FRI, Univerza v Ljubljani

Statistika in analiza podatkov

Aleksandar Jurišić

Ljubljana, februar 2010

različica: 1. junij 2010 / 18 : 34

Kazalo

Predstavitev

Aleksandar Jurišić

FRI, Jadranska 21, soba 5

e-pošta: ajurisic@valjhun.fmf.uni-lj.si

WWW: http://lkrv.fri.uni-lj.si/~ajurisic

asistent:

Gregor Jerše (gregor.jerse@fmf.uni-lj.si).

Statistika in analiza podatkov [70028]

Cilj predmeta:

predstaviti kategorije statistike in njihovo vlogo v poslovanju.

Kratka vsebina:

Proučevanje množičnih pojavov. Statistično opazovanje. Urejevanje statističnega gradiva. Frekvenčne distribucije. Časovne vrste. Statistične metode. Vzorčenje. Nacionalni in mednarodni informacijski sistemi in baze podatkov: kazalci družbeno ekonomskega razvoja, poslovne in tržne informacije; znanstvenotehnološke in referalne(bibliografske) baze podatkov. Analiza in primerjalna analiza podatkov. Statistična raziskovanja. Modeli, predvidevanje in napovedovanje razvoja.

Obveznosti študenta

Uspešno opravljena *kolokvija*.
Če ni šlo, je potrebno opraviti
pisni izpit iz reševanja nalog.

Ko ima enkrat pozitivno oceno iz reševanja nalog, mora v tekočem letu opraviti še *izpit iz teorije*
(lahko je pisni ali ustni -
odvisno od števila prijavljenih).

Viri

Pri predavanjih se bomo pretežno opirali na naslednje knjige/skripte:

L. Gonick in W. Smith, *The Cartoon guide to Statistics*, 1993.

D. S. Moore (Purdue University), *Statistika: znanost o podatkih* (5. izdaja prevedena v slovenščino leta 2007).

W. Mendenhall in T. Sincich, *Statistics for engineering and the sciences*, 4th edition, Prentice Hall, 1995.

M. Hladnik: *Verjetnost in statistika*. Založba FE in FRI, Ljubljana 2002.

A. Ferligoj: *Osnove statistike na prosojnicah*. Samozaložba, Ljubljana 1995.

Obstaja obilna literatura na spletu in v knjižnicah.

Gradiva bodo dosegljiva preko internetne učilnice (moodle).

Pri delu z dejanskimi podatki se bomo v glavnem naslonili na prosti statistični program **R**.

Program je prosto dostopen na:

http://www.r-project.org/

proti koncu semestra pa morda tudi Minitab.

UVOD

1. Igre na srečo

Ste se kdaj vprašali, zakaj so igre na srečo,
ki so za nekatere rekreacija ali pa droga, tako dober posel za igralnice?

Vsak uspešen posel mora iz uslug,
ki jih ponuja, kovati napovedljive dobičke.
To velja tudi v primeru, ko so te usluge igre na srečo.

Posamezni hazarderji lahko zmagajo ali pa izgubijo.
Nikoli ne morejo vedeti, če se bo njihov obisk
igralnice končal z dobičkom ali z izgubo.

Igralnica pa ne kocka, pač pa dosledno dobiva
in država lepo služi na račun loterij ter drugih oblik iger na srečo.

Presenetljivo je, da lahko skupni rezultat več 1000 naključnih izidov poznamo s skoraj popolno gotovostjo. Igralnici ni potrebno obtežiti kock, označiti kart ali spremeniti kolesa rulete. Ve, da ji bo na dolgi rok vsak stavljeni euro prinesel približno 5 centov dobička.

Splača se ji torej osredotočiti na brezplačne predstave ali poceni avtobusne vozovnice, da bi privabili več gostov in tako povečali število stavljenega denarja. Posledica bo večji dobiček.

Igralnice niso edine, ki se okoriščajo z dejstvom,
da so velikokratne ponovitve slučajnih izidov napovedljive.

Na primer, čeprav zavarovalnica ne ve, kateri
od njenih zavarovancev bodo umrli v prihodnjem letu,
lahko precej natančno napove, koliko jih bo umrlo.
Premije življenjskih zavarovanj postavi v skladu
s tem znanjem, ravno tako kot igralnica določi glavne dobitke.

2. Kaj je statistika?

Skozi življenje se prebijamo z odločitvami,
ki jih naredimo na osnovi nepopolnih informacij ...

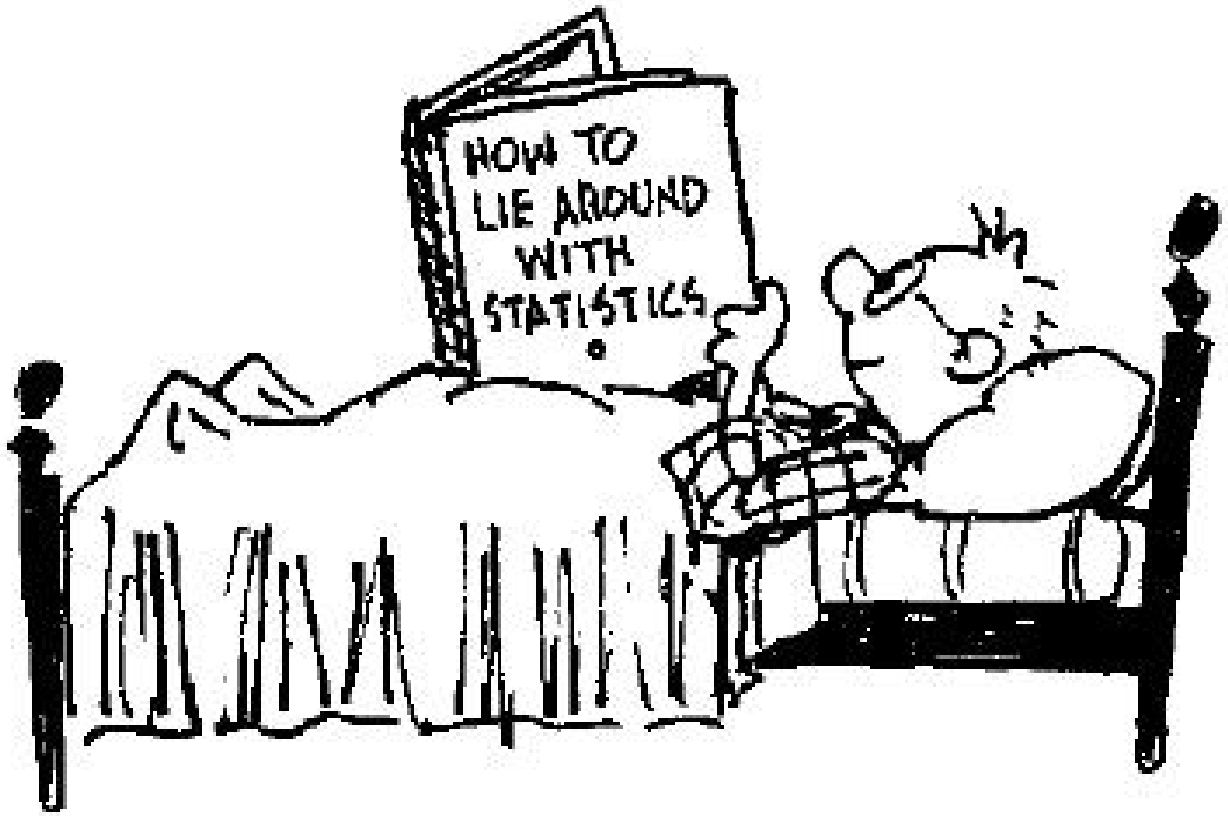
Pogled od zunaj

Števila so me pogosto begala,
še posebej, če sem imel pred seboj
neko njihovo razvrstitev,
tako da je tem primeru obveljala misel,
ki so jo pripisali Diaraeliju,
z vso pravico in močjo:

"Obstajajo tri vrste laži:

laži,
preklete laži in
statistika."

iz Autobiografije Marka Twaina


HOW TO
LIE ABOUND
WITH
STATISTICS

Okvirni načrt za statistiko

Opisna statistika

ena spremenljivka

– mere centralne tendence

– mere razpršenosti

– mere oblike

dve spremenljivki

– mere asociacije

Inferenčna (analitična) statistika

– točkovno in intervalno ocenjevanje

– 1- in 2- vzorčno testiranje hipotez

– kontingenčne tabele

– regresija

Statistika

preučuje podatke, jih

– zbira,

– klasificira,

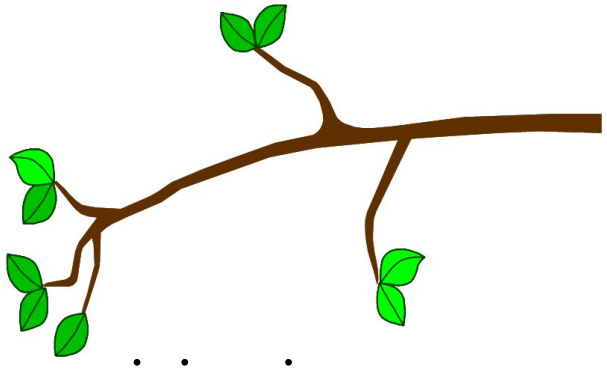
– povzema,

– organizira,

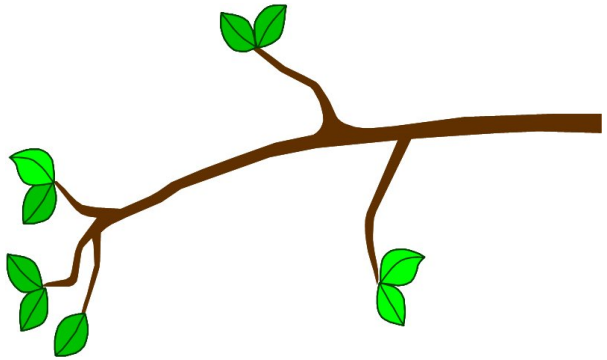
– analizira in

– interpretira.

Glavni veji statistike



Opisna statistika se ukvarja z organiziranjem, povzemanjem in opisovanjem zbirk podatkov (reduciranje podatkov na povzetke)

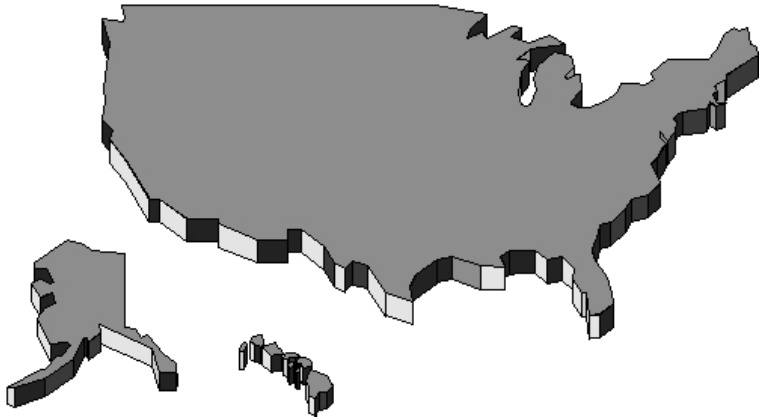


Analitična statistika jemlje vzorce podatkov in na osnovi njih naredi zaključke (inferenčnost) o populaciji (ekstrapolacija).

Tipi podatkovnih množic

- *Populacija*

– vsi objekti,
ki jih opazujemo



Primer:

vsi registrirani glasovalci

- *Vzorec*

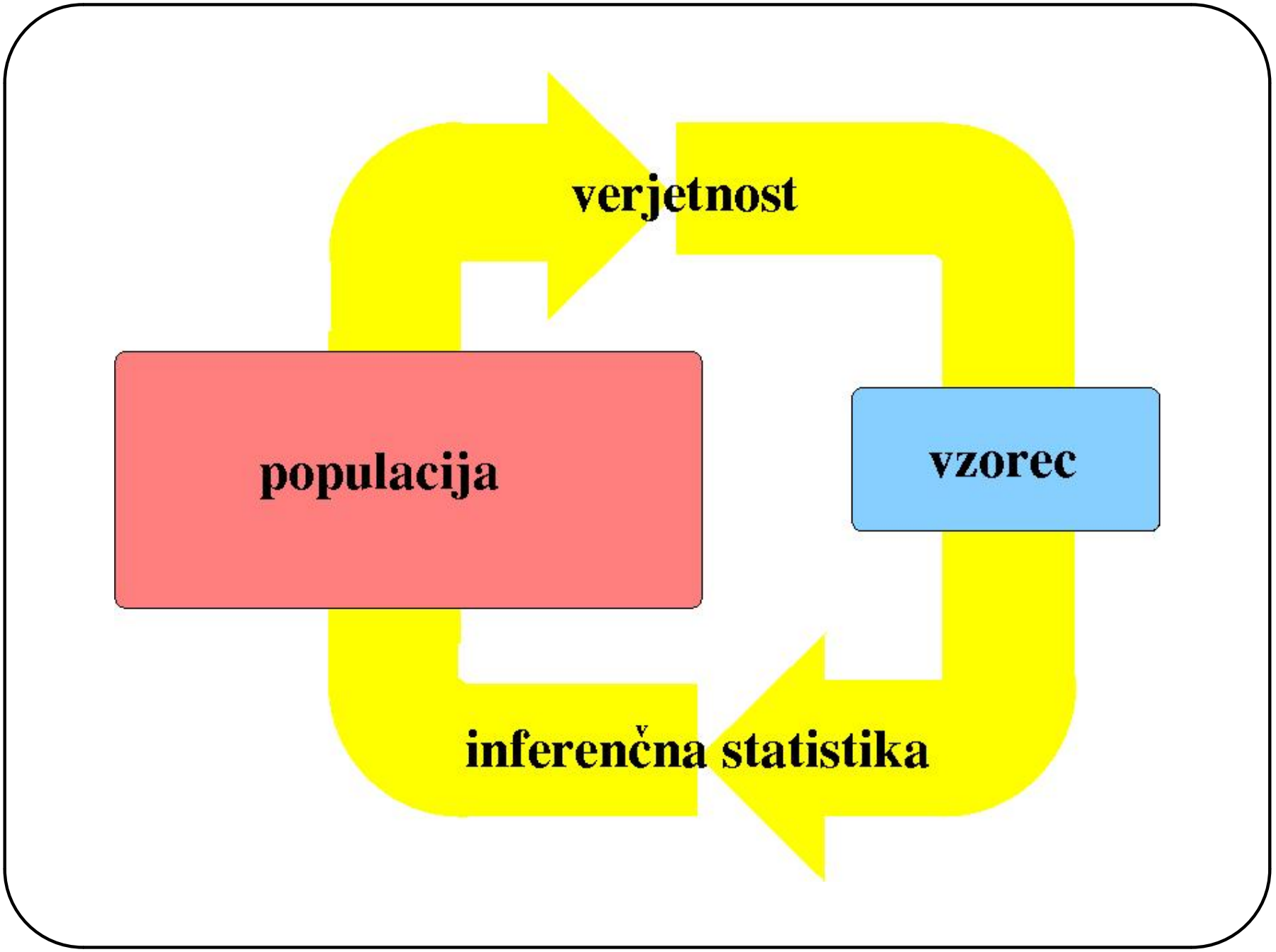
– podmnožica populacije

Primer:

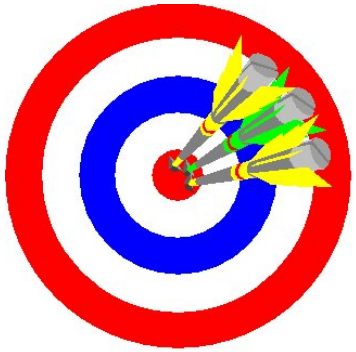
100 registriranih glasovalcev

Populacija je podatkovna množica, ki ji je namenjena naša pozornost.

Vzorec je podmnožica podatkov, ki so izbrani iz polulacije (po velikosti bistveno manjši od populacije).


verjetnost
populacija
vzorec
inferenčna statistika

Tipi podatkov

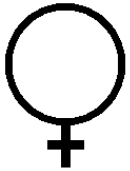


– *kvantitativni*
(numerični)

predstavljajo kvantiteto

ali količino nečesa.

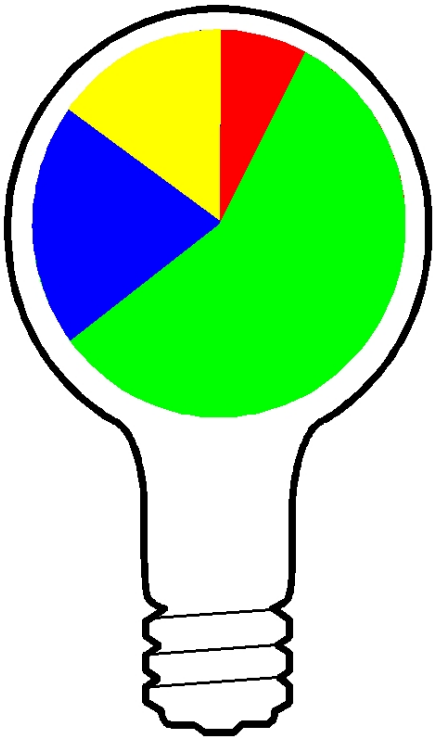
– *kvalitativni* (kategorije) ni kvantitativnih interpretacij.



Kvantitativni (numerični)

- **interval**
  - poljubna ničla
  - Enaki intervali predstavljajo enake količine.

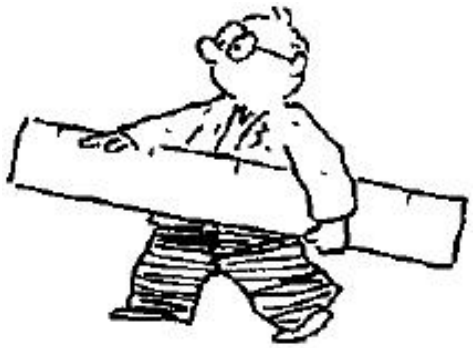
- **razmerje**
  - smiselna točka nič
  - Operacije seštevanje, odštevanje, množenje in deljenje so smiselne.



Kvalitativni (kategorični)

- **nominalni**
  – kategorije brez odgovarjajočega vrstnega reda / urejenosti

- **ordinalni/številski**
  – kategorije z urejenostjo

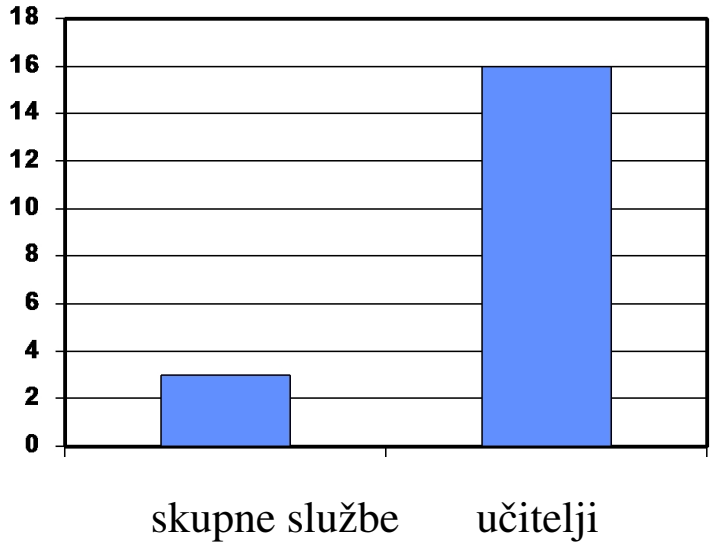


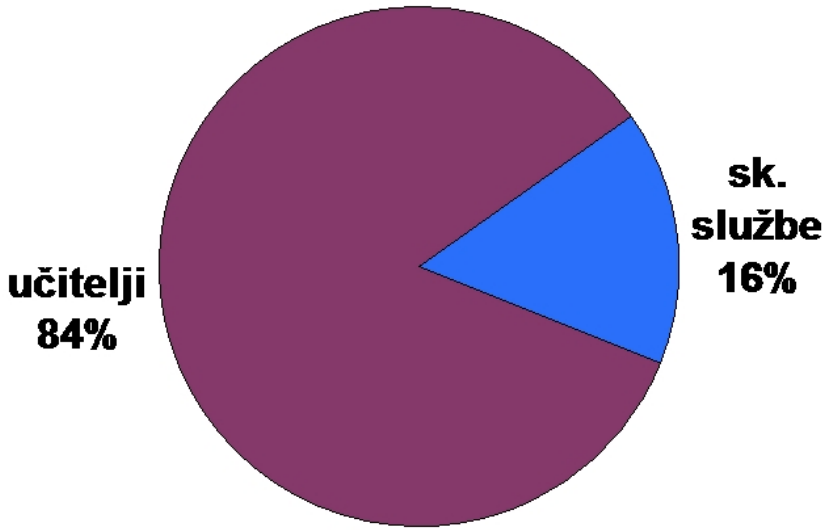
Oddelek sistemskih inženirjev

kategorija	frekvenca	relativna frekvenca
vrsta zaposlenih	število zaposlenih	delež
učitelji	16	0,8421
skupne službe	3	0,1579
skupaj	19	1,0000

Grafična predstavitev kvantitativnih podatkov

- **stolpčni graf**
  – poligonski diagram
- **strukturni krog**
  – pogača, kolač





Grafična predstavitev kvantitativnih podatkov

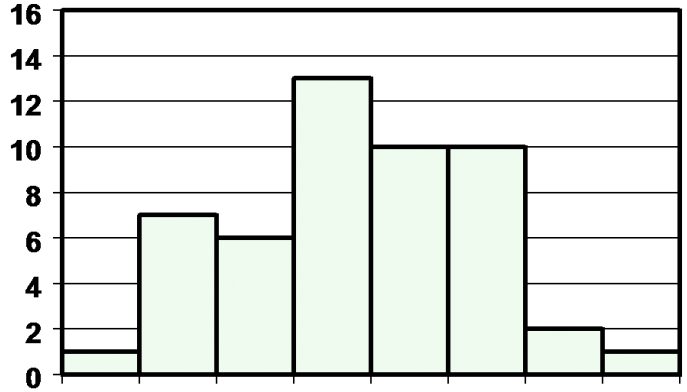
– runs plot (X, Y plot)

– zaporedje (dot plot)

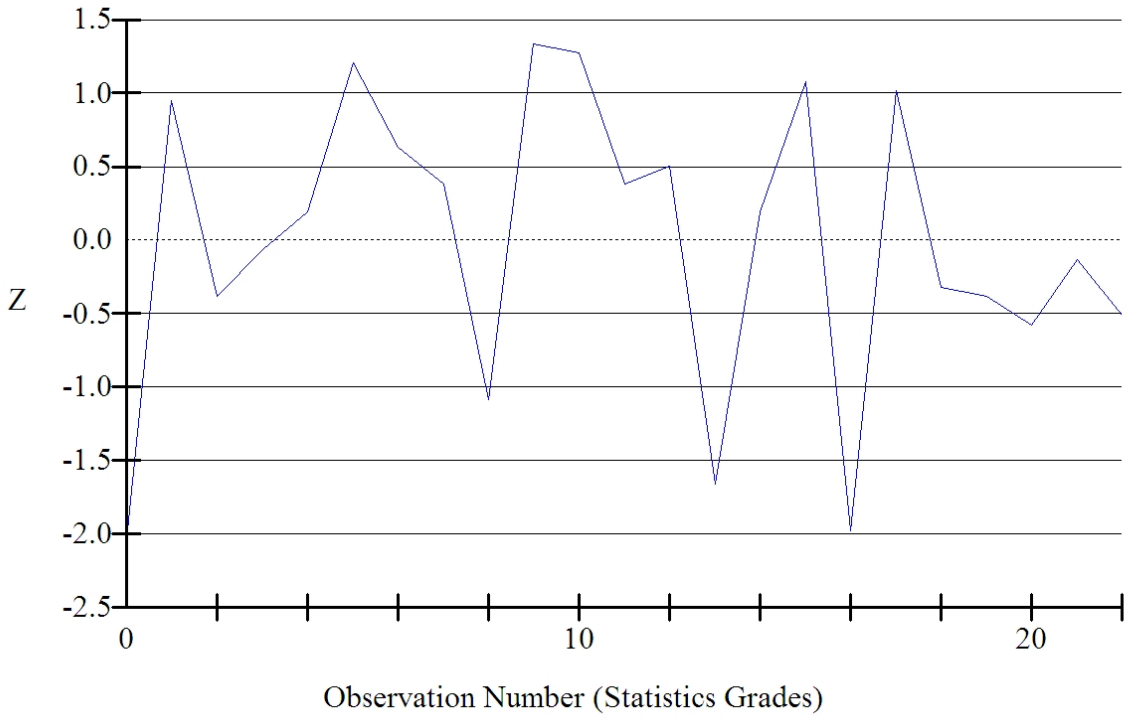
– steblo-list predstavitev (angl. stem-and-leaf)

– histogrami

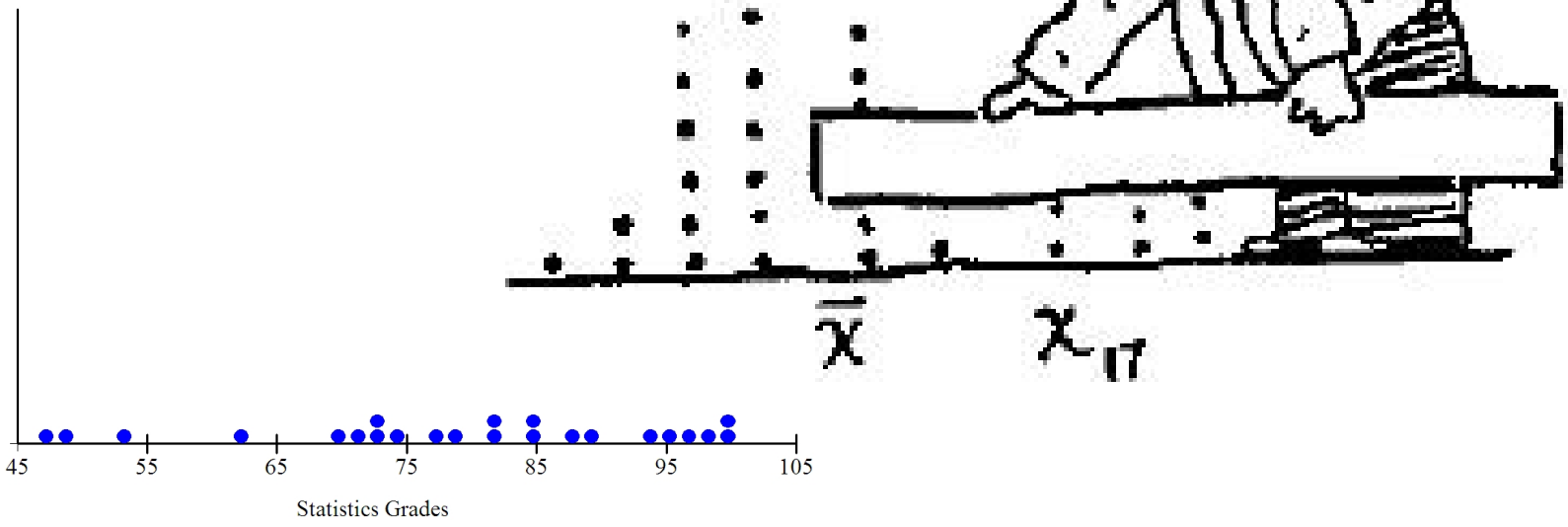
– škatla z brki (box plot)



Runs Chart

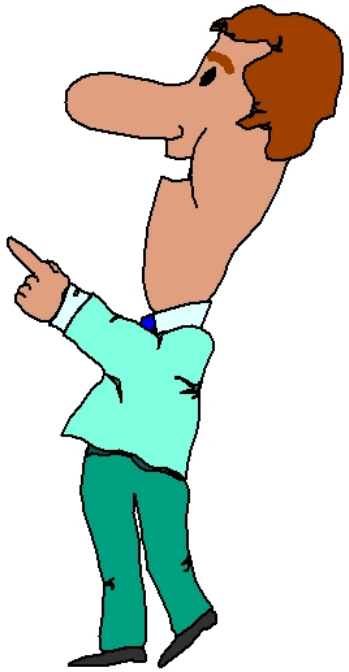

1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
Z
0
10
20
Observation Number (Statistics Grades)

Dot Plot



Urejeno zaporedje/ranžirana vrsta

Urejeno zaporedje je zapis
podatkov v vrsto po
njihovi numerični velikosti
(ustreznemu mestu pravimo *rang*).



Primer zaporedja podatkov (nal. 2.48, str.64)

(a) Konstruiraj urejeno zaporedje.

(b) Nariši steblo-list diagram.

(c) Naredi histogram.

88	103	113	122	132
92	108	114	124	133
95	109	116	124	133
97	109	116	124	135
97	111	117	128	136
97	111	118	128	138
98	112	119	128	138
98	112	120	131	142
100	112	120	131	146
100	113	122	131	150

Koraki za konstrukcijo steblo-list predstavitve

1. Razdeli vsako opazovanje-podatke na dva dela: **stebla** (angl. stem) in **listi** (angl. leaf).
2. Naštej stebla po vrsti v stolpec, tako da začneš pri najmanjšem in končaš pri največjem.
3. Upoštevaj vse podatke in postavi liste za vsak dogodek/meritev v ustrezno vrstico/steblo.
4. Preštej frekvence za vsako steblo.

stebla/listi

08	8
09	2 5 7 7 7 8 8
10	0 0 3 8 9 9
11	1 1 2 2 2 3 3 4 6 6 7 8 9
12	0 0 2 2 4 4 4 8 8 8
13	1 1 1 2 3 3 5 6 8 8
14	2 6
15	0

Steblo-list diagram

stebla	listi	rel. ν	ν
08	8	1	2%
09	2 5 7 7 7 8 8	7	14%
10	0 0 3 8 9 9	6	12%
11	1 1 2 2 2 3 3 4 6 6 7 8 9	13	26%
12	0 0 2 2 4 4 4 8 8 8	10	20%
13	1 1 1 2 3 3 5 6 8 8	10	20%
14	2 6	2	4%
15	0	1	2%
		50	100%

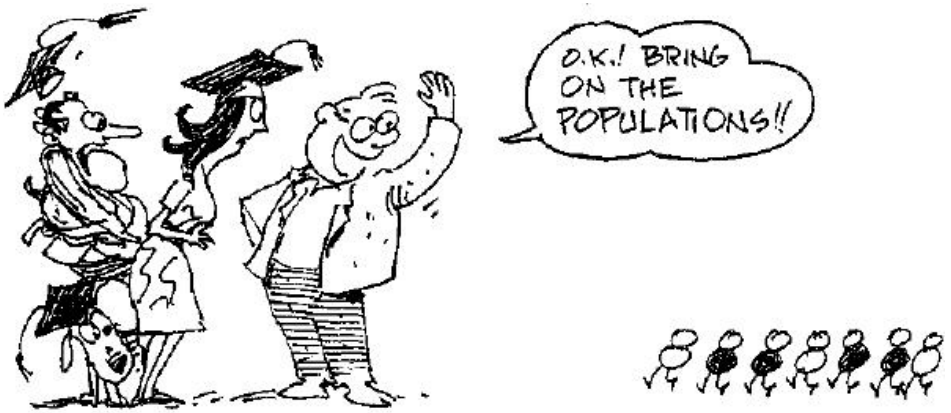
Histogrami

(1) kako zgradimo histogram

(2) število razredov

(3) frekvenca

(4) procenti



(a) Kako zgradimo histogram

(a) Izračunaj **razpon** podatkov.

(b) Razdeli razpon na *5 do 20 razredov* enake širine.

(c) Za vsak razred preštej število vzorcev,
ki spadajo v ta razred.
To število imenujemo **frekvenca razreda**.

(d) Izračunaj vse **relativne frekvence razredov**.

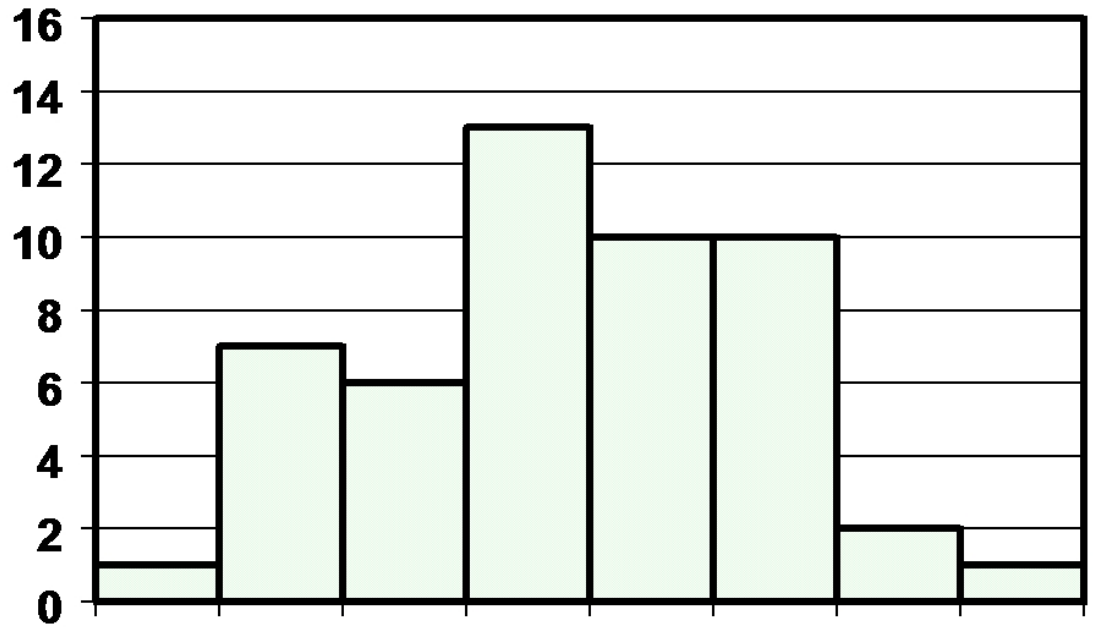
(b) Pravilo za določanje števila razredov v histogramu

število vzorcev v množici podatkov	število razredov
manj kot 25	5 ali 6
$25 - 50$	$7 - 14$
več kot 50	$15 - 20$

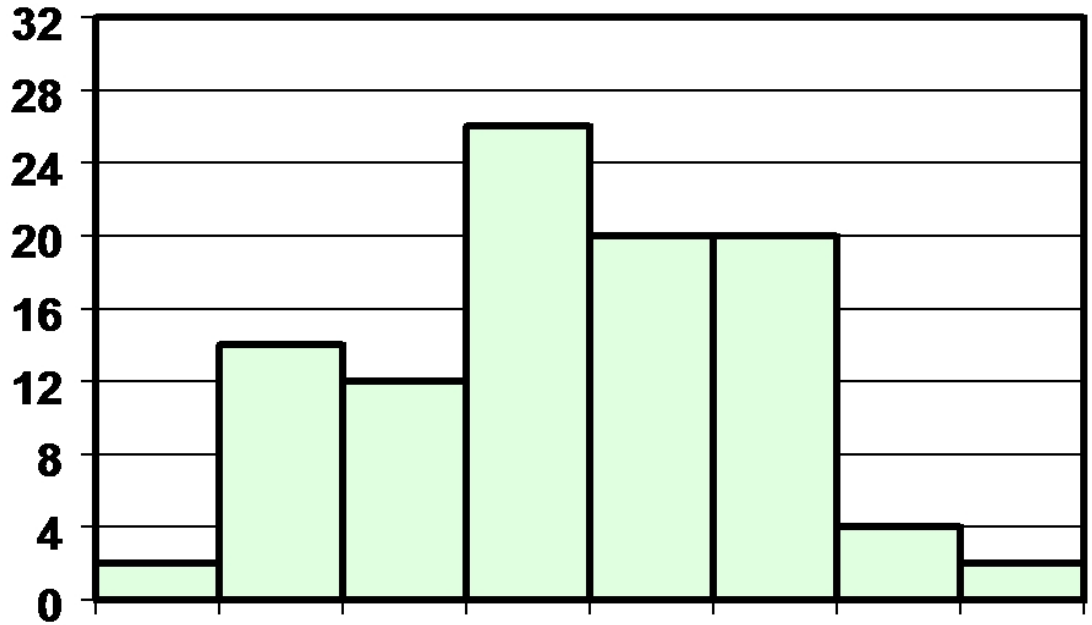
(c,d) Frekvenčna porazdelitev

razred	interval razreda	frekvenca	relativna frekvenca
1	80 – 90	1	2%
2	90 – 100	7	14%
3	100 – 110	6	12%
4	110 – 120	13	26%
5	120 – 130	10	20%
6	130 – 140	10	20%
7	140 – 150	2	4%
8	150 – 160	1	2%

Frekvenčni histogram



Procentni histogram



Mere za lokacijo in razpršenost

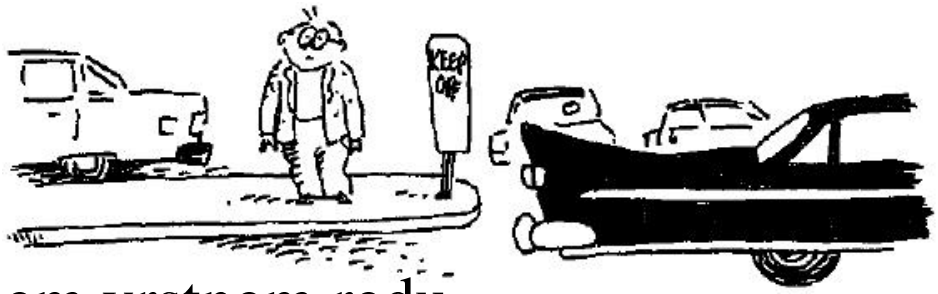
- srednje vrednosti
- razpon (min/max)
- centili, kvartili
- varianca
- standardni odklon
- Z-vrednosti

Modus

Modus (oznaka M_0) množice podatkov
je tista vrednost,
ki se pojavi z
največjo frekvenco.

Mediana (M_e)

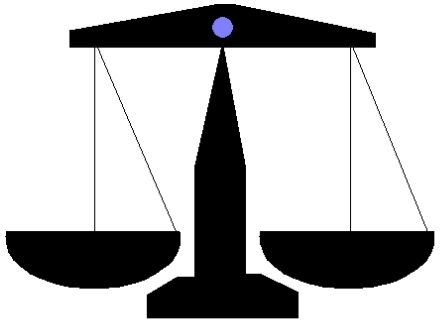
Da bi prišli do *mediane* (oznaka M_e)
za neko množico podatkov,
naredimo naslednje:



1. Podatke uredimo
   po velikosti v naraščujočem vrstnem redu,
2. Če je število podatkov liho,
   potem je mediana podatek na sredini,
3. Če je število podatkov sodo, je mediana
   enaka povprečju dveh podatkov na sredini.

Oznake: mediana populacije: μ mediana vzorca: m

Povprečja



Povrečje populacije: $$\mu = \frac{1}{N}(y_1 + \cdots + y_N) = \frac{\sum_{i=1}^{N} y_i}{N}$$

Povrečje vzorca: $$\overline{y} = \frac{1}{n}(y_1 + \cdots + y_n) = \frac{\sum_{i=1}^{n} y_i}{n}$$

Razpon ali variacijski razmik

Razpon je razlika med
največjo in najmanjšo meritvijo
v množici podatkov.

Centili

100p-ti centil ($p \in [0, 1]$)
je definiran kot število,
od katerega ima 100p %
meritev manjšo ali enako
numerično vrednost.

Določanje 100p-tega centila:

Izračunaj vrednost $p(n + 1)$
in jo zaokroži na najbližje celo število.
Naj bo to število enako i.

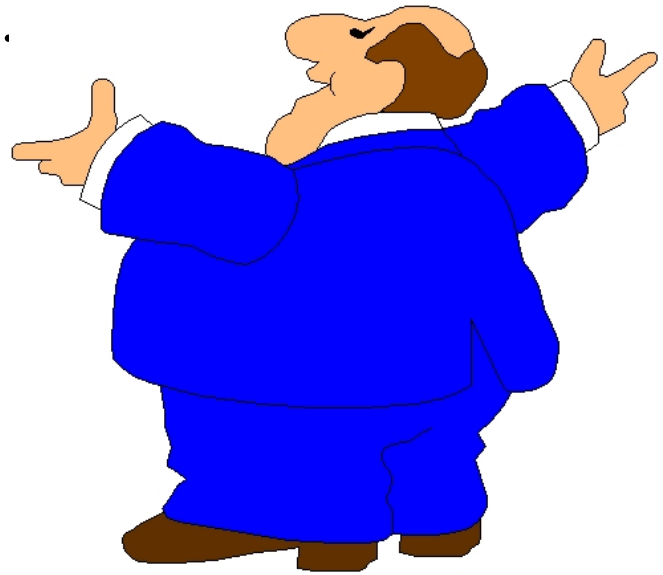
Izmerjena vrednost z i-tim rangom je 100p-ti centil.

... Centili

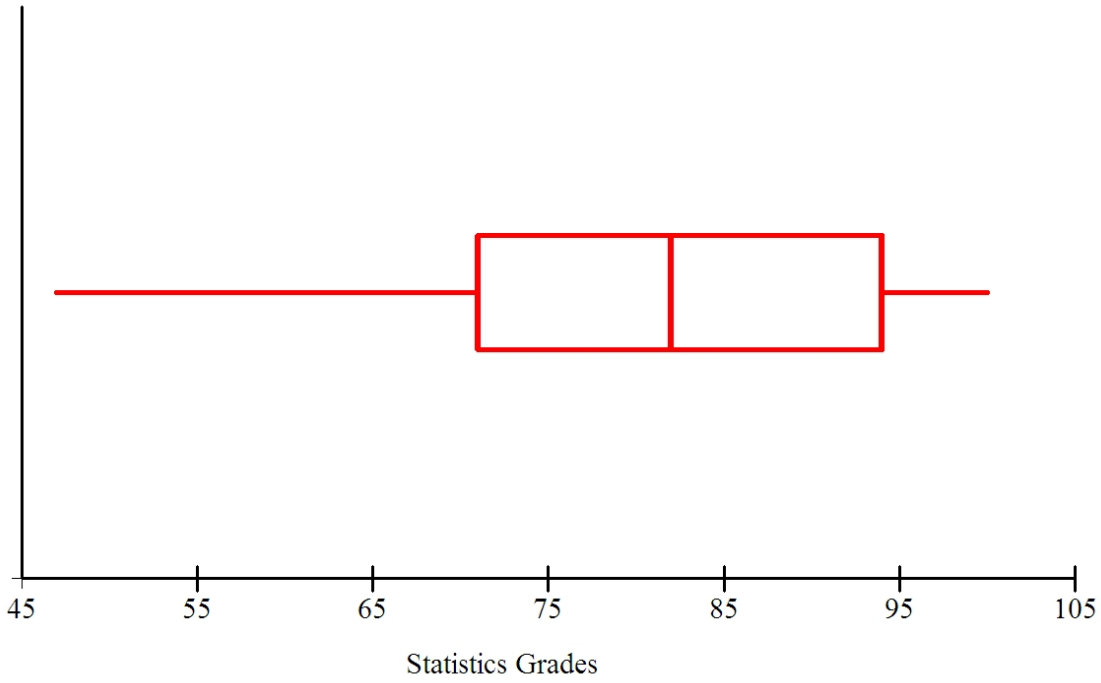
25. centil se imenuje tudi **1. kvartil**.

50. centil se imenuje **2. kvartil** ali **mediana**.

75. centil se imenuje tudi **3. kvartil**.



Škatla z brki (angl. box plot)



Mere razpršenosti

varianca

– kvadrat pričakovanega odklona (populacije)

– vsota kvadratov odklonov deljena s stopnjo prostosti (vzorca)

standardni odklon (deviacija)

– pozitivni kvadratni koren variance

koeficient variacije

– standardni odklon deljen s povprečjem

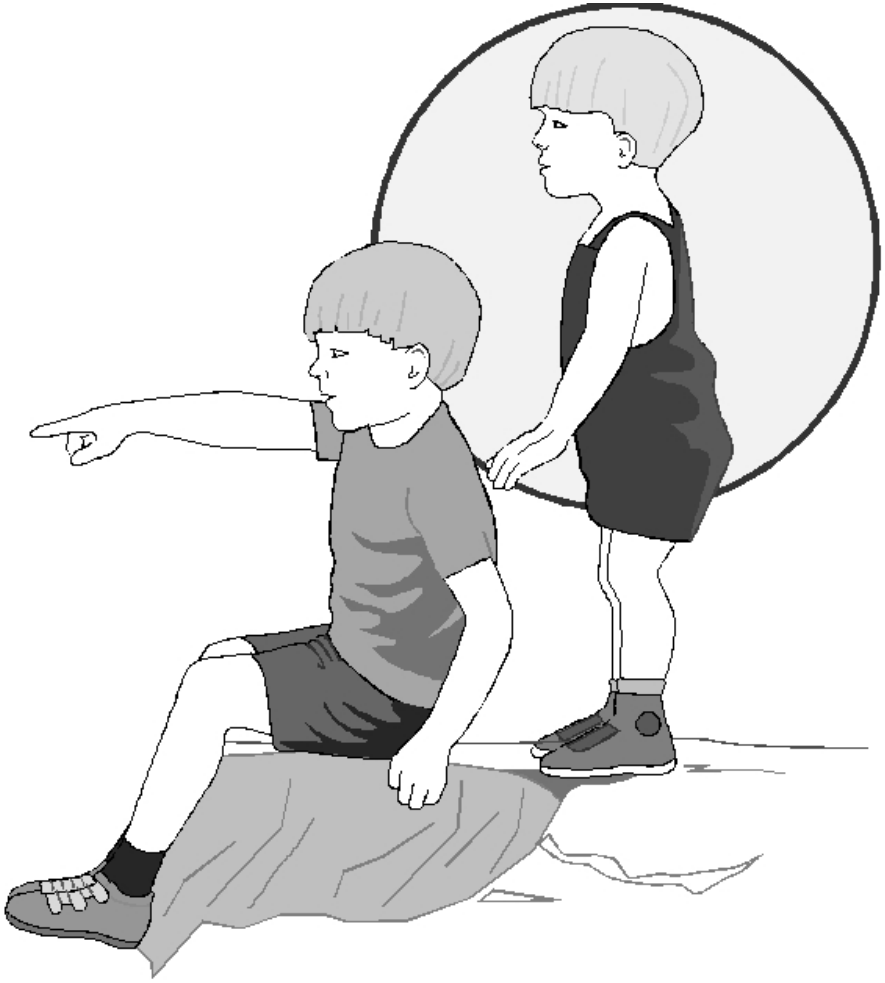
... Mere razpršenosti

	populacija	vzorec
varianca	σ^2	$S^2,\ s^2$
	$D,\ V$	
standardni odklon	σ	$S,\ \ s$

Za vzorec smo vzeli osebje na FRI.

Zabeležili smo naslednje število otrok:

1	2	2
1	2	5
1	2	



Varianca in standardni odklon

Varianca populacije (končne populacije z N elementi):

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

Varianca vzorca (n meritvami):

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1}$$

Standardni odklon je pozitivno predznačen kvadratni koren variance.

Sredine

$a_1, \ldots, a_n \geq 0$

Aritmetična: $$A_n = \frac{a_1 + \cdots + a_n}{n}$$

Geometrična: $$G_n = \sqrt[n]{a_1 \cdot \ldots \cdot a_n}$$

Harmonična: $$H_n = \frac{n}{\frac{1}{a_1} + \cdots + \frac{1}{a_n}}$$

Kvadratna: $$K_n = \sqrt{\frac{a_1^2 + \cdots + a_n^2}{n}}$$


NNGH!
WHO THE @#$% CAN REMEMBER HOW TO TAKE SQUARE ROOTS?!!

Sredine: $H_2 \leq G_2 \leq A_2 \leq K_2$

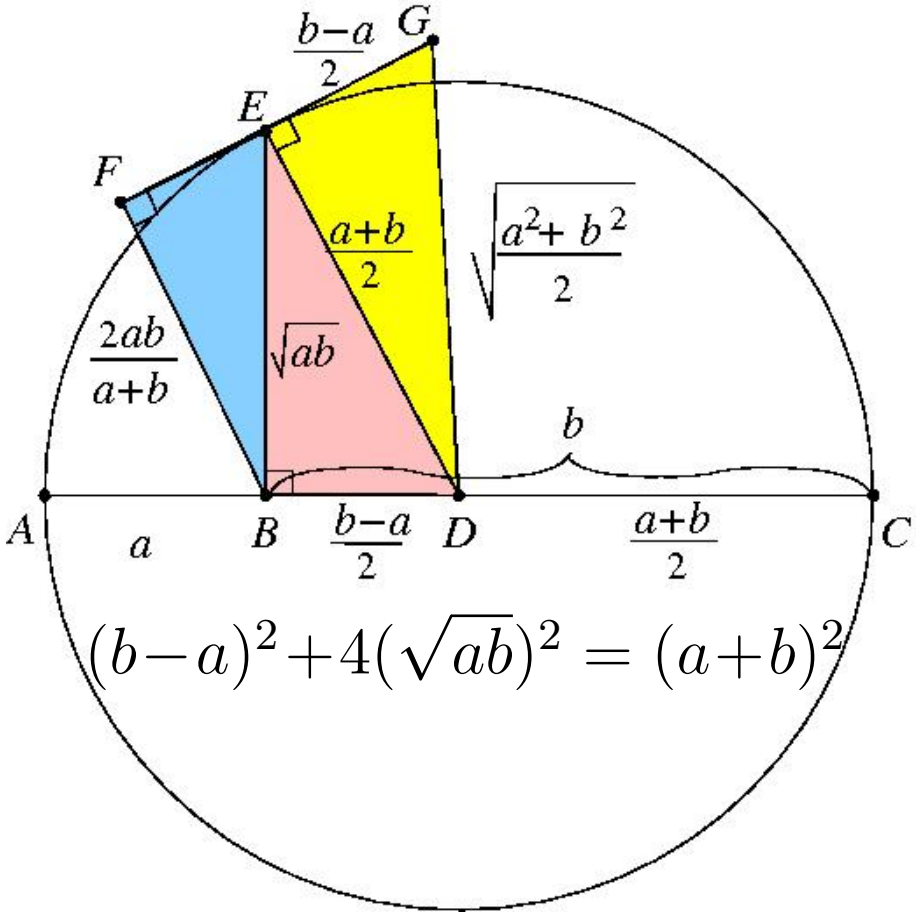
$a, b \geq 0$

$$H_2 = \frac{2}{\frac{1}{a} + \frac{1}{b}}$$

$$G_2 = \sqrt{ab}$$

$$A_2 = \frac{a+b}{2}$$

$$K_2 = \sqrt{\frac{a^2+b^2}{2}}$$



Sidney H. Kung

(iz R.B. Nelsenove knjige “Dokazi brez besed”)

... Sredine

Potenčna (stopnje k): $$P_{n,k} = \sqrt[k]{\frac{a_1^k + \cdots + a_n^k}{n}}$$

Velja:

$$H_n = P_{n,-1}, \quad G_n = \lim_{k\to 0} P_{n,k} \quad A_n = P_{n,1} \text{ in } K_n = P_{n,2}$$

ter

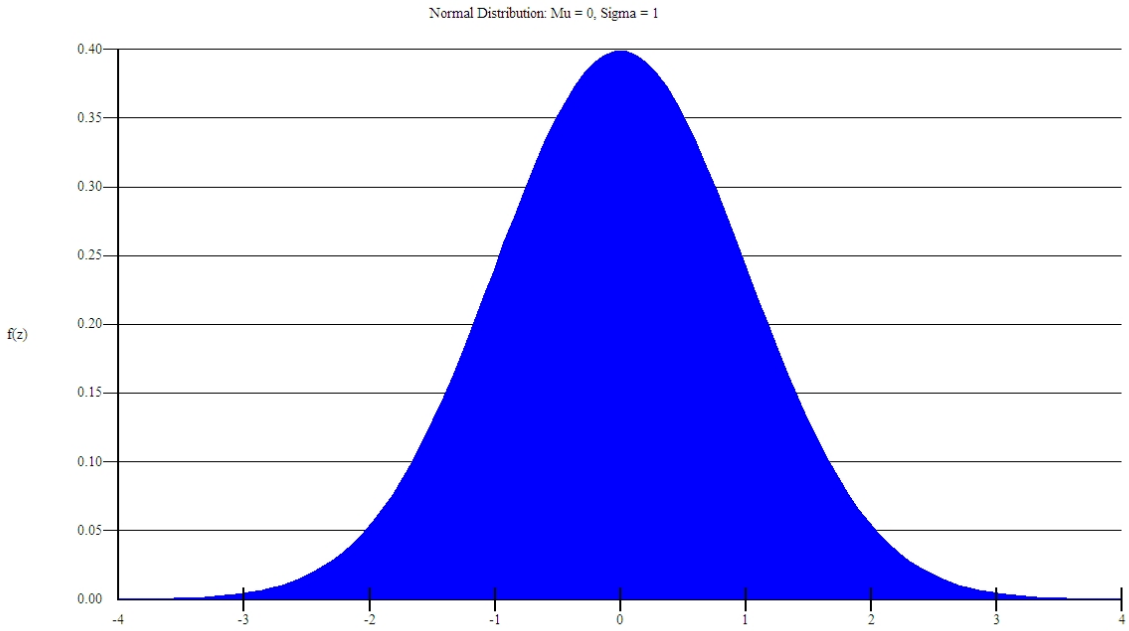
$$H_n \leq G_n \leq A_n \leq K_n$$

oziroma za $k \leq m$:

$$P_{n,k} \leq P_{n,m}$$

Normalna porazdelitev

Veliko podatkovnih množic ima porazdelitev približno zvonaste oblike (unimodalna oblika - ima en sam vrh):



Empirična pravila

Če ima podatkovna množica porazdelitev
približno zvonaste oblike,
potem veljajo naslednja pravila (angl. rule of thumb),
ki jih lahko uporabimo za opis podatkovne množice:

1. Približno 68,3% vseh meritev leži na razdalji
   $1\times$ *standardnega odklona* od njihovega povprečja.
2. Približno 95,4% meritev leži na razdalji do
   $2\times$ *standardnega odklona* od njihovega povprečja.
3. Skoraj vse meritve (99,7%) ležijo na razdalji
   $3\times$ *standardnega odklona* od njihovega povprečja.

Mere oblike

Če je spremenljivka približno normalno porazdeljena, potem jo statistični karakteristiki povprečje in standardni odklon zelo dobro opisujeta.

V primeru unimodalne porazdelitve spremenljivke, ki pa je bolj asimetrična in bolj ali manj sploščena (koničasta), pa je potrebno izračunati še stopnjo *asimetrije* in *sploščenosti (koničavosti)*.

Centralni momenti

ℓ-ti centralni moment je

$$m_\ell = \frac{(y_1 - \mu)^\ell + \cdots + (y_n - \mu)^\ell}{n}.$$

$m_1 = 0, m_2 = \sigma^2, \ldots$

Koeficient asimetrije (s centralnimi momenti):

$$g_1 = m_3/m_2^{3/2}.$$

Mere asimetrije

Razlike med srednjimi vrednostimi so tem večje,

čim bolj je porazdelitev asimetrična:

$$KA_{M_0} = (\mu - M_0)/\sigma,$$

$$KA_{M_e} = 3(\mu - M_e)/\sigma.$$

Mera sploščenosti (kurtosis)

Koeficient sploščenosti (s centralnimi momenti)

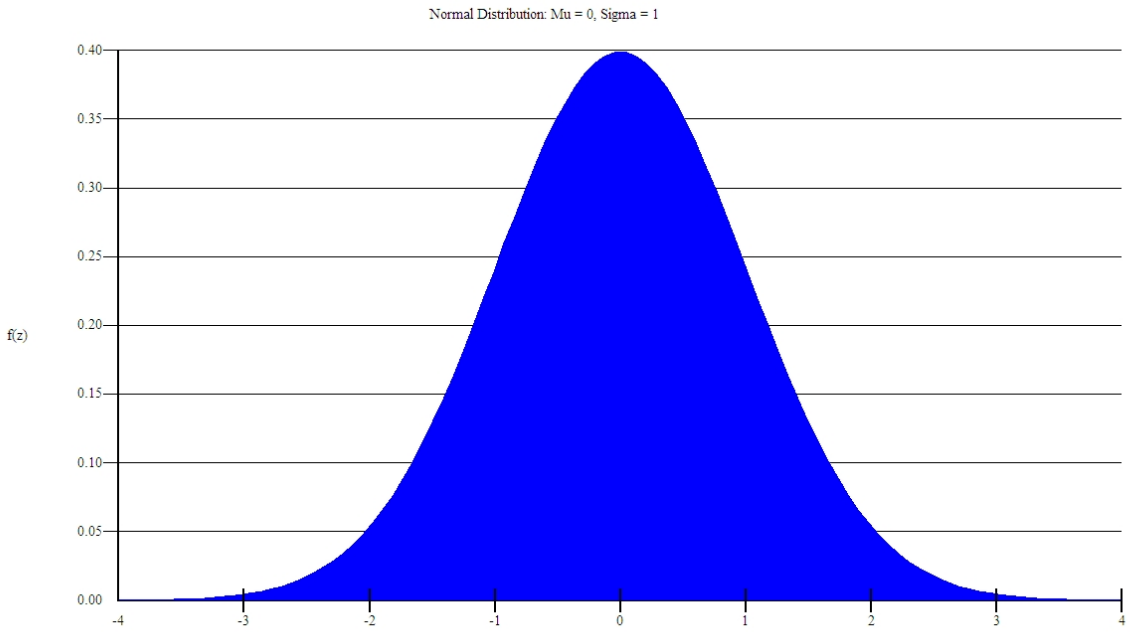
$$K = g_2 = m_4/m_2^2 - 3$$

– $K = 3$ (ali 0)
normalna porazdelitev zvonaste-oblike (*mesokurtic*),

– $K < 3$ (ali negativna)
bolj kopasta kot normalna porazdelitev, s krajšimi repi (*platykurtic*),

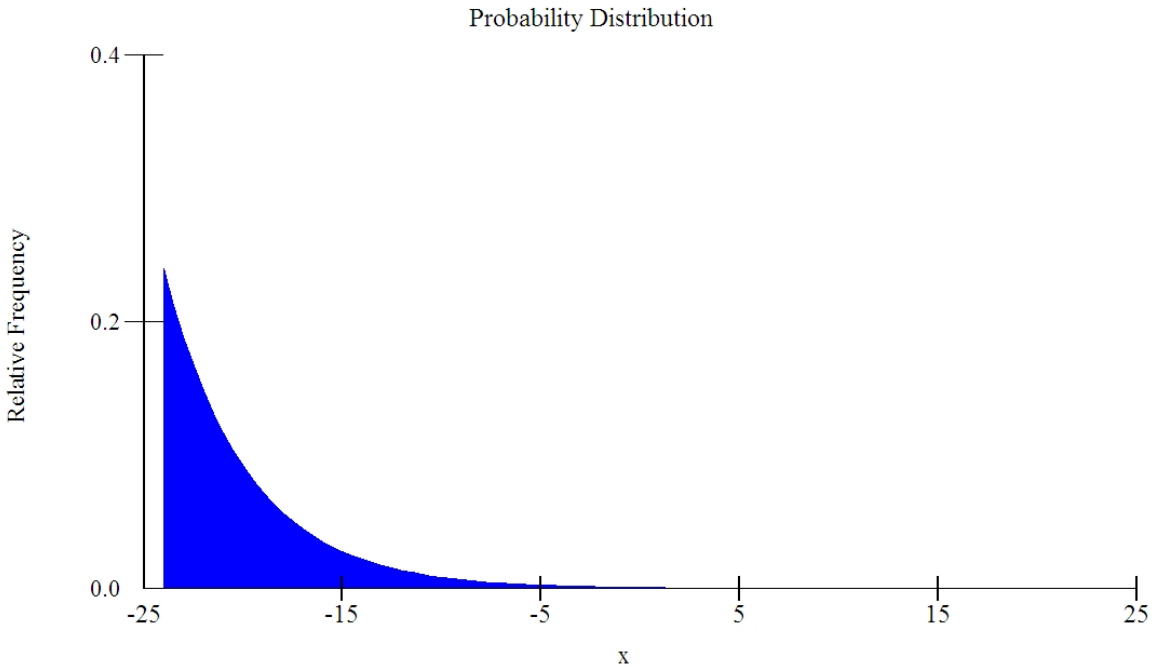
– $K > 3$ (ali pozitivna)
bolj špičasta kot normalna porazdelitev, z daljšimi repi (*leptokurtic*).

Normalna porazdelitev



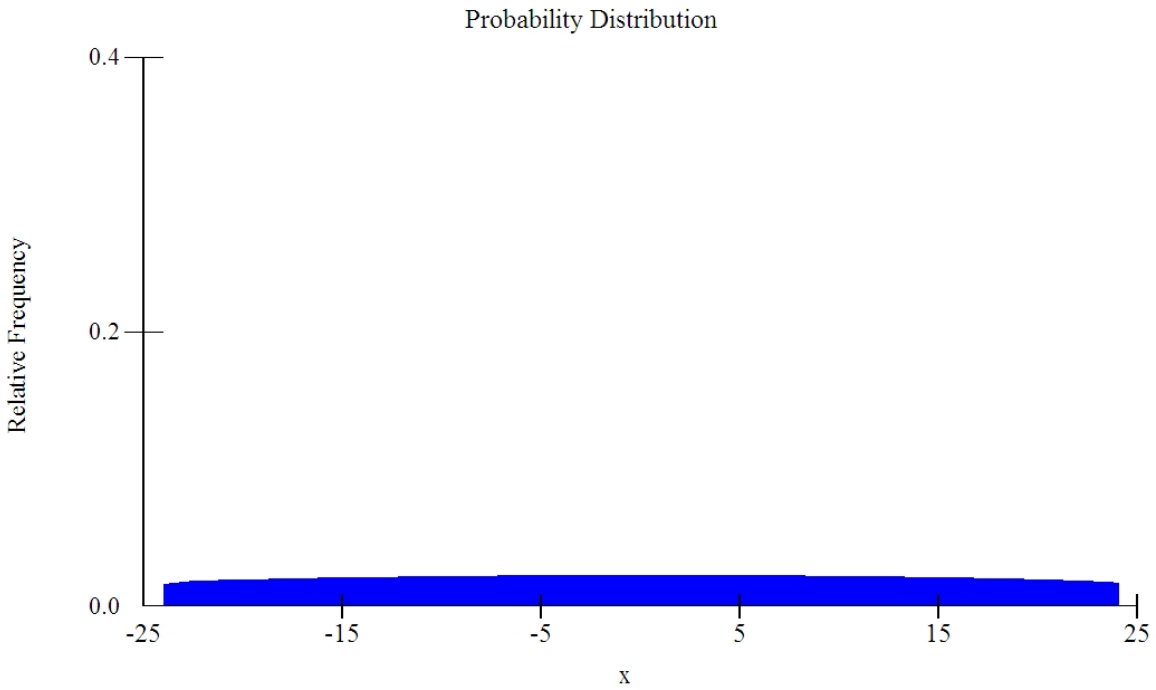
asimetričnost= 0, sploščenost= 3 (mesokurtic).

Asimetrična v desno



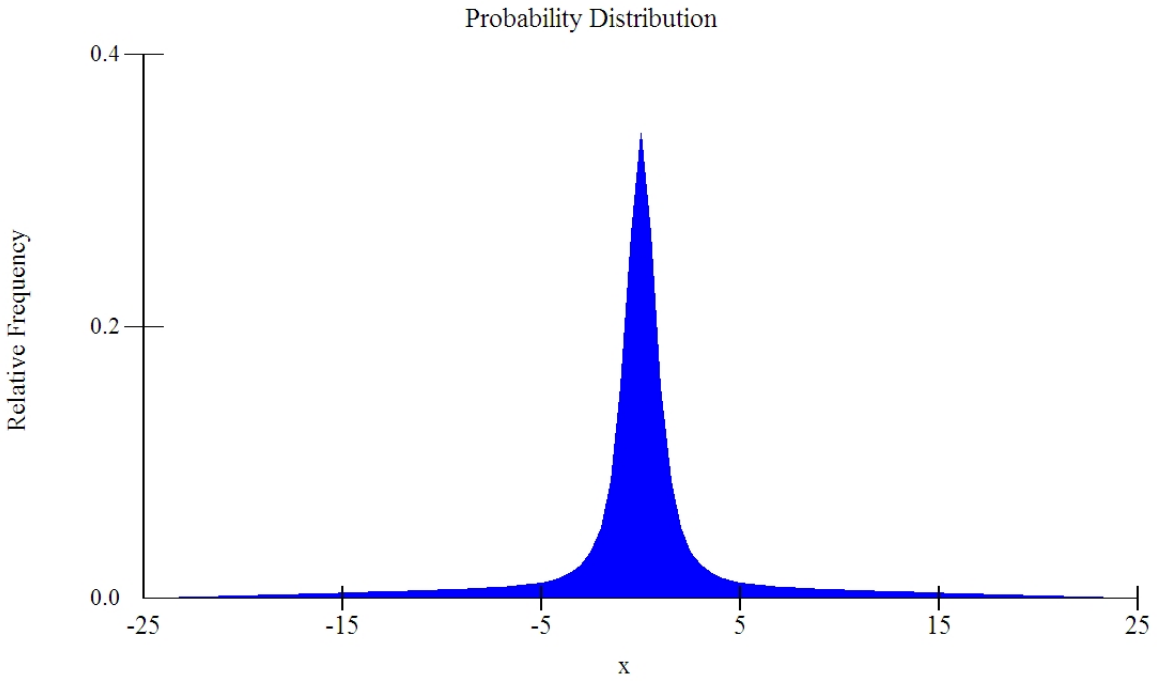
asimetričnost$= 1,99,$ sploščenost$= 8,85.$

Kopasta porazdelitev



asimetričnost$= 0$, sploščenost$= 1,86$ (platykurtic).

Špičasta porazdelitev



asimetričnost$= -1,99,$ sploščenost$= 8,85$ (leptokurtic).

Standardizacija

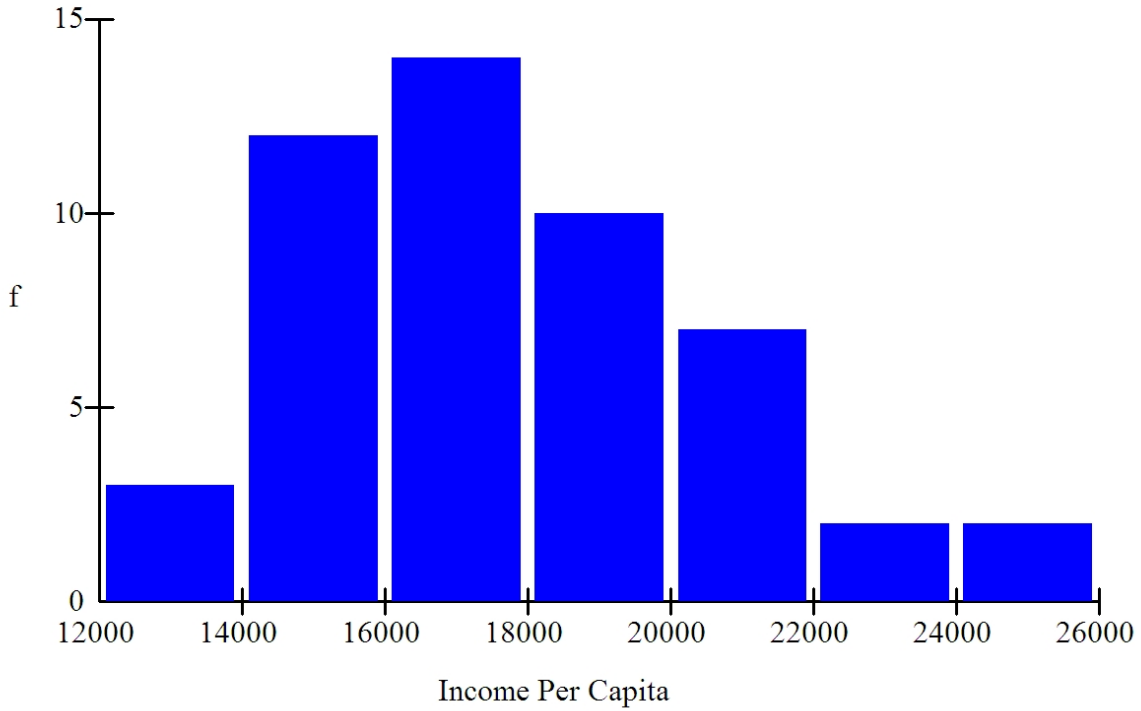
Vsaki vrednosti x_i spremenljivke X odštejemo njeno povprečje μ in delimo z njenim standardnim odklonom σ:

$$z_i = \frac{x_i - \mu}{\sigma}.$$

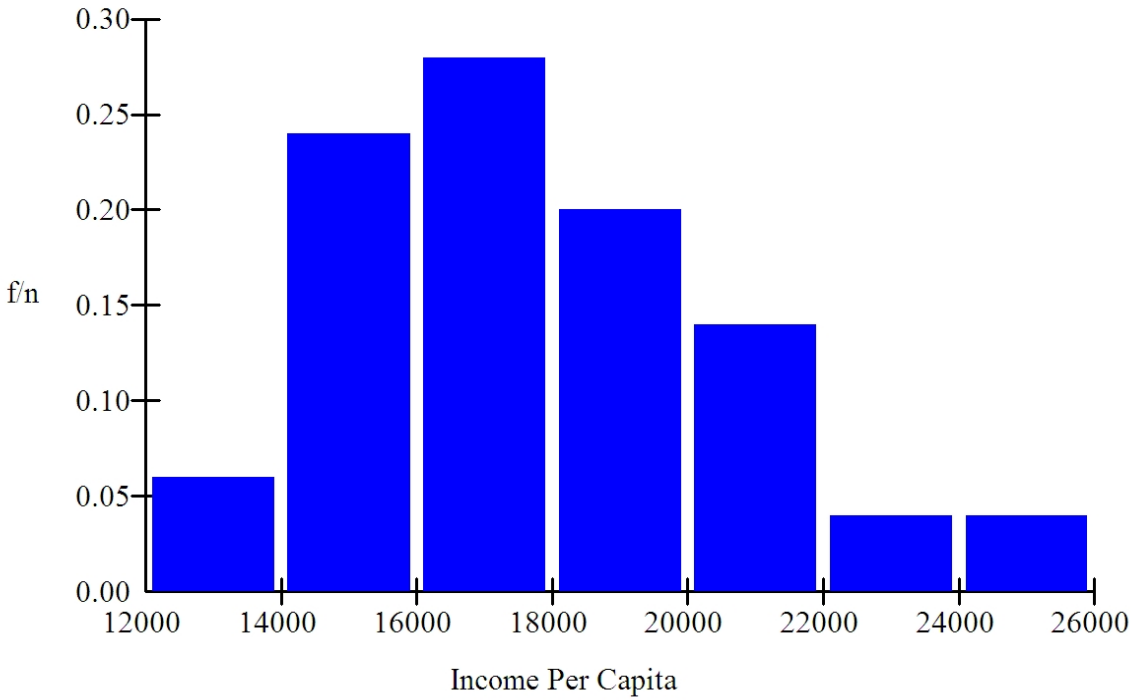
Za novo spremenljivko Z bomo rekli, da je **standardizirana**, z_i pa je **standardizirana vrednost**.

Potem je $\mu(Z) = 0$ in $\sigma(Z) = 1$.

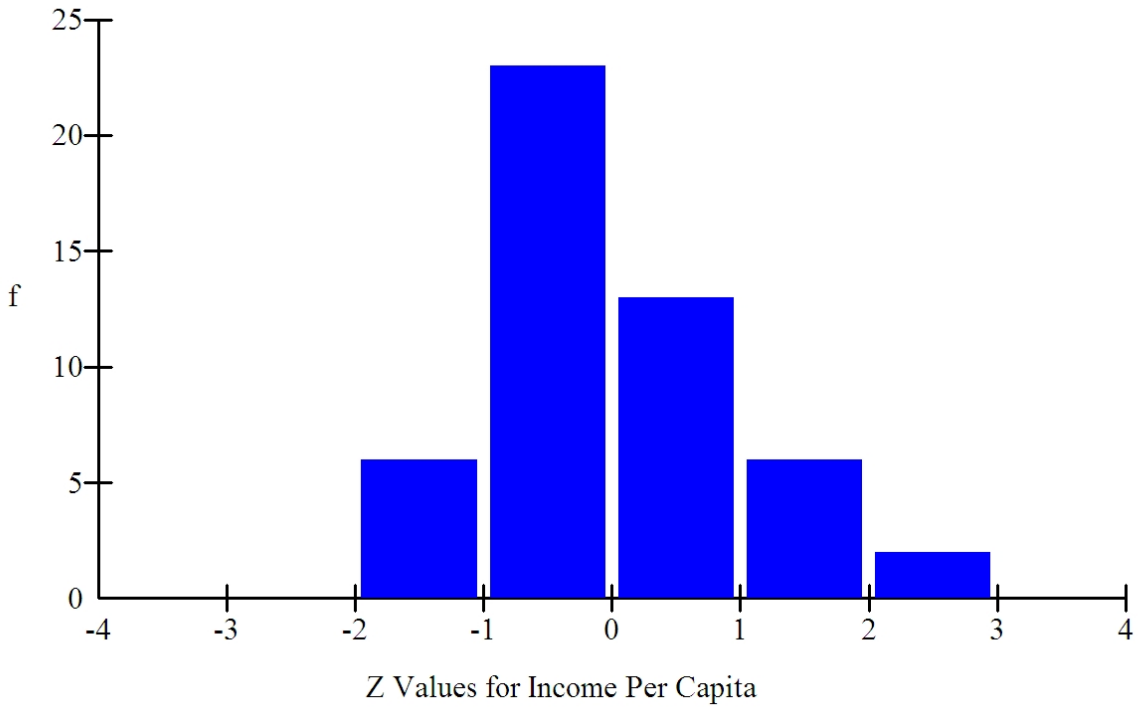
Frekvenčni histogram



Relativni frekvenčni histogram



Histogram standardiziranih Z-vrednosti



3. Zamenjalna šifra

Tomaž Pisanski, Skrivnostno sporočilo
Presek V/1, 1977/78, str. 40-42.

```
YHW?HD+CVODHVTHVO-!JVG:CDCYJ(JV/-V?HV(
-T?HVW-4YC4(?-DJV/-(?S-VO3CWC%J(-V4-DC
V!CW-?CVNJDJVD-?+-VO3CWC%J(-VQW-DQ-VJ+
V?HVDWHN-V3C:CODCV!H+?-DJVD-?+CV3JO-YC
```

(črko Č smo zamenjali s C, črko Ć pa z D)

Imamo $26! = 403291461126656355840000000$
možnosti z direktnim preizkušanjem,
zato v članku dobimo naslednje nasvete:

(0) Relativna frekvenca črk in presledkov v slovenščini: presledek 173,

E	A	I	O	N	R	S	L	J	T	V	D
89	84	74	73	57	44	43	39	37	37	33	30

K	M	P	U	Z	B	G	"C	H	"S	C	"Z	F
29	27	26	18	17	15	12	12	9	9	6	6	1

(1) Na začetku besed so najpogostejše črke
N, S, K, T, J, L.

(2) Najpogostejše končnice pa so
E, A, I, O, U, R, N.

(3) Ugotovi, kateri znaki zagotovo predstavljajo samoglasnike in kateri soglasnike.

(4) V vsaki besedi je vsaj en samoglasnik
ali samoglasniški R.

(5) V vsaki besedi z dvema črkama je ena
črka samoglasnik, druga pa soglasnik.

(6) detektivska sreča

```
(0)  V  -  C  D  J  ?  H  W  O  (  +  3
    23 19 16 12 11 10  9  7  6  6  5  4

     Y  4  !  /  Q  :  %  T  N  S  G
     4  3  3  2  2  2  2  2  2  1  1
```

Zaključek `V -->  ' '` (drugi znaki z visoko frekvenco ne morejo biti).

Dve besedi se ponovita: `03CWC%J(-`,
opazimo pa tudi eno sklanjatev:
`D-?+-` ter `D-?+C`.

Torej nadaljujemo z naslednjim tekstom:

```
YHW?HD+C ODH TH O-!J G:CDCYJ(J /- ?H
(-T?H W-4YD4(?-DJ /-(?S- 03CWC%J(- 4-DC
!CW-?C NJDJ D-?+- 03CWC%J(- QW-DQ- J+
?H DWHN- 3C:C0DC !H+?-DJ D-?+C 3J0-YC
```

(3) Kanditati za samoglasnike e,a,i,o so znaki z visokimi frekvancami.
Vzamemo:

$$\{e,a,i,o\} = \{-,C,J,H\}$$

(saj D izključi -,H,J,C in ? izključi -,H,C,
znaki -,C,J,H pa se ne izključujejo)

Razporeditev teh znakov kot samoglasnikov izgleda prav verjetna.
To potrdi tudi gostota končnic, gostota parov je namreč:

AV	CV	HV	JV	VO	?H	−D	DC	JM	W−	DJ	UC	CW	−?	VD
7	5	5	5	4	4	4	3	3	3	3	3	3	3	3

(5) Preučimo besede z dvema črkama:

```
   Samoglasnik na koncu
1) da ga na pa ta za (ha ja la)
2) "ce je le me ne se "se te ve "ze (he)
3) bi ji ki mi ni si ti vi
4) bo do (ho) jo ko no po so to
5) ju mu tu (bu)
6) r"z rt
```

```
   Samoglasnik na za"cetku
1) ar as                (ah aj au)
2) en ep                 (ej eh)
3) in iz ig
4) on ob od os on        (oh oj)
5) uk up u"s ud um ur    (uh ut)
```

in opazujemo besedi: /- ?H

ter besedi: J+ ?H.

J+ ima najmanj možnosti, + pa verjetno ni črka n, zato nam ostane samo še:

```
J+ ?H          DWHN-
/- ?H
iz te         (ne gre zaradi: D-?+C)
ob ta(e,o)    (ne gre zaradi: D-?+C)
od te         (ne gre zaradi: D-?+C)
```

tako da bo potrebno nekaj spremeniti in preizkusiti še naslednje:
on bo; on jo; in so; in se; in je; in ta; en je; od tu ...

(6) Če nam po dolgem premisleku ne uspe najti rdeče niti, bo morda potrebno iskati napako s prijatelji (tudi računalniški program z metodo lokalne optimizacije ni zmogel problema zaradi premajhne dolžine tajnopisa, vsekakor pa bi bilo problem mogoče rešiti s pomočjo elektronskega slovarja).

Tudi psihološki pristop pomaga, je svetoval Martin Juvan

in naloga je bila rešena (poskusite sami!).

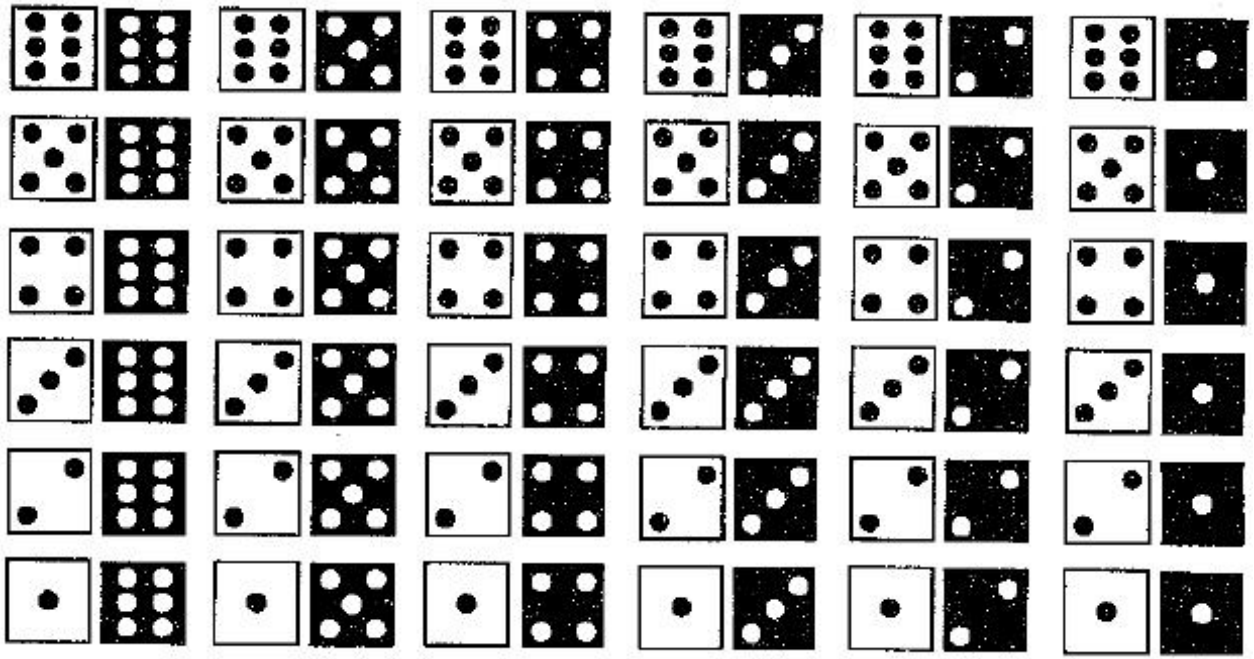
Podobna naloga je v angleščini dosti lažja, saj je v tem jeziku veliko členov THE, A in AN, vendar pa zato običajno najprej izpustimo presledke iz teksta, ki ga želimo spraviti v tajnopis. V angleščini imajo seveda črke drugačno gostoto kot v slovenščini. Razdelimo jih v naslednjih pet skupin:

1. E, z verjetnostjo okoli 0,120,
2. T, A, O, I, N, S, H, R, vse z verjetnostjo med 0,06 in 0,09,
3. D, L, obe z verjetnostjo okoli 0,04,
4. C, U, M, W, F, G, Y, P, B, vse z verjetnostjo med 0,015 in 0,028,
5. V, K, J, X, Q, Z, vse z verjetnostjo manjšo od 0,01.

Najbolj pogosti pari so (v padajočem zaporedju): TH, HE, IN, ER, AN, RE, ED, ON, ES, ST, EN, AT, TO, NT, HA, ND, OU, EA, NG, AS, OR, TI, IS, ET, IT, AR, TE, SE, HI in OF,

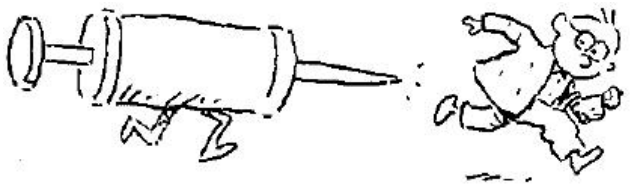
Najbolj pogoste trojice pa so (v padajočem zaporedju): THE, ING, AND, HER, ERE, ENT, THA, NTH, WAS, ETH, FOR in DTH.

KOMBINATORIKA (ponovitev)



Funkcije/preslikave

Funkcija f iz množice A v množico B je predpis, ki vsakemu elementu iz množice A priredi natanko določen element iz množice B, oznaka $f : A \longrightarrow B$.



Funkcija $f : A \longrightarrow B$ je:

- **injektivna** (angl. one to one) če za $\forall\, x, y \in A$

$$x \neq y \;\Rightarrow\; f(x) \neq f(y),$$

- **surjektivna** (angl. on to), če za $\forall\, b \in B$

$$\exists\, a \in A, \text{ tako da je } f(a) = b.$$

Injektivni in surjektivni funkciji pravimo **bijekcija**.

Množicama med katerima obstaja bijekcija pravimo **bijektivni** množici.

Bijektivni množici imata enako število elementov
(npr. končno, števno neskončno, itd).

Trditev: *Če sta množici A in B končni ter je $f : A \longrightarrow B$ funkcija*
iz injektivnosti funkcije f sledi surjektivnost, in obratno,
iz surjektivnosti funkcije f sledi injektivnost.

Permutacije

Permutacija elementov $1, \ldots, n$ je bijekcija, ki slika iz množice $\{1, \ldots, n\}$ v množico $\{1, \ldots, n\}$.

Npr. permutacija kart je običajno premešanje kart (spremeni se vrstni red, karte pa ostanejo iste).

Število permutacij n elementov, tj. razvrstitev n-tih različnih elementov, je enako $n! := 1 \cdot 2 \cdot \ldots \cdot n$ (oziroma definirano rekurzivno $n! = (n-1)!n$ in $0! = 1$).

Permutacijo lahko opišemo z zapisom:

$$\pi = \begin{pmatrix} 1 & 2 & \dots & n \\ a_1 & a_2 & \dots & a_n \end{pmatrix}$$

kjer je $\{1, 2, \dots, n\} = \{a_1, a_2, \dots, a_n\}$.

To pomeni $\pi(1) = a_1,\ \pi(2) = a_2, \dots,\ \pi(n) = a_n$.

Primer: $n = 11$,

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix}$$

Naj bo A neka množica. Permutacije množice A med seboj množimo po naslednjem pravilu: $\pi = \pi_1 \circ \pi_2$ je permutacija množice A, ki preslika $a \in A$ v $\pi_2(\pi_1(a))$.

Primer:

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix}$$

$$\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 8 & 2 & 1 & 3 & 10 & 9 & 4 & 5 & 7 & 6 & 11 \end{pmatrix}$$

Potem je

$$\pi = \pi_1 \circ \pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 1 & 3 & 10 & 6 & 2 & 8 & 4 & 7 & 11 & 9 & 5 \end{pmatrix}$$

Cikel je permutacija, za katero je

$$\pi(a_1) = a_2,\ \pi(a_2) = a_3,\ \ \ldots, \pi(a_r) = a_1,$$

ostale elementi pa so fiksni (tj. $\pi(a) = a$).

Na kratko jo zapišemo z $(a_1\, a_2 \ldots\, a_r)$.

Trditev: *Vsako permutacijo lahko zapišemo kot produkt disjunktnih ciklov.*

Primer:

$$\pi_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 4 & 5 & 10 & 2 & 1 & 7 & 9 & 11 & 6 & 8 \end{pmatrix}$$

$$\pi_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 8 & 2 & 1 & 3 & 10 & 9 & 4 & 5 & 7 & 6 & 11 \end{pmatrix}$$

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 1 & 3 & 10 & 6 & 2 & 8 & 4 & 7 & 11 & 9 & 5 \end{pmatrix}$$

Potem je

$$\begin{aligned} \pi_1 &= (1\,3\,5\,2\,4\,10\,6)\,(8\,9\,11), \\ \pi_2 &= (1\,8\,5\,10\,6\,9\,7\,4\,3), \\ \pi &= (2\,3\,10\,9\,11\,5\,2)\,(4\,6\,8\,7) \end{aligned}$$

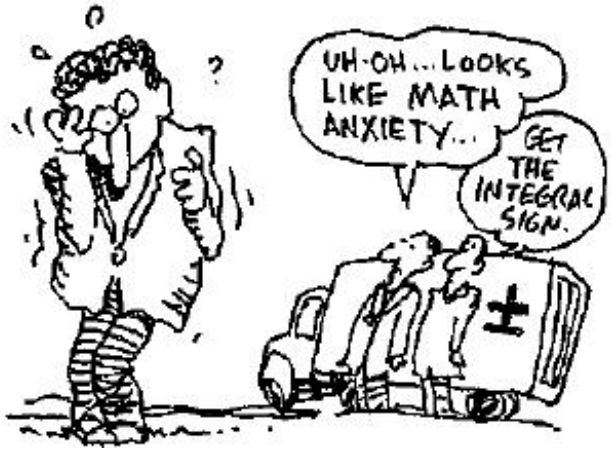
Transpozicija je cikel dolžine 2. Vsak cikel pa je produkt transpozicij:

$$(a_1\, a_2\, a_3\, \ldots\, a_r) = (a_1\, a_2) \circ (a_2\, a_3) \circ \cdots \circ (a_{r-1}\, a_r),$$

torej je tudi vsaka permutacija produkt transpozicij.

Seveda ta produkt ni nujno enolično določen,
vseeno pa velja:

Trditev: *Nobena permutacija se ne da zapisati kot produkt sodega števila in kot produkt lihega števila permutacij.*



Dokaz: Naj bodo $x_1, x_2, \ldots, x_n$ različna realna števila. Poglejmo si produkt: $P = \prod_{i<j}(x_i - x_j)$. Izberimo indeksa a in b, $a < b$, in poglejmo v katerih razlikah se pojavita:

$x_1 - x_a, \ldots, x_{a-1} - x_a,$		$x_a - x_{a+1}, \ldots, x_a - x_{b-1},$	$x_a - x_b,$	$x_a - x_{b+1}, \ldots, x_a - x_n,$
$x_1 - x_b, \ldots, x_{a-1} - x_b,$	$x_a - x_b,$	$x_{a+1} - x_b, \ldots, x_{b-1} - x_b,$		$x_b - x_{b+1}, \ldots, x_b - x_n.$

Razliko $x_a - x_b$ smo navedli dvakrat, a se v produktu P pojavi samo enkrat.
Če na množici indeksov opravimo transpozicijo $(a\,b)$,
razlika $x_a - x_b$ preide v razliko $x_b - x_a$, torej zamenja predznak,
razlike iz prvega in zadnjega stolpca se med seboj zamenjajo,
razlike iz srednjega stolpca pa tudi zamenjajo predznake
(vendar je le-teh sodo mnogo in zato ne vplivajo na produkt P).

Sedaj pa napravimo na množici indeksov permutacijo π.
V tem primeru je produkt

$$P_\pi = \prod_{i<j} (x_{\pi(i)} - x_{\pi(j)}).$$

enak $\pm P$. Če uporabimo sodo število transpozicij, potem je $P_\pi = P$, sicer pa $P_\pi = -P$. ■

Glede na sodo oziroma liho število transpozicij imenujemo permutacijo **soda** oziroma **liha** permutacija.

Permutacije s ponavljanjem

Permutacije s ponavljanjem so nekakšne permutacije, pri katerih pa ne ločimo elementov v skupinah s $k_1, k_2, \ldots, k_r$ elementi - zato delimo število vseh permutacij s številom njihovih vrstnih redov, tj. permutacij:

$$\frac{n!}{k_1!k_2!\cdots k_r!}.$$

Kombinacije

Binomski koeficient oz. število **kombinacij**, tj. število m-elementnih podmnožic množice moči n, je

$$\binom{n}{m} = \frac{n \cdot (n-1) \cdots (n-m+1)}{1 \cdot 2 \cdots m} = \frac{n!}{m!(n-m)!},$$

saj lahko prvi element izberemo na n načinov,
drugi na $n-1$ načinov, . . . ,
zadnji na $n-m+1$ načinov,
ker pa vrstni red izbranih elementov ni pomemben,
dobljeno število še delimo s številom permutacij.

Binomski obrazec - ponovitev

Trditev: *Za binomske simbole velja*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \quad in \quad \binom{n}{m} + \binom{n}{m+1} = \binom{n+1}{m+1}.$$

Dokaz: Po definiciji je desna enakost ekvivalentna z

$$\frac{n!}{m!(n-m)!} + \frac{n!}{(m+1)!(n-m-1)!} = \frac{(n+1)!}{(m+1)!(n-m)!},$$

oziroma po množenju z $m!(n-m-1)!/n!$ z

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

Prvi del trditve dokažemo s kombinatoriko (štetje)
ali matematično indukcijo. ■

Pascalov trikotnik - ponovitev

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

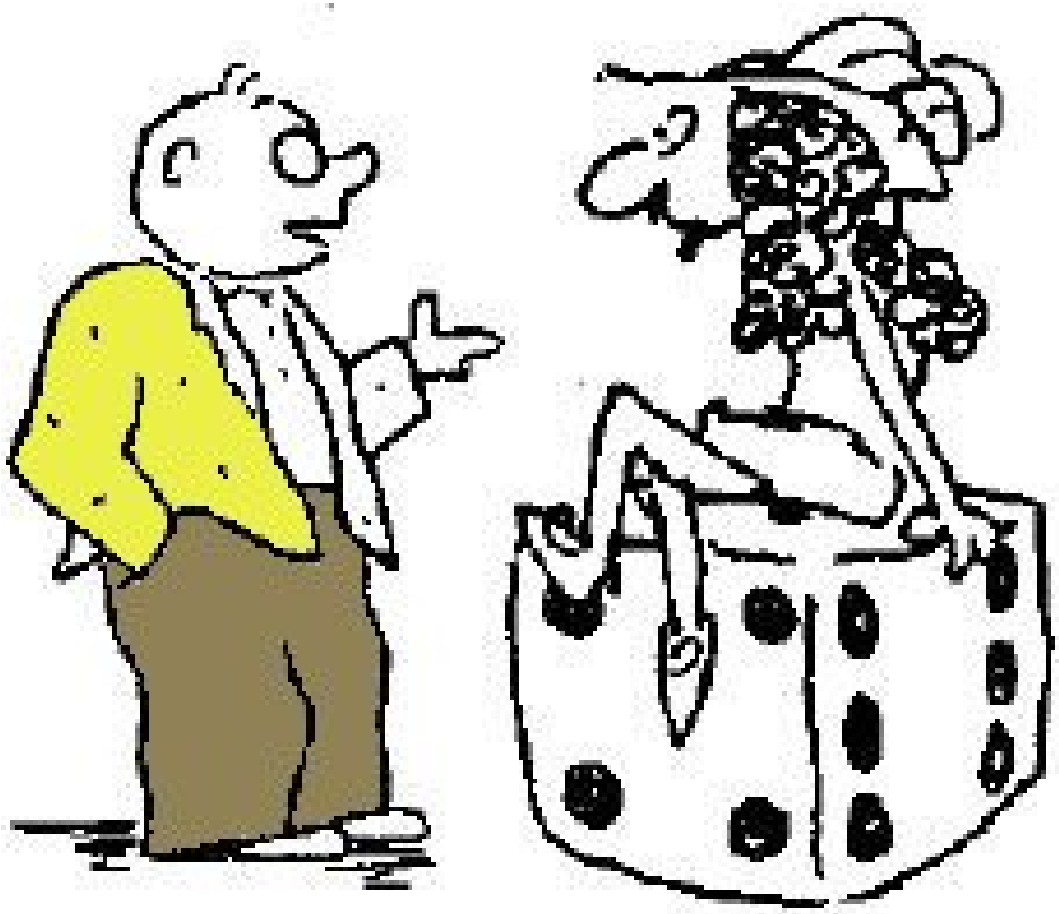
1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

To so v obliki trikotnika zapisani binomski simboli,
vsaka vrstica pa ustreza enemu binomskemu obrazcu.

I. VERJETNOST



I.1. Poskusi, dogodki in verjetnost

Začetki verjetnosti

Ahil in Ajaks kockata, amfora, okrog 530 pr.n.š, Eksekias, Vatikan

Naključnost so poznale že stare kulture: Egipčani, Grki, ...
a je niso poskušale razumeti –
razlagale so jo kot voljo bogov.

Leta 1662 je plemič Chevalier de Mere zastavil matematiku Blaise Pascalu vprašanje:

zakaj določene stave prinašajo dobiček druge pa ne.

Le-ta si je o tem začel dopisovati s Fermatom
in iz tega so nastali začetki verjetnostnega računa.

. . . začetki

Prvo tovrstno razpravo je napisal že leta 1545 italijanski kockar in matematik Cardano, a ni bila širše znana.

Tudi leta 1662 je angležJohn Graunt sestavil na osnovi podatkov prve zavarovalniške tabele.

Leta 1713 je Jakob Bernoulli objavil svojo knjigo *Umetnost ugibanja* s katero je verjetnostni račun postal resna in splošno uporabna veda.

Njegov pomen je še utrdil Laplace, ko je pokazal njegov pomen pri analizi astronomskih podatkov (1812).

Leta 1865 je avstrijski menih Gregor Mendel uporabil verjetnostno analizo pri razlagi dednosti v genetiki.

V 20. stoletju se je uporaba verjetnostnih pristopov razširila skoraj na vsa področja.

Poskus

Verjetnostni račun obravnava zakonitosti, ki se pokažejo v velikih množicah enakih ali vsaj zelo podobnih pojavov. Predmet verjetnostnega računa je torej empirične narave in njegovi osnovni pojmi so povzeti iz izkušnje. Osnovni pojmi v verjetnostnem računu so: poskus, dogodek in verjetnost dogodka.

Poskus je realizacija neke množice skupaj nastopajočih dejstev (kompleksa pogojev). Poskus je torej vsako dejanje, ki ga opravimo v natanko določenih pogojih.

Primeri:

- met igralne kocke,
- iz kupa 20 igralnih kart izberemo eno karto.

Dogodki

Pojav, ki v množico skupaj nastopajočih dejstev ne spada in se lahko v posameznem poskusu zgodi ali pa ne, imenujemo **dogodek**.

Primeri:

- v poskusu meta igralne kocke je na primer dogodek, da vržemo 6 pik;
- v poskusu, da vlečemo igralno karto iz kupa 20 kart, je dogodek, da izvlečemo rdečo barvo.

Za poskuse bomo privzeli, da jih lahko neomejeno velikokrat ponovimo. Dogodki se bodo nanašali na isti poskus.

Poskuse označujemo z velikimi črkami iz konca abecede, npr. X, Y, X_1. Dogodke pa označujemo z velikimi črkami iz začetka abecede, npr. A, C, E_1.

Vrste dogodkov

Dogodek je lahko:

- **gotov** dogodek – G: ob vsaki ponovitvi poskusa se zgodi.

  Primer: dogodek, da vržemo 1, 2, 3, 4, 5, ali 6 pik pri metu igralne kocke;

- **nemogoč** dogodek – N: nikoli se ne zgodi.

  Primer: dogodek, da vržemo 7 pik pri metu igralne kocke;

- **slučajen** dogodek: včasih se zgodi, včasih ne.

  Primer: dogodek, da vržemo 6 pik pri metu igralne kocke.

Računanje z dogodki

Dogodek A je **poddogodek** ali **način** dogodka B, kar zapišemo $A \subset B$, če se vsakič, ko se zgodi dogodek A, zagotovo zgodi tudi dogodek B.

Primer: Pri metu kocke je dogodek A, da pade šest pik,
način dogodka B, da pade sodo število pik.

Če je dogodek A način dogodka B in sočasno dogodek B način dogodka A, sta dogodka **enaka**: $(A \subset B) \wedge (B \subset A) \iff A = B$.

Vsota dogodkov A in B, označimo jo z $A \cup B$ ali $A + B$,
se zgodi, če se zgodi **vsaj** eden od dogodkov A in B.

Primer: Vsota dogodka A, da vržemo sodo število pik,
in dogodka B, da vržemo liho število pik, je gotov dogodek.

Velja: $A \cup B = B \cup A$; $A \cup N = A$; $A \cup G = G$; $A \cup A = A$

$$B \subset A \iff A \cup B = A \,;\; A \cup (B \cup C) = (A \cup B) \cup C$$

. . . Računanje z dogodki

Produkt dogodkov A in B, označimo ga z $A \cap B$ ali AB, se zgodi, če se zgodita A in B **hkrati**.

Primer: Produkt dogodka A, da vržemo sodo število pik,
in dogodka B, da vržemo liho število pik, je nemogoč dogodek.

Velja: $A \cap B = B \cap A$; $A \cap N = N$; $A \cap G = A$; $A \cap A = A$

$$B \subset A \iff A \cap B = B \ ; \ A \cap (B \cap C) = (A \cap B) \cap C$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \ ; \ A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Dogodku A **nasproten** dogodek $\overline{A}$ imenujemo negacijo dogodka A.

Primer: Nasproten dogodek dogodku, da vržemo sodo število pik,
je dogodek, da vržemo liho število pik.

Velja: $A \cap \overline{A} = N$; $A \cup \overline{A} = G$; $\overline{N} = G$; $\overline{\overline{A}} = A$

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \ ; \ \overline{A \cap B} = \overline{A} \cup \overline{B}$$

. . . Računanje z dogodki

Dogodka A in B sta **nezdružljiva**, če se ne moreta zgoditi hkrati, njun produkt je torej nemogoč dogodek, $A \cap B = N$.

Primer: Dogodka, A – da pri metu kocke pade sodo število pik in B – da pade liho število pik, sta nezdružljiva.

Poljuben dogodek in njegov nasprotni dogodek sta vedno nezdružljiva. Ob vsaki ponovitvi poskusa se zagotovo zgodi eden od njiju, zato je njuna vsota gotov dogodek: $(A \cap \overline{A} = N) \ \wedge \ (A \cup \overline{A} = G)$.

Če lahko dogodek A izrazimo kot vsoto nezdružljivih in mogočih dogodkov, rečemo, da je A **sestavljen** dogodek. Dogodek, ki ni sestavljen, imenujemo **osnoven** ali **elementaren** dogodek.

Primer: Pri metu kocke je šest osnovnih dogodkov: E_1, da pade 1 pika, E_2, da padeta 2 piki, . . . , E_6, da pade 6 pik. Dogodek, da pade sodo število pik je sestavljen dogodek iz treh osnovnih dogodkov (E_2, E_4 in E_6).

. . . Računanje z dogodki

Množico dogodkov $S = \{A_1, A_2, \ldots, A_n\}$ imenujemo **popoln sistem dogodkov**, če se v vsaki ponovitvi poskusa zgodi natanko eden od dogodkov iz množice S.

To pomeni, da so vsi mogoči

$$A_i \neq N,$$

paroma nezdružljivi

$$A_i \cap A_j = N \quad i \neq j$$

in njihova vsota je gotov dogodek

$$A_1 \cup A_2 \cup \ldots \cup A_n = G.$$

Primer: Popoln sistem dogodkov pri metu kocke sestavljajo na primer osnovni dogodki ali pa tudi dva dogodka: dogodek, da vržem sodo število pik, in dogodek, da vržem liho število pik.

I.2. Definicija verjetnosti

Opišimo najpreprostejšo verjetnostno zakonitost. Denimo, da smo n–krat ponovili dan poskus in da se je k–krat zgodil dogodek A. Ponovitve poskusa, v katerih se A zgodi, imenujemo ugodne za dogodek A, število

$$f(A) = \frac{k}{n}$$

pa je **relativna frekvenca** (pogostost) dogodka A v opravljenih poskusih.

Statistični zakon, ki ga kaže izkušnja, je:

> *Če poskus X dolgo ponavljamo, se relativna frekvenca slučajnega dogodka ustali in sicer skoraj zmeraj toliko bolj, kolikor več ponovitev poskusa napravimo.*

Statistična definicija verjetnosti

To temeljno zakonitost so empirično preverjali na več načinov. Najbolj znan je poskus s kovanci, kjer so določali relativno frekvenco grba ($f(A)$):

- Buffon je v 4040 metih dobil $f(A) = 0,5069$,
- Pearson je v 12000 metih dobil $f(A) = 0,5016$,
- Pearson je v 24000 metih dobil $f(A) = 0,5005$.

Ti poskusi kažejo, da se relativna frekvenca grba pri metih kovanca običajno ustali blizu $0,5$. Ker tudi drugi poskusi kažejo, da je ustalitev relativne frekvence v dovolj velikem številu ponovitev poskusa splošna zakonitost, je smiselna naslednja *statistična definicija verjetnosti*:

> *Verjetnost dogodka A v danem poskusu je število $P(A)$, pri katerem se navadno ustali relativna frekvenca dogodka A v velikem številu ponovitev tega poskusa.*

Osnovne lastnosti verjetnosti

1. *Ker je relativna frekvenca vedno nenegativna, je verjetnost* $P(A) \geq 0$.

2. $P(G) = 1$, $P(N) = 0$ *in* $A \subset B \Rightarrow P(A) \leq P(B)$.

3. *Naj bosta dogodka* A *in* B *nezdružljiva. Tedaj velja*

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

Klasična definicija verjetnosti

Pri določitvi verjetnosti si pri nekaterih poskusih in dogodkih lahko pomagamo s *klasično definicijo verjetnosti*:

Vzemimo, da so dogodki iz popolnega sistema dogodkov $\{E_1, E_2, \ldots, E_s\}$ *enako verjetni:*
$P(E_1) = P(E_2) = \ldots = P(E_s) = p.$
Tedaj je $P(E_i) = 1/s \quad i = 1, \ldots, s.$
Če je nek dogodek A *sestavljen iz* r *dogodkov iz tega popolnega sistema dogodkov, potem je njegova verjetnost* $P(A) = r/s.$

Primer: Izračunajmo verjetnost dogodka A,
da pri metu kocke padejo manj kot 3 pike.
Popolni sistem enako verjetnih dogodkov sestavlja 6 dogodkov.
Od teh sta le dva ugodna za dogodek A (1 in 2 piki).

Zato je verjetnost dogodka A enaka $\frac{2}{6} = \frac{1}{3}$.

Geometrijska verjetnost

V primerih, ko lahko osnovne dogodke predstavimo kot ‘enakovredne’ točke na delu premice (ravnine ali prostora),
določimo verjetnost sestavljenega dogodka kot razmerje dolžin (ploščin, prostornin) dela, ki ustreza ugodnim izidom,
in dela, ki ustreza vsem možnim izidom.

Še dve lastnosti verjetnosti

4. Za dogodka A in B velja:

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

Primer: Denimo, da je verjetnost, da študent naredi izpit iz Sociologije $P(S) = 2/3$. Verjetnost, da naredi izpit iz Politologije je $P(P) = 5/9$. Če je verjetnost, da naredi vsaj enega od obeh izpitov $P(S \cup P) = 4/5$, kolikšna je verjetnost, da naredi oba izpita?

$$P(S \cap P) = P(S) + P(P) - P(S \cup P) =$$

$$= \frac{2}{3} + \frac{5}{9} - \frac{4}{5} = 0,42.$$

Za dogodke A, B in C velja:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C). \end{aligned}$$

Kako lahko to pravilo posplošimo še na več dogodkov?

Namig: Pravilo o vključitvi in izključitvi za množice $A_1, A_2, \ldots, A_n$:

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{i=1}^{n} |A_i| - \sum_{1 \le i_1 < i_2 \le n} |A_{i_1} \cap A_{i_2}|$$

$$+ \sum_{1 \le i_1 < i_2 < i_3 \le n} |A_{i_1} \cap A_{i_2} \cap A_{i_3}| - \cdots + (-1)^{n-1} |A_1 \cap A_2 \cap \ldots \cap A_n|.$$

. . . Še dve lastnosti verjetnosti

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

Primer: Iz kupa 32 kart slučajno povlečemo 3 karte. Kolikšna je verjetnost, da je med tremi kartami vsaj en as (dogodek A)?

Pomagamo si z nasprotnim dogodkom. Nasprotni dogodek $\overline{A}$ dogodka A je, da med tremi kartami ni asa. Njegova verjetnost po klasični definiciji verjetnosti je določena s kvocientom števila vseh ugodnih dogodkov v popolnem sistemu dogodkov s številom vseh dogodkov v tem sistemu dogodkov. Vseh dogodkov v popolnem sistemu dogodkov je $\binom{32}{3}$, ugodni pa so tisti, kjer zbiramo med ne–asi, t.j. $\binom{28}{3}$. Torej je

$$P(\overline{A}) = \frac{\binom{28}{3}}{\binom{32}{3}} = 0,66 \; ; \; P(A) = 1 - P(\overline{A}) = 1 - 0,66 = 0,34.$$

. . . Še dve lastnosti verjetnosti

Posledica. *Če so dogodki A_i, $i \in I$ paroma nezdružljivi, velja*

$$P\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} P(A_i).$$

Velja tudi za števno neskončne množice dogodkov.

Aksiomi Kolmogorova

Dogodek predstavimo z množico zanj ugodnih izidov; gotov dogodek G ustreza univerzalni množici; nemogoč dogodek pa prazni množici.

Neprazna družina dogodkov $\mathcal{D}$ je **algebra**, če velja:

- $A \in \mathcal{D} \Rightarrow \overline{A} \in \mathcal{D}$,
- $A, B \in \mathcal{D} \Rightarrow A \cup B \in \mathcal{D}$.

Pri neskončnih množicah dogodkov moramo drugo zahtevo posplošiti

- $A_i \in \mathcal{D}, i \in I \Rightarrow \bigcup_{i \in I} A_i \in \mathcal{D}$.

Dobljeni strukturi rečemo σ**-algebra**.

. . . Aksiomi Kolmogorova

Naj bo $\mathcal{D}$ σ-algebra v G. **Verjetnost na G** je preslikava $P : \mathcal{D} \to \mathbb{R}$ z lastnostmi:

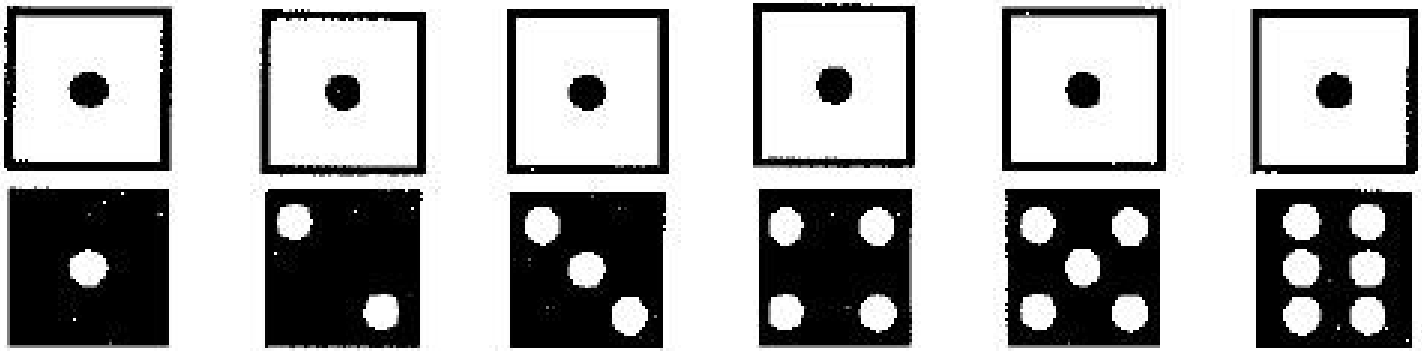
1. $P(A) \geq 0$,
2. $P(G) = 1$,
3. Če so dogodki A_i, $i \in I$ paroma nezdružljivi, je

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i).$$

Trojica $(G, \mathcal{D}, P)$ določa **verjetnostni prostor**.

Iz teh treh aksiomov lahko izpeljemo vse ostale lastnosti verjetnosti (Hladnik, str. 12).

I.3. Pogojna verjetnost



Intriga (po Kvarkadabri)

Ne podcenjujmo vpliva najrazličnejših rubrik v popularnih časopisnih prilogah, kjer nas domnevni “strokovnjaki” zasipajo z nasveti vseh vrst,

- rubrike krojijo mnenja ljudi in spreminjajo navade celotnih nacij,

- sprožajo obsežne polemike tako med širšimi množicami
  kot tudi v ozki strokovni javnosti.

Na področju zdravja in prehrane tako burne odzive seveda pričakujemo, povsem nekaj drugega pa je, če jih sproži preprosto *matematično vprašanje*.

Revija *Parade* - kot prilogo jo vsako nedeljo dodajo več kot 400 ameriškim časopisom in doseže okoli 70 milijonov bralcev, že dolgo izhaja rubrika z imenom "Vprašajte Marilyn." Ureja jo **Marilyn vos Savant**.

Sredi 80ih jo je *Guinnessova knjiga rekordov* razglasila za rekorderko z najvišjim inteligenčnim količnikom na planetu.

V svoji rubriki zdaj že več kot 20 let odgovarja na najrazličnejša vprašanja bralcev in rešuje njihove težave.

Med vsemi vprašanji, ki jih je kdaj obravnavala, ima prav posebno mesto na prvi pogled zelo preprost problem, ki ji ga je 9. septembra 1990 zastavil gospod Craig F. Whitaker:

Dve kozi in avtomobil

"Vzemimo, da sodelujete v nagradni igri,
kjer vam ponudijo na izbiro troje vrat.

Za enimi se skriva avto, za drugima dvema pa koza.

Recimo, da izberete vrata številka 3,
voditelj igre, ki ve, kaj se nahaja za posameznimi vrati,
pa nato odpre vrata številka 1, *za katerimi se pokaže koza.*

Nato vas vpraša: 'Bi se sedaj raje odločili za vrata številka 2*?'*

Zanima me, ali se tekmovalcu splača zamenjati izbor vrat?*"*

Poudariti je potrebno, da mora gostitelj nagradne igre vsakič postopati enako. Ne more enkrat ponuditi zamenjavo (npr. takrat, ko vidi, da nastopajoči kaže na vrata za katerimi se skriva avto), drugič pa ne (npr. takrat, ko nastopajoči kaže na vrata za katerimi je koza).

Vprašanja se je prijelo ime "*problem Montyja Halla*", po imenu voditelja popularne ameriške televizijske oddaje Pogodimo se (Let's Make a Deal),

v kateri je voditelj Monty Hall goste izzival, da so sprejemali ali zavračali najrazličnejše ponudbe, ki jim jih je zastavljal.

Marilyn je bralcu v svoji rubriki odgovorila, da se nam vrata vsekakor splača zamenjati, saj se tako verjetnost, da bomo zadeli avto,
poveča za dvakrat. Tole je njen odgovor:

Seveda se splača zamenjati vrata.
Prva vrata imajo le 1/3 verjetnosti za zmago,
medtem ko imajo druga verjetnost 2/3.

Namig predlanskega asistenta Borisa Cergola

A ... dobimo avto

V ... na začetku izberemo prava vrata

1) Si ne premislimo:

$$P(A) = P(V) = 1/3$$

2) Si premislimo:

$$\begin{aligned} P(A) &= P(A/V) * P(V) + P(A/V^c) * P(V^c) \\ &= 0 * 1/3 + 1 * 2/3 \\ &= 2/3. \end{aligned}$$

Namig

Najlažje si vse skupaj predstavljate takole.

Predpostavimo, da je na voljo milijon vrat in vi izberete prva.

Nato voditelj, ki ve, kaj se nahaja za posameznimi vrati,
odpre vsa vrata razen vrat številka 777777.

V tem primeru bi zelo hitro zamenjali svoj izbor, kajne?

Se najinteligentnejša ženska na planetu moti?

Sledila je ploha kritik (več kot 10.000 pisem jeznih bralcev,
med katerimi je bilo ogromno učiteljev matematike).

Skoraj 1000 pisem je bilo podpisanih z imeni (dr. nazivi, napisana na papirju z glavo katere od ameriških univerz - www.marilynvossavant.com).

Marylin bralce zavaja, saj se verjetnost za zadetek
nikakor ne more spremeniti, če vmes zamenjamo izbor vrat.

Neki profesor matematike je bil zelo neposreden:
“Udarili ste mimo! ... Kot profesionalni matematik sem
zelo zaskrbljen nad pomanjkanjem matematičnih veščin v širši javnosti.
Prosim, da se opravičite in ste v prihodnosti bolj pazljivi.”

Drugi je Marylin celo obtožil, da je ona sama koza.

Polemika je pristala celo na naslovnici New York Timesa, v razpravo so se vključila tudi nekatera znana imena iz sveta matematike.

O odgovoru vos Savantove, da naj tekmovalec zamenja vrata, so razpravljali tako na hodnikih Cie kot v oporiščih vojaških pilotov ob Perzijskem zalivu. Analizirali so ga matematiki z MIT in računalniški programerji laboratorijev Los Alamos v Novi Mehiki.

Poleg žaljivih pisem, ki so njen odgovor kritizirala, je Marilyn vseeno prejela tudi nekaj pohval. Profesor s prestižnega MIT:

“Seveda imate prav. S kolegi v službi smo se poigrali s problemom in moram priznati, da je bila večina, med njimi sem bil tudi sam, sprva prepričana, da se motite!”

Eksperimentalna ugotovitev

Marilyn se kritik ni ustrašila - navsezadnje je objektivno izmerljivo po inteligenčnem količniku pametnejša od vseh svojih kritikov,

zato je v eni od svojih naslednjih kolumn vsem učiteljem v državi zadala nalogo, da to preprosto igrico igrajo s svojimi učenci v razredu (seveda ne s pravimi kozami in avtomobilom) in ji pošljejo svoje rezultate.

Te je nato tudi objavila in seveda so se povsem skladali z njenim nasvetom, da se v tem konkretnem primeru bistveno bolj splača spremeniti izbiro vrat.

Kdo ima prav?

Razprava o problemu Montyja Halla spada na področje,
ki mu matematiki pravijo **pogojna verjetnost**.

Najbolj preprosto rečeno je to veda, ki se ukvarja s tem, kako prilagoditi verjetnost za posamezne dogodke, ko se pojavijo novi podatki.

Bistvo zapleta, ki je izzval tako obsežno in čustveno nabito reakcijo bralcev, je v tem, da so bralci večinoma spregledali ključni podatek.

Zelo pomembno je namreč dejstvo, da voditelj igre vnaprej ve,
za katerimi vrati je avtomobil.

Ko v drugem delu odpre vrata, za katerimi se pokaže koza,
vnaprej ve, da za temi vrati ni avtomobila.

Če voditelj te informacije ne bi imel
in bi vrata odpiral povsem naključno tako kot igralec,
se verjetnost za zadetek ob spremembi vrat res ne bi povečala.

Potem bi držale ugotovitve več 1000 bralcev,
ki so poslali jezna pisma na uredništvo revije,
da Marilyn ne pozna osnov matematike.

Matematična intuicija nam namreč pravi,
da je verjetnost, da bo avto za enimi ali za drugimi vrati,
ko so dvoja še zaprta, enaka.

To je seveda res, če zraven ne bi bilo še voditelja, ki ve več kot mi.

Najlažje nejasnost pojasnimo, če analiziramo dogajanje izza kulis,
od koder ves čas vidimo, za katerimi vrati je avto in kje sta kozi.

Če tekmovalec že v prvo izbere vrata, za katerimi je avto,
bo voditelj odprl katera koli od preostalih dveh vrat
in zamenjava bo tekmovalcu v tem primeru le škodila.

Ampak to velja le za primer, če v prvo izbere vrata,
za katerimi je avto, verjetnost za to pa je 1/3.

Če pa v prvo tekmovalec izbere vrata, za katerimi je koza,
bo voditelj moral odpreti edina preostala vrata,
za katerimi se nahaja koza.

V tem primeru se bo tekmovalcu zamenjava vrat v vsakem primeru
obrestovala in bo tako z gotovostjo zadel avto.

Če v prvo tekmovalec izbere kozo, se mu vedno splača zamenjati,
če pa v prvo izbere avto, se mu zamenjava ne izplača.

Verjetnost, da v prvo izbere kozo, je 2/3,
medtem ko je verjetnost, da izbere avto, le 1/3.

Če se tekmovalec odloči za strategijo zamenjave,
je zato verjetnost, da zadane avtomobil, 2/3,
če zamenjavo zavrne, pa je verjetnost pol manjša, tj. 1/3.

Če se torej drži strategije zamenjave vrat, ko mu jo voditelj ponudi,
bo tako vedno, ko v prvo izbere kozo, ob zamenjavi vrat dobil avto,
kar ga do dobitka pripelje v $2\times$ večjem številu primerov,
kot sicer. Verjetnost za zadetek se mu tako s 33% poveča na 66%.

Če vam ni takoj jasno, se ne sekirajte preveč. Tudi mnogi matematiki
so potrebovali kar nekaj časa, da so si razjasnili ta problem.

Definicija pogojne verjetnosti

Opazujemo dogodek A ob poskusu X, ki je realizacija kompleksa pogojev K. Verjetnost dogodka A je tedaj $P(A)$.

Kompleksu pogojev K pridružimo mogoč dogodek B, tj. $P(B) > 0$. Realizacija tega kompleksa pogojev $K' = K \cap B$ je poskus X' in verjetnost dogodka A v tem poskusu je $P_B(A)$, ki se z verjetnostjo $P(A)$ ujema ali pa ne.

Pravimo, da je poskus X' poskus X s pogojem B in verjetnost $P_B(A)$ **pogojna verjetnost** dogodka A glede na dogodek B, kar zapišemo takole:

$$P_B(A) = P(A/B).$$

Pogojna verjetnost $P(A/B)$ v poskusu X' je verjetnost dogodka A v poskusu X s pogojem B.

Pogosto pogojno verjetnost pišejo tudi $P(A/B)$.

. . . Pogojna verjetnost

Denimo, da smo n–krat ponovili poskus X in da se je ob tem k_B–krat zgodil dogodek B. To pomeni, da smo v n ponovitvah poskusa X napravili k_B–krat poskus X'. Dogodek A se je zgodil ob poskusu X' le, če se je zgodil tudi B, t.j. $A \cap B$. Denimo, da se je dogodek $A \cap B$ zgodil ob ponovitvi poskusa $k_{A\cap B}$–krat. Potem je relativna frekvenca dogodka A v opravljenih ponovitvah poskusa X':

$$f_B(A) = f(A/B) = \frac{k_{A\cap B}}{k_B} = \frac{k_{A\cap B}/n}{k_B/n} = \frac{f(A \cap B)}{f(B)}$$

oziroma

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

Pogojna verjetnost P_B ima prav take lastnosti kot brezpogojna. Trojica $(B, \mathcal{D}_B, P_B)$, $\mathcal{D}_B = \{A \cap B \mid A \in \mathcal{D}\}$ je zopet verjetnostni prostor.

. . . Pogojna verjetnost

Primer: Denimo, da je v nekem naselju 900 polnoletnih prebivalcev. Zanima nas struktura prebivalcev po spolu (M – moški, Ž – ženski spol) in po zaposlenosti (Z – zaposlen(a), N – nezaposlen(a)). Podatke po obeh spremenljivkah uredimo v dvorazsežno frekvenčno porazdelitev, ki jo imenujemo tudi **kontingenčna tabela**:

spol \ *zap.*	Z	N	
M	460	40	500
Ž	240	160	400
	700	200	900

. . . Pogojna verjetnost

Poglejmo, kolikšna je verjetnost, da bo slučajno izbrana oseba moški pri pogoju, da je zaposlena.

$$P(Z) = \frac{700}{900} \quad , \quad P(M \cap Z) = \frac{460}{900}$$

$$P(M/Z) = \frac{P(M \cap Z)}{P(Z)} = \frac{460 \cdot 900}{900 \cdot 700} = \frac{460}{700}$$

ali neposredno iz kontingenčne tabele

$$P(M/Z) = \frac{460}{700}.$$

. . . Pogojna verjetnost

Iz formule za pogojno verjetnost sledi:

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

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

Torej velja:

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

Dogodka A in B sta **neodvisna**, če velja

$$P(A/B) = P(A).$$

Zato za neodvisna dogodka A in B velja $P(A \cap B) = P(A) \cdot P(B)$.

Za nezdružljiva dogodka A in B velja $P(A/B) = 0$.

. . . Pogojna verjetnost

Primer: Iz posode, v kateri imamo 8 belih in 2 rdeči krogli, $2\times$ na slepo izberemo po eno kroglo. Kolikšna je verjetnost dogodka, da je prva krogla bela (B_1) in druga rdeča (R_2).

1. Če po prvem izbiranju izvlečeno kroglo ne vrnemo v posodo (odvisnost), je:

$$P(B_1 \cap R_2) = P(B_1) \cdot P(R_2/B_1) =$$

$$= \frac{8}{10} \cdot \frac{2}{9} = 0,18.$$

2. Če po prvem izbiranju izvlečeno kroglo vrnemo v posodo (neodvisnost), je:

$$P(B_1 \cap R_2) = P(B_1) \cdot P(R_2/B_1) =$$

$$= P(B_1) \cdot P(R_2) = \frac{8}{10} \cdot \frac{2}{10} = 0,16.$$

. . . Pogojna verjetnost

Dogodka A in B sta neodvisna, če je $P(A/B) = P(A/\overline{B})$.

Nadalje velja

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

(Tudi slednje pravilo lahko posplošimo naprej.)

Dogodki A_i, $i \in I$ so **neodvisni**, če je $P(A_j) = P(A_j/\bigcap_{i=1}^{j-1} A_i)$, $j \in I$.

Za neodvisne dogodke A_i, $i \in I$ velja

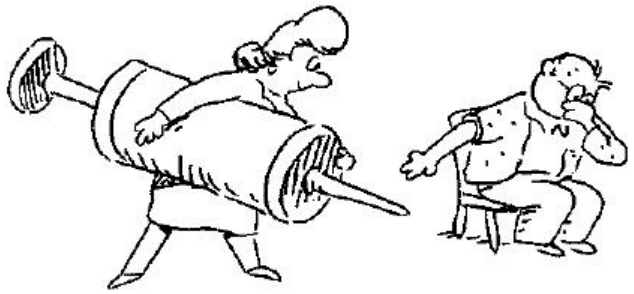
$$P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i).$$

Naloga iz pogojne verjetnosti

Redko nalezljivo bolezen dobi ena oseba na 1000.

Imamo dober, a ne popoln test za to bolezen:

če ima neka oseba to bolezen, potem test to pokaže v 99% primerih, vendar pa test napačno označi tudi 2% zdravih pacientov za bolane.



V Tvojem primeru je bil test pravkar **pozitiven**.

Kakšna je verjetnost, da si zares dobili nalezljivo bolezen?

. . . Naloga iz pogojne verjetnosti

Delamo z naslednjimi dogodki:

A: pacient je dobil nalezljivo bolezen,

B: pacientov test je bil pozitiven.

Izrazimo informacijo o učinkovitosti testov:

$P(A) = 0,001$ (en pacient na 1000 se naleze),

$P(B/A) = 0,99$ (test pravilno označi okuženega),

$P(B/\overline{A}) = 0,02$ (test napačno označi zdravega).

Zanima nas $P(A/B)$ (verjetnost, da smo se nalezli, če je test pozitiven).

Obrazec za razbitje in večstopenjski poskusi

Naj bo H_i, $i \in I$ **razbitje** gotovega dogodka: $\bigcup_{i\in I} H_i = G$,
hkrati pa naj bodo dogodki paroma nezdružljivi: $H_i \cap H_j = N, i \neq j$.

Zanima nas verjetnost dogodka A, če poznamo verjetnost $P(H_i)$, in pogojno verjetnost $P(A/H_i)$ za $i \in I$:

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

Ker so tudi dogodki $A \cap H_i$ paroma nezdružljivi, velja:

$$P(A) = \sum_{i\in I} P(A \cap H_i) = \sum_{i\in I} P(H_i)P(A/H_i).$$

Na stvar lahko pogledamo tudi kot na večstopenjski poskus:
v prvem koraku se zgodi natanko eden od dogodkov H_i,
ki ga imenujemo hipoteza
(hipoteze sestavljajo popoln sistem dogodkov).

Šele izidi na prejšnjih stopnjah določajo,
kako bo potekal poskus na naslednji stopnji.

Omejimo se na poskus z dvema stopnjama.

Naj bo A eden izmed mogočih dogodkov na drugi stopnji.
Včasih nas zanima po uspešnem izhodu tudi druge stopnje,
verjetnost tega, da se je na prvi stopnji zgodil dogodek H_i.

Odgovor dobimo iz zgornjega obrazca
in mu pravimo **Bayesov obrazec**:

$$P(H_k/A) = \frac{P(H_k) \cdot P(A/H_k)}{\sum_{i \in I} P(H_i) \cdot P(A/H_i)}.$$

Rev. T. Bayes

Leta 2001 je bila na vrsti že 300 letnica rojstva
angleškega matematika Bayesa.

Zgled. Trije lovci so hkrati ustrelili na divjega prašiča in ga ubili.
Ko so prišli do njega, so našli v njem eno samo kroglo.
Kolikšne so verjetnosti, da je vepra ubil

(a) prvi,

(b) drugi,

(b) tretji

lovec, če poznamo njihove verjetnosti, da zadanejo: $0,2$; $0,4$ in $0,6$?
Na ta način jim namreč lahko pomagamo pri pošteni delitvi plena
(kajti ne smemo pozabiti, da imajo vsi v rokah nevarno orožje).

Sestavimo popoln sistem dogodkov in uporabimo dejstvo,
da so lovci med seboj neodvisni, torej

$$P(A * B * C) = P(A) * P(B) * P(C).$$

To nam zna pomagati pri računanju verjetnosti hipotez.

	.2	.4	.6					
	prvi	drugi	tretji		P(H_i)	st.kr.	P(E/H_i)	P(E*H_i)
H1	1	1	1	,2*,4*,6	=0,048	3	0	0
H2	0	1	1	,8*,4*,6	=0,192	2	0	0
H3	1	0	1	,2*,6*,6	=0,072	2	0	0
H4	1	1	0	,2*,4*,4	=0,032	2	0	0
H5	1	0	0	,2*,6*,4	=0,048	1	1	0,048
H6	0	1	0	,8*,4*,4	=0,128	1	1	0,128
H7	0	0	1	,8*,6*,6	=0,288	1	1	0,288
H8	0	0	0	,8*,6*,4	=0,192	0	0	0
vsota					=1,000			0,464

$$P(\text{ena krogla je zadela}) = 0,048 + 0,128 + 0,288 = 0,464 = P(E).$$

Ostale verjetnosti računamo za preiskus:

$$\begin{aligned} P(\text{nobena krogla ni zadela}) &= 0,192 = P(N'), \\ P(\text{dve krogli sta zadeli}) = 0,192 + 0,072 + 0,032 &= 0,296 = P(D), \\ P(\text{tri krogle so zadele}) &= 0,048 = P(T). \end{aligned}$$

Vsota teh verjetnosti je seveda enaka 1.

Končno uporabimo Bayesov obrazec:

$$P(H_5/E) = \frac{P(H_5 * E)}{P(E)} = \frac{0,048}{0,464} = 0,103 \quad = \quad P(\text{prvi je zadel}),$$

$$P(H_6/E) = \frac{P(H_6 * E)}{P(E)} = \frac{0,128}{0,464} = 0,276 \quad = \quad P(\text{drugi je zadel}),$$

$$P(H_7/E) = \frac{P(H_7 * E)}{P(E)} = \frac{0,288}{0,464} = 0,621 \quad = \quad P(\text{tretji je zadel}).$$

Tudi vsota teh verjetnosti pa je enaka 1.

Delitev plena se opravi v razmerju 10,3 : 27,6 : 62,1 = 3 : 8 : 18

(in ne 2 : 4 : 6 oziroma 16,6 : 33,3 : 50,

kot bi kdo utegnil na hitro pomisliti).

Bonus vprašanje:

Kako bi si razdelili plen, če bi v divjim prašiču našli dve krogli?

I.4. Bernoullijevo zaporedje neodvisnih poskusov

O zaporedju neodvisnih poskusov

$$X_1, X_2, \ldots, X_n, \ldots$$

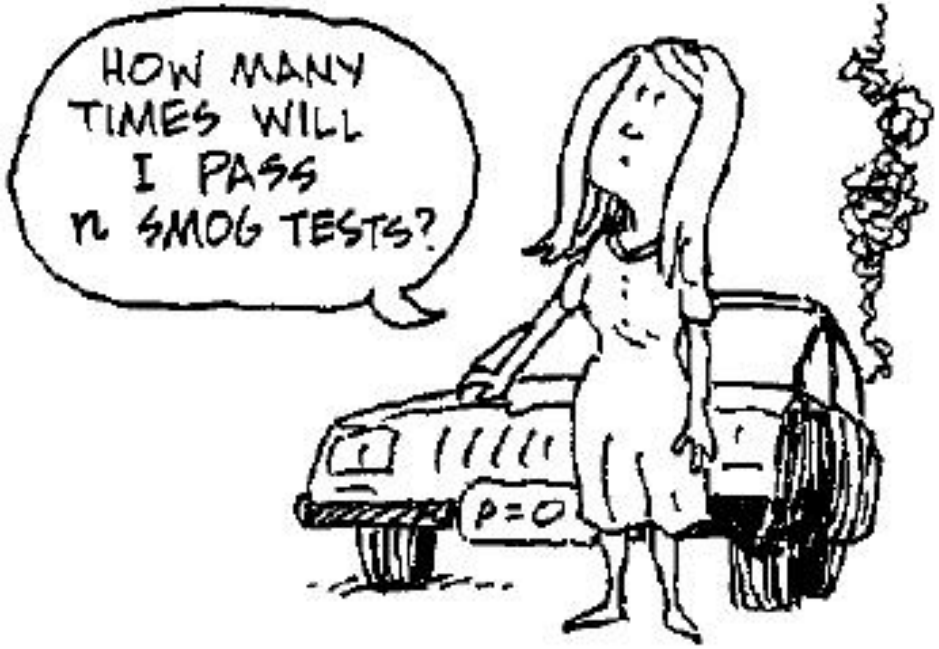
govorimo tedaj, ko so verjetnosti izidov v enem poskusu neodvisne od tega, kaj se zgodi v drugih poskusih.

Jakob Bernoulli um 1687

Zaporedje neodvisnih poskusov se imenuje **Bernoullijevo zaporedje**, *če se more zgoditi v vsakem poskusu iz zaporedja neodvisnih poskusov le dogodek A z verjetnostjo $P(A) = p$ ali dogodek $\overline{A}$ z verjetnostjo $P(\overline{A}) = 1 - P(A) = 1 - p = q$.*

Primer:

Primer Bernoullijevega zaporedja poskusov je met kocke,
kjer ob vsaki ponovitvi poskusa pade šestica (dogodek A)
z verjetnostjo $P(A) = p = 1/6$
ali ne pade šestica (dogodek $\overline{A}$)
z verjetnostjo $P(\overline{A}) = 1 - p = q = 5/6$.



. . . Bernoullijevo zaporedje neodvisnih poskusov

V Bernoullijevem zaporedju neodvisnih poskusov nas zanima, kolikšna je verjetnost, da se v n zaporednih poskusih zgodi dogodek A natanko k–krat. To se lahko zgodi na primer tako, da se najprej zgodi k–krat dogodek A in nato v preostalih $(n-k)$ poskusih zgodi nasprotni dogodek $\overline{A}$:

$$P\left(\bigcap_{i=1}^{k}(X_i = A) \cap \bigcap_{i=k+1}^{n}(X_i = \overline{A})\right) = \prod_{i=1}^{k} P(A) \cdot \prod_{i=k+1}^{n} P(\overline{A}) = p^k \cdot q^{n-k}.$$

Dogodek, da se dogodek A v n zaporednih poskusih zgodi natanko k–krat, se lahko zgodi tudi na druge načine in sicer je teh toliko, na kolikor načinov lahko izberemo k poskusov iz n poskusov. Teh je $\binom{n}{k}$. Ker so ti načini nezdružljivi med seboj, je verjetnost tega dogodka enaka

$$P_n(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Tej zvezi pravimo **Bernoullijev obrazec**.

Primer: Iz posode, v kateri imamo 8 belih in 2 rdeči krogli, na slepo izberemo po eno kroglo in po izbiranju izvlečeno kroglo vrnemo v posodo. Kolikšna je verjetnost, da v petih poskusih izberemo 3–krat belo kroglo?

Dogodek A je, da izvlečem belo kroglo. Potem je

$$p = P(A) = \frac{8}{10} = 0,8 \quad \text{in} \quad q = 1 - p = 1 - 0,8 = 0,2$$

Verjetnost, da v petih poskusih izberemo 3–krat belo kroglo, je:

$$P_5(3) = \binom{5}{3} 0,8^3 (1 - 0,8)^{5-3} = 0,205.$$

Računanje $P_n(k)$

Uporaba rekurzije: $P_n(0) = q^n$

$$P_n(k) = \frac{(n-k+1)p}{kq} P_n(k-1), \qquad \text{za } k = 1, \ldots$$

Stirlingov obrazec:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Poissonov obrazec: za majhne verjetnosti, tj. p blizu 0:

$$P_n(k) \approx \frac{(np)^k e^{-np}}{k!}.$$

Laplaceov točkovni obrazec:

$$P_n(k) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}.$$

Računanje $P_n(k)$

Program R: Vrednost $P_n(k)$ dobimo z ukazom

`dbinom(`k`,size=`n`,prob=`p`)`

```
> dbinom(50,size=1000,prob=0.05)
[1] 0.05778798
```

Izpeljava rekurzivne zveze

$$\frac{P_n(k)}{P_n(k-1)} = \frac{\binom{n}{k} p^k q^{n-k}}{\binom{n}{k-1} p^{k-1} q^{n-k+1}} =$$

$$= \frac{n!\,(k-1)!(n-k+1)!\,p}{k!(n-k)!\,n!\,q} = \frac{(n-k+1)p}{kq}$$

Torej je res

$$P_n(k) = \frac{(n-k+1)p}{kq} P_n(k-1), \qquad \text{za } k = 1, \ldots$$

Bernoullijev zakon velikih števil

IZREK 1 (J. Bernoulli, 1713) *Naj bo k frekvenca dogodka A v n neodvisnih ponovitvah danega poskusa, v katerem ima dogodek A verjetnost p. Tedaj za vsak $\varepsilon > 0$ velja*

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1.$$

Ta izrek opravičuje statistično definicijo verjetnosti.

I.5. Slučajne spremenljivke in porazdelitve

Denimo, da imamo poskus, katerega izidi so števila (npr. pri metu kocke so izidi števila pik). Se pravi, da je poskusom prirejena neka količina, ki more imeti različne vrednosti. Torej je spremenljivka. Katero od mogočih vrednosti zavzame v določeni ponovitvi poskusa, je odvisno od slučaja. Zato ji rečemo **slučajna spremenljivka**.

Da je slučajna spremenljivka znana, je potrebno vedeti

1. kakšne vrednosti more imeti (*zaloga vrednosti*) in
2. kolikšna je verjetnost vsake izmed možnih vrednosti ali intervala vrednosti.
   Predpis, ki določa te verjetnosti, imenujemo **porazdelitveni zakon**.

. . . Slučajne spremenljivke

Slučajne spremenljivke označujemo z velikimi tiskanimi črkami iz konca abecede, vrednosti spremenljivke pa z enakimi malimi črkami. Tako je npr. $(X = x_i)$ dogodek, da slučajna spremenljivka X zavzame vrednost x_i.

Porazdelitveni zakon slučajne spremenljivke X je poznan, če je mogoče za vsako realno število x določiti verjetnost

$$F(x) = P(X < x).$$

$F(x)$ imenujemo **porazdelitvena funkcija**.

Najpogosteje uporabljamo naslednji vrsti slučajnih spremenljivk:

1. **diskretna** slučajna spremenljivka, pri kateri je zaloga vrednosti neka števna (diskretna) množica
2. **zvezna** slučajna spremenljivka, ki lahko zavzame vsako realno število znotraj določenega intervala.

Lastnosti porazdelitvene funkcije

1. *Funkcija F je definirana na vsem $\mathbb{R}$ in velja $0 \leq F(x) \leq 1, x \in \mathbb{R}$.*
2. *Funkcija F je nepadajoča $x_1 < x_2 \Longrightarrow F(x_1) \leq F(x_2)$.*
3. *$F(-\infty) = 0$ in $F(\infty) = 1$.*
4. *Funkcija je v vsaki točki zvezna od leve $F(x-) = F(x)$.*
5. *Funkcija ima lahko v nekaterih točkah skok.*
   Vseh skokov je največ števno mnogo.
6. *$P(x_1 \leq X < x_2) = F(x_2) - F(x_1)$.*
7. *$P(x_1 < X < x_2) = F(x_2) - F(x_1+)$.*
8. *$P(X \geq x) = 1 - F(x)$.*
9. *$P(X = x) = F(x+) - F(x)$.*

Diskretne slučajne spremenljivke

Zaloga vrednosti diskretne slučajne spremenljivke X je števna množica $\{x_1, x_2, \ldots, x_m, \ldots\}$. Torej je lahko tudi števno neskončna, kot npr. množici naravnih ali celih števil: $\mathbb{N}$, $\mathbb{Z}$.
Dogodki

$$X = x_k \quad k = 1, 2, \ldots$$

sestavljajo popoln sistem dogodkov. Označimo verjetnost posameznega dogodka s

$$P(X = x_i) = p_i.$$

Vsota verjetnosti vseh dogodkov je enaka 1:

$$p_1 + p_2 + \cdots + p_m + \cdots = 1.$$

Verjetnostna tabela

Verjetnostna tabela prikazuje diskretno slučajno spremenljivko s tabelo tako, da so v prvi vrstici zapisane vse vrednosti x_i, pod njimi pa so pripisane pripadajoče verjetnosti:

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_m & \cdots \\ p_1 & p_2 & \cdots & p_m & \cdots \end{pmatrix}.$$

Porazdelitvena funkcija je v tem primeru

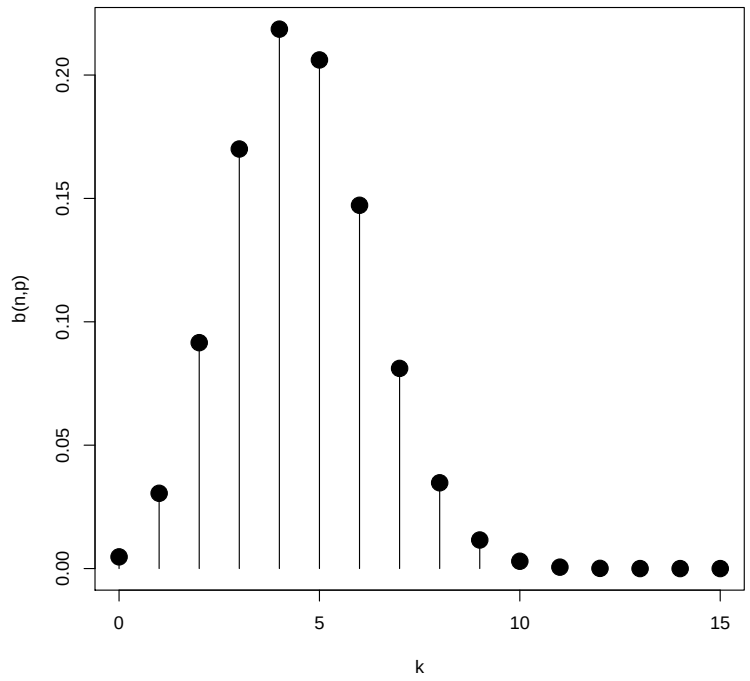
$$F(x_k) = P(X < x_k) = \sum_{i=1}^{k-1} p_i.$$

Enakomerna diskretna porazdelitev

Končna diskretna slučajna spremenljivka se porazdeljuje **enakomerno**, če so vse njene vrednosti enako verjetne. Primer take slučajne spremenljivke je število pik pri metu kocke

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Binomska porazdelitev




```
> h <- dbinom(0:15,size=15,prob=0.3)
> plot(0:15,h,type="h",xlab="k",ylab="b(n,p)")
> points(0:15,h,pch=16,cex=2)
```

Binomska porazdelitev ima zalogo vrednosti $\{0, 1, \ldots, n\}$ in verjetnosti, ki jih računamo po Bernoullijevem obrazcu:

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

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

Binomska porazdelitev je natanko določena z dvema podatkoma – parametroma: n in p. Če se slučajna spremenljivka X porazdeljuje binomsko s parametroma n in p, zapišemo:

$$X : \ B(n, p).$$

Binomska porazdelitev / Primer

Naj bo slučajna spremenljivka X določena s številom fantkov v družini s 4 otroki. Denimo, da je enako verjetno, da se v družini rodi fantek ali deklica:

$$P(F) = p = \frac{1}{2} \quad , \quad P(D) = q = \frac{1}{2}.$$

Spremenljivka X se tedaj porazdeljuje binomsko $B(4, \frac{1}{2})$ in njena verjetnostna shema je:

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1/16 & 4/16 & 6/16 & 4/16 & 1/16 \end{pmatrix}.$$

Npr.

$$P(X = 2) = P_4(2) = \binom{4}{2} \left(\frac{1}{2}\right)^2 (1 - \frac{1}{2})^{4-2} = \frac{6}{16}$$

Porazdelitev obravnavane slučajne spremenljivke je simetrična.
Pokazati se da, da je binomska porazdelitev simetrična, le če je $p = 0,5$.

Poissonova porazdelitev $P(\lambda)$

Poissonova porazdelitev izraža verjetnost števila dogodkov, ki se zgodijo v danem časovnem intervalu, če vemo, da se ti dogodki pojavijo s poznano povprečno frekvenco in neodvisno od časa, ko se je zgodil zadnji dogodek. Poissonovo porazdelitev lahko uporabimo tudi za število dogodkov v drugih intervalih, npr. razdalja, prostornina,...

Ima zalogo vrednosti $\{0, 1, 2, \ldots\}$,
njena verjetnostna funkcija pa je

$$p_k = P(\#\texttt{dogodkov} = k) = \lambda^k \frac{e^{-\lambda}}{k!},$$

kjer je $\lambda > 0$ dani parameter – in predstavlja pričakovano pogostost nekega dogodka.

$$p_{k+1} = \frac{\lambda}{k+1} p_k, \quad p_0 = e^{-\lambda}.$$

Vidimo, da zaloga vrednosti te slučajne spremenljivke ni omejena, saj je verjetnost, da se v nekem časovnem obdobju zgodi mnogo uspehov različna od nič.

To je bistvena razlika v primerjavi z binomsko porazdelitvijo, kjer število uspehov seveda ne more presegati števila Bernoullijevih poskusov n.

Primer. Posebno pomembna je ta porazdelitev v teoriji množične strežbe. Če se dogodek pojavi v povprečju 3-krat na minuto in nas zanima kolikokrat se bo zgodil v četrt ure, potem uporabimo za model Poissonovo porazdelitev z $\lambda = 15 \cdot 3 = 45$.

Naštejmo še nekaj primerov, ki jih dobro opišemo (modeliramo)
s Poissonovo porazdelitvijo:

- število dostopov do omrežnega strežnika na minuto
  (pod predpostavko homogenosti),
- število telefonskih klicev na bazni postaji na minuto,
- število mutacij v danem intervalu RNK
  po določeni količini sprejete radiacije,
- število vojakov, ki so umrli vsako leto za posledicami
  konjske brce v vsaki divizijo Pruske konjenice,
  (iz knjige Ladislausa Josephovicha Bortkiewicza, 1868–1931).

Zveza med to in binomsko porazdelitvijo

Iz analize se spomnimo naslednje zveze

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Če lahko definiramo binomsko porazdelitev tako, da je $p = \lambda/n$, potem lahko izračunamo limito verjetnosti P za velike n na naslednji način:

$$\begin{aligned}
\lim_{n\to\infty} P(X=k) &= \lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} \\
&= \lim_{n\to\infty}\frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty}\underbrace{\left[\frac{n!}{n^k\,(n-k)!}\right]}_{F}\left(\frac{\lambda^k}{k!}\right)\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}}\underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1}.
\end{aligned}$$

Za oceno faktorja F si najprej oglejmo njegov logaritem:

$$\lim_{n\to\infty} \log(F) = \log(n!) - k\log(n) - \log[(n-k)!].$$

Z uporabo Stirlingove aproksimacije je

$$\lim_{n\to\infty} \log(n!) \to n\log(n) - n,$$

izraz za $\log(F)$ pa lahko naprej poenostavimo v

$$\lim_{n\to\infty} \log(F) = [n\log(n) - n] - [k\log(n)] - [(n-k)\log(n-k) - (n-k)]$$

$$= (n-k)\log\left(\frac{n}{n-k}\right) - k = \underbrace{-\left(1-\frac{k}{n}\right)}_{\to -1}\underbrace{\log\left(1-\frac{k}{n}\right)^{n}}_{\to -k} - k = k - k = 0.$$

Torej je $\lim_{n\to\infty} F = e^0 = 1$. Od tod pa sledi, da je porazdelitev v limiti enaka

$$\frac{\lambda^k e^{-\lambda}}{k!}, \quad \text{ki ima sedaj obliko Poissonove porazdelitve.}$$

Pascalova porazdelitev $P(m, p)$

Pascalova porazdelitev ima
zalogo vrednosti $\{m, m+1, m+2, \ldots\}$,
verjetnostna funkcija pa je

$$p_k = \binom{k-1}{m-1} p^m q^{k-m},$$

kjer je $0 < p < 1$ dani parameter
– verjetnost dogodka A v posameznem poskusu.

Opisuje porazdelitev števila poskusov potrebnih,
da se dogodek A zgodi m–krat.

Za $m = 1$, porazdelitvi $G(p) = P(1, p)$ pravimo **geometrijska** porazdelitev. Opisuje porazdelitev števila poskusov potrebnih, da se dogodek A zgodi prvič.

Primer: Če mečemo kovanec toliko časa, da pade grb, in z X označimo število potrebnih metov, vključno z zadnjim, potem je slučajna spremenljivka X geometrijsko porazdeljena.

Če z X označimo število metov, vključno z zadnjim, do m-tega grba, potem dobimo **negativno binomsko** slučajno spremenljivko X : NegBin(m, p) in

$$P(X = k) = \binom{k-1}{m-1} p^m (1-p)^{k-m} \quad \text{za } k \geq m.$$

Hipergeometrijska porazdelitev $H(n; M, N)$

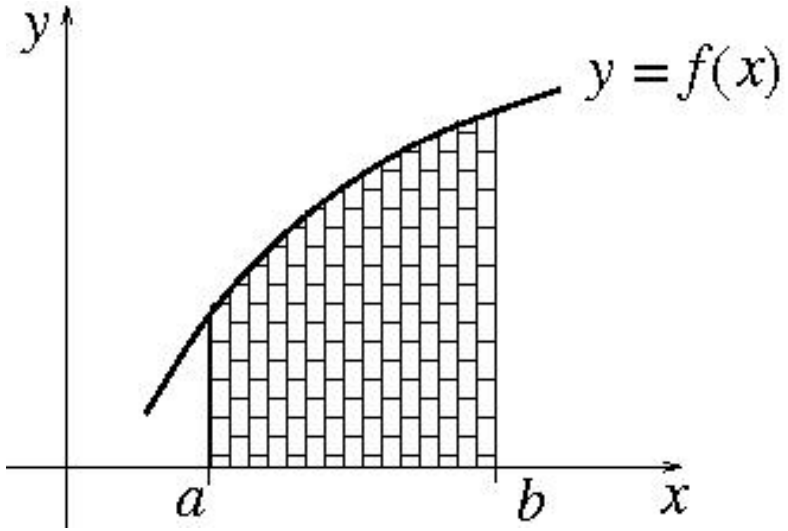
Hipergeometrijska porazdelitev ima zalogo vrednosti $\{0, 1, 2, \ldots\}$, verjetnostna funkcija pa je

$$p_k = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

kjer so $k \leq n \leq \min(M, N - M)$ dani parametri.
Opisuje verjetnost dogodka, da je med n izbranimi kroglicami natanko k belih, če je v posodi M belih in $N - M$ črnih kroglic in izbiramo n–krat brez vračanja.

Ponovitev: integrali

$$\int_a^b f(x)dx$$



Določeni integral predstavlja ploščino pod krivuljo.

Naj bo funkcija $y = f(x)$ zvezna na $[a, b]$ in nenegativna.
Ploščina lika med krivuljo $f(x) \geq 0$, in abscisno osjo
na intervalu $[a, b]$ je enaka določenemu integralu

$$\int_a^b f(x)dx.$$

Lastnosti določenega integrala:

1) $\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$

2) Če je $f(x) \leq 0 \ \forall x \in [a,b]$,

je vrednost integrala negativna.

3) Naj bo $c \in [a,b]$

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

4) Naj bo $f(x) \geq g(x),\ x \in [a,b]$,

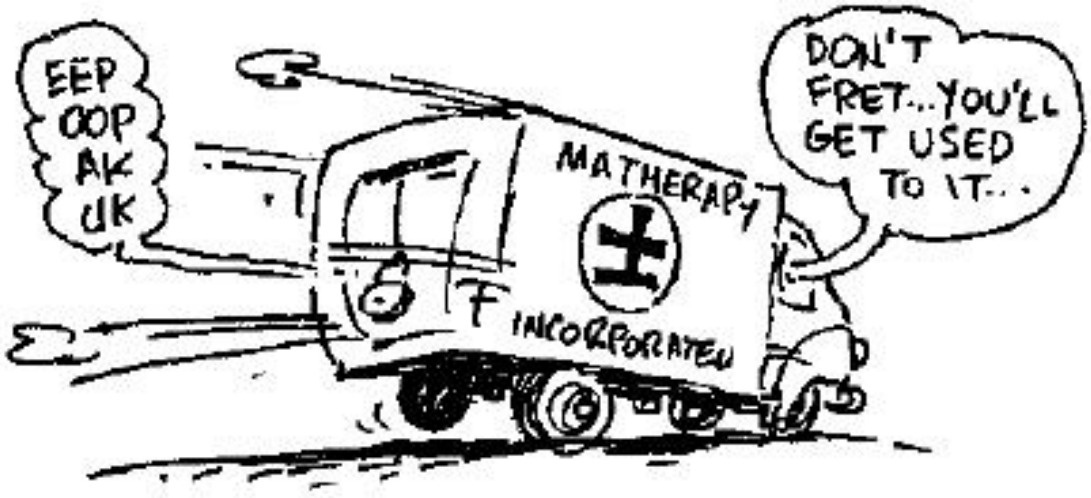
potem velja $\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx.$

Saj vas razumem!

Potem pa uporabimo še ∞ za mejo pri integriranju.

Brez preplaha!

Iščemo le celotno ploščino pod krivuljo,
od enega konca do drugega,
le da konca pravzaprav sploh ni.



Zvezne slučajne spremenljivke

Slučajna spremenljivka X je **zvezno porazdeljena**, če obstaja taka integrabilna funkcija p, imenovana **gostota verjetnosti**, da za vsak $x \in \mathbb{R}$ velja:

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(t)\,dt,$$

kjer $p(x) \geq 0$. To verjetnost si lahko predstavimo tudi grafično v koordinatnem sistemu, kjer na abscisno os nanašamo vrednosti slučajne spremenljivke, na ordinatno pa gostoto verjetnosti $p(x)$. Verjetnost je tedaj predstavljena kot ploščina pod krivuljo, ki jo določa $p(x)$. *Velja*

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{in} \quad P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} p(t)\,dt$$

ter $p(x) = F'(x)$.

Enakomerna porazdelitev zvezne slučajne spremenljivke

Verjetnostna gostota enakomerno porazdeljene zvezne slučajne spremenljivke je:

$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \leq X \leq b \\ 0 & \text{drugod.} \end{cases}$$

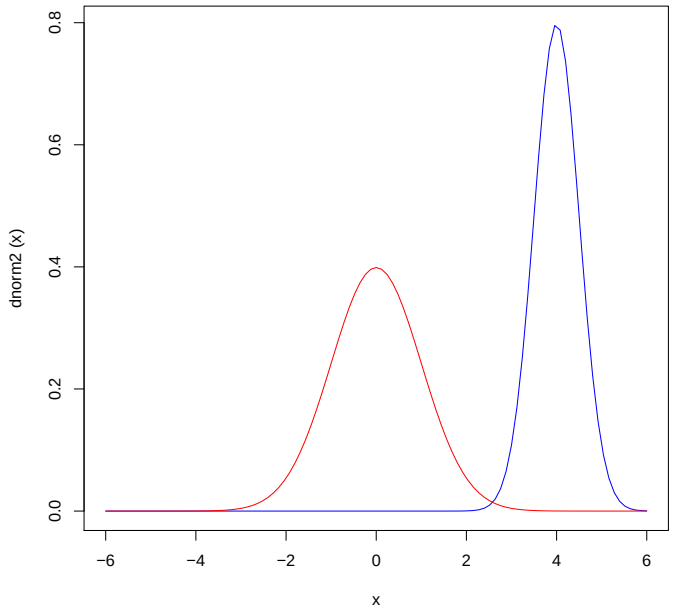
Grafično si jo predstavljamo kot pravokotnik nad intervalom (a, b) višine $\dfrac{1}{b-a}$.

Normalna ali Gaussova porazdelitev

Leta 1738 je Abraham De Moivre (1667-1754) objavil aproksimacijo binomske porazdelitve, ki je normalna krivulja.

Leta 1809 je Karl Frederic Gauss (1777-1855) raziskoval matematično ozadje planetarnih orbit, ko je izpeljal normalno porazdelitveno funkcijo.

... Normalna porazdelitev




```
> d2 <- function(x){dnorm(x,mean=4,sd=0.5)}
> curve(d2,-6,6,col="blue")
> curve(dnorm,-6,6,col="red",add=TRUE)
```

Zaloga vrednosti **normalno porazdeljene** slučajne spremenljivke so vsa realna števila, gostota verjetnosti pa je:

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

Normalna porazdelitev je natanko določena z parametroma: μ in σ.

Če se slučajna spremenljivka X porazdeljuje normalno s parametroma μ in σ zapišemo:

$$X : N(\mu, \sigma).$$

Laplaceov intervalski obrazec

Zanima nas, kolikšna je verjetnost $P_n(k_1, k_2)$, da se v Bernoullijevem zaporedju neodvisnih poskusov v n zaporednih poskusih zgodi dogodek A vsaj k_1–krat in manj kot k_2–krat. Označimo

$$x_k = \frac{k - np}{\sqrt{npq}} \quad \text{in} \quad \Delta x_k = x_{k+1} - x_k = \frac{1}{\sqrt{npq}}.$$

Tedaj je, če upoštevamo Laplaceov točkovni obrazec,

$$P_n(k_1, k_2) = \sum_{k=k_1}^{k_2-1} P_n(k) = \frac{1}{\sqrt{2\pi}} \sum_{k=k_1}^{k_2-1} e^{-\frac{1}{2}x_k^2} \Delta x_k.$$

Za (zelo) velike n lahko vsoto zamenjamo z integralom

$$P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi}} \int_{x_{k_1}}^{x_{k_2}} e^{-\frac{1}{2}x^2} dx.$$

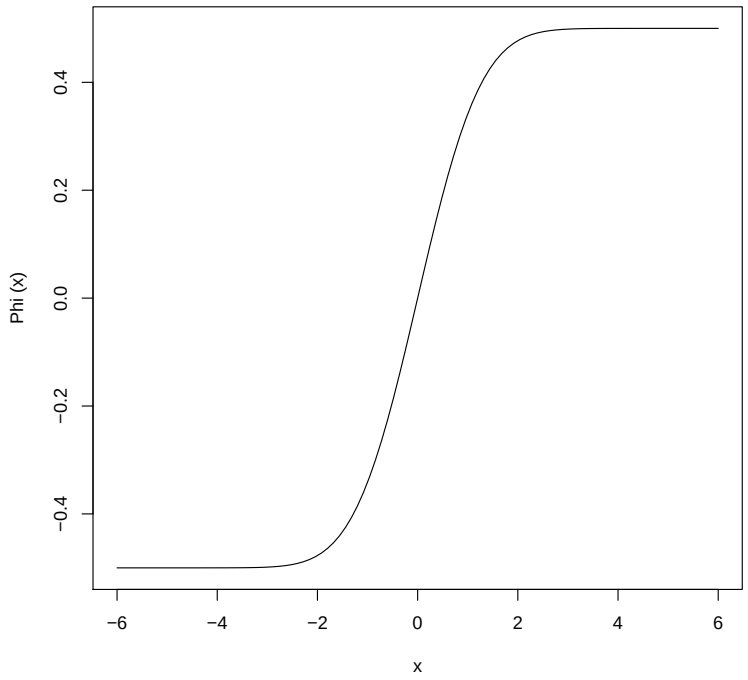
Funkcija napake $\Phi(x)$

Funkcija napake imenujemo funkcijo

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2}\, dt.$$

Funkcija napake je liha, zvezno odvedljiva, strogo naraščajoča funkcija. $\Phi(-\infty) = -\frac{1}{2}$, $\Phi(0) = 0$, $\Phi(\infty) = \frac{1}{2}$ in $P_n(k_1, k_2) \approx \Phi(x_{k_2}) - \Phi(x_{k_1})$.

Vrednosti funkcije napake najdemo v tabelah ali pa je vgrajena v statističnih programih.




```
> Phi <- function(x){pnorm(x)-0.5}
> curve(Phi,-6.6)
```

```
> x2 <- (50 - 1000*0.05)/sqrt(1000*0.05*0.95)
> x1 <- (0 - 1000*0.05)/sqrt(1000*0.05*0.95)
> pnorm(x2)-pnorm(x1)
[1] 0.5
```

. . . Normalna porazdelitev

$$
\begin{aligned}
F(x) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}\, dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{1}{2}s^2}\, ds \\
&= \frac{1}{2} + \Phi(\frac{x-\mu}{\sigma}).
\end{aligned}
$$

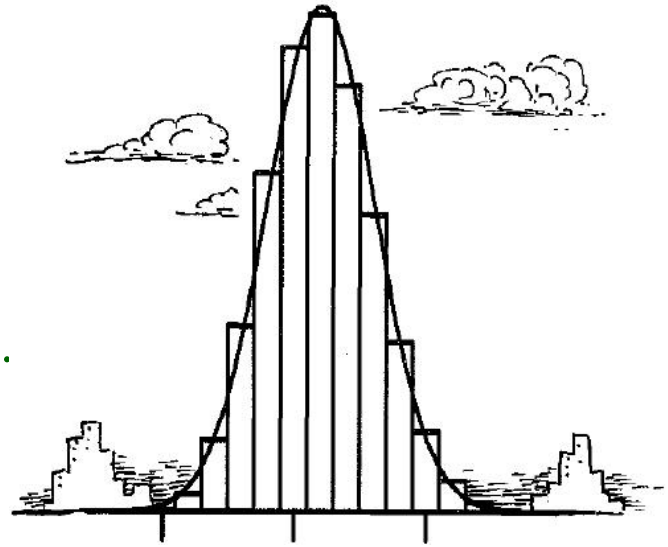
$$P(x_1 \leq X < x_2) = \Phi\left(\frac{x_2-\mu}{\sigma}\right) - \Phi\left(\frac{x_1-\mu}{\sigma}\right)$$

Porazdelitev $N(0,1)$ je **standardizirana normalna porazdelitev**.

Spremenljivko $X: \ N(\mu, \sigma)$ *pretvorimo z*

$$z = \frac{x - \mu}{\sigma}$$

v standardizirano spremenljivko $Z: \ N(0,1)$.



Iz Laplaceovega obrazca izhaja

$$B(n,p) \approx N(np, \sqrt{npq}).$$

Porazdelitev Poissonovega toka, eksponentna

Gostota **eksponentne porazdelitve** je enaka

$$p(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

porazdelitvena funkcija pa

$$F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}.$$

Porazdelitev gama

Naj bosta $b, c > 0$. Tedaj ima **porazdelitev Gama** $\Gamma(b, c)$ gostoto:

$$p(x) = \frac{c^b}{\Gamma(b)} x^{b-1} e^{-cx}, \ 0 < x$$

in $p(x) = 0$ za $x \leq 0$.

Funkcijo Gama lahko definiramo z določenim integralom za $\Re[z] > 0$ (Eulerjeva integralna forma)

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt = 2 \int_0^\infty e^{-t^2} t^{2z-1} \, dt,$$

Torej je $\Gamma(1) = 1$ in

$$\Gamma(z) = \int_0^1 \left[\ln \frac{1}{t}\right]^{z-1} dt.$$

(Je povsod analitična z izjemo $z = 0, -1, -2, \ldots$ in nima ničel.

Glej npr. `http://en.wikipedia.org/wiki/Gamma_function` in `http://en.wikipedia.org/wiki/Gamma_distribution`.)

Integracija po delih (po realnem argumentu) nam da
(za $v = t^n$ in $du = e^{-t}dx$ velja $dv = nt^{n-1}$ in $u = e^{-t}$):

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$$

$$= \left[-t^{x-1}e^{-t}\right]_0^\infty + \int_0^\infty (x-1)t^{x-2}e^{-t}\,dt$$

$$= (x-1)\int_0^\infty t^{x-2}e^{-t}\,dt = (x-1)\Gamma(x-1).$$

Za naravno število x ($n = 1, 2, 3, ...$), dobimo

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)...1 = (n-1)!$$

torej se Γ funkcija zreducira v 'faktorijel'.

Porazdelitev hi-kvadrat

Porazdelitev **hi-kvadrat** je poseben primer porazdelitve Gama:

$$\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2})$$

($n \in \mathbb{N}$ je število prostostnih stopenj)
in ima gostoto

$$p(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad \text{kjer je} x > 0.$$

Ernst Abbe
(1840-1905)

Leta 1863 jo je prvi izpeljal nemški fizik Ernst Abbe, ko je preučeval porazdelitev vsote kvadratov napak.

Leta 1878 je Ludwig Boltzmann izpeljal hi-kvadrat porazdelitev z dvema in tremi prostostnimi stopnjami, ko je študiral kinetično energijo molekul.

Karl Pearson (1875-1937) je demonstriral uporabnost hi-kvadrat porazdelitve statistikom.

Cauchyeva porazdelitev

z gostoto

$$p(x) = \frac{a}{\pi} \frac{1}{1 + a^2(x-b)^2}, \quad -\infty < x < \infty,\ a > 0$$

ima porazdelitveno funkcijo

$$F(x) = \frac{a}{\pi} \int_{-\infty}^{x} \frac{1}{1 + a^2(x-b)^2} \mathrm{d}x = \frac{1}{\pi} \operatorname{arc\,tg}(a(x-b)) + \frac{1}{2}$$

Porazdelitve v R-ju

V R-ju so za delo s pomembnejšimi porazdelitvami na voljo funkcije:

`d`*ime* – gostota porazdelitve *ime* $p_{ime}(x)$

`p`*ime* – porazdelitvena funkcija *ime* $F_{ime}(q)$

`q`*ime* – obratna funkcija: $q = F_{ime}(p)$

`r`*ime* – slučajno zaporedje iz dane porazdelitve

Za *ime* lahko postavimo: `unif` – zvezna enakomerna, `binom` – binomska, `norm` – normalna, `exp` – eksponentna, `lnorm` – logaritmičnonormalna, `chisq` – porazdelitev χ^2, ...

Opis posamezne funkcije in njenih parametrov dobimo z ukazom `help`. Na primer `help(rnorm)`.

I.6. Slučajni vektorji in neodvisnost slučajnih spremenljivk

Slučajni vektor je n-terica slučajnih spremenljivk $X = (X_1, \ldots, X_n)$.
Opišemo ga s porazdelitveno funkcijo ($x_i \in \mathbb{R}$)

$$F(x_1, \ldots, x_n) = P(X_1 < x_1, \ldots, X_n < x_n),$$

pri čemer slednja oznaka pomeni $P\big(\{X_1 < x_1\} \cap \ldots \cap \{X_n < x_n\}\big)$,
in za katero velja: $0 \leq F(x_1, \ldots, x_n) \leq 1$

Funkcija F je za vsako spremenljivko naraščajoča in od leve zvezna.

$F(-\infty, \ldots, -\infty) = 0$ in $F(\infty, \ldots, \infty) = 1$.

Funkciji $F_i(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty)$ pravimo
robna porazdelitvena funkcija spremenljivke X_i.

Slučajni vektorji – primer

Naj bo

$$A(x,y) = \{(u,v) \in \mathbb{R}^2 : u < x \wedge v < y\}$$

(levi spodnji kvadrant glede na (x,y)).

Naj porazdelitvena funkcija opisuje verjetnost, da je slučajna točka (X,Y) v množici $A(x,y)$

$$F(x,y) = P(X < x, Y < y) = P((X,Y) \in A(x,y)).$$

Tedaj je verjetnost, da je slučajna točka (X,Y) v pravokotniku $[a,b) \times [c,d)$ enaka

$$P((X,Y) \in [a,b) \times [c,d)) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

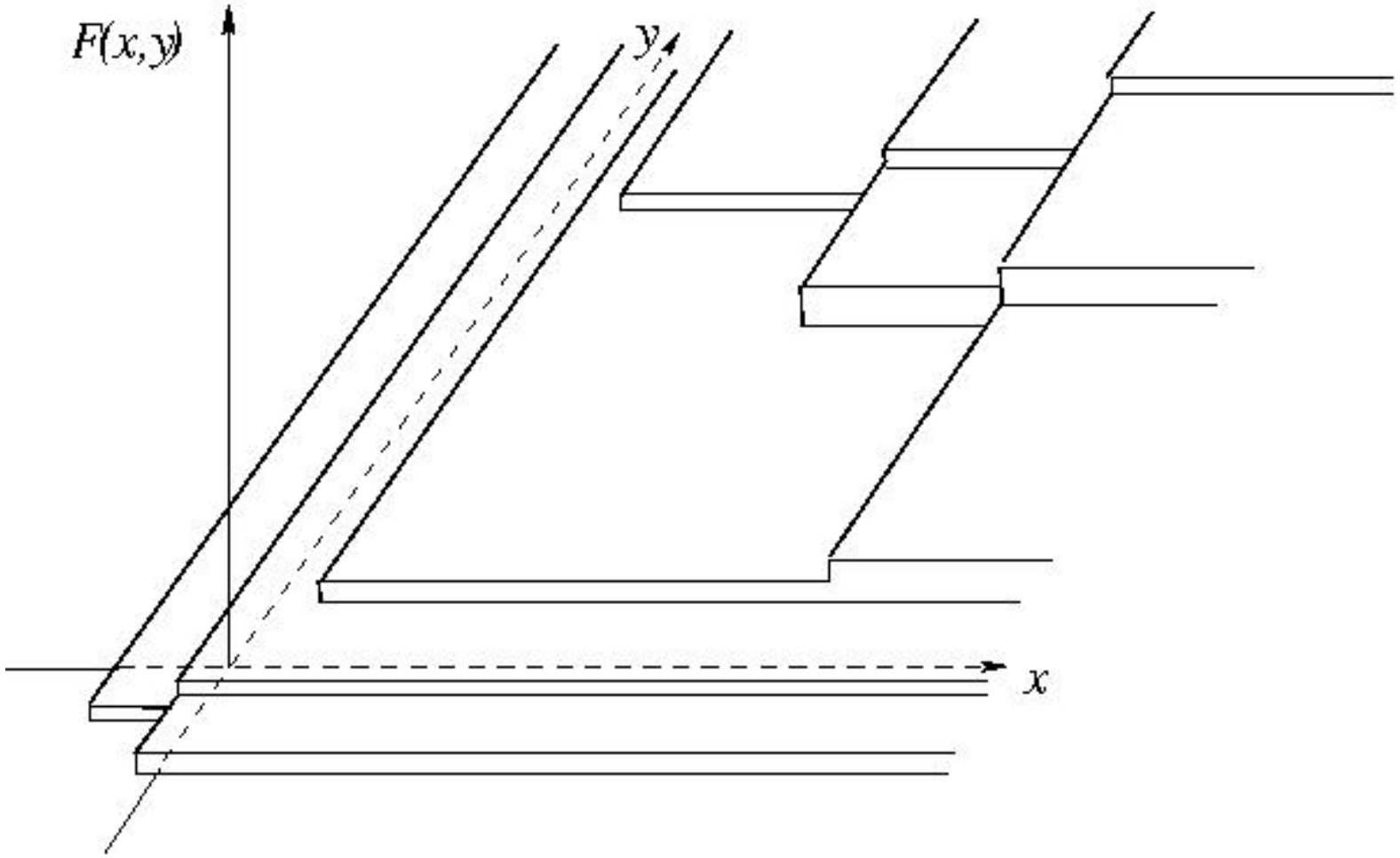
Diskretne večrazsežne porazdelitve

Zaloga vrednosti je kvečjemu števna množica. Opišemo jo z **verjetnostno funkcijo** $p_{k_1,\ldots,k_n} = P(X_1 = x_{k_1}, \ldots, X_n = x_{k_n})$.

Za $n = 2$, $X : \{x_1, x_2, \ldots, x_k\}$, $Y : \{y_1, \ldots, y_m\}$ in $P(X = x_i, Y = y_j)$, sestavimo **verjetnostno tabelo**:

$X \setminus Y$	y_1	y_2	$\ldots$	y_m	X
x_1	p_{11}	p_{12}	$\ldots$	p_{1m}	p_1
x_2	p_{21}	p_{22}	$\ldots$	p_{2m}	p_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	p_{k1}	p_{k2}	$\ldots$	p_{km}	p_k
Y	q_1	q_2	$\ldots$	q_m	1

$p_i = P(X = x_i) = \sum_{j=1}^{m} p_{ij}$ in $q_j = P(Y = y_j) = \sum_{i=1}^{k} p_{ij}$



Porazdelitvena funkcija $F(x, y)$, v primeru, ko sta spremenljivki X in Y diskretni.

Diskretne večrazsežne porazdelitve – polinomska

Polinomska porazdelitev $P(n; p_1, p_2, \ldots, p_r)$, $\sum p_i = 1$, $\sum k_i = n$ je določena s predpisom

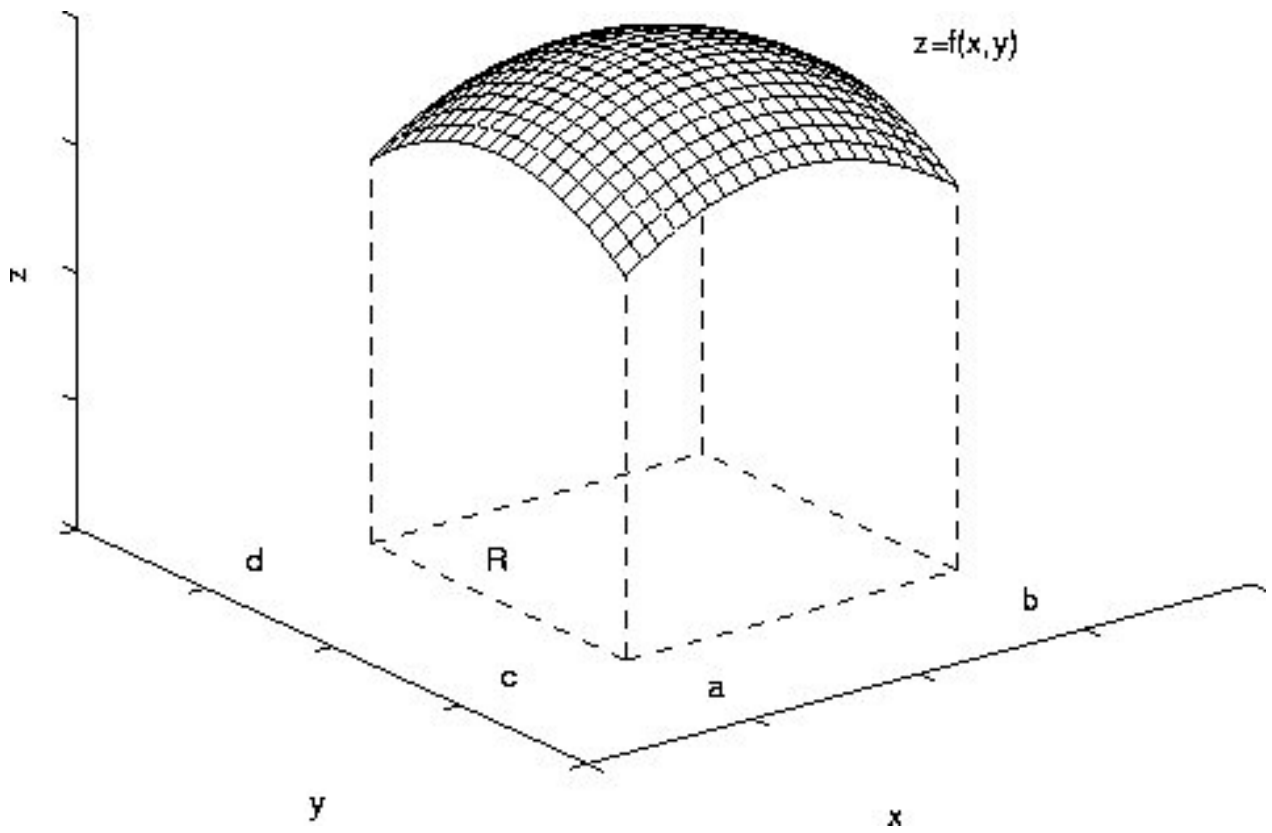
$$P(X_1 = k_1, \ldots, X_r = k_r) = \frac{n!}{k_1! \cdots k_r!} p_1^{k_1} \cdots p_r^{k_r}.$$

Keficient šteje permutacije s ponavljanjem.

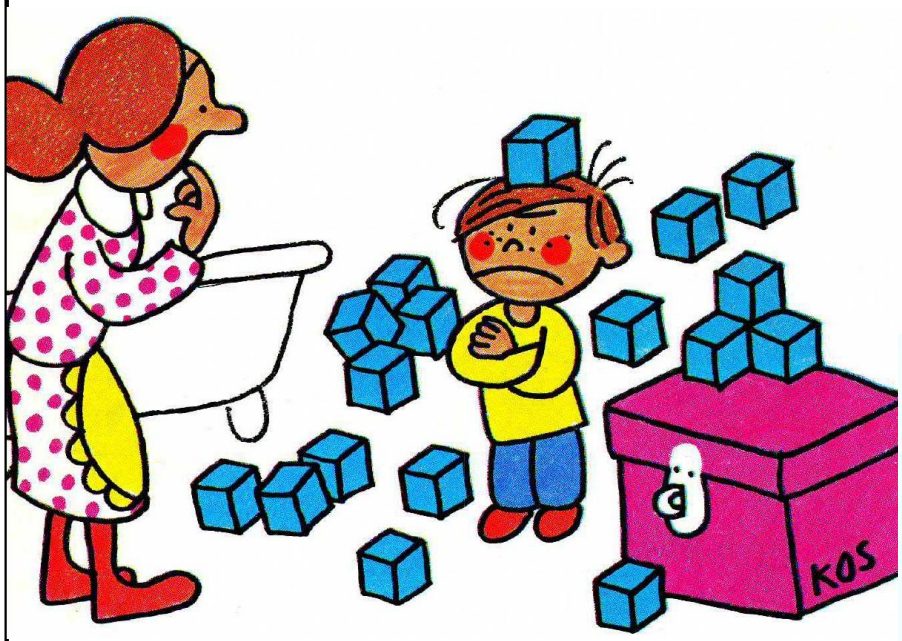
`http://en.wikipedia.org/wiki/Multinomial_distribution`

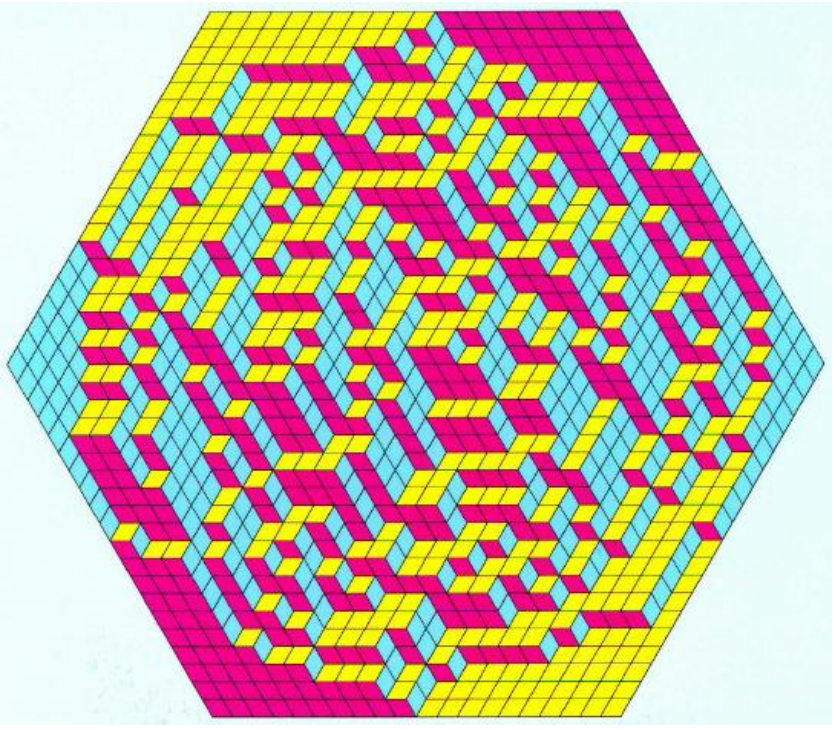
Za $r = 2$ dobimo binomsko porazdelitev, tj. $B(n, p) = P(n; p, q)$.

Ponovitev: Dvojni integral



predstavlja prostornino pod ploskvijo.


KOS



... dvojni integral

Naj bo funkcija $z = f(x, y) \geq 0$ zvezna na nekem območju R v ravnini $\mathbb{R}^2$ (npr. kar $[a, b] \times [c, d]$).

Ploščina telesa med ploskvijo podano z $z = f(x, y)$, in ravnino $z = 0$ je enaka dvojnemu integralu

$$\iint_R f(x, y)\, dxdy,$$

ki ga izračunamo s pomočjo dvakratnega integrala

$$\int_c^d \left(\int_a^b f(x, y)\, dx \right) dy = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx.$$

Lastnosti dvojnega integrala

1) Če je $f(x,y) \leq 0 \ \ \forall (x,y) \in R$, je vrednost dvojnega integrala negativna.

2) Naj bo območje $R = R_1 \cup R_2$, kjer je $R_1 \cap R_2 = \emptyset$. Potem velja

$$\iint_R f(x,y)\,dxdy = \iint_{R_1} f(x,y)\,dxdy + \iint_{R_2} f(x,y)\,dxdy.$$

3) Naj bo $f(x,y) \leq g(x,y)$, za vse točke $(x,y) \in R$, potem velja

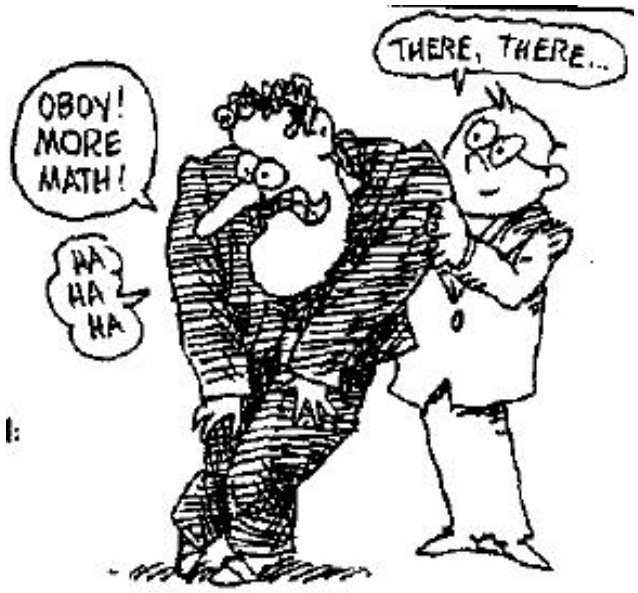
$$\iint_R f(x,y)\,dxdy \leq \iint_R g(x,y)\,dxdy.$$

Več o dvojnih integralih najdete npr. na:

`http://www.math.oregonstate.edu/home/programs/undergrad/CalculusQuestStudyGuides/vcalc/255doub/255doub.html`

Računanje dvojnih integralov na pravokotnem območju se prevede na dva običajna (enkratna) integrala.

Kot bomo videli kasneje na primerih, pa je težje izračunati dvojni integral na območju, ki ni pravokotno, ampak je omejeno s poljubnimi krivuljami.



Zvezne večrazsežne porazdelitve

Slučajni vektor $X = (X_1, X_2, \ldots, X_n)$ je **zvezno porazdeljen**, če obstaja integrabilna funkcija (**gostota verjetnosti**) $p(x_1, x_2, \ldots, x_n) \geq 0$
z lastnostjo

$$F(x_1, x_2, x_3, \ldots, x_n) =$$

$$= \int_{-\infty}^{x_1} \left(\int_{-\infty}^{x_2} \left(\ldots \left(\int_{-\infty}^{x_n} p(t_1, t_2, \ldots, t_n)\, dt_n \right) \ldots \right) dt_2 \right) dt_1$$

in

$$F(\infty, \infty, \infty, \ldots, \infty) = 1.$$

Zvezne dvorazsežne porazdelitve

$$F(x,y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} p(u,v)\, dv \right) du,$$

$$P((X,Y) \in [a,b) \times [c,d)) = \int_{a}^{b} \left(\int_{c}^{d} p(u,v)\, dv \right) du.$$

Kjer je p zvezna je

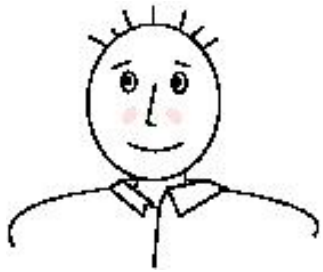
$$\frac{\partial F}{\partial x} = \int_{-\infty}^{y} p(x,v)\, dv \quad \text{in} \quad \frac{\partial^2 F}{\partial x \partial y} = p(x,y).$$

Robni verjetnostni gostoti sta

$$p_X(x) = F_X'(x) = \int_{-\infty}^{\infty} p(x,y)\, dy,$$

$$p_Y(y) = F_Y'(y) = \int_{-\infty}^{\infty} p(x,y)\, dx.$$

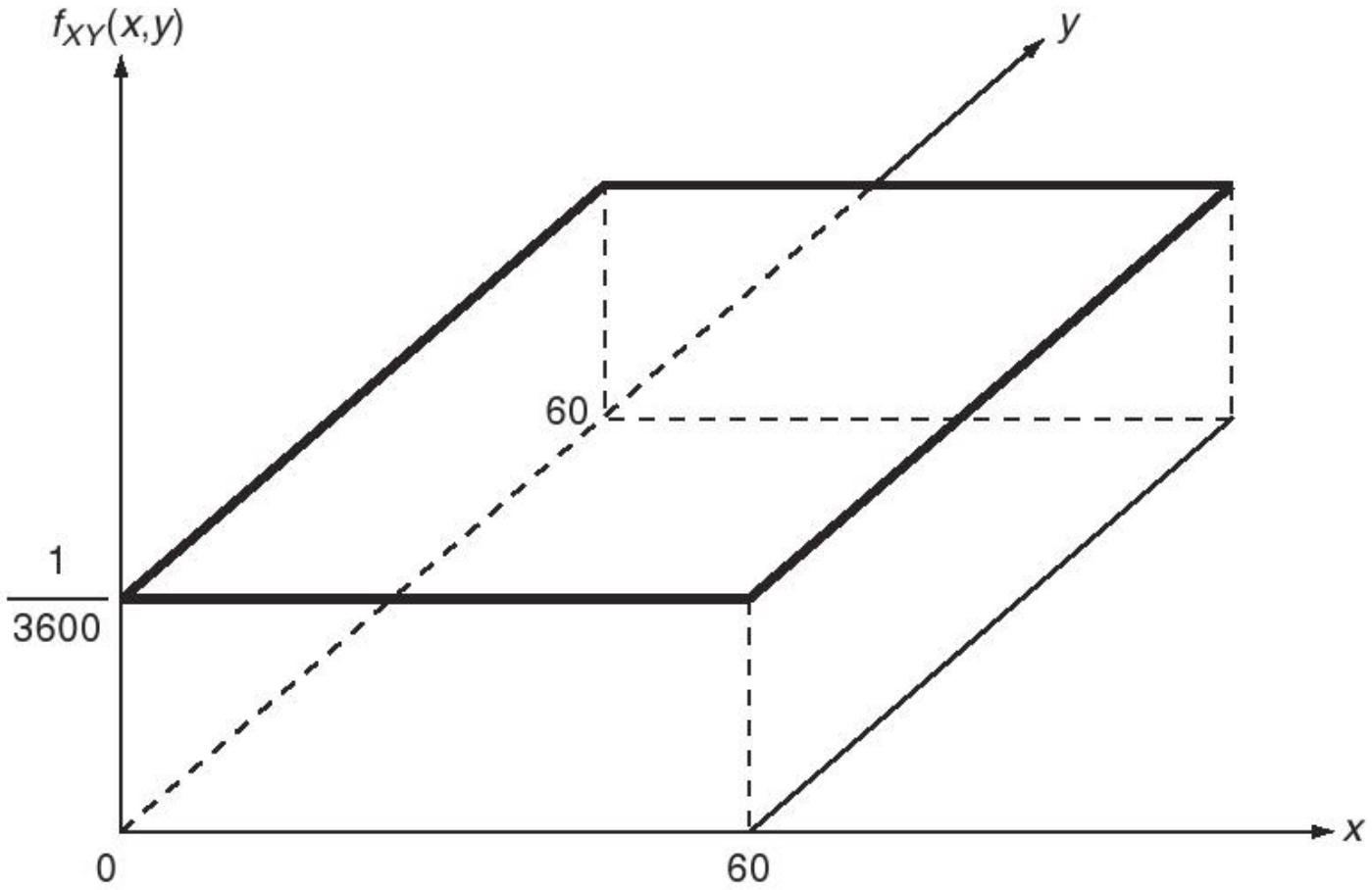
Naloga



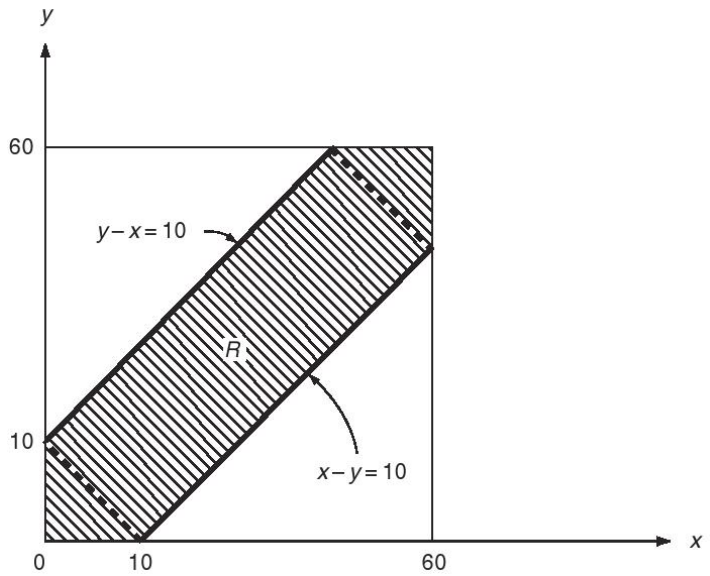
Dekle in fant se želita srečati na določenem mestu med 9-o in 10-o uro, pri čemer noben od njiju ne bo čakal drugega dlje od 10-ih minut.

Če je vsak čas med 9-o in 10-o za vsakega od njiju enako verjeten, in sta njuna časa prihodov neodvisna, poišči verjetnost, da se bosta srečala.

Naj bo čas prihoda fanta X minut po 9-i, pravtako pa naj bo čas prihoda dekleta Y minut po 9-i.


$f_{XY}(x,y)$
y
60
1
3600
0
60
x

Ploskev, ki jo določa gostota porazdelitve, je ravnina,
ker pa je prostornina pod njo enaka 1,
je oddaljena od ravnine $z = 0$ za 1/3600.



Prostornina, ki jo iščemo,
se nahaja nad področjem R,
ki je določeno z $|X - Y| \leq 10$,
torej je verjetnost srečanja enaka:

$$P(|X - Y| \leq 10) = \frac{(2 \times 5 \times 10 + 10\sqrt{2} \cdot 50\sqrt{2})}{3600} = \frac{11}{36}.$$

Pri bolj zapletenih gostotah verjetnosti, moramo dejansko izračunati integral

$$F(x, y) = \iint_R p(x, y)\, dydx.$$

Za vajo izračunajmo obe robni verjetnostni gostoti. Očitno velja:

$$F(x,y) = 0 \text{ za } (x,y) < (0,0) \quad \text{in} \quad F(x,y) = 1 \text{ za } (x,y) > (60,60).$$

Sedaj pa za $(0,0) \leq (x,y) \leq (60,60)$ velja

$$F(x,y) = \int_0^y \int_0^x \left(\frac{1}{3600}\right) dydx = \frac{xy}{3600}.$$

in

$$p_X(x) = F'_X(x) = \int_0^{60} \left(\frac{1}{3600}\right) dy = \frac{1}{60} \quad \text{za} \quad 0 \leq y \leq 60,$$

$$p_Y(y) = F'_Y(y) = \int_0^{60} \left(\frac{1}{3600}\right) dx = \frac{1}{60} \quad \text{za} \quad 0 \leq x \leq 60,$$

za vse ostale x in y pa je $p_X(x) = 0$ ter $p_Y(x) = 0$, torej sta X in Y obe enakomerno porazdeljeni slučajni spremenljivki na intervalu $[0,60]$.

Večrazsežna normalna porazdelitev

V dveh razsežnostih $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ ima gostoto

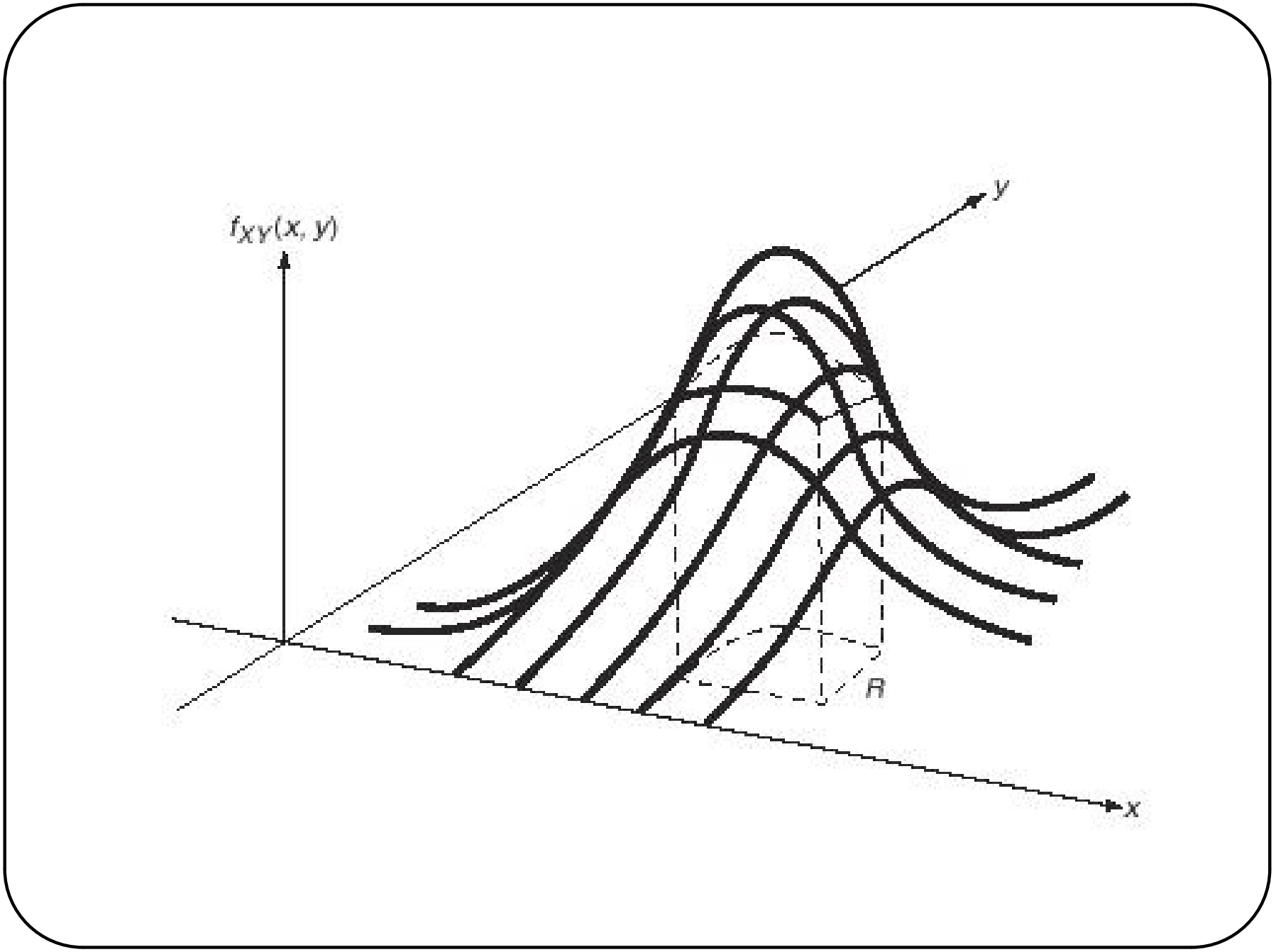
$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)}.$$

V splošnem pa jo zapišemo v matrični obliki

$$p(\boldsymbol{x}) = \sqrt{\frac{\det A}{(2\pi)^n}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T A(\boldsymbol{x}-\boldsymbol{\mu})},$$

kjer je A simetrična pozitivno definitna matrika.

Vse robne porazdelitve so normalne.


$f_{XY}(x, y)$
y
R
x

Naloga

Pri študiju upora Y strukturnega elementa in sile X, ki deluje nanj, smatramo za slučajni spremenljivki. Verjetnost napake n_f je definirana z $P(Y \leq X)$. Predpostavimo, da je

$$p(x,y) = abe^{-(ax+by)} \quad \text{za} \quad (x,y) > 0$$

in $p(x,y) = 0$ sicer, pri čemer sta a in b poznani pozitivni števili. Želimo izračunati n_f, tj.

$$F(x,y) = \iint_R p(x,y)\, dydx,$$

kjer je območje R določeno s pogojem $Y \leq X$. Ker slučajni spremenljivki X in Y zavzameta samo pozitivne vrednosti, velja

$$n_f = \int_0^\infty \int_y^\infty a\, b\, e^{-(ax+by)}\, dx\, dy = \int_0^\infty \int_0^x a\, b\, e^{-(ax+by)}\, dy\, dx.$$

Tu izračunajmo prvi integral, ki smo ga pričeli računati že na prejšnjih predavanjih (bodite pozorni na to, da so sedaj meje popravljene).
Upoštevamo $a\,dx = d(ax) = -d(-ax - by)$:

$$\int_0^\infty \int_y^\infty a\,b\,e^{-(ax+by)}\,dx\,dy = -b\int_0^\infty \left(\int_y^\infty e^{-(ax+by)}\,d(-ax-by)\right) dy$$

$$= -b\int_0^\infty \left(e^{-(ax+by)}\Big|_{x=y}^{\infty}\right) dy = b\int_0^\infty e^{-y(a+b)}\,dy$$

$$= \frac{-b}{a+b}\int_0^\infty e^{-y(a+b)}\,d\big(-y(a+b)\big) = \frac{-b}{a+b}\left(e^{-y(a+b)}\Big|_{y=0}^{\infty}\right) = \frac{b}{a+b}.$$

Vaša domača naloga pa je, da za vajo izračunate drugega (rok 12. nov. 09).

Neodvisnost slučajnih spremenljivk

Podobno kot pri dogodkih:

Slučajne spremenljivke $X_1, X_2, X_3, \ldots, X_n$ so med seboj **neodvisne**, če za poljubne vrednosti $x_1, x_2, x_3, \ldots, x_n \in \mathbb{R}$ velja

$$F(x_1, x_2, x_3, \ldots, x_n) = F_1(x_1) \cdot F_2(x_2) \cdot F_3(x_3) \cdots F_n(x_n),$$

kjer je F porazdelitvena funkcija vektorja, F_i pa so porazdelitvene funkcije njegovih komponent.

Če sta

$$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix} \text{ in } Y : \begin{pmatrix} y_1 & y_2 & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}$$

diskretni slučajni spremenljivki in p_{ij} verjetnostna funkcija slučajnega vektorja (X, Y), potem sta X in Y neodvisni natanko takrat, ko je $p_{ij} = p_i q_j$ za vsak par i, j.

. . . Neodvisnost slučajnih spremenljivk

Če sta X in Y zvezno porazdeljeni slučajni spremenljivki z gostotama p_X in p_Y ter je p gostota zvezno porazdeljenega slučajnega vektorja (X, Y), potem sta X in Y neodvisni natanko takrat, ko za vsak par x, y velja $p(x, y) = p_X(x) \cdot p_Y(y)$.

Primer: Naj bo dvorazsežni slučajni vektor (X, Y) z normalno porazdelitvijo $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$. Če je $\rho = 0$ je

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}((\frac{x-\mu_x}{\sigma_x})^2+(\frac{y-\mu_y}{\sigma_y})^2)} = p_X(x) \cdot p_Y(y).$$

Torej sta komponenti X in Y neodvisni.

. . . Neodvisnost slučajnih spremenljivk

Zvezno porazdeljeni slučajni spremenljivki X in Y sta neodvisni natanko takrat, ko lahko gostoto verjetnosti slučajnega vektorja (X, Y) zapišemo v obliki $p(x, y) = f(x) \cdot g(y)$.

Naj bosta zvezno porazdeljeni slučajni spremenljivki X in Y tudi neodvisni ter A in B poljubni (Borelovi) podmnožici v $\mathbb{R}$. Potem sta neodvisna tudi dogodka $X \in A$ in $Y \in B$.

Trditev velja tudi za diskretni slučajni spremenljivki X in Y.

Pogosto pokažemo odvisnost spremenljivk X in Y tako, da najdemo množici A in B, za kateri je

$$P(X \in A, Y \in B) \neq P(X \in A) \cdot P(Y \in B).$$

I.7. Funkcije slučajnih spremenljivk/vektorjev in pogojne porazdelitve

Funkcije slučajnih spremenljivk

Naj bo $X : G \to \mathbb{R}$ slučajna spremenljivka in $f : \mathbb{R} \to \mathbb{R}$ neka realna funkcija. Tedaj je njun kompozitum $Y = f \circ X$ določen s predpisom $Y(e) = f(X(e))$, za vsak $e \in G$, določa novo preslikavo $Y : G \to \mathbb{R}$.

Kdaj je tudi Y slučajna spremenljivka na $(G, \mathcal{D}, P)$?

V ta namen mora biti za vsak $y \in \mathbb{R}$ množica

$$(Y < y) = \{e \in G : Y(e) < y\} = \{e \in G : X(e) \in f^{-1}(-\infty, y)\}$$

dogodek – torej v $\mathcal{D}$.

Če je to res, imenujemo Y **funkcija slučajne spremenljivke** X in jo zapišemo kar $Y = f(X)$. Njena porazdelitvena funkcija je

$$F_Y(y) = P(Y < y).$$

Borelove množice

Vprašanje: kakšna mora biti množica A,
da je množica

$$X^{-1}(A) = \{e \in G : X(e) \in A\}$$

v $\mathcal{D}$?

Zadoščajo množice A, ki so ali intervali,
ali števne unije intervalov, ali števni preseki
števnih unij intervalov – **Borelove množice**.

Emile Borel

Kdaj je $f^{-1}(-\infty, y)$ Borelova množica?
Vsekakor je to res, ko je f zvezna funkcija.
V nadaljevanju nas bodo zanimali samo taki primeri.

Primer: zvezne strogo naraščajoče funkcije

Naj bo $f : \mathbb{R} \to \mathbb{R}$ zvezna in strogo naraščajoča funkcija.
Tedaj je taka tudi funkcija f^{-1} in velja

$$\begin{aligned} f^{-1}(-\infty, y) &= \{x \in \mathbb{R} : f(x) < y\} = \{x \in \mathbb{R} : x < f^{-1}(y)\} \\ &= (-\infty, f^{-1}(y)) \end{aligned}$$

in potemtakem tudi $F_Y = F_X \circ f^{-1}$, o čemer se prepričamo takole

$$F_Y(y) = P(Y < y) = P(f(X) < y) = P(X < f^{-1}(y)) = F_X(f^{-1}(y))$$

Če je X porazdeljena zvezno z gostoto $p(x)$, je $F_Y(y) = \int_{-\infty}^{f^{-1}(y)} p(x)\,dx$
in, če je f odvedljiva, še $p_Y(y) = p(f^{-1}(y))f^{-1}(y)'$.

Če funkcija ni monotona, jo razdelimo na intervale monotonosti.

Primer: kvadrat normalno porazdeljene spremenljivke

Naj bo $X : N(0,1)$ in $Y = X^2$.

Tedaj je $F_Y(y) = P(Y < y) = P(X^2 < y) = 0$ za $y \leq 0$; in za $y > 0$

$$F_Y(y) = P(|X| < \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

in ker/če je $p_X(x)$ soda funkcija

$$p_Y(y) = p_X(\sqrt{y})\frac{1}{2\sqrt{y}} + p_X(-\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{y}}p_X(\sqrt{y})$$

Vstavimo še standardizirano normalno porazdelitev

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}} & y > 0 \\ 0 & y \leq 0 \end{cases}$$

pa dobimo porazdelitev $\chi^2(1)$.

Funkcije in neodvisnost

Če sta X in Y neodvisni slučajni spremenljivki ter f in g zvezni funkciji na $\mathbb{R}$, sta tudi $U = f(X)$ in $V = g(Y)$ neodvisni slučajni spremenljivki.

V to se prepričamo takole. Za poljubna $u, v \in \mathbb{R}$ velja

$$\begin{aligned} P(U < u, V < v) &= P(f(X) < u, g(Y) < v) \\ &= P(X \in f^{-1}(-\infty, u), Y \in g^{-1}(-\infty, v)) \\ &\quad (X \text{ in } Y \text{ sta neodvisni}) \\ &= P(X \in f^{-1}(-\infty, u)) \cdot P(Y \in g^{-1}(-\infty, v)) \\ &\quad (\text{in naprej}) \\ &= P(f(X) < u) \cdot P(g(Y) < v) \\ &= P(U < u) \cdot P(V < v). \end{aligned}$$

Funkcije slučajnih vektorjev

Imejmo slučajni vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) : G \to \mathbb{R}^n$ in zvezno vektorsko preslikavo $f = (f_1, f_2, \ldots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$. Tedaj so $Y_j = f_j(X_1, X_2, \ldots, X_n)$, $j = 1, \ldots, m$ slučajne spremenljivke – komponente slučajnega vektorja $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$.

Pravimo tudi, da je $\boldsymbol{Y}$ **funkcija slučajnega vektorja** $\boldsymbol{X}$, tj. $\boldsymbol{Y} = f(\boldsymbol{X})$.

Porazdelitve komponent dobimo na običajen način

$$F_{Y_j}(y) = P(Y_j < y) = P(f_j(\boldsymbol{X}) < y) = P(\boldsymbol{X} \in f_j^{-1}(-\infty, y))$$

in, če je $\boldsymbol{X}$ zvezno porazdeljen z gostoto $p(x_1, x_2, \ldots, x_n)$, potem je

$$F_{Y_j}(y) = \int \int \ldots \int_{f_j^{-1}(-\infty, y)} p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n.$$

Primer: vsota

Naj bo $Z = X + Y$, kjer je (X, Y) zvezno porazdeljen slučajni vektor z gostoto $p(x, y)$. Tedaj je

$$\begin{aligned} F_Z(z) &= P(Z < z) = P(X + Y < z) = \\ &= \int\int_{x+y<z} p(x,y)\,dx\,dy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p(x,y)\,dy \end{aligned}$$

in $p_Z(z) = F_Z'(z) = \int_{-\infty}^{\infty} p(x, z-x)\,dx = \int_{-\infty}^{\infty} p(z-y, y)\,dy.$

Če sta spremenljivki X in Y neodvisni dobimo naprej zvezo

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)\, p_Y(z-x)\,dx.$$

Gostota $p_Z = p_X * p_Y$ je **konvolucija** funkcij p_X in p_Y.

. . . Primer: vsota

Če je $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, je vsota $Z = X + Y$ zopet normalno porazdeljena $Z : N(\mu_x + \mu_y, \sqrt{\sigma_x^2 + 2\rho\sigma_x\sigma_y + \sigma_y^2})$.

Če sta $X : \chi^2(n)$ in $Y : \chi^2(m)$ neodvisni slučajni spremenljivki, je tudi njuna vsota $Z = X + Y$ porazdeljena po tej porazdelitvi $Z : \chi^2(n + m)$.

Dosedanje ugotovitve lahko združimo v naslednjo:

Če so $X_1, X_2, \ldots, X_n$ neodvisne standardizirano normalne slučajne spremenljivke, je slučajna spremenljivka $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ porazdeljena po $\chi^2(n)$.

Primer: transformacije

Naj bo sedaj $f : (x, y) \mapsto (u, v)$ transformacija slučajnega vektorja (X, Y) v slučajni vektor (U, V) določena z zvezama $u = u(x, y)$ in $v = v(x, y)$, torej je $U = u(X, Y)$ in $V = v(X, Y)$.

Porazdelitveni zakon za nov slučajni vektor (U, V) je

$$\begin{aligned} F_{U,V}(u, v) &= P(U < u, V < v) = P((U, V) \in A(u, v)) = \\ &= P((X, Y) \in f^{-1}(A(u, v)). \end{aligned}$$

Pri zvezno porazdeljenem slučajnem vektorju (X, Y) z gostoto $p(x, y)$ je

$$F_{U,V}(u, v) = \iint_{f^{-1}(A(u,v))} p(x, y)\, dx\, dy.$$

. . . Primer: transformacije

Če je f bijektivna z zveznimi parcialnimi odvodi, lahko nadaljujemo

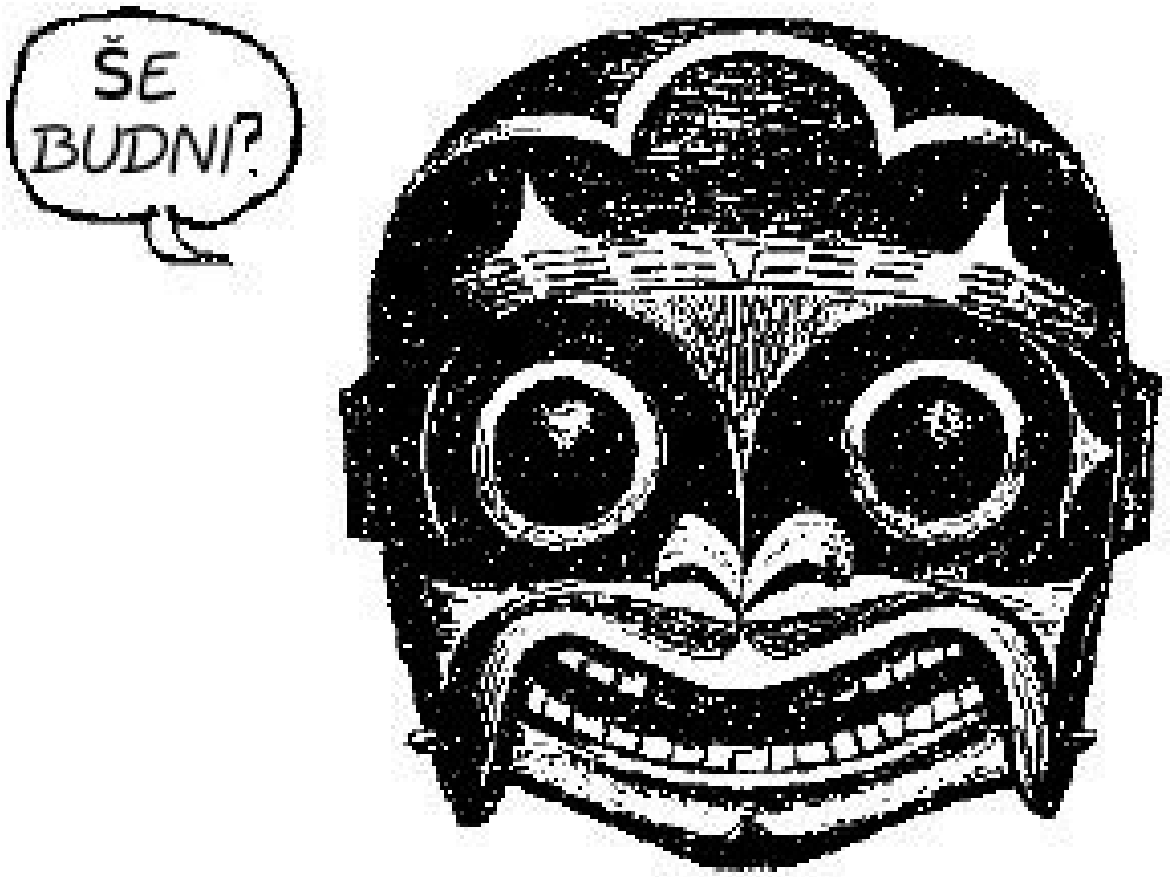
$$F_{U,V}(u,v) = \iint_{A(u,v)} p\big(x(u,v), y(u,v)\big)\, |J(u,v)|\, du\, dv,$$

kjer je (glej učbenik `http://rkb.home.cern.ch/rkb/titleA.html`)

$$J(u,v) = \frac{\partial(u,v)}{\partial(x,y)} = \det \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix}.$$

Jacobijeva determinanta (glej `http://en.wikipedia.org/wiki/Jacobian` za kakšen primer). Za gostoto $q(u,v)$ vektorja (U,V) dobimo od tu

$$q(u,v) = p\big(x(u,v), y(u,v)\big)\, |J(u,v)|.$$


ŠE
BUDNI?

Zgled:

$$\Omega = \{(x, y) \,|\, 0 < x \leq 1,\ 0 < y \leq 1\}.$$

Naj bo

$$r = \sqrt{-2\log(x)}, \qquad \varphi = 2\pi y,$$
$$u = r\cos\varphi, \qquad v = r\sin\varphi.$$

Potem po pravilu za odvajanje posrednih funkcij in definiciji Jacobijeve matrike velja

$$\frac{\partial(u,v)}{\partial(x,y)} = \left(\frac{\partial(u,v)}{\partial(r,\varphi)}\right)\left(\frac{\partial(r,\varphi)}{\partial(x,y)}\right) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix} \begin{pmatrix} \frac{-1}{rx} & 0 \\ 0 & 2\pi \end{pmatrix}.$$

Jacobijeva determinanta je

$$\det\left(\frac{d\boldsymbol{u}}{d\boldsymbol{x}}\right)=\det\left(\frac{\partial(u,v)}{\partial(x,y)}\right)=\det\left(\frac{\partial(u,v)}{\partial(r,\varphi)}\right)\det\left(\frac{\partial(r,\varphi)}{\partial(x,y)}\right)=r\frac{-2\pi}{rx}=\frac{-2\pi}{x}$$

in

$$d^2\boldsymbol{x}=\left|\det\left(\frac{d\boldsymbol{x}}{d\boldsymbol{u}}\right)\right|d^2\boldsymbol{u}=\left|\det\left(\frac{d\boldsymbol{u}}{d\boldsymbol{x}}\right)\right|^{-1}d^2\boldsymbol{u}=\frac{x}{2\pi}\,d^2\boldsymbol{u}=\frac{e^{-\frac{u^2+v^2}{2}}}{2\pi}\,d^2\boldsymbol{u}.$$

Od tod zaključimo, da za neodvisni slučajni spremenljivki x in y, ki sta enakomerno porazdeljeni med 0 in 1, zgoraj definirani slučajni spremenljivki u in v pravtako neodvisni in porazdeljeni normalno.

Pogojne porazdelitve

Naj bo B nek mogoč dogodek, tj. $P(B) > 0$. Potem lahko vpeljemo **pogojno porazdelitveno funkcijo**

$$F(x \mid B) = P(X < x \mid B) = \frac{P(X < x, B)}{P(B)}.$$

V diskretnem primeru je: $p_{ik} = P(X = x_i, Y = y_k)$, $B = (Y = y_k)$ in $P(B) = P(Y = y_k) = q_k$. Tedaj je pogojna porazdelitvena funkcija

$$\begin{aligned} F_X(x \mid y_k) &= F_X(x \mid Y = y_k) = P(X < x \mid Y = y_k) = \\ &= \frac{P(X < x, Y = y_k)}{P(Y = y_k)} = \frac{1}{q_k} \sum_{x_i < x} p_{ik} \end{aligned}$$

Vpeljimo **pogojno verjetnostno funkcijo** s $p_{i|k} = \frac{p_{ik}}{q_k}$.

Tedaj je $F_X(x \mid y_k) = \sum_{x_i < x} p_{i|k}$.

Zvezne pogojne porazdelitve

Postavimo $B = (y \leq Y < y + h)$ za $h > 0$ in zahtevajmo $P(B) > 0$.

$$F_X(x \mid B) = P(X < x \mid B) = \frac{P(X < x,\ y \leq Y < y + h)}{P(y \leq Y < y + h)} =$$

$$= \frac{F(x, y + h) - F(x, y)}{F_Y(y + h) - F_Y(y)}.$$

Če obstaja limita (za $h \to 0$)

$$F_X(x|y) = F_X(x \mid Y = y) = \lim_{h \to 0} \frac{F(x, y + h) - F(x, y)}{F_Y(y + h) - F_Y(y)},$$

jo imenujemo **pogojna porazdelitvena funkcija**
slučajne spremenljivke X glede na dogodek $(Y = y)$.

Gostota zvezne pogojne porazdelitve

Naj bosta gostoti $p(x,y)$ in $p_Y(y)$ zvezni ter $p_Y(y) > 0$. Tedaj je

$$F_X(x|y) = \lim_{h\to 0} \frac{\frac{F(x,y+h)-F(x,y)}{h}}{\frac{F_Y(y+h)-F_Y(y)}{h}} = \frac{\frac{\partial F}{\partial y}(x,y)}{F'_Y(y)} = \frac{1}{p_Y(y)} \int_{-\infty}^{x} p(u,y)du$$

oziroma, če vpeljemo **pogojno gostoto**

$$p_X(x|y) = \frac{p(x,y)}{p_Y(y)},$$

tudi $F_X(x|y) = \int_{-\infty}^{x} p_X(u|y)\, du$.

V primeru dvorazsežne normalne porazdelitve dobimo

$$p_X(x|y) : N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x\sqrt{1-\rho^2}).$$

I.8. Momenti in kovarianca

Matematično upanje

Matematično upanje $\mathsf{E}X$ (pričakovana vrednost) je posplošitev povprečne vrednosti diskretne spremenljivke X : $\begin{pmatrix} x_1 & x_2 & \cdots x_n \\ p_1 & p_2 & \cdots p_n \end{pmatrix}$,

tj.

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{n} x_i k_i = \sum_{i=1}^{n} x_i f_i,$$

od koder izhaja

$$\mathsf{E}X = \sum_{i=1}^{n} x_i p_i.$$

Diskretna slučajna spremenljivka X z verjetnostno funkcijo p_k ima matematično upanje $\mathsf{E}X = \sum_{i=1}^{\infty} x_i p_i$, če je

$$\sum_{i=1}^{\infty} |x_i| p_i < \infty.$$

Zvezna slučajna spremenljivka X z gostoto $p(x)$ ima matematično upanje $\mathsf{E}X = \int_{-\infty}^{\infty} x p(x)\, dx$, če je

$$\int_{-\infty}^{\infty} |x| p(x)\, dx < \infty.$$

Primeri slučajnih spremenljivk, za katere matematično upanje ne obstaja:

Diskretna: $x_k = (-1)^{k+1} 2^k / k$ in $p_k = 2^{-k}$.

Zvezna: $X : p(x) = \frac{1}{\pi(1+x^2)}$ – Cauchyeva porazdelitev.

Lastnosti matematičnega upanja

Naj bo a realna konstanta. Če je $P(X = a) = 1$, $\mathsf{E}X = a$.

Slučajna spremenljivka X ima matematično upanje natanko takrat, ko ga ima slučajna spremenljivka $|X|$. Velja $|\mathsf{E}X| \leq \mathsf{E}|X|$.
Za diskretno slučajno spremenljivko je $\mathsf{E}|X| = \sum_{i=1}^{\infty} |x_i| p_i$,
za zvezno pa $\mathsf{E}|X| = \int_{-\infty}^{\infty} |x| p(x)\, dx$.

Velja splošno: matematično upanje funkcije $f(X)$ obstaja in je enako za diskretno slučajno spremenljivko $\mathsf{E}f(X) = \sum_{i=1}^{\infty} f(x_i) p_i$, za zvezno pa $\mathsf{E}f(X) = \int_{-\infty}^{\infty} f(x) p(x)\, dx$, če ustrezni izraz absolutno konvergira.

Naj bo a realna konstanta. Če ima slučajna spremenljivka X matematično upanje, potem ga ima tudi spremenljivka aX in velja $\mathsf{E}(aX) = a\mathsf{E}X$.

Če imata slučajni spremenljivki X in Y matematično upanje, ga ima tudi njuna vsota $X + Y$ in velja $\mathsf{E}(X + Y) = \mathsf{E}X + \mathsf{E}Y$.

. . . Lastnosti matematičnega upanja

Za primer dokažimo zadnjo lastnost za zvezne slučajne spremenljivke.

Naj bo p gostota slučajnega vektorja (X, Y) in $Z = X + Y$.
Kot vemo, je $p_Z(z) = \int_{-\infty}^{\infty} p(x, z - x)\, dx$.

Pokažimo najprej, da Z ima matematično upanje.

$$\mathsf{E}|X + Y| = \mathsf{E}|Z| = \int_{-\infty}^{\infty} |z| p_Z(z) dz = \int_{-\infty}^{\infty} |z| \big(\int_{-\infty}^{\infty} p(x, z - x)\, dx\big) dz$$
$$= \int_{-\infty}^{\infty} \big(\int_{-\infty}^{\infty} |x + y| p(x, y)\, dx\big) dy$$
$$\leq \int_{-\infty}^{\infty} \big(\int_{-\infty}^{\infty} |x| p(x, y)\, dx\big) dy + \int_{-\infty}^{\infty} \big(\int_{-\infty}^{\infty} |y| p(x, y)\, dx\big) dy$$
$$= \int_{-\infty}^{\infty} |x| p_X(x) dx + \int_{-\infty}^{\infty} |y| p_Y(y) dy = \mathsf{E}|X| + \mathsf{E}|Y| < \infty$$

Sedaj pa še zvezo

$$\mathsf{E}(X + Y) = \mathsf{E}Z = \int_{-\infty}^{\infty} z p_Z(z)\, dz = \int_{-\infty}^{\infty} z \big(\int_{-\infty}^{\infty} p(x, z - x)\, dx\big) dz$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) p(x, y) dx dy$$
$$= \int_{-\infty}^{\infty} \big(\int_{-\infty}^{\infty} x p(x, y)\, dx\big) dy + \int_{-\infty}^{\infty} \big(\int_{-\infty}^{\infty} y p(x, y)\, dx\big) dy$$
$$= \int_{-\infty}^{\infty} x p_X(x)\, dx + \int_{-\infty}^{\infty} y p_Y(y)\, dy = \mathsf{E}X + \mathsf{E}Y$$

. . . Lastnosti matematičnega upanja

Torej je matematično upanje E **linearen funkcional**, tj.

$$\mathsf{E}(aX + bY) = a\mathsf{E}X + b\mathsf{E}Y.$$

Z indukcijo posplošimo to na
poljubno končno število členov

$$\mathsf{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n)$$

$$= a_1\mathsf{E}X_1 + a_2\mathsf{E}X_2 + \cdots + a_n\mathsf{E}X_n$$

. . . Lastnosti matematičnega upanja

Če obstajata matematični upanji $\mathsf{E}X^2$ in $\mathsf{E}Y^2$, obstaja tudi matematično upanje produkta $\mathsf{E}XY$ in velja ocena $\mathsf{E}|XY| \leq \sqrt{\mathsf{E}X^2\mathsf{E}Y^2}$.

Enakost velja natanko takrat, ko velja $Y = \pm\sqrt{\mathsf{E}Y^2/\mathsf{E}X^2}X$
z verjetnostjo 1.

Če sta slučajni spremenljivki, ki imata matematično upanje, neodvisni, obstaja tudi matematično upanje njunega produkta in velja
$\mathsf{E}XY = \mathsf{E}X \cdot \mathsf{E}Y$.

Opomba: obstajajo tudi odvisne spremenljivke, za katere velja gornja zveza.
Spremenljivki, za kateri velja $\mathsf{E}XY \neq \mathsf{E}X \cdot \mathsf{E}Y$ imenujemo **korelirani**.

Disperzija

Disperzija ali **varianca** $\mathsf{D}X$ slučajne spremenljivke, ki ima matematično upanje, je določena z izrazom

$$\mathsf{D}X = \mathsf{E}(X - \mathsf{E}X)^2$$

Disperzija je vedno nenegativna, $\mathsf{D}X \geq 0$, je pa lahko tudi neskončna.

Velja zveza

$$\mathsf{D}X = \mathsf{E}X^2 - (\mathsf{E}X)^2$$

Naj bo a realna konstanta. Če je $P(X = a) = 1$, je $\mathsf{D}X = 0$.

$\mathsf{D}aX = a^2\mathsf{D}X$

Če obstaja $\mathsf{D}X$ in je a realna konstanta, obstaja tudi $\mathsf{E}(X - a)^2$ in velja $\mathsf{E}(X - a)^2 \geq \mathsf{D}X$. Enakost velja natanko za $a = \mathsf{E}X$.

Količino $\sigma X = \sqrt{\mathsf{D}X}$ imenujemo **standardna deviacija** ali **standardni odklon**.

Standardizirane spremenljivke

Slučajno spremenljivko X **standardiziramo** s transformacijo

$$X_S = \frac{X - \mu}{\sigma},$$

kjer sta $\mu = \mathsf{E}X$ in $\sigma = \sqrt{\mathsf{D}X}$.

Za X_S velja $\mathsf{E}X_S = 0$ in $\mathsf{D}X_S = 1$.

$$\mathsf{E}X_S = \mathsf{E}\frac{X - \mu}{\sigma} = \frac{\mathsf{E}(X - \mu)}{\sigma} = \frac{\mu - \mu}{\sigma} = 0.$$

$$\mathsf{D}X_S = \mathsf{D}\frac{X - \mu}{\sigma} = \frac{\mathsf{D}(X - \mu)}{\sigma^2} = \frac{\sigma^2 - 0}{\sigma^2} = 1.$$

Matematična upanje in disperzije porazdelitev

porazdelitev	$\mathsf{E}X$	$\mathsf{D}X$
binomska $B(n,p)$	np	npq
Poissonova $P(\lambda)$	λ	λ
Pascalova $P(m,p)$	m/p	mq/p^2
geometrijska $G(p)$	$1/p$	q/p^2
enakomerna zv. $E(a,b)$	$(a+b)/2$	$(b-a)^2/12$
normalna $N(\mu,\sigma)$	μ	σ^2
gama $\Gamma(b,c)$	b/c	b/c^2
hi-kvadrat $\chi^2(n)$	n	$2n$

Kovarianca

Kovarianca $\mathsf{Cov}(X,Y)$ slučajnih spremenljivk X in Y je določena z izrazom

$$\mathsf{Cov}(X,Y) = \mathsf{E}\big((X-\mathsf{E}X)(Y-\mathsf{E}Y)\big) = \mathsf{E}XY - \mathsf{E}X\mathsf{E}Y$$

Velja: $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$ (simetričnost) in

$\mathsf{Cov}(aX+bY,Z) = a\,\mathsf{Cov}(X,Z) + b\,\mathsf{Cov}(Y,Z)$ (bilinearnost).

Če obstajata $\mathsf{D}X$ in $\mathsf{D}Y$, obstaja tudi $\mathsf{Cov}(X,Y)$ in velja

$$|\mathsf{Cov}(X,Y)| \leq \sqrt{\mathsf{D}X\mathsf{D}Y} = \sigma X \sigma Y.$$

Enakost velja natanko takrat, ko je

$$Y - \mathsf{E}Y = \pm\frac{\sigma Y}{\sigma X}(X-\mathsf{E}X)$$

z verjetnostjo 1.

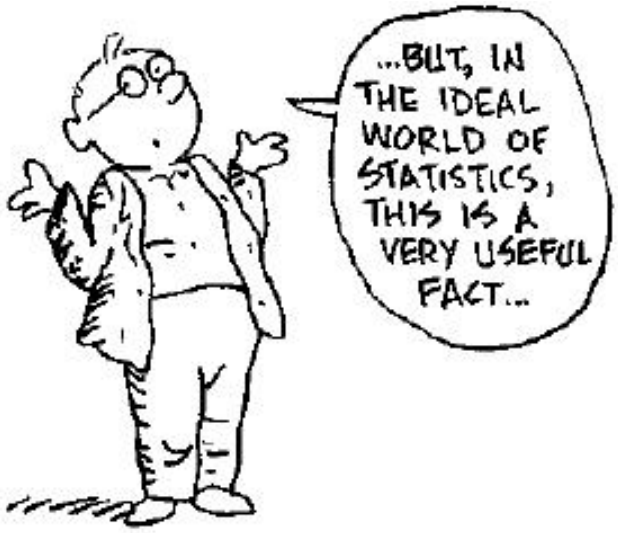
Spremenljivki X in Y sta nekorelirani natanko takrat, ko je $\mathsf{Cov}(X,Y) = 0$.

Če imata spremenljivki X in Y končni disperziji, jo ima tudi njuna vsota $X+Y$ in velja

$$\mathsf{D}(X+Y) = \mathsf{D}X + \mathsf{D}Y + 2\mathsf{Cov}(X,Y).$$

Če pa sta spremenljivki nekorelirani, je enostavno

$$\mathsf{D}(X+Y) = \mathsf{D}X + \mathsf{D}Y.$$



Zvezo lahko posplošimo na

$$\mathsf{D}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{D}X_i + \sum_{i\neq j} \mathsf{Cov}(X_i, X_j)$$

in za paroma nekorelirane spremenljivke

$$\mathsf{D}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{D}X.$$

Korelacijski koeficient

Korelacijski koeficient slučajnih spremenljivk X in Y je določen z izrazom

$$r(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sigma X \sigma Y} = \frac{\mathsf{E}((X-\mathsf{E}X)(Y-\mathsf{E}Y))}{\sigma X \sigma Y}.$$

Za $(X,Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $r(X,Y) = \rho$.
Torej sta normalno porazdeljeni slučajni spremenljivki X in Y neodvisni natanko takrat, ko sta nekorelirani.

Velja še: $-1 \leq r(X,Y) \leq 1$

$r(X,Y) = 0$ natanko takrat, ko sta X in Y nekorelirani.

$r(X,Y) = 1$ natanko takrat, ko je $Y = \frac{\sigma Y}{\sigma X}(X - \mathsf{E}X) + \mathsf{E}Y$ z verjetnostjo 1;

$r(X,Y) = -1$ natanko takrat, ko je $Y = -\frac{\sigma Y}{\sigma X}(X - \mathsf{E}X) + \mathsf{E}Y$ z verjetnostjo 1. Torej, če je $|r(X,Y)| = 1$, obstaja med X in Y linearna zveza z verjetnostjo 1.

Pogojno matematično upanje

Pogojno matematično upanje je matematično upanje pogojne porazdelitve:

Diskretna slučajna spremenljivka X ima pri pogoju $Y = y_k$ pogojno verjetnostno funkcijo $p_{i|k} = p_{ik}/q_k$, $i = 1, 2, \ldots$ in potemtakem pogojno matematično upanje

$$\mathsf{E}(X|y_k) = \sum_{i=1}^{\infty} x_i p_{i|k} = \frac{1}{q_k} \sum_{i=1}^{\infty} x_i p_{ik}.$$

Slučajna spremenljivka

$$\mathsf{E}(X|Y): \begin{pmatrix} \mathsf{E}(X|y_1) & \mathsf{E}(X|y_2) & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}$$

ima enako matematično upanje kot spremenljivka X:

$$\begin{aligned} \mathsf{E}(\mathsf{E}(X|Y)) &= \sum_{k=1}^{\infty} q_k \mathsf{E}(X|y_k) = \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} x_i p_{ik} \\ &= \sum_{i=1}^{\infty} x_i \sum_{k=1}^{\infty} p_{ik} = \sum_{i=1}^{\infty} x_i p_i = \mathsf{E}X. \end{aligned}$$

Pogojno matematično upanje zvezne spremenljivke

Zvezna slučajna spremenljivka X ima pri pogoju $Y = y$
pogojno verjetnostno gostoto $p(x|y) = p(x,y)/p_Y(y)$, $x \in \mathbb{R}$
in potemtakem pogojno matematično upanje

$$\mathsf{E}(X|y) = \int_{-\infty}^{\infty} xp(x|y)\,dx = \frac{1}{p_Y(y)} \int_{-\infty}^{\infty} xp(x,y)\,dx.$$

Slučajna spremenljivka $\mathsf{E}(X|Y)$ z gostoto $p_Y(y)$ ima enako matematično upanje kot spremenljivka X

$$\mathsf{E}(\mathsf{E}(X|Y)) = \int_{-\infty}^{\infty} \mathsf{E}(X|y)p_Y(y)\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} xp_X(x)dx = \mathsf{E}X.$$

Regresijska funkcija

Preslikavo $x \mapsto \mathsf{E}(Y|x)$ imenujemo **regresija** slučajne spremenljivke Y glede na slučajno spremenljivko X.

Primer: Naj bo $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$.
Tedaj je, kot vemo $p_X(x|y) : N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y), \sigma_x\sqrt{1-\rho^2})$.
Torej je pogojno matematično upanje

$$\mathsf{E}(X|y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)$$

in prirejena spremenljivka

$$\mathsf{E}(X|Y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y - \mu_y).$$

Na podoben način vpeljemo regresijo slučajne spremenljivke X glede na slučajno spremenljivko Y. Za dvorazsežno normalno porazdelitev dobimo

$$\mathsf{E}(Y|X) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(X - \mu_x).$$

Obe regresijski funkciji sta **linearni**.

Kovariančna matrika

Matematično upanje slučajnega vektorja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ je vektor $\mathsf{E}\boldsymbol{X} = (\mathsf{E}X_1, \mathsf{E}X_2, \ldots, \mathsf{E}X_n)$.

Primer: Za $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\mathsf{E}(X, Y) = (\mu_x, \mu_y)$.

Matematično upanje slučajne spremenljivke Y, ki je linearna kombinacija spremenljivk $X_1, X_2, \ldots, X_n$, je potem

$$\mathsf{E}Y = \mathsf{E}(\textstyle\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i \mathsf{E}X_i$$

Za disperzijo spremenljivke Y pa dobimo $\mathsf{D}Y = \mathsf{E}(Y - \mathsf{E}Y)^2 =$

$$\mathsf{E}\Big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j)\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathsf{Cov}(X_i, X_j) = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a},$$

kjer je $\mathsf{Cov}(X_i, X_j) = \mathsf{E}((X_i - \mathsf{E}X_i)(X_j - \mathsf{E}X_j))$ kovarianca spremenljivk X_i in X_j, $\boldsymbol{K} = [\mathsf{Cov}(X_i, X_j)]$ **kovariančna matrika** vektorja X, ter $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$.

Lastnosti kovariančne matrike

Kovariančna matrika $\boldsymbol{K} = [K_{ij}]$ je *simetrična*: $K_{ij} = K_{ji}$.

Diagonalne vrednosti so disperzije spremenljivk: $K_{ii} = \mathsf{D}X_i$.

Ker je $\boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a} = \mathsf{D}Y \geq 0$, je pozitivno semidefinitna matrika.

Naj bo $\boldsymbol{a}$, $||\boldsymbol{a}|| = 1$ lastni vektor, ki pripada lastni vrednosti λ kovariančne matrike $\boldsymbol{K}$, tj. $\boldsymbol{K}\boldsymbol{a} = \lambda \boldsymbol{a}$. Tedaj je $0 \leq \mathsf{D}Y = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a} = \lambda$,
kar pomeni, da so vse lastne vrednosti kovariančne matrike nenegativne.

Če je kaka lastna vrednost enaka 0, je vsa verjetnost skoncentrirana na neki hiperravnini – porazdelitev je *izrojena*. To se zgodi natanko takrat, ko kovariančna matrika $\boldsymbol{K}$ ni obrnljiva, oziroma ko je $\det \boldsymbol{K} = 0$.

Primer: Za $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je $\boldsymbol{K} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$.

Ker je $|\rho| < 1$, je $\det \boldsymbol{K} = \sigma_x^2 \sigma_y^2 (1 - \rho^2) > 0$ in je potemtakem porazdelitev vedno neizrojena. Za $N(\boldsymbol{\mu}, \boldsymbol{A})$ je $\boldsymbol{K} = \boldsymbol{A}^{-1}$.

. . . Lastnosti kovariančne matrike

Poglejmo še, kako se spremeni kovariančna matrika pri linearni transformaciji vektorja $X' = \boldsymbol{A}X$, kjer je $\boldsymbol{A}$ poljubna matrika reda $n \times n$.

Vemo, da je $\mathsf{D}(\boldsymbol{a}^T X) = \boldsymbol{a}^T \boldsymbol{K} \boldsymbol{a}$.

Tedaj je, če označimo kovariančno matriko vektorja X' s $\boldsymbol{K}'$,

$$\begin{aligned} \boldsymbol{a}^T \boldsymbol{K}' \boldsymbol{a} &= \mathsf{D}(\boldsymbol{a}^T X') = \mathsf{D}(\boldsymbol{a}^T \boldsymbol{A} X) = \mathsf{D}((\boldsymbol{A}^T \boldsymbol{a})^T X) \\ &= (\boldsymbol{A}^T \boldsymbol{a})^T \boldsymbol{K} (\boldsymbol{A}^T \boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T \boldsymbol{a} \end{aligned}$$

in potemtakem

$$\boldsymbol{K}' = \boldsymbol{A} \boldsymbol{K} \boldsymbol{A}^T.$$

Višji momenti

Višji momenti so posplošitev pojmov matematičnega upanja in disperzije.

Moment reda $k \in \mathbb{N}$ *glede na točko* $a \in \mathbb{R}$ imenujemo količino

$$m_k(a) = \mathsf{E}((X-a)^k).$$

Moment obstaja, če obstaja matematično upanje $\mathsf{E}(|X-a|^k) < \infty$.
Za $a = 0$ dobimo **začetni moment** $z_k = m_k(0)$;
za $a = \mathsf{E}X$ pa **centralni moment** $m_k = m_k(\mathsf{E}X)$.
Primera: $\mathsf{E}X = z_1$ in $\mathsf{D}X = m_2$.

Če obstaja moment $m_n(a)$, potem obstajajo tudi vsi momenti $m_k(a)$ za $k < n$.

Če obstaja moment z_n, obstaja tudi moment $m_n(a)$ za vse $a \in \mathbb{R}$.

$$m_n(a) = \mathsf{E}((X-a)^n) = \sum_{k=0}^{n} \binom{n}{k} (-a)^{n-k} z_k.$$

. . . Višji momenti

Posebej za centralni moment velja

$$m_n = m_n(z_1) = \sum_{k=0}^{n} \binom{n}{k} (-z_1)^k z_{n-k}$$

$m_0 = 1, m_1 = 0, m_2 = z_2 - z_1^2, m_3 = z_3 - 3z_2z_1 + 2z_1^3, \ldots$

Asimetrija spremenljivke X imenujemo količino $A(X) = \dfrac{m_3}{\sigma^3}$.

Sploščenost spremenljivke X imenujemo količino $K(X) = \dfrac{m_4}{\sigma^4} - 3$,

kjer je $\sigma = \sqrt{m_2}$.

Za simetrično glede na $z_1 = \mathsf{E}X$ porazdeljene spremenljivke so vsi lihi centralni momenti enaki 0.

. . . Višji momenti

Za $X : N(\mu, \sigma)$ so $m_{2k+1} = 0$ in $m_{2k} = (2k-1)!!\sigma^{2k}$.
Zato sta tudi $A(X) = 0$ in $K(X) = 0$.

Če sta spremenljivki X in Y neodvisni, je $m_3(X+Y) = m_3(X)+m_3(Y)$.

Za binomsko porazdeljeno spremenljivko $X : B(n,p)$ je
$m_3(X) = npq(q-p)$ in dalje $A(X) = \frac{q-p}{\sqrt{npq}}$.

Kadar spremenljivka nima momentov, uporabljamo kvantile.

Kvantil reda $p \in (0,1)$ je vsaka vrednost $x \in \mathbb{R}$, za katero velja
$P(X \leq x) \geq p$ in $P(X \geq x) \geq 1-p$ oziroma $F(x) \leq p \leq F(x+)$.
Kvantil reda p označimo z x_p. Za zvezno spremenljivko je $F(x_p) = p$.

Kvantil $x_{\frac{1}{2}}$ imenujemo **mediana**; $x_{\frac{i}{4}}$, $i = 0,1,2,3,4$ so **kvartili**.

Kot nadomestek za standardni odklon uporabljamo **kvartilni razmik**
$\frac{1}{2}(x_{\frac{3}{4}} - x_{\frac{1}{4}})$.

I.9. Karakteristične funkcije in limitni izreki

Karakteristična funkcija

Naj bo Z kompleksna slučajna spremenljivka, tj. $Z = X + iY$ za slučajni spremenljivki X in Y.

Njeno upanje izračunamo z

$$E(Z) = E(X) + iE(Y),$$

disperzijo pa z

$$D(Z) = E(|Z - E(Z)|^2) = D(X) + D(Y),$$

Kompleksna funkcija realne slučajne spremenljivke je kompleksna slučajna spremenljivka, npr. e^{iX}.

... Karakteristična funkcija

Karakteristična funkcija realne slučajne spremenljivke X je kompleksna funkcija $\varphi_X(t)$ realne spremenljivke t določena z zvezo $\varphi_X(t) = \mathsf{E}e^{itX}$.

Karakteristične funkcije vedno obstajajo in so močno računsko orodje.

Posebej pomembni lastnosti sta:

Če obstaja začetni moment z_n, je karakteristična funkcija n-krat odvedljiva v vsaki točki in velja $\varphi_X^{(k)}(0) = i^k z_k$.

Za neodvisni spremenljivki X in Y je $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

Pojem karakteristične funkcije lahko posplošimo tudi na slučajne vektorje.

Reprodukcijska lastnost normalne porazdelitve

Vsaka linearna kombinacija *neodvisnih*
in *normalno* porazdeljenih slučajnih spremenljivk
je tudi sama **normalno** porazdeljena.

Če so slučajne spremenljivke $X_1, \ldots, X_n$ neodvisne in normalno porazdeljene $N(\mu_i, \sigma_i)$, potem je njihova vsota tudi normalno porazdeljena:

$$N\left(\sum \mu_i, \sqrt{\sum \sigma_i^2}\right).$$

Da ne bi vsota povprečij rastla z n, nadomestimo vsoto spremenljivk X_i z njihovim provprečjem $\overline{X}$ in dobimo

$$N\left(\overline{\mu}, \sqrt{\sum \left(\frac{\sigma_i}{n}\right)^2}\right).$$

Če privzamemo $\mu_i = \mu$ in $\sigma_i = \sigma$, dobimo $N\left(\mu, \sigma/\sqrt{n}\right)$.

Limitni izreki

Zaporedje slučajnih spremenljivk X_n **verjetnostno konvergira** k slučajni spremenljivki X, če za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

ali enakovredno

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Zaporedje slučajnih spremenljivk X_n **skoraj gotovo konvergira** k slučajni spremenljivki X, če velja

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

. . . Limitni izreki

Če zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X, potem za vsak $\varepsilon > 0$ velja

$$\lim_{m\to\infty} P(|X_n - X| < \varepsilon \quad \text{za vsak} \quad n \geq m) = 1.$$

Od tu izhaja:

če konvergira skoraj gotovo $X_n \to X$,
potem konvergira tudi verjetnostno $X_n \to X$.

Šibki in krepki zakon velikih števil

Naj bo $X_1, \ldots, X_n$ zaporedje spremenljivk, ki imajo matematično upanje. Označimo $S_n = \sum_{k=1}^{n} X_k$ in

$$Y_n = \frac{S_n - \mathsf{E}S_n}{n} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mathsf{E}X_k) = \frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}X_k.$$

Pravimo, da za zaporedje slučajnih spremenljivk X_k velja:

- **šibki zakon velikih števil**, če gre verjetnostno $Y_n \to 0$, tj., če $\forall \varepsilon > 0$ velja
$$\lim_{n\to\infty} P\left(\left|\frac{S_n - \mathsf{E}S_n}{n}\right| < \varepsilon\right) = 1;$$
- **krepki zakon velikih števil**, če gre skoraj gotovo $Y_n \to 0$, tj., če velja
$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1.$$

Če za zaporedje $X_1, \ldots, X_n$ velja krepki zakon, velja tudi šibki.

Neenakost Čebiševa

Če ima slučajna spremenljivka X končno disperzijo, tj. $\mathsf{D}X < \infty$, velja za vsak $\varepsilon > 0$ **neenakost Čebiševa**

$$P(|X - \mathsf{E}X| \geq \varepsilon) \leq \frac{\mathsf{D}X}{\varepsilon^2}.$$

Dokaz: Pokažimo jo za zvezne spremenljivke

$$P(|X - \mathsf{E}X| \geq \varepsilon) = \int_{|x-\mathsf{E}X|\geq\varepsilon} p(x)\,dx = \frac{1}{\varepsilon^2}\int_{|x-\mathsf{E}X|\geq\varepsilon} \varepsilon^2 p(x)\,dx$$

$$\leq \frac{1}{\varepsilon^2}\int_{-\infty}^{\infty} (x-\mathsf{E}X)^2 p(x)\,dx = \frac{\mathsf{D}X}{\varepsilon^2}. \quad \blacksquare$$

Neenakost Čebiševa – posledice

(**Markov**) Če gre za zaporedje slučajnih spremenljivk X_i izraz

$$\frac{\mathsf{D}S_n}{n^2} \to 0,$$

ko gre $n \to \infty$, velja za zaporedje šibki zakon velikih števil.

(**Čebišev**) Če so slučajne spremenljivke X_i paroma nekorelirane in so vse njihove disperzije omejene z isto konstanto C, tj.

$$\mathsf{D}X_i < C \quad \text{za vsak } i,$$

velja za zaporedje šibki zakon velikih števil.

Dokaz Bernoullijevega izreka

Za Bernoullijevo zaporedje X_i so spremenljivke paroma neodvisne, $\mathsf{D}X_i = pq$, $S_n = k$. Pogoji izreka Čebiševa so izpolnjeni in dobimo:

(**Bernoulli** 1713) Za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1.$$

Še nekaj izrekov

(**Hinčin**) Če so neodvisne slučajne spremenljivke X_i enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = a$ za vsak i,
potem velja zanje šibki zakon velikih števil, tj. za vsak $\varepsilon > 0$ je

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - a\right| < \varepsilon\right) = 1.$$

(**Kolmogorov**) Če so slučajne spremenljivke X_i neodvisne, imajo končno disperzijo in velja $\sum_{n=1}^{\infty} \frac{\mathsf{D}S_n}{n^2} < \infty$, potem velja krepki zakon velikih števil:

$$P\left(\lim_{n\to\infty} \frac{S_n - \mathsf{E}S_n}{n} = 0\right) = 1.$$

. . . Še nekaj izrekov

(**Kolmogorov**) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene in imajo matematično upanje $\mathsf{E}X_i = \mu$, potem velja krepki zakon velikih števil

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

(**Borel** 1909) Za Bernoullijevo zaporedje velja

$$P\left(\lim_{n\to\infty} \frac{k}{n} = p\right) = 1.$$

Centralni limitni izrek (CLI)

Leta 1810 je Pierre Laplace (1749-1827) študiral anomalije orbit Jupitra in Saturna, ko je izpeljal razširitev De Moivrovega limitnega izreka,

"Vsaka vsota ali povprečje, če je število členov dovolj veliko,
je približno normalno porazdeljena."

Centralni limitni zakon

Opazujmo sedaj zaporedje standardiziranih spremenljivk

$$Z_n = \frac{S_n - \mathsf{E}S_n}{\sigma(S_n)}, \qquad \text{kjer je } S_n = X_1 + \cdots + X_n.$$

Za zaporedje slučajnih spremenljivk X_i velja **centralni limitni zakon**, če porazdelitvene funkcije za Z_n gredo proti porazdelitveni funkciji standardizirane normalne porazdelitve, to je, če za vsak $x \in \mathbb{R}$ velja

$$\lim_{n\to\infty} P\left(\frac{S_n - \mathsf{E}S_n}{\sigma(S_n)} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

Osnovni centralni limitni izrek (CLI) Če so slučajne spremenljivke X_i neodvisne, enako porazdeljene s končnim matematičnim upanjem in končno disperzijo, potem zanje velja centralni limitni zakon.

Skica dokaza centralnega limitnega izreka

Naj bo $Z_i = \dfrac{X_i - \mu}{\sigma}$. Potem je

$$M_Z(t) = 1 - \frac{t^2}{2!} + \frac{t^3}{3!} E(Z_i^3) + \cdots$$

Za $Y_n = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \dfrac{1}{\sqrt{n}} \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma} = \dfrac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$ velja

$$M_n(t) = \left[M_Z\left(\frac{t}{\sqrt{n}}\right) \right]^n = \left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\, k + \cdots \right)^n,$$

kjer je $k = E(Z_i^3)$.

... Skica dokaza CLI

$$\log M_n(t) = n \log\left(1 - \frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}}\,k + \cdots\right)$$

Za $x = \left(-\frac{t^2}{2\,n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right)$ velja

$$\log M_n(t) = n \log(1+x) = n(x - \frac{x^2}{2} + \cdots) =$$

$$n\left[\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right) - \frac{1}{2}\left(-\frac{t^2}{2n} + \frac{t^3}{3!\,n^{3/2}} + \cdots\right)^2 + \cdots\right]$$

in od tod končno še

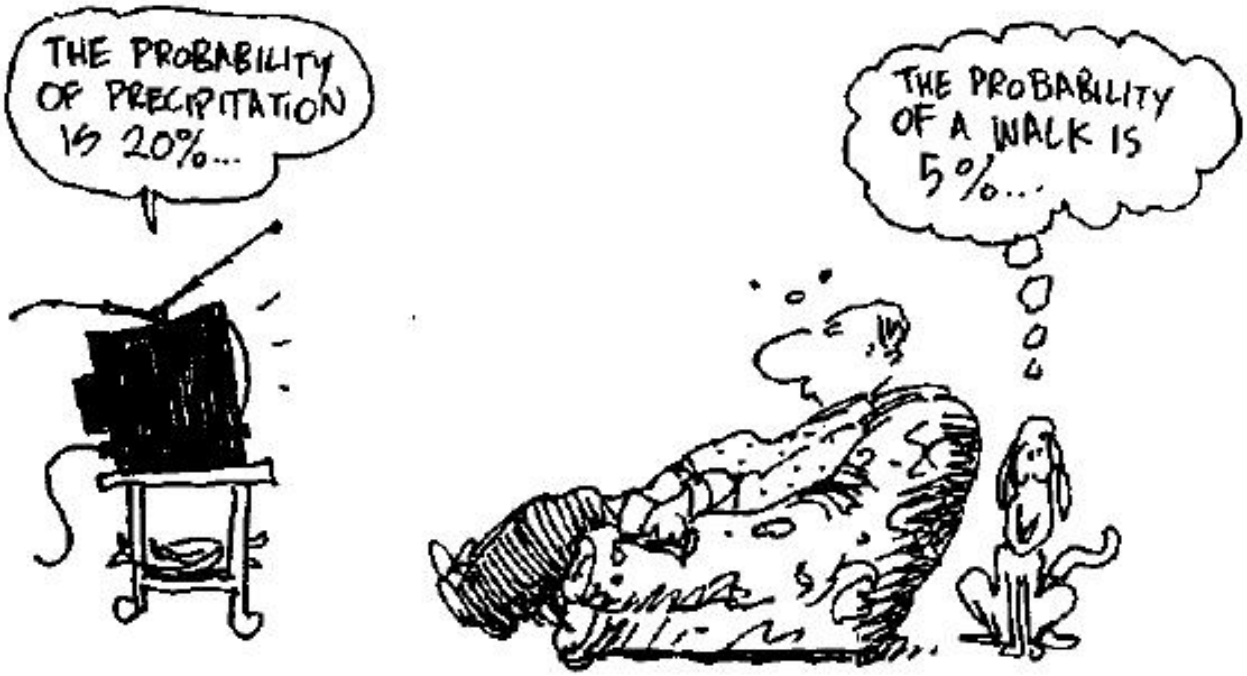
$$\lim_{n\to\infty} \log M_n(t) = -\frac{t^2}{2} \qquad \text{oziroma} \qquad \lim_{n\to\infty} M_n(t) = e^{-t^2/2}.$$

... Skica dokaza CLI

Iz konvergence karakterističnih funkcij φ_{Y_n}
proti karakteristični funkciji standardizirano normalne porazdelitve
lahko sklepamo po obratnem konvergenčnem izreku,
da tudi porazdelitvene funkcije za Y_n konvergirajo
proti porazdelitveni funkciji standardizirano normalne porazdelitve.

Torej velja centralni limitni zakon. ■

I.10. Uporaba

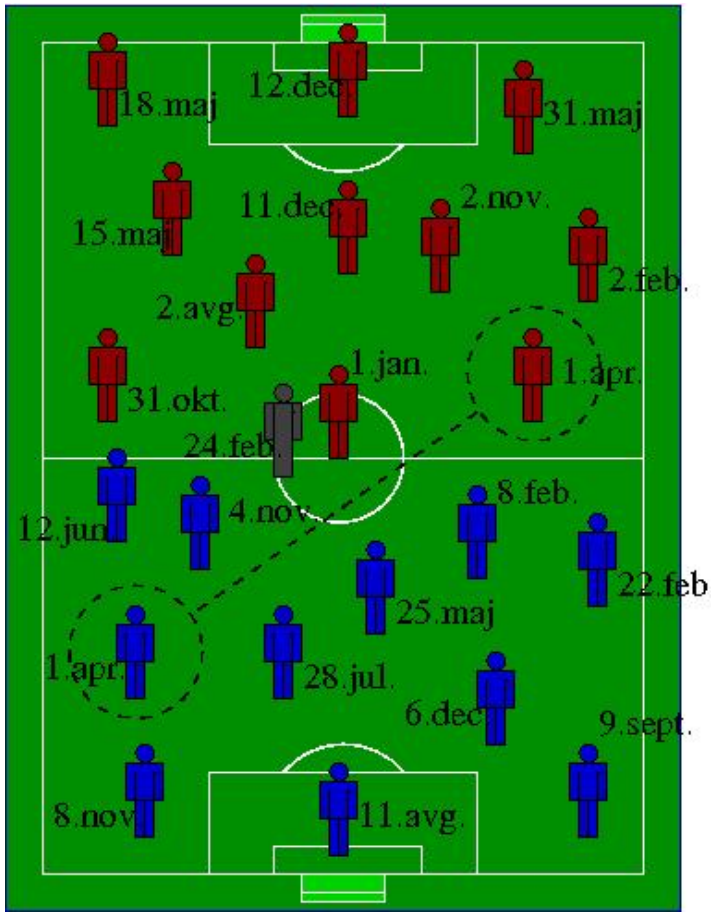

THE PROBABILITY OF PRECIPITATION IS 20%...
THE PROBABILITY OF A WALK IS 5%...

Kakšno naključje!!! Mar res?

Na nogometni tekmi sta
na igrišču dve enajsterici
in sodnik, skupaj
23 oseb.

Kakšna je verjetnost,
da imata **dve osebi**
isti rojstni dan?

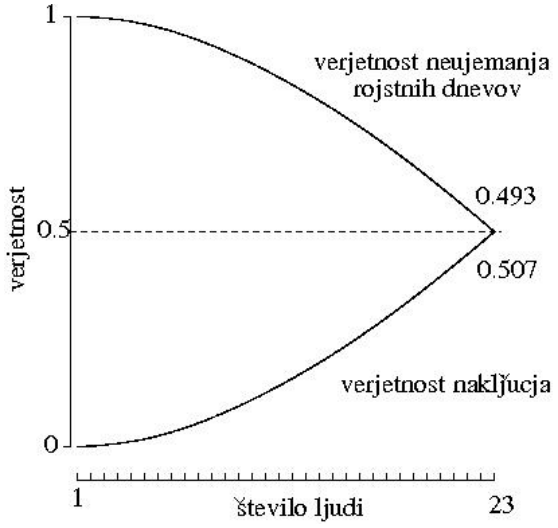
Ali je ta verjetnost lahko
večja od **0,5**?



Ko vstopi v sobo k-ta oseba, je verjetnost, da je vseh k rojstnih dnevov različnih enaka:

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365-k+1}{365} =$$

$$= \begin{cases} 0,493; & \text{če je } k=22 \\ 0,507; & \text{če je } k=23 \end{cases}$$



**V poljubni skupini 23-ih ljudi je verjetnost,
da imata vsaj dva skupni rojstni dan $> 1/2$.**

Čeprav je 23 majhno število, je med 23 osebami 253 različnih parov.
To število je veliko bolj povezano z iskano verjetnostjo.

Testirajte to na zabavah z več kot 23 osebami.

Organizirajte stave in dolgoročno boste gotovo na boljšem,
na velikih zabavah pa boste zlahka zmagovali.

Napad s pomočjo paradoksa rojstnih dnevov

(angl. Birthday Attack)

To seveda ni paradoks, a vseeno ponavadi zavede naš občutek.

Ocenimo še splošno verjetnost.

Mečemo k žogic v n posod in gledamo,
ali sta v kakšni posodi vsaj dve žogici.

Poiščimo spodnjo mejo za verjetnost zgoraj opisanega dogodka:

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)=\prod_{i=1}^{k-1}\left(1-\frac{i}{n}\right)$$

Iz Taylorjeve vrste

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots$$

ocenimo $1 - x \approx e^{-x}$ in dobimo

$$\prod_{i=1}^{k-1}\left(1 - \frac{i}{n}\right) \approx \prod_{i=1}^{k-1} e^{\frac{-i}{n}} = e^{\frac{-k(k-1)}{2n}}.$$

Torej je verjetnost trčenja

$$1 - e^{\frac{-k(k-1)}{2n}}.$$

Potem velja

$$e^{\frac{-k(k-1)}{2n}} \approx 1-\varepsilon$$

oziroma

$$\frac{-k(k-1)}{2n} \approx \log(1-\varepsilon), \quad \text{tj.} \quad k^2-k \approx 2n\log\frac{1}{1-\varepsilon}$$

in če ignoriramo $-k$, dobimo končno

$$\boxed{k \approx \sqrt{2n\log\frac{1}{1-\varepsilon}}\,.}$$

Za $\varepsilon = 0,5$ je

$$k \approx 1,17\sqrt{n},$$

kar pomeni, da, če zgostimo nekaj več kot $\sqrt{n}$ elementov,
je bolj verjetno, da pride do trčenja kot da ne pride do trčenja.

V splošnem je k proporcionalen s $\sqrt{n}$.

Raba v kriptografiji

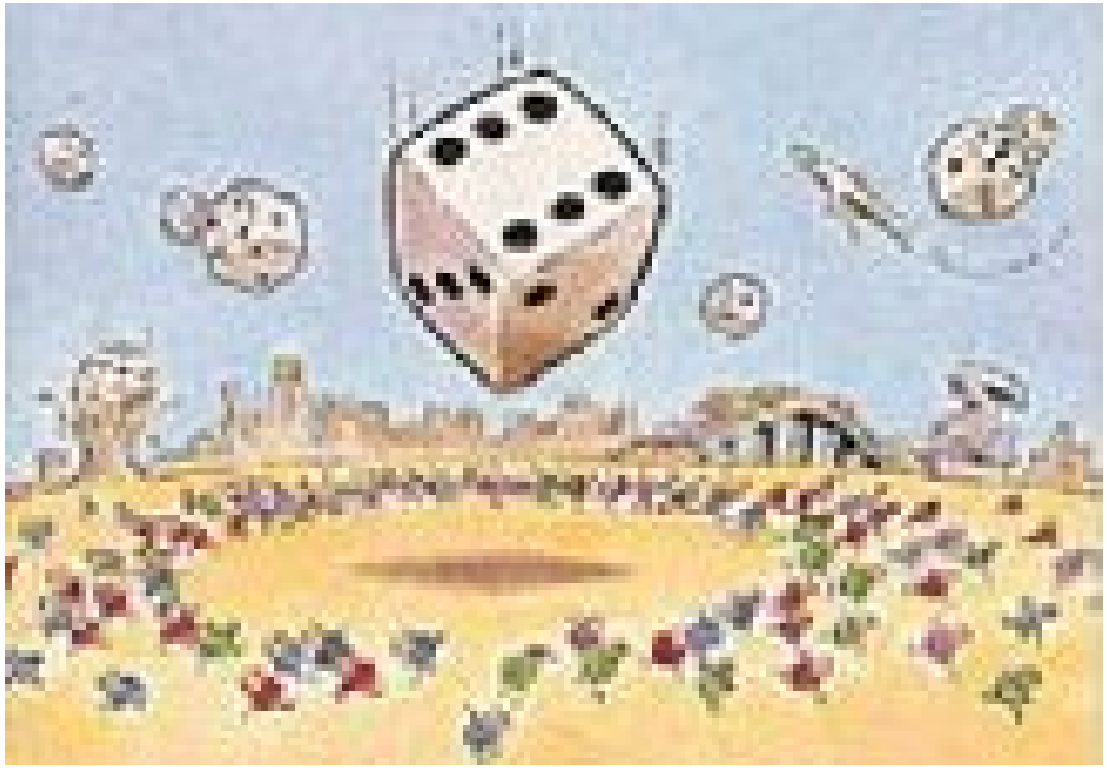
Napad s pomočjo paradoksa rojstnih dnevov s tem določi spodnjo mejo za velikost zaloge vrednosti zgoščevalnih funkcij, ki jih uporabljamo v kriptografiji in računalniški varnosti.

40-bitna zgostitev ne bi bila varna, saj bi prišli do trčenja z nekaj več kot 2^{20} (se pravi milijon) naključnimi zgostitvami z verjetnostjo vsaj $1/2$.

V praksi je priporočena najmanj 128-bitna zgostitev in standard za shema digitalnega podpisa (160 bitov) to vsekakor upošteva.

Podobno si lahko pomagamo tudi pri napadih na DLP in še kje.

II. STATISTIKA



II.1. Osnovni pojmi

Statistika je veda, ki proučuje množične pojave.

Ljudje običajno besedo **statistika** povezujejo z zbiranjem in urejanjem podatkov o nekem pojavu, izračunom raznih značilnosti iz teh podatkov, njih predstavitvijo in razlago.

To je najstarejši del statistike in ima svoje začetke že v antiki – z nastankom večjih združb (držav) se je pojavila potreba po poznavanju stanja – 'računovodstvo', astronomija, ...

Sama beseda ***statistika*** naj bi izvirala iz latinske besede ***status*** – v pomenu država. Tej veji statistike pravimo *opisna statistika*.

Druga veja, *inferenčna statistika*, poskuša spoznanja iz zbranih podatkov posplošiti (razširiti, podaljšati, napovedati, ...) in oceniti kakovost teh pospošitev.

Statistiko lahko razdelimo tudi na *uporabno* in *teoretično* (računalniško in matematično) statistiko.

... Osnovni pojmi

(Statistična) enota – posamezna proučevana stvar ali pojav.
Primer: redni študent na Univerzi v Ljubljani v študijskem letu 2008/09.

Populacija – množica vseh proučevanih enot; pomembna je natančna opredelitev populacije (npr. časovno in prostorsko).
Primer: vsi redni študentje na UL v študijskem letu 2008/09

Vzorec – podmnožica populacije, na osnovi katere ponavadi sklepamo o lastnostih celotne populacije.
Primer: vzorec 300 slučajno izbranih rednih študentov na UL v l. 2008/09.

Spremenljivka – lastnost enot; označujemo jih npr. z X, Y, X_1.
Vrednost spremenljivke X na i–ti enoti označimo z x_i.
Primer: spol, uspeh iz matematike v zadnjem razredu srednje šole, izobrazba matere in višina mesečnih dohodkov staršev študenta.

. . . Osnovni pojmi

Posamezne spremenljivke in odnose med njimi opisujejo ustrezne porazdelitve.

Parameter – značilnost populacije; običajno jih označujemo z malimi grškimi črkami.

Statistika – značilnost vzorca; običajno jih označujemo z malimi latinskimi črkami. Vrednost statistike je lahko za različne vzorce različna.

Eno izmed osnovnih vprašanj statistike je,
kako z uporabo ustreznih statistik oceniti
vrednosti izbranih parametrov.

Vrste spremenljivk

Vrste spremenljivk glede na vrsto vrednosti:

1. **opisne** (ali atributivne) spremenljivke – vrednosti lahko opišemo z imeni razredov (npr. poklic, uspeh);
2. **številske** (ali numerične) spremenljivke – vrednosti lahko izrazimo s števili (npr. starost).

. . . Vrste spremenljivk

Vrste spremenljivk glede na vrsto merske lestvice:

1. **imenske** (ali nominalne) spremenljivke – vrednosti lahko le razlikujemo med seboj: dve vrednosti sta enaki ali različni (npr. spol);
2. **urejenostne** (ali ordinalne) spremenljivke – vrednosti lahko uredimo od najmanjše do največje (npr. uspeh);
3. **razmične** (ali intervalne) spremenljivke – lahko primerjamo razlike med vrednostima dvojic enot (npr. temperatura);
4. **razmernostne** spremenljivke – lahko primerjamo razmerja med vrednostima dvojic enot (npr. starost).
5. **absolutne** spremenljivke – štetja (npr. število prebivalcev).

. . . Vrste spremenljivk

dovoljene transformacije	*vrsta lestvice*	*primeri*
$f(x) = x$ (identiteta)	absolutna	štetje
$f(x) = a.x, a > 0$ podobnost	razmernostna	masa temperatura (K)
$f(x) = a.x + b, a > 0$	razmična	temperatura (C,F) čas (koledar)
$x \geq y \Leftrightarrow f(x) \geq f(y)$ strogo naraščajoča	urejenostna	šolske ocene, kakovost zraka, trdost kamnin
f je povratno enolična	imenska	barva las, narodnost

. . . Vrste spremenljivk

Vrste spremenljivk so urejene od tistih z najslabšimi merskimi lastnostmi do tistih z najboljšimi. Urejenostne spremenljivke zadoščajo lastnostim, ki jih imajo imenske spremenljivke; in podobno razmernostne spremenljivke zadoščajo lastnostim, ki jih imajo razmične, urejenostne in imenske spremenljivke.

$$\text{absolutna} \subset \text{razmernostna} \subset \text{razmična} \subset \text{urejenostna} \subset \text{imenska}$$

Posamezne statistične metode predpostavljajo določeno vrsto spremenljivk. Največ učinkovitih statističnih metod je razvitih za številske spremenljivke.

V teoriji merjenja pravimo, da je nek stavek *smiseln*, če ohranja resničnost/lažnost pri zamenjavi meritev z enakovrednimi (glede na dovoljene transformacije) meritvami.

Frekvenčna porazdelitev

Število vseh možnih vrednosti proučevane spremenljivke je lahko preveliko za pregledno prikazovanje podatkov. Zato sorodne vrednosti razvrstimo v skupine. Posamezni skupini priredimo ustrezno reprezentativno vrednost, ki je nova vrednost spremenljivke. Skupine vrednosti morajo biti določene **enolično**: vsaka enota s svojo vrednostjo je lahko uvrščena v natanko eno skupino vrednosti.

Frekvenčna porazdelitev spremenljivke je ***tabela***, ki jo določajo ***vrednosti ali skupine vrednosti*** in njihove ***frekvence***.

Če je spremenljivka vsaj urejenostna, vrednosti (ali skupine vrednosti) uredimo od najmanjše do največje.

Skupine vrednosti številskih spremenljivk imenujemo *razredi*.

. . . Frekvenčna porazdelitev

x_{min} in x_{max} – *najmanjša* in *največja* vrednost spremenljivke X.

$x_{i,min}$ in $x_{i,max}$ – *spodnja* in *zgornja meja* i-tega razreda.

Meje razredov so določene tako, da velja $x_{i,max} = x_{i+1,min}$.

Širina i-tega razreda je $d_i = x_{i,max} - x_{i,min}$.
Če je le mogoče, vrednosti razvrstimo v razrede enake širine.

Sredina i-tega razreda je $x_i = \frac{x_{i,min}+x_{i,max}}{2}$ in je značilna vrednost – predstavnik tega razreda.

Kumulativa (ali nakopičena frekvenca) je frekvenca do spodnje meje določenega razreda. Velja $F_{i+1} = F_i + f_i$, kjer je F_i kumulativa in f_i frekvenca v i-tem razredu.

Slikovni prikazi

Stolpčni prikaz: Na eni osi prikažemo (urejene) razrede. Nad vsakim naredimo stolpec/črto višine sorazmerne frekvenci razreda.

Krožni prikaz: Vsakemu razredu priredimo krožni izsek s kotom $\alpha_i = \frac{f_i}{n} 360$ stopinj.

Histogram: drug poleg drugega rišemo stolpce – pravokotnike, katerih ploščina je sorazmerna frekvenci v razredu. Če so razredi enako široki, je višina sorazmerna tudi frekvenci.

Poligon: v koordinatnem sistemu zaznamujemo točke (x_i, f_i), kjer je x_i sredina i–tega razreda in f_i njegova frekvenca. K tem točkam dodamo še točki $(x_0, 0)$ in $(x_{k+1}, 0)$, če je v frekvenčni porazdelitvi k razredov. Točke zvežemo z daljicami.

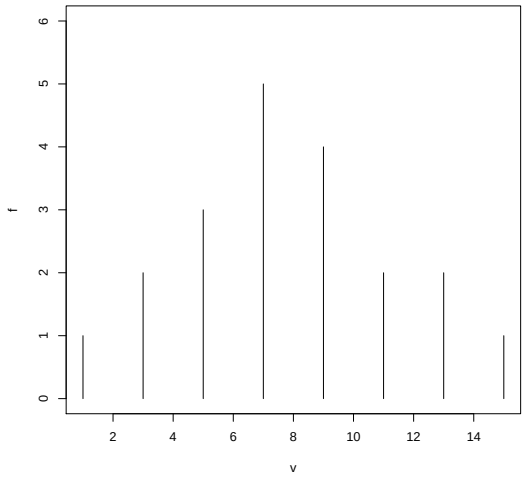
Ogiva: grafična predstavitev kumulativne frekvenčne porazdelitve s poligonom, kjer v koordinatni sistem nanašamo točke $(x_{i,min}, F_i)$.

Nekaj ukazov v R-ju

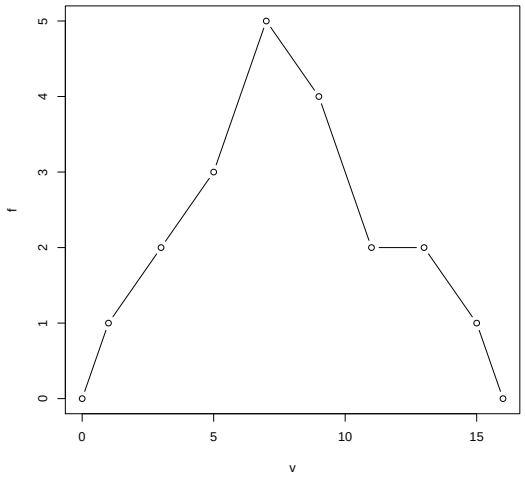
```
> X <- c(5,11,3,7,5,7,15,1,13,11,9,9,3,13,9,7,7,5,9,7)
> n <- length(X)
> t <- tabulate(X)
> t
 [1] 1 0 2 0 3 0 5 0 4 0 2 0 2 0 1
> v <- (1:max(X))[t>0]
> f <- t[t>0]
> rbind(v,f)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
v    1    3    5    7    9   11   13   15
f    1    2    3    5    4    2    2    1
> plot(v,f,type="h")
> plot(c(0,v,16),c(0,f,0),type="b",xlab="v",ylab="f")
> pie(f,v)
> plot(c(0,v,16),c(0,cumsum(f)/n,1),col="red",type="s",
  xlab="v",ylab="f")
> x <- sort(rnorm(100,mean=175,sd=30))
> y <- (1:100)/100
> plot(x,y,main="Normalna porazdelitev, n=100",type="s")
> curve(pnorm(x,mean=175,sd=30),add=T,col="red")
```


. . . Slikovni prikazi

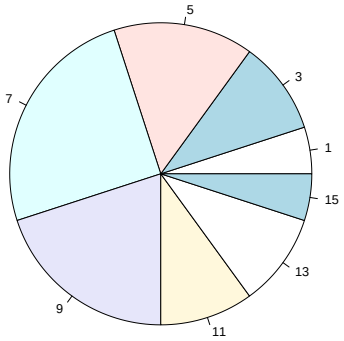
stolpci



poligon



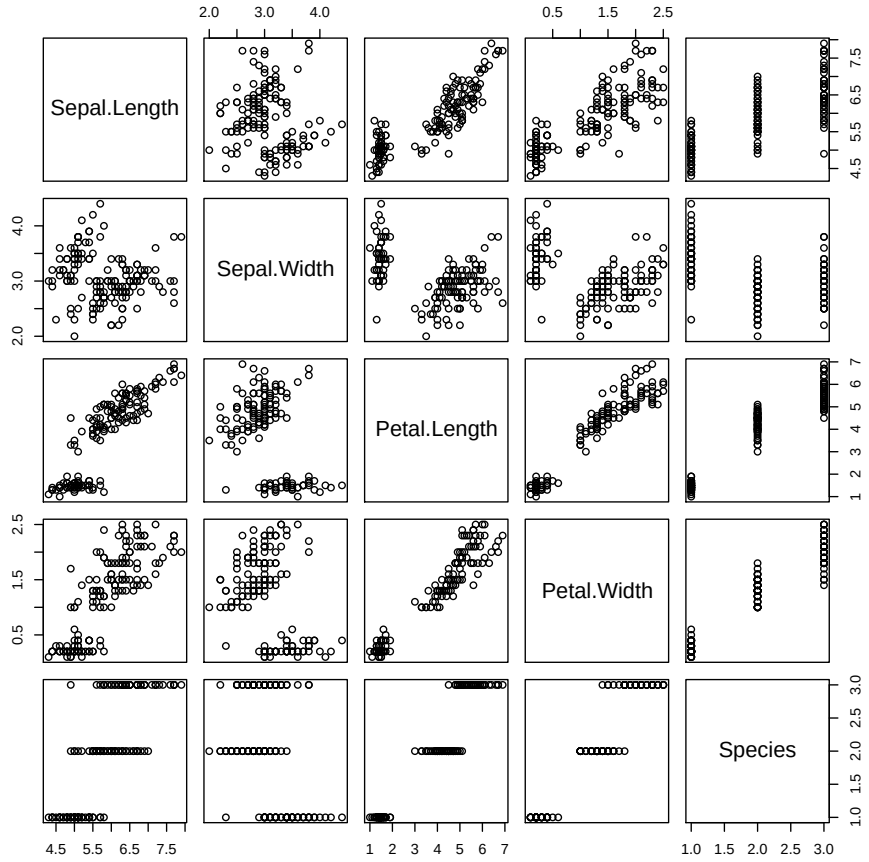
strukturni krog



Še nekaj ukazov v R-ju

```
> x <- rnorm(1000,mean=175,sd=30)
> mean(x)
[1] 175.2683
> sd(x)
[1] 30.78941
> var(x)
[1] 947.9878
> median(x)
[1] 174.4802
> min(x)
[1] 92.09012
> max(x)
[1] 261.3666
> quantile(x,seq(0,1,0.1))
       0%       10%       20%       30%
 92.09012 135.83928 148.33908 158.53864
      40%       50%       60%       70%
166.96955 174.48018 182.08577 191.29261
      80%       90%      100%
200.86309 216.94009 261.36656

> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  92.09  154.20  174.50  175.30  195.50  261.40
> hist(x,freq=F)
> curve(dnorm(x,mean=175,sd=30),add=T,col="red")
```


Fisherjeve oziroma Andersonove perunike (Iris data)




```
> data()
> data(iris)
> help(iris)
> summary(iris)
  Sepal.Length    Sepal.Width
 Min.   :4.300   Min.   :2.000
 1st Qu.:5.100   1st Qu.:2.800
 Median :5.800   Median :3.000
 Mean   :5.843   Mean   :3.057
 3rd Qu.:6.400   3rd Qu.:3.300
 Max.   :7.900   Max.   :4.400
  Petal.Length    Petal.Width
 Min.   :1.000   Min.   :0.100
 1st Qu.:1.600   1st Qu.:0.300
 Median :4.350   Median :1.300
 Mean   :3.758   Mean   :1.199
 3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
> pairs(iris)
```

Parni prikaz.

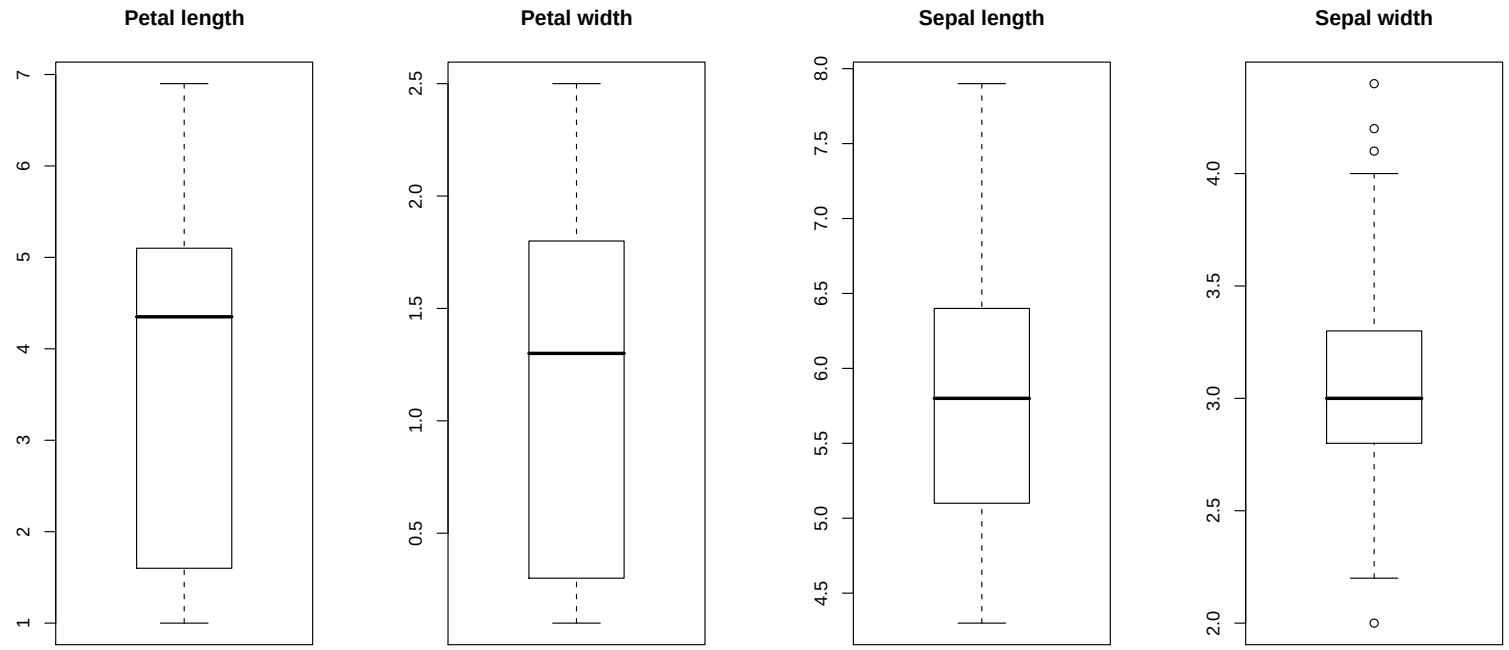
Škatle in Q-Q-prikazi

Škatle (box-and-whiskers plot; grafikon kvantilov) `boxplot`:
škatla prikazuje notranja kvartila razdeljena z mediansko črto.
Daljici – brka vodita do robnih podatkov, ki sta največ za 1,5 dolžine škatle oddaljena od nje. Ostali podatki so prikazani posamično.

Q-Q-prikaz `qqnorm` je namenjen prikazu normalnosti porazdelitve danih n podatkov. Podatke uredimo in prikažemo pare točk sestavljene iz vrednosti k-tega podatka in pričakovane vrednosti k-tega podatka izmed n normalno porazdeljenih podatkov. Če sta obe porazdelitvi normalni, ležijo točke na premici. Premica `qqline` nariše premico skozi prvi in tretji kvartil.

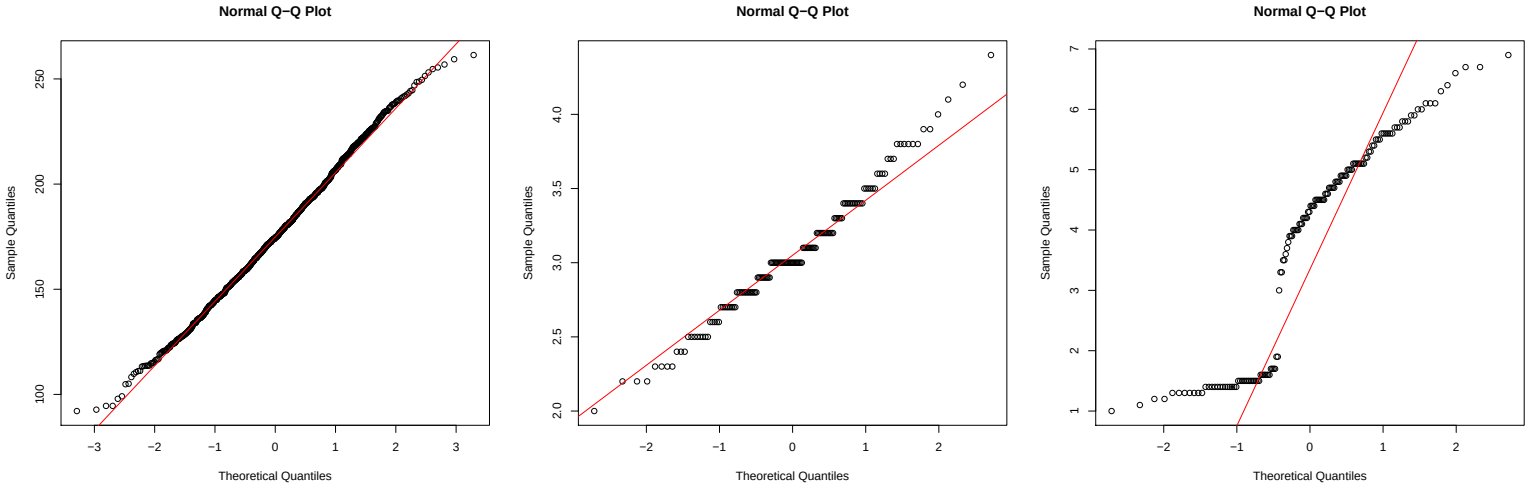
Obstaja tudi splošnejši ukaz `qqplot`, ki omogoča prikaz povezanosti poljubnega para porazdelitev. S parametrom `datax=T` zamenjamo vlogo koordinatnih osi.

Škatle



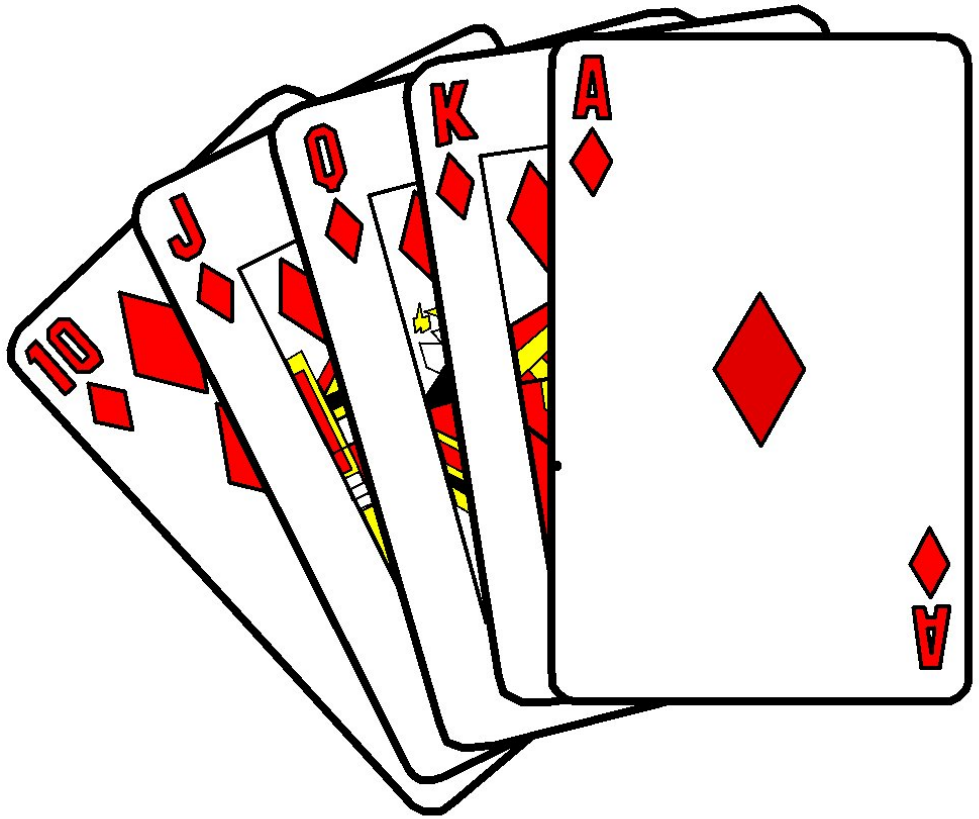

```
> par(mfrow=c(1,2))
> boxplot(iris$Petal.Length,main="Petal length")
> boxplot(iris$Petal.Width,main="Petal width")
> boxplot(iris$Sepal.Length,main="Sepal length")
> boxplot(iris$Sepal.Width,main="Sepal width")
> par(mfrow=c(1,1))
```

Q-Q-prikaz



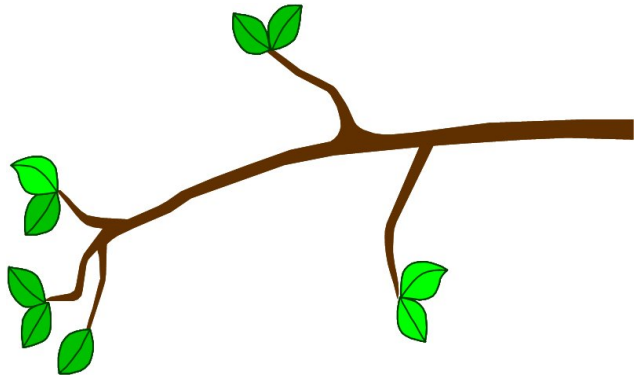

```
> qqnorm(x)
> qqline(x,col="red")
> qqnorm(iris$Sepal.Width)
> qqline(iris$Sepal.Width,col="red")
> qqnorm(iris$Petal.Length)
> qqline(iris$Petal.Length,col="red")
```

II.2. Vzorčenje



. . . Vzorčenje

Analitična statistika je veja statistike, ki se ukvarja z uporabo vzorčnih podatkov, da bi z njimi naredili zaključek (inferenco) o populaciji.



Zakaj vzorčenje?

- cena
- čas
- destruktivno testiranje

Glavno vprašanje statistike je:

kakšen mora biti vzorec, da lahko iz podatkov zbranih na njem veljavno sklepamo o lastnostih celotne populacije.

. . . Vzorčenje

Kdaj vzorec dobro predstavlja celo populacijo?

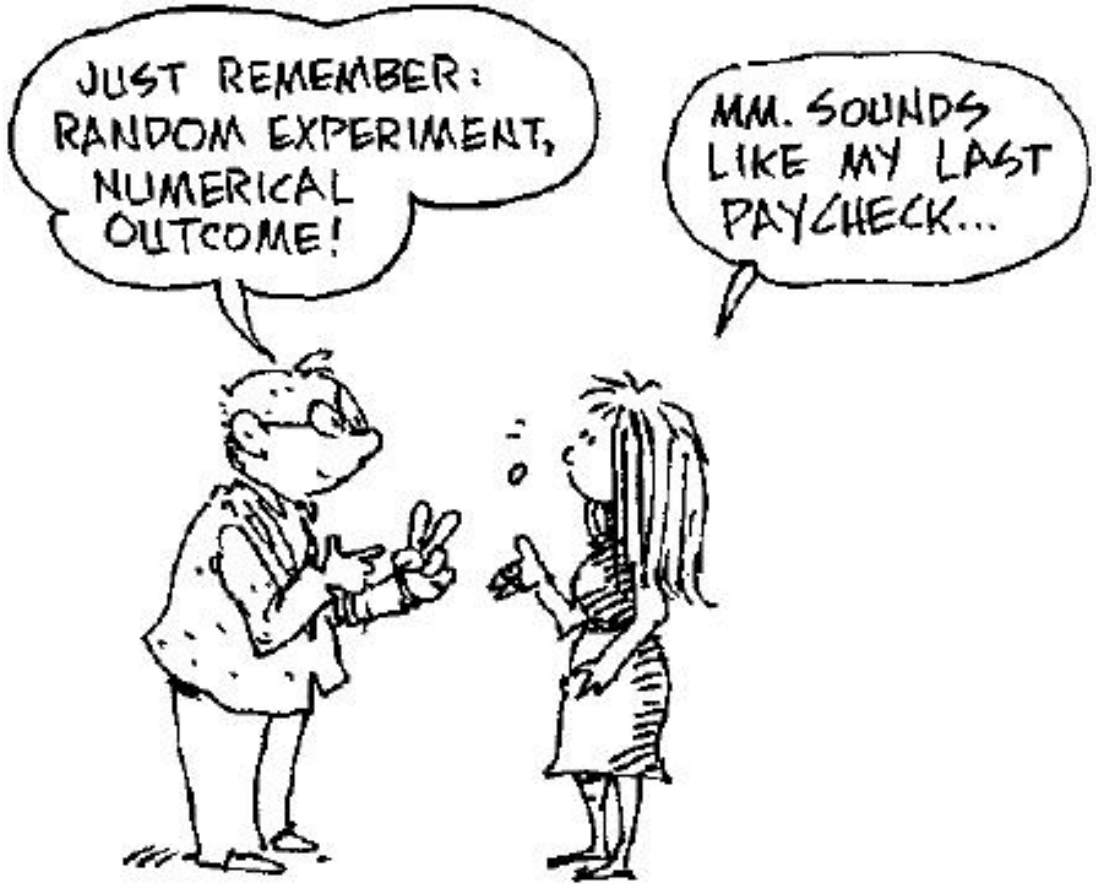
Preprost odgovor je:

- vzorec mora biti izbran *nepristransko*,
- vzorec mora biti *dovolj velik*.

Recimo, da merimo spremenljivko X, tako da n-krat naključno izberemo neko enoto in na njej izmerimo vrednost spremenljivke X. Postopku ustreza slučajni vektor

$$(X_1, X_2, \ldots, X_n),$$

ki mu rečemo *vzorec*. Število n je *velikost* vzorca.


JUST REMEMBER: RANDOM EXPERIMENT, NUMERICAL OUTCOME!
MM. SOUNDS LIKE MY LAST PAYCHECK...

. . . Vzorčenje

Ker v vzorcu merimo isto spremenljivko in posamezna meritev
ne sme vplivati na ostale, lahko predpostavimo:

1. vsi členi X_i vektorja imajo ***isto*** porazdelitev, kot spremenljivka X,
2. členi X_i so med seboj ***neodvisni***.

Takemu vzorcu rečemo *enostavni slučajni vzorec*.

Večina statistične teorije temelji na predpostavki,
da imamo opravka enostavnim slučajnim vzorcem.

Če je populacija končna, lahko dobimo enostavni slučajni vzorec,
tako da slučajno izbiramo (z vračanjem) enote z enako verjetnostjo.

Z vprašanjem, kako sestaviti dobre vzorce v praksi,
se ukvarja posebno področje statistike – *teorija vzorčenja*.

Načini vzorčenja

- ocena
  - priročnost
- naključno
  - enostavno: pri enostavnem naključnem vzorčenju je vsak član populacije izbran/vključen z *enako verjetnostjo.*
  - deljeno: razdeljen naključni vzorec dobimo tako, da razdelimo populacijo na disjunktne množice oziroma dele (razrede) in nato izberemo enostavne naključne vzorce za vsak del posebej.
  - grozdno: takšno vzorčenje je enostavno naključno vzorčenje skupin ali klastrov/grozdov elementov.

Osnovni izrek statistike

Spremenljivka X ima na populaciji G porazdelitev $F(x) = P(X < x)$. Toda tudi vsakemu vzorcu ustreza neka porazdelitev.

Za realizacijo vzorca $(x_1, x_2, x_3, \ldots, x_n)$ in $x \in \mathbb{R}$ postavimo

$$K(x) = \left|\{x_i : x_i < x, i = 1, \ldots, n\}\right| \quad \text{in} \quad V_n(x) = K(x)/n.$$

Slučajni spremenljivki $V_n(x)$ pravimo *vzorčna porazdelitvena funkcija*. Ker ima, tako kot tudi $K(x)$, $n+1$ možnih vrednosti $k/n,\ k = 0, \ldots, n$, je njena verjetnostna funkcija $B(n, F(x))$

$$P(V_n(x) = k/n) = \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.$$

. . . Osnovni izrek statistike

Če vzamemo n neodvisnih Bernoullijevih spremenljivk

$$Y_i(x) : \begin{pmatrix} 1 & 0 \\ F(x) & 1 - F(x) \end{pmatrix},$$

velja

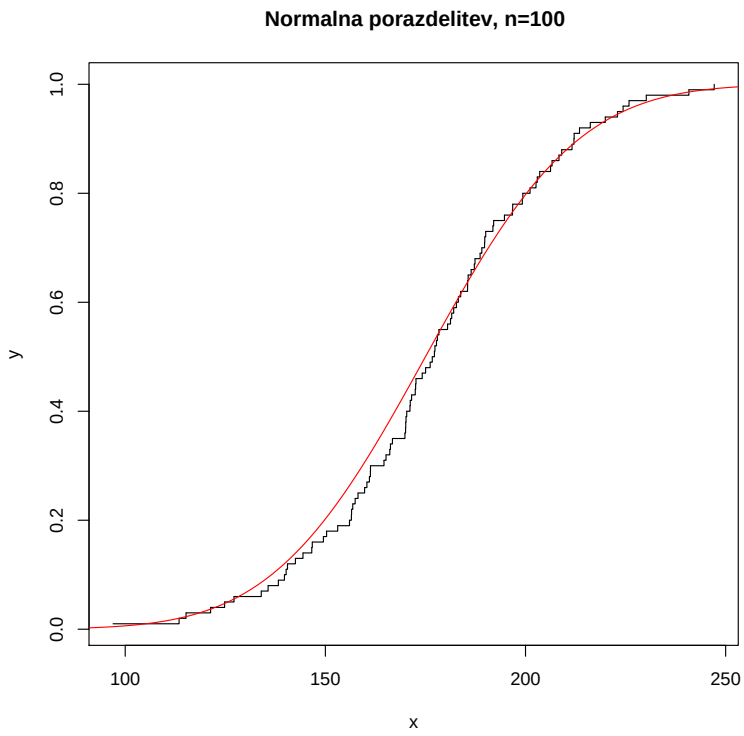
$$V_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i(x).$$

Krepki zakon velikih števil tedaj zagotavlja, da za vsak x velja

$$P\left(\lim_{n \to \infty} V_n(x) = F(x) \right) = 1.$$

To je v bistvu Borelov zakon, da relativna frekvenca dogodka $(X < x)$ skoraj gotovo konvergira proti verjetnosti tega dogodka.

. . . Osnovni izrek statistike



Velja pa še več. $V_n(x)$ je stopničasta funkcija, ki se praviloma dobro prilega funkciji $F(x)$.

Odstopanje med $V_n(x)$ in $F(x)$ lahko izmerimo s slučajno spremenljivko

$$D_n = \sup_{x \in \mathbb{R}} \left| V_n(x) - F(x) \right|$$

za $n = 1, 2, 3, \ldots$ Zanjo lahko pokažemo *osnovni izrek statistike*

$$P\left(\lim_{n \to \infty} D_n = 0 \right) = 1.$$

Torej se z rastjo velikosti vzorca $V_n(x)$ enakomerno vse bolje prilega funkciji $F(x)$ – vse bolje povzema razmere na celotni populaciji.

Vzorčne ocene

Najpogostejša parametra, ki bi ju radi ocenili sta:

sredina populacije μ glede na izbrano lastnost – matematično upanje spremenljivke X na populaciji; in

povprečni odklon od sredine σ – standardni odklon spremenljivke X na populaciji.

Statistike/ocene za te parametre so izračunane iz podatkov vzorca.
Zato jim tudi rečemo *vzorčne ocene*.

Sredinske mere

Kot sredinske mere se pogosto uporabljajo:

Vzorčni modus – najpogostejša vrednost (smiselna tudi za imenske).

Vzorčna mediana – srednja vrednost, glede na urejenost, (smiselna tudi za urejenostne).

Vzorčno povprečje – povprečna vrednost (smiselna za vsaj razmične)

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

Vzorčna geometrijska sredina – (smiselna za vsaj razmernostne)

$$G(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Mere razpršenosti

Za oceno populacijskega odklona uporabljamo *mere razpršenosti*.

Vzorčni razmah $= \max_i x_i - \min_i x_i.$

Vzorčna disperzija $s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

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

ter ustrezna *vzorčna odklona* s_0 in s.

Porazdelitve vzorčnih statistik

Denimo, da je v populaciji N enot in
da iz te populacije slučajno izbiramo n enot
v enostavni slučajni vzorec ali na kratko slučajni vzorec
(vsaka enota ima enako verjetnost, da bo izbrana v vzorec, tj. $1/N$).

Če hočemo dobiti slučajni vzorec, moramo izbrane
enote pred ponovnim izbiranjem vrniti v populacijo
(vzorec s ponavljanjem).

Če je velikost vzorca v primerjavi s populacijo majhna,
se ne pregrešimo preveč, če imamo za slučajni vzorec tudi vzorec,
ki nastane s slučajnim izbiranjem brez vračanja.

Predstavljajmo si, da smo iz populacije izbrali vse možne vzorce.
Dobili smo populacijo vseh možnih vzorcev.
Teh je v primeru enostavnih slučajnih vzorcev *s ponavljanjem* N^n;
kjer je N število enot v populaciji in n število enot v vzorcu.

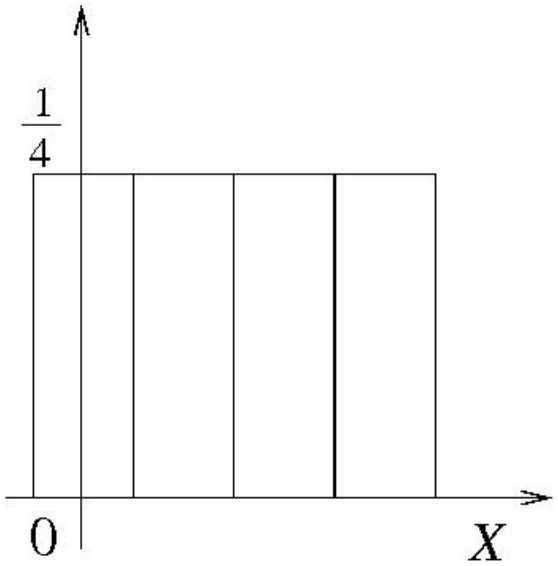
Število slučajnih vzorcev *brez ponavljanja* pa je

$\binom{N}{n}$, če *ne* upoštevamo vrstnega reda izbranih enot v vzorcu;
$\binom{N+n-1}{n}$, če upoštevamo vrstni red.

Primer: Vzemimo populacijo z $N = 4$ enotami, ki imajo naslednje vrednosti spremenljivke X:

$$0, 1, 2, 3$$

Grafično si lahko porazdelitev spremenljivke X predstavimo s histogramom:



in izračunamo populacijsko aritmetično sredino in varianco:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{3}{2},$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{5}{4}.$$

Sedaj pa tvorimo vse možne vzorce velikosti $n = 2$ s ponavljanjem, in na vsakem izračunajmo vzorčno aritmetično sredino $\overline{X}$:

