smanjiti α i β moguće je povećanjem duljine uzorka, jer tada varijance odgovarajućih statistika padaju.

Primjer:

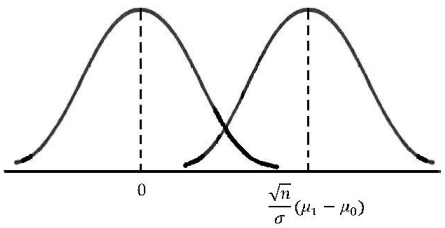
$$X_i \sim \mathrm{N}(\mu,\sigma) \Longrightarrow \overline{X} = \frac{1}{n}\sum_i \sim \mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$$



Za $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$

$\longrightarrow$ povećanje duljine uzorka vodi ka smanjenju varijance

$\longrightarrow$ u praksi radimo sa $U = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$



H_0 točno $\Rightarrow U \sim \mathrm{N}(0,1)$

H_1 točno $\Rightarrow$

$$\mathbb{E}(U) = \frac{\sqrt{n}}{\sigma}\mathbb{E}(\overline{x}-\mu_0) = \frac{\sqrt{n}}{\sigma}(\mu_1-\mu_0) \Rightarrow$$

$$U \sim \mathrm{N}(\frac{\sqrt{n}}{\sigma}(\mu_1-\mu_0), 1).$$

Sljedeća tvrdnja je potrebna kod izgradnje mnogih uobičajenih testova.

Tvrdnja: Neka su $X_1, \ldots, X_n$ nekorelirane slučajne varijable, $\mu = \mathbb{E}(X_i), \sigma^2 = \mathrm{Var}(X_i)$.

Tada je $\mathbb{E}(\frac{1}{n-1}\sum(X_i-\overline{X})^2)=\sigma^2$.

Dokaz: Bez smanjenja općenitosti stavimo $\mu=0$, tj. $\mathbb{E}(X_i)=0$, pa računjamo:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})^2 &= \sum_i X_i^2 - 2\sum_i X_i\overline{X} + \sum_i \overline{X}^2 = \\ &= \sum_i X_i^2 - n\overline{X}^2 = \\ &= \sum_i X_i^2 - n(\frac{1}{n}\sum_i X_i)^2 = \\ &= \sum_i X_i^2 - \frac{1}{n}(\sum_i X_i)^2 = \\ &= \sum_i X_i^2 - \frac{1}{n}\left(\sum_i X_i^2 + \sum_{i\neq j} X_iX_j\right).\end{aligned}$$

Odatle je

$$\begin{aligned}\mathbb{E}(\sum_{i=1}^{n}(X_i-\overline{X})^2) &= (1-\frac{1}{n})\sum_i \mathbb{E}(X_i^2) - \frac{1}{n}\sum_{i\neq j}\underbrace{\mathbb{E}(X_i\cdot X_j)}_{0} = \\ &= \frac{n-1}{n}n\sigma^2 = (n-1)\sigma^2,\end{aligned}$$

i konačno

$$\Rightarrow \sigma^2 = \frac{1}{n-1}\mathbb{E}(\sum_{i=1}^{n}\left(x_i-\overline{x})^2\right).$$

5.3 Testovi za srednju vrijednost te varijancu

- Neka je X slučajna varijabla koja predstavlja populaciju od interesa, te označimo $\mu=\mathbb{E}(X), \sigma^2=\mathrm{Var}(X)$.
- Neka su $X_1, X_2, \ldots, X_n$ nezavisne slučajne varijable raspodijeljene kao i X (što je i precizna definicija *slučajnog uzorka* uzetog iz populacije). Radi se, dakle, o jednom slučajnom vektoru. Skup podataka (brojeva) $x_1, x_2, \ldots, x_n$ dobivenih mjerenjem ili opažanjem samo je jedna od mogućih realizacija tog slučajnog vektora.
- Statistike prikladne za tvrdnje o srednjaku i varijanci:

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i, \quad S^2=\sum_{i=1}^{n}\frac{1}{n}(X_i-\overline{X})^2.$$

- Neki momenti (dobivaju se korištenjem nezavisnosti, a služe kod određivanja odgovarajućih zakona razdiobe):

$$\mathbb{E}(\overline{X})=\mu, \mathrm{Var}(\overline{X})=\frac{\sigma^2}{n},$$

$$\mathbb{E}(S^2) = \frac{n-1}{n}\sigma^2.$$

- Razdiobe vjerojatnosti:
  - $\overline{X} \sim \mathrm{N}(\mu, \sigma/\sqrt{n})$ i to egzaktno ako je $X \sim \mathrm{N}(\mu, \sigma)$, ili približno ako je n dovoljno velik.
  - $nS^2/\sigma^2 \sim \chi^2(n-1)$ ako je $X \sim \mathrm{N}(\mu, \sigma)$.
- **Test za srednjak uz poznatu varijancu:** Pretpostavljamo da je μ_0 zadano te σ^2 poznato. Slijede standardni koraci u izgradnji testa.

  1) $H_0 : \mu = \mu_0$
  $H_1 : \mu \neq \mu_0$, ili $H_1' : \mu > \mu_0$, ili $H_1'' : \mu < \mu_0$.
  2) Statistika
  $$U = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$
  3) Nul-distribucija: Ako vrijedi H_0, onda je $U \sim \mathrm{N}(0,1)$ i to egzaktno ili približno, kako je napisano gore.
  4) Za alternativnu hipotezu H_1 biramo dvostrano kritično područje, za H_1' područje na desnoj strani, a za H_1'' na lijevoj strani razdiobe. Za odabrani novo značajnosti α, kritična vrijednost $u_{1-\alpha}$ se određuje iz tablica normalne razdiobe prema $P(U < u_{1-\alpha}) = 1 - \alpha$.
  5) Iz uzorka se izračuna u (konkretna, realizirana vrijednost statistike U) i donese odluka. Npr. ako je test dvostrani, onda
     * $|u| > u_{1-\alpha/2} \Rightarrow$ odbacujemo H_0 i prihvaćamo hipotezu H_1,
     * $|u| < u_{1-\alpha/2} \Rightarrow$ ne možemo odbaciti H_0. To nipošto ne znači da je ta hipoteza testom potvrđena kao valjana, već samo to da podaci ne daju dovoljno argumenata za njeno odbacivanje (na zadanom nivou značajnosti).
- **Studentova ili t-razdioba**
  - Neka su $X \sim \mathrm{N}(0,1)$ te $U \sim \chi^2(n)$ nezavisne slučajne varijable. Gosset je, pod pseudonimom Student, 1908. godine odredio funkciju gustoće slučajne varijable $T = \frac{X}{\sqrt{U/n}}$, koja se po njemu naziva Studentova ili t-razdioba (s n stupnjeva slobode), dok se pripadna varijabla zove Studentova varijabla. Pišemo $T \sim t(n)$.

– Neka su $X_1, X_2, \ldots, X_n$ nezavisne slučajne varijable, $X_i \sim \mathrm{N}(\mu, \sigma)$. Neka je

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}).$$

Tada varijabla

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}}$$

ima Studentovu razdiobu s $n-1$ stupnjeva slobode, $T \sim t(n-1)$.
Napomena: Bitno je da razdioba statistike T ne ovisi o nepoznatom σ.

– t-razdioba je simetrična s obzirom na $y-$os.

– Kada $n \to \infty$ t-razdioba teži k $\mathrm{N}(0,1)$. U praksi se za $n > 30$ umjesto Studentove koristi standardna normalna razdioba, $\mathrm{N}(0,1)$.

- **Test za srednjak uz nepoznatu varijancu:** Pretpostavljamo da je μ_0 zadano, ali da je σ^2 nepoznato. Slijede standardni koraci u izgradnji testa.

1) $H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$, ili $H_1' : \mu > \mu_0$, ili $H_1'' : \mu < \mu_0$.

2) Za dobiti prikladnu statistiku umjesto nepoznate standardne devijacije ($\mathrm{Var}(\overline{X})^{1/2}$) stavimo *procjenu*, $\frac{S}{\sqrt{n-1}}$. Tako dobivamo

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n-1}}} = \frac{\overline{X} - \mu}{\frac{1}{\sqrt{n(n-1)}}\sqrt{\sum(X_i - \overline{X})^2}}.$$

3) Ako vrijedi H_0 te ako su varijable X_i normalno raspodijeljene, $X_i \sim \mathrm{N}(\mu_0, \sigma)$, onda je $T \sim t(n-1)$.
Ako vrijedi H_0 i n je *dovoljno velik* onda je *približno* ispunjeno $T \sim \mathrm{N}(0,1)$, i to bez pretpostavke da su X_i Gaussove varijable.

4) i 5) Radi se kao i u prethodnom slučaju, ali se gledaju tablice za t-razdiobu, odnosno normalnu razdiobu ako je $n > 30$.

- **Test jednakosti dvaju srednjaka za *velike* uzorke:** Pretpostavimo da su $\{X_{11}, X_{12}, \ldots, X_{1n}\}$, $\{X_{21}, X_{22}, \ldots, X_{2m}\}$ dva slučajna uzorka iz međusobno *nezavisnih* populacija. Neka je $\mathbb{E}(X_{1i}) = \mu_1$, $\mathbb{E}(X_{2i}) = \mu_2$, $\mathrm{Var}(X_{1i}) = \sigma_1$, $\mathrm{Var}(X_{2i}) = \sigma_2$.

1) $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$, ili $\mu_1 > \mu_2$, ili $\mu_1 < \mu_2$.

2) i 3) Stavimo $\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$, $\overline{X_2} = \frac{1}{m}\sum_{i=1}^{m} X_{2i}$. Za *velike* n imat će $\overline{X_1}, \overline{X_2}, \overline{X_1} - \overline{X_2}$ *približno* normalnu raspodjelu. Dakle, bit će

$$\frac{\overline{X_1} - \overline{X_2}}{\left[\mathrm{Var}(\overline{X_1} - \overline{X_2})\right]^{1/2}} \sim \mathrm{N}(0,1).$$

Zbog nezavisnosti je $\mathrm{Var}(\overline{X_1} - \overline{X_2}) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$, što zamijenimo procjenom

$$\mathrm{Var}(\overline{X_1} - \overline{X_2}) = \frac{S_{X_1}^2}{n-1} + \frac{S_{X_2}^2}{m-1}.$$

Dakle, ako vrijedi H_0 onda približno vrijedi

$$U = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_{X_1}^2}{n-1} + \frac{S_{X_2}^2}{m-1}}} \sim \mathrm{N}(0,1).$$

4) i 5) Standardno.

- **Test jednakosti dvaju srednjaka za nezavisne, normalne populacije s nepoznatim, ali *istim* varijancama:** Kao i gore, pretpostavimo da su $\{X_{11}, X_{12}, \ldots, X_{1n}\}$, $\{X_{21}, X_{22}, \ldots, X_{2m}\}$ dva slučajna uzorka iz međusobno nezavisnih populacija $\mathrm{N}(\mu_1, \sigma_1^2)$ i $\mathrm{N}(\mu_2, \sigma_2^2)$, redom.

  1) $H_0 : \mu_1 = \mu_2$
  $H_1 : \mu_1 \neq \mu_2$, ili $\mu_1 > \mu_2$, ili $\mu_1 < \mu_2$.

2) i 3) Stavimo $\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$, $\overline{X_2} = \frac{1}{m}\sum_{i=1}^{m} X_{2i}$. Ako je H_0 točno, onda statistika

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{1}{m+n-2}\left[\sum_{i=1}^{n}\left(X_{1i} - \overline{X_1}\right)^2 + \sum_{i=1}^{m}\left(X_{2i} - \overline{X_2}\right)^2\right]\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

ima Studentovu t-razdiobu s $m + n - 2$ stupnja slobode.

4) i 5) kao i prije
Pojašnjenje: Ako je H_0 točno onda je $\overline{X_1} - \overline{X_2} \sim \mathrm{N}\left(0, \sigma\sqrt{1/n + 1/m}\right)$, dok se za procjenu (nepoznatog) σ^2 iskoristi jednakost:

$$\mathbb{E}\left(\sum_{i=1}^{n}(X_{1i} - \overline{X_1})^2 + \sum_{i=1}^{m}(X_{2i} - \overline{X_2})^2\right) = (n-1)\sigma^2 + (m-1)\sigma^2 = (n+m-2)\sigma^2.$$

- **Test za varijancu:** Pretpostavljamo da imamo nezavisni uzorak iz normalne razdiobe $\mathrm{N}(\mu, \sigma)$. Želimo provjeriti da li varijanca populacije odstupa od zadane vrijednosti σ_0^2 ?

1) $H_0 : \sigma = \sigma_0$
   $H_1 : \sigma \neq \sigma_0$, ili $\sigma > \sigma_0$, ili $\sigma < \sigma_0$.

2) i 3) Ako vrijedi H_0 onda statistika

$$U = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma_0} \right)^2$$

ima χ^2-razdiobu s $n-1$ stupnjeva slobode.

4) Za tri alternativne hipoteze iz točke 1), kritično područje se postavlja na oba kraja, na desni te na lijevi kraj razdiobe, redom.

- **Test za jednakost dviju varijanci:** Ako su $X_1 \sim \chi^2(n)$ i $X_2 \sim \chi^2(m)$ dvije međusobno nezavisne slučajne varijable, onda varijabla

$$F = \frac{\frac{1}{n} X_1}{\frac{1}{m} X_2}$$

ima tzv. Fisherovu ili F-razdiobu s (n, m) stupnjeva slobode.
Kao i gore, pretpostavimo da imamo dva slučajna uzorka $\{X_{11}, X_{12}, \ldots, X_{1n}\}$, $\{X_{21}, X_{22}, \ldots, X_{2m}\}$ iz međusobno nezavisnih populacija $\mathrm{N}(\mu_1, \sigma_1^2), \mathrm{N}(\mu_2, \sigma_2^2)$.

1) $H_0 : \sigma = \sigma_0$
   $H_1 : \sigma \neq \sigma_0$, ili $\sigma > \sigma_0$, ili $\sigma < \sigma_0$.

2) i 3) Ako je H_0 istinito onda statistika

$$F = \frac{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_{1i} - \overline{X_1} \right)^2}{\frac{1}{m-1} \sum_{i=1}^{m} \left(X_{2i} - \overline{X_2} \right)^2} = \frac{\frac{n}{n-1} S_{X_1}^2}{\frac{m}{m-1} S_{X_2}^2}$$

ima F-razdiobu s $(n-1, m-1)$ stupnjeva slobode.

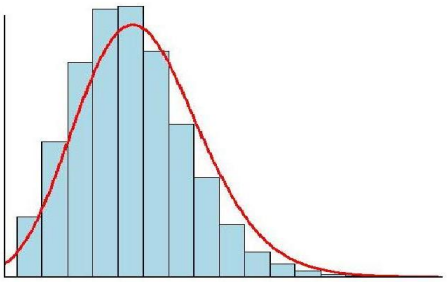
4) i 5) idu standardno, no treba paziti na kritično područje. Tablice F-razdiobe obično sadrže samo kritične vrijednosti na desnom kraju razdiobe. Vjerojatnosti na lijevom kraju se mogu dobiti iz formule:

$$F_p(m, n) = \frac{1}{F_{1-p}(n, m)},$$

pri čemu je s F_p označena inverzna vrijednost kumulativne F-razdiobe, tj. ako je X slučajna varijabla s F-razdiobom, onda vrijedi $P\{X < F_p\} = p$.

5.4 Testiranje uspješnosti prilagodbe teorijske razdiobe empirijskoj razdiobi čestina

Ideja: Izračunati teorijske te empirijske čestine za svaku klasu, usporediti ih i odgovarajućim testom procjeniti slaganje teorijske i empirijske distribucije.
U tu svrhu može se koristiti **Pearson-ov** χ^2-**test**:



1) H_0 : Uzorak potječe iz zadane razdiobe
   H_1 : Uzorak ne potječe iz zadane razdiobe

2) Statistika U glasi:

$$U = \sum_{i=1}^{k} \frac{(f_{t_i} - f_{e_i})^2}{f_{t_i}},$$

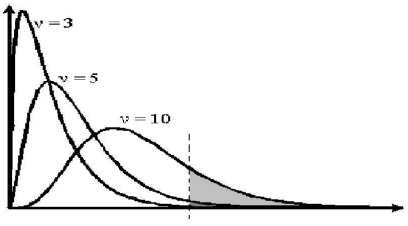
   pri čemu $t \ldots$ označava teorijsku, $e \ldots$ empirijsku, $i \ldots$ i-tu klasu, a $k-$ broj klasa, dok f označava *apsolutne* čestine (koliko je podataka stvarno osmotreno, te koliko bi ih teorijski trebalo biti osmotreno u svakoj klasi).

3) Teorem (Pearson): Neka su podaci međusobno nezavisni, tj. neka je uzorak slučajan. Tada, ako je H_0 točno, te ako je svaka teorijska klasa dovoljno brojna (u praksi, barem 5 podataka), onda je razdioba statistike U bliska χ^2 razdiobi s $\nu = k - p - 1$ stupnjeva slobode, pri čemu je p broj parametara teorijske razdiobe.

   Za postizanje dovoljno brojne klase moguća su dva pristupa; združiti krajnje razrede ili pak unaprijed odrediti razrede tako da teorijske čestine budu dovoljno velike.

4) Zadamo α. Pitamo se koje vrijednosti od U podupiru H_1 ?

   $\Rightarrow$ Kritično područje je *uvijek* na desnoj strani.
   $u_{1-\alpha}$ *odredimo preko* $\chi^2(k-p-1)$ *i tablica*



5) $u > u_{1-\alpha}$ odbacujemo H_0 i prihvaćamo H_1
   $u < u_{1-\alpha}$ ne možemo odbaciti H_0

5.5 Test nezavisnosti u tablici kontingencije

Tablica kontingencije ili povezanosti sadrži čestine događaja razvrstanih u klase pomoću (najčešće) dva ili više kriterija. Određuje empirijsku združenu razdiobu vjerojatnosti.

	B_1	$\dots$	B_s	$\sum$
A_1	f_{11}	$\dots$	f_{1s}	$f_{1.}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	f_{r1}	$\dots$	f_{rs}	$f_{r.}$
$\sum$	$f_{.1}$	$\dots$	$f_{.s}$	N

$N-$ ukupan broj podataka
$f_{ij}-$ združene čestine
$f_{.j}, f_{i.}-$ marginalne čestine
A_i- događaji iz I klase
B_j- događaji iz II klase

Pomoću tablice kontingencije možemo provjeriti da li su neki događaji (odnosno odgovarajuće slučajne varijable) nezavisni. Naime, ako su događaji A_i, B_i međusobno nezavisni onda $\forall i, j$ vrijedi

$$P(A_i \cap B_j) = \underbrace{P(A_i)}_{p_{i.}} \cdot \underbrace{P(B_j)}_{p_{.j}}.$$

Uz osmotrene čestine f_{ij}, te pretpostavku da su A_i i B_i nezavisni, ML metoda daje sljedeće procjene marginalnih vjerojatnosti:

$$\hat{p}_{i.} = \frac{f_{i.}}{N}, \quad \hat{p}_{.j} = \frac{f_{j.}}{N}$$

χ^2 test:

1) $H_0 : \frac{f_{ij}}{N} = \frac{f_{i.}}{N}\frac{f_{j.}}{N}$
H_1 : nije H_0

2) i 3) Statistika testa:

$$U = \sum_{i,j} \frac{(f_{ij} - f_{i.} \cdot f_{.j} \cdot \frac{1}{N})^2}{f_{i.} \cdot f_{j.} \cdot \frac{1}{N}} \sim \chi^2\left((r-1)(s-1)\right).$$

- Situacija je analogna prilagodbi teorijske razdiobe empirijskoj razdiobi.
- Broj stupnjeva slobode $= rs - (r+s-2) - 1 = r(s-1) - (s-1) = (r-1)(s-1)$; pri čemu je rs ukupan broj podataka, a $r+s-2$ broj parametara određenih ML metodom.

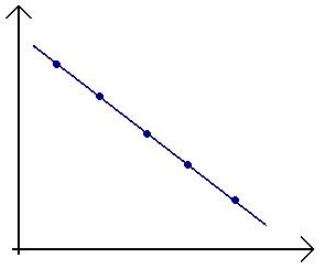
Poglavlje 6

MEÃŘUSOBNA ZAVISNOST SLUČAJNIH VARIJABLI

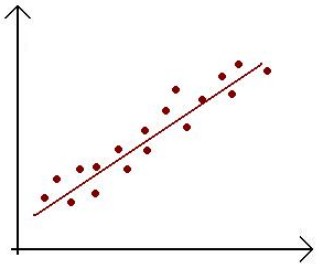
Neka imamo više skupina podataka. Povezanost može biti:

- *Funkcijska*, što znači da je svakom x pridružen jedan određeni y, $y = f(x)$,
- *Stohastička* što znači da za svaki x postoji određena neizvjesnost oko vrijednosti odgovarajućeg y-na.

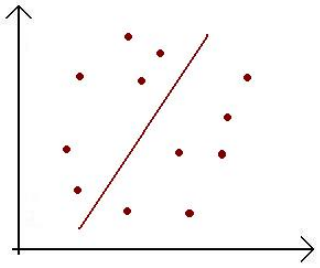
U geofizici najčešća je stohastička povezanost, budući da na mjerene veličine utječu razni, nepotpuno poznati, čimbenici.



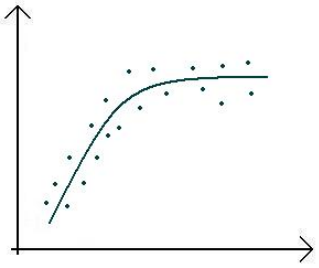
funkcijska, linearna



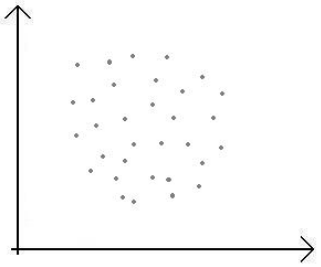
stohastička, linearna jaka



stohastička, linearna slaba



stohastička, nelinearna



nema veze

Nameću se dva pitanja:

1. Postoji li stohastička veza i koliko je čvrsta? $\longrightarrow$ *teorija korelacije*,

2. Ako postoji, kojeg je oblika? $\longrightarrow$ *teorija regresije*.

6.1 Linearna korelacija i regresija - klasični pristup

Koeficijent korelacije. Neka su X i Y slučajne varijable, $\mu_x = \mathbb{E}(X)$, $\mu_y = \mathbb{E}(Y)$. Kovarijanca

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left((X - \mu_x)\,(Y - \mu_y)\right)$$

govori o sklonosti slučajnih varijabli da simultano poprimaju vrijednsti veće od svojih srednjaka ili pak manje od njih. Kovarijanca ovisi o jedinicama mjere. Normiranjem se dobije *koeficijent korelacije*:

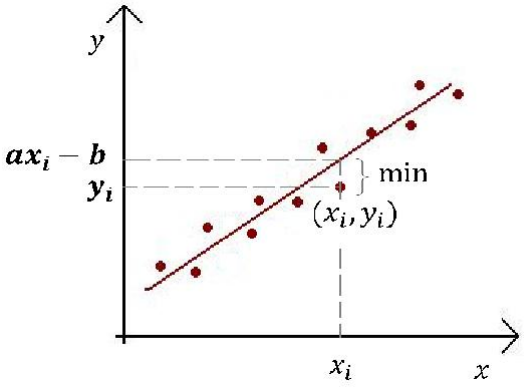
$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

pri čemu vrijedi $-1 \leq \rho_{xy} \leq 1$.

Linearna regresija. Imamo dva uzorka, $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$. Tražimo pravac $Y = aX + b$ tako da je

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \longrightarrow \min$$

.



$$\frac{\partial}{\partial a} = 0 \Rightarrow -2\sum_{i=1}^{n}(y_i - ax_i - b)\,x_i = 0$$

$$\frac{\partial}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0$$

Odatle je

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} ax_i = \overline{y} - a\overline{x}.$$

Uvrštavanjem dobivamo:

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - (\overline{y} - a\overline{x})\sum_{i=1}^{n} x_i = 0,$$

te zatim:

$$a = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2},$$

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X).}$$

6.2 Obična linearna regresija - geometrijska interpretacija

6.2.1 Veza između uzorka (slučajnih varijabli) i vektora

Sa X označimo uzorak (skup mjerenih vrijednosti) $X = \{x_1, \ldots, x_n\}$. Za uzorak definiramo *srednju vrijednost*

$$\overline{X} = \frac{1}{n}\sum x_i = \mathbb{E}(X)$$

(zadnju jednakost možemo pisati ako uzorak shvatimo kao diskretnu slučajnu varijablu, X, sa zakonom razdiobe $P\{x = x_i\} = \frac{1}{n}$).

Uzorke X i Y možemo shvatiti i kao vektore

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n.$$

Tada vrijedi::

- Drugi mješoviti moment oko nule u vezi je sa skalarnim produktom
$$\mathbb{E}(XY) = \frac{1}{n}\sum x_i y_i = \frac{1}{n}(X, Y).$$
- Drugi moment oko nule u vezi je s duljinom vektora
$$\mathbb{E}(X^2) = \frac{1}{n}\sum x_i^2 = \frac{1}{n}\|X\|.^2$$

Ako je $\overline{X} = \overline{Y} = 0$ (tj. ako su uzorci centrirani) onda je

$$\mathbb{E}(X \cdot Y) = \mathrm{Cov}(X, Y) = \frac{1}{n}(X, Y),$$

$$\mathbb{E}(X^2) = \mathrm{Var}(X) = \frac{1}{n}\|X\|^2.$$

Odatle: X i Y su nekorelirani $\iff \mathrm{Cov}(X, Y) = 0 \iff (X, Y) = 0 \iff X \perp Y$.

GEOMETRIJA	STATISTIKA (uz $\overline{X} = \overline{Y} = 0$)
$X, Y \in \mathbb{R}^n$	X, Y - uzorci
$\frac{1}{n}(X, Y)$	$\mathrm{Cov}(X, Y)$
$\frac{1}{n}\|X\|^2$	$\mathrm{Var}(X)$
$\frac{1}{\sqrt{n}}\|X\|$	standardna devijacija (σ_x)
$X \perp Y$	$\mathrm{Cov}(X, Y) = 0$
$\cos \sphericalangle(X, Y)$	koeficijent korelacije (ρ_{xy})
Pitagorin poučak	rastav varijance

6.2.2 Primjena na linearnu regresiju

Neka su zadanci uzorci (slučajne varijable) X ("ulaz") i Y ("izlaz"). Cilj je procjeniti Y pomoću linearne funkcije od X, tj. naći $a, b \in \mathbb{R}$ tako da razlika $\varepsilon = Y - (bX + a)$ bude što manja (u nekom smislu). Radi jednostavnosti, umjesto s X i Y radimo s odstupanjima $X - \overline{X}$ i $Y - \overline{Y}$. Tada je $a = 0$.

Pretpostavljamo, dakle, da je $\overline{X} = \overline{Y} = 0$, te tražimo $b \in \mathbb{R}$ tako da $\varepsilon = Y - bX$ bude što manje u smislu *najmanjih kvadrata*, tj.

$$\mathrm{Var}(\varepsilon) \longrightarrow \min.$$

Ako su X i Y uzorci duljine n, $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, onda je

$$\mathrm{Var}(\varepsilon) = \mathbb{E}\left(\varepsilon^2\right) = \mathbb{E}\left((Y - bX)^2\right) = \frac{1}{n}\sum_{i=1}^{n}(y_i - bx_i)^2.$$

Brojevi $y_i - bx_i$ su komponente vektora $y - bx$, a suma kvadrata komponenti je kvadrat duljine vektora. Iz slike je vidljivo da je:

$$\begin{aligned} \|\varepsilon\|^2 \longrightarrow \min \quad &\Longleftrightarrow \quad \varepsilon \perp X, \\ \mathrm{Var}(\varepsilon) \longrightarrow \min \quad &\Longleftrightarrow \quad X \text{ i } \varepsilon \text{ nisu korelirani.} \end{aligned}$$

Riječima: Procjena $Y-$na pomoću $X-$a se može popravljati dokle god pogreška (ostatak) "ima veze" s X.

Račun:

a)

$$\begin{aligned} Y &= bX + \varepsilon \qquad /(\cdot, X), \\ (Y, X) &= b\|X\|^2 + \underbrace{(\varepsilon, X)}_{0}, \end{aligned}$$

$$b = \frac{(Y,X)}{\|X\|^2} = \frac{\frac{1}{n}\sum x_i y_i}{\frac{1}{n}\sum x_i^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

b) Odredimo MSE (srednju kvadratnu pogrešku), tj. $\mathrm{Var}(\varepsilon)$

$$\begin{aligned} n \cdot \mathrm{Var}(\varepsilon) = \|Y - bX\|^2 &= (Y - bX, Y - bX) = \\ &= (Y - bX, Y) - \underbrace{(\underbrace{Y - bX}_{\varepsilon}, bX)}_{0} = \\ &= \|Y\|^2 - b(X,Y) = \\ &= \|Y\|^2 - \frac{(Y,X)}{\|X\|^2}(X,Y) = \\ &= \|Y\|^2\Big(1 - \underbrace{\frac{(X,Y)^2}{\|X\|^2\|Y\|^2}}_{r^2 \in [0,1]}\Big). \end{aligned}$$

Dobilo smo:

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y)(1 - r^2),$$

pri čemu je

$$r = \frac{(X,Y)}{\|X\|\|Y\|} = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

koeficijent korelacije.

c) Rastav varijance:

$$Y = bX + \varepsilon, \quad \text{pri čemu je} \quad bX \perp \varepsilon.$$

Pitagorin poučak daje:

$$\|Y\|^2 = \|bX\|^2 + \|\varepsilon\|^2,$$

odnosno

$$\underbrace{\mathrm{Var}(Y)}_{1} = \underbrace{\mathrm{Var}(bX)}_{r^2} + \underbrace{\mathrm{Var}(\varepsilon)}_{1-r^2},$$

što je *jednadžba analize varijance*. Vidimo da r^2 daje udio varijance "izlaza" Y koji je opisan regresijom, tj. koji se može pripisati "ulazu" X.

6.3 Višestruka linearna regresija

Neka su zadani su uzorci (slučajne varijable) X_1, X_2, Y, te pretpostavimo da je $\overline{X_2} = \overline{X_1} = \overline{Y} = 0$. Tražimo $b_1, b_2 \in \mathbb{R}$ tako da vrijedi

$$\mathrm{Var}\,(Y - b_1X_1 - b_2X_2) \longrightarrow \min,$$

odnosno

$$\|Y - b_1X_1 - b_2X_2\|^2 \longrightarrow \min.$$

U igri su tri vektora u n dimenzionalnom prostoru, koji međutim razapinju jedan 3−dimenzionalni prostor. Kada brojevi b_1 i b_2 variraju, vektor $b_1X_1 + b_2X_2$ "šeta" ravninom $\Pi(X_1, X_2)$ razapetom vektorima X_1 i X_2. Traži se dakle vektor (točka) u ravnini Π koja je najmanje udaljena od Y. To je ortogonalna projekcija vektora (tocke) Y na ravninu Π.

Zaključak:

$$\|\varepsilon\|^2 \longrightarrow \min \quad \Longleftrightarrow \quad \varepsilon \perp \Pi(X_1, X_2),$$

$$\|\varepsilon\|^2 \longrightarrow \min \quad \Longleftrightarrow \quad \varepsilon \perp X_1, \varepsilon \perp X_2,$$

odnosno

$$\mathrm{Var}(\varepsilon) \longrightarrow \min \quad \Longleftrightarrow \quad \varepsilon \text{ nije korelirano ni sa } X_1, \text{ ni sa } X_2.$$

Račun:

$$Y = b_1X_1 + b_2X_2 + \varepsilon \qquad /(\cdot, X_1), (\cdot, X_2),$$

$$(Y, X_1) = b_1\|X_1\|^2 + b_2(X_2, X_1) + \underbrace{(\varepsilon, X_1)}_{0} \qquad / \cdot \|X_2\|^2,$$

$$(Y, X_2) = b_1(X_1, X_2) + b_2\|X_2\|^2 + \underbrace{(\varepsilon, X_2)}_{0} \qquad / \cdot (X_2, X_1),$$

$$b_1 = \frac{(Y, X_1\|X_2\|^2 - (Y, X_2)(X_2, X_1)}{\|X_1\|^2\|X_2\|^2 - (X_1, X_2)(X_2, X_1)} \,:\, \frac{\|X_1\|^2\|X_2\|^2}{\|X_1\|^2\|X_2\|^2},$$

$$b_1 = \frac{\frac{(Y,X_1)}{\|X_1\|^2} - \frac{(X_2,X_1)}{\|X_1\|\|X_2\|} \cdot \frac{(Y,X_2)}{\|X_1\|\|X_2\|}}{1 - \frac{(X_1,X_2)}{\|X_1\|\|X_2\|} \cdot \frac{(X_2,X_1)}{\|X_1\|\|X_2\|}}.$$

Analogni izraz vrijedi i za b_2. Uz oznake:

$$s_y \equiv \frac{1}{\sqrt{n}}\|Y\|, \quad s_{x_1} \equiv \frac{1}{\sqrt{n}}\|X_1\|, \quad s_{x_2} \equiv \frac{1}{\sqrt{n}}\|X_2\|$$

konačno dobivamo:

$$b_1 = \frac{s_y}{s_{x_1}} \cdot \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2},$$
$$b_2 = \frac{s_y}{s_{x_2}} \cdot \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}.$$

Odredimo MSE:

$$n\mathrm{Var}(\varepsilon) = (\varepsilon,\varepsilon) = (Y - b_1X_1 - b_2X_2, Y - b_1X_1 - b_2X_2) =$$

$$= \|Y\|^2 - b_1(X_1,Y) - b_2(X_2,Y) =$$

$$= \|Y\|^2 - \frac{s_y}{s_{x_1}} \cdot \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}(X_1,Y) - \frac{s_y}{s_{x_2}} \cdot \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}(X_2,Y) =$$

$$= \|Y\|^2 \left(1 - \frac{\frac{1}{\sqrt{n}}\|Y\|}{\frac{1}{\sqrt{n}}\|X_1\|}\frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}\frac{(X_1,Y)}{\|Y\|\|Y\|} - \frac{\frac{1}{\sqrt{n}}\|Y\|}{\frac{1}{\sqrt{n}}\|X_2\|}\frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}\frac{(X_2,Y)}{\|Y\|\|Y\|}\right) =$$

$$= \|Y\|^2 \left(1 - \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}r_{1y} - \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}r_{2y}\right) =$$

$$= \|Y\|^2 \left(1 - \frac{r_{1y}^2 - r_{12}r_{1y}r_{2y} + r_{2y}^2 - r_{12}r_{1y}r_{2y}}{1 - r_{12}^2}\right).$$

Odatle je:

$$n\mathrm{Var}(\varepsilon) = n\mathrm{Var}(Y)(1 - \frac{r_{1y}^2 - 2r_{12}r_{1y}r_{2y} + r_{2y}^2}{1 - r_{12}^2}),$$

odnosno

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y)(1 - r_{y.12}^2),$$

pri čemu smo sa

$$r_{y.12} = \sqrt{\frac{r_{1y}^2 - 2r_{12}r_{1y}r_{2y} + r_{2y}^2}{1 - r_{12}^2}}$$

označili *koeficijent višestruke korelacije.*

Rastav varijance:

$$Y = b_1X_1 + b_2X_2 + \varepsilon, \quad \text{pri čemu je} \quad b_1X_1 + b_2X_2 \perp \varepsilon.$$

Pitagorin poučak daje:

$$\|Y\|^2 = \|b_1X_1 + b_2X_2\|^2 + \|\varepsilon\|^2,$$

odnosno

$$\underbrace{\mathrm{Var}(Y)}_{1} = \underbrace{\mathrm{Var}(b_1X_1 + b_2X_2)}_{r^2_{y.12}} + \underbrace{\mathrm{Var}(\varepsilon)}_{1-r^2_{y.12}},$$

što je *jednadžba analize varijance*. Vidimo da $r^2_{y.12}$ daje postotak varijance "izlaza" Y koji je opisan regresijom, tj. koji se može pripisati "ulazima" X_1, X_2.

Napomena:

a) Ako su ulazi potpuno korelirani (kolinearni, tj. ako je $r_{12} = 1$) postupak se "raspada". Koeficijenti b_1, b_2, te koeficijent višestruke korelacije postaju neodređeni.

b) Metoda *višestruke linearne regresije* je puno osjetljivija od obične linearne regresije, što naročito dolazi do izražaja ako su ulazi X_1 i X_2 visoko korelirani. Treba dobro paziti i procjeniti da li je unošenje novih ulaza opravdano, tj. svesti broj ulaza na minimum!

6.4 Koeficijent parcijalne korelacije

Neka su X_1 i Y slučajne varijable i pretpostavimo da koristeći linearnu regresiju procjenjujemo Y pomoću X_1. Pitamo se da li se procjena može poboljšati uvođenjem nove varijable X_2.

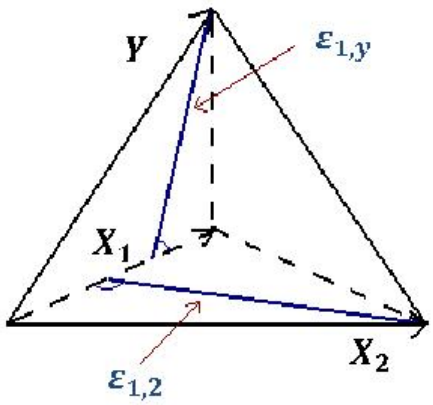
Treba vidjeti kolika je korelacija između X_2 i Y kada se iz obje varijable isključi "utjecaj" od X_1. Ako je ta korelacija mala onda varijabla X_2 ne nosi novu informaciju, tj. onu koja već nije sadržana u X_1. Promotrimo linearne regesije:

$$\varepsilon_{1,y} = Y - b_1^{'}X_1,$$

$$\varepsilon_{1,2} = X_2 - b_2^{'}X_1.$$

Koeficijent parcijalne korelacije izmedju Y i X_2 kada se ukloni utjecaj od X_1 je obični koeficijent korelacije između ostataka $\varepsilon_{1,y}$ i $\varepsilon_{1,2}$. Račun daje:

$$r_{y2.1} = \frac{r_{y2} - r_{y1} \cdot r_{21}}{\sqrt{1 - r^2_{y1}}\sqrt{1 - r^2_{12}}}.$$



Sa slike vidimo da je $\sphericalangle(\varepsilon_{1,2}, \varepsilon_{1,y})$ zapravo kut između ravnina $\Pi(X_2, X_1)$ i $\Pi(X_1, Y)$, tj. između stranica piramide koje se sastaju u bridu X_1. Ako je taj kut blizu $90°$, tj. $r_{y2.1}$ je malo onda projekcija od Y na $\Pi(X_1, X_2)$ pada blizu vektora X_1, tj. vidi se da je udio od X_2 u višestrukoj regresiji $Y = b_1X_1 + b_2X_2 + \varepsilon$ malen, što povlači da ulaz X_2 nije bitan. Posljedično, procjena $b_1X_1 + b_2X_2$ je bliska procjeni temeljenoj samo na X_1.

6.5 Slučaj $\overline{X} \neq 0$, $\overline{Y} \neq 0$

a) Neka je X uzorak (slučajna varijabla). Odredimo broj $a \in \mathbb{R}$ tako da vrijedi:

$$\mathbb{E}\left((X-a)^2\right) \longrightarrow \min$$

(pokušavamo X opisati *konstantom*, $X = a + \varepsilon$), tj. za $X = \{x_1, \ldots, x_n\}$ gledamo

$$\mathbb{E}\left((X-a)^2\right) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = \frac{1}{n}\|X - a \cdot \mathbf{1}\|^2 \longrightarrow \min,$$

pri čemu smo uveli oznaku

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Od prije znamo da vrijedi:

$$\mathbb{E}\left((X-a)^2\right) \longrightarrow \min \quad \Longleftrightarrow \quad X - a \cdot \mathbf{1} \perp \mathbf{1}.$$

Odatke je

$$(X, \mathbf{1}) - a(\mathbf{1}, \mathbf{1}) = 0,$$

$$\sum_{i=1}^{n} x_i - a \cdot n = 0,$$

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

Zaključak: Srednje kvadratno odstupanje je najmanje ako se računa od srednjaka.

b) Za X, Y (uzorci ili slučajne varijable) takve da $\overline{X} \neq 0$, $\overline{Y} \neq 0$ gledamo

$$Y = bX + a + \varepsilon, \qquad \mathrm{Var}(\varepsilon) \longrightarrow \min.$$

$$\mathbb{E}\Big((\underbrace{Y - bX}_{Y'} - a)^2\Big) \longrightarrow \min$$

$$\Longrightarrow \quad a = \overline{Y'} = \overline{Y - bX} = \overline{Y} - b\overline{X},$$

$$\Longrightarrow \quad Y = bX + \overline{Y} - b\overline{X} + \varepsilon,$$

$$\Longrightarrow \quad Y - \overline{Y} = b(X - \overline{X}) + \varepsilon$$

Zaključak: Ako radimo s odstupanjima, $X - \overline{X}$, $Y - \overline{Y}$, smijemo staviti $a = 0$. Dobiveni b i ε ostaju *isti*!.

6.6 BND i linearna regresija

- Neka je $(X, Y) \sim \mathrm{BND}\,(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, ρ je koeficijent korelacije, a $X \sim \mathrm{N}\,(\mu_X, \sigma_X)$, $Y \sim \mathrm{N}\,(\mu_Y, \sigma_Y)$ su marginalne razdiobe.

- Združena gustoća za slučaj $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ glasi:

$$f(x, y; \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{-\frac{1}{2\sqrt{(1-\rho^2)}}(x^2 - 2\rho xy + y^2)}.$$

  Ako je $\rho \neq 0$ izolinije su elipse, a ako je $\rho = 0$ izolinije su kružnice.

- Uvjetne gustoće za slučaj $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(y - \rho x)^2},$$

  tj.

$$f_{Y|X=x}(y) \sim \mathrm{N}\Big(\rho x, \underbrace{\sqrt{1-\rho^2}}_{\text{ne ovisi o } x}\Big).$$

  U općem slučaju, uvjetna razdioba je i dalje normalna, pri čemu srednja vrijednost i standardna devijacija glase:

$$\mu_{Y|X=x} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

$$\sigma_{Y|X=x} = \sigma_Y\sqrt{1-\rho^2}.$$

- Procijenimo vrijednost varijable Y ako znamo da je $X = x$, pri čemu je $x \in \mathbb{R}$ fiksan broj.

- Poznanje vrijednosti od x smanjuje neizvjesnost oko Y (ako je $\rho \neq 0$).
- Uvjetna distribucija je jednodmenzionalna normalna razdioba.
- Optimalna, u smislu najmanjih kvadrata, procjena od Y uz uvjet $X = x$, je srednja vrijednost pripadne uvjetne razdiobe. Sve srednje vrijednosti leže na pravcu $y = \rho x$, ako su varijable X i Y centrirane i normirane, odnosno $y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ u općem slučaju. Ujedno je to i pravac regresije između X i Y.

- Zaključak: Označimo s $\hat{y}$ optimalnu (u smislu najmanjih kvadrata) procjenu $Y-$na pomoću $X-$a. Ako vektor (X, Y) ima bivarijantnu normalnu razdiobu, onda je *linearna* procjena ujedno i optimalna i ona glasi:

$$\frac{\hat{y} - \mu_Y}{\sigma_Y} = \rho\frac{x - \mu_X}{\sigma_X},$$

  odnosno $\hat{y} = \rho x$ ako su varijable X i Y centrirane i normirane.

- Uočimo da se za $x > \mu_X$ tjeme uvjetne razdiobe postiže prije $y = x$, a za $x < mu_X$ poslije. Odatle dolazi naziv regresija.

6.7 Testiranje značajnosti koeficijenta korelacije

- *Pretpostavka* : $(X, Y) \sim$ BND.
- $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ - *nezavisni* uzorak, raspodijeljen kao i (X, Y)
- Hipoteze:
  $H_0 : \rho = 0$
  $H_1 : \rho \neq 0$ ili $H_1' : \rho > 0$ ili $H_1'' : \rho < 0$.
- Koeficijent korelacije uzorka (*procjenitelj*):

$$r = \frac{\frac{1}{n}\sum(x_1 - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_1 - \bar{x})^2}\sqrt{\frac{1}{n}\sum(y_i - \bar{y}^2}}.$$

- Ako vrijedi pretpostavka $H_0 : \rho = 0$, onda slučajna varijabla

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

  ima $t-$razdiobu s $n - 2$ stupnja slobode. Ako je $t > t_{1-\alpha}$ odbacujemo H_0 (uz $H_1 : \rho > 0$).

- Za provjeru hipoteze $H_0 : \rho = 0$, zgodno je napraviti graf $r = f(t, n)$ tj.

$$r = \frac{t}{\sqrt{t^2 + n - 2}}$$

gdje je $t = t_{1-\alpha}(n)$:

- Slučaj kada koeficijent korelacije populacije (tj. "stvarni" koeficijent korlacije) nije nula:

  - Hipoteze:

    $H_0 : \rho = \rho_0$

    $H_1 : \rho > \rho_0$ ili $H_1' : \rho < \rho_0$ ili $H_1'' : \rho \neq \rho_0$

  - Za $\rho_0 > 0$, slučajna varijabla ima kompliciranu razdiobu $\Longrightarrow$

$$\text{transformacija } r \longmapsto u = \tfrac{1}{2} \ln \tfrac{1+r}{1-r}$$

  - Kada $n \to \infty$ ($n > 50$), razdioba od U teži k normalnoj razdiobi

$$\mathrm{N}\left(\frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)}, \frac{1}{\sqrt{n-3}}\right).$$

- Napomena: Kod procjene pouzdanosti koeficijenta korelacije bitnija je fizika nego statistika!

Poglavlje 7

POČETNA ANALIZA VREMENSKIH NIZOVA U KLIMATOLOGIJI

7.1 Spearmanov test ranga

Neka su polazni podaci nizovi $\{x_1, \ldots, x_n\}$ i $\{y_1, \ldots, y_n\}$, a njihovi rangovi $\{m_1, \ldots, m_n\}$ i $\{n_1, \ldots, n_n\}$.
Spearmanov koeficijent korelacije rangova, r_s je obični koeficijent korelacije između $\{m_i\}$ i $\{n_i\}$.

- uočimo da je: $1 \leqslant m_i, n_i \leqslant n$
- definiramo $D_i = m_i - n_i, i = 1, \ldots, n$, tada je
$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$
- napomena - ako je veza između X i Y MONOTONA ($x_i > x_j \implies y_i > y_j$) $\iff r_s = 1$ (obični $r = 1 \iff$ veza je linearna).
- r_s je jak i otporan (nije osjetljiv na ekstreme)

Primjena na ispitivanje relativne homogenosti: $\longrightarrow$ test otkriva nagli skok ili polagani trend na jednoj od postaja.

Vrijedi: Ako su nizovi relativno homogeni te ako je $n > 8$, statistika
$$t = r_s \cdot \sqrt{\frac{n-2}{1 - r_s^2}}$$

ima t razdiobu s $n-2$ stupnjeva slobode. Kritično područje se postavlja s obe strane razdiobe.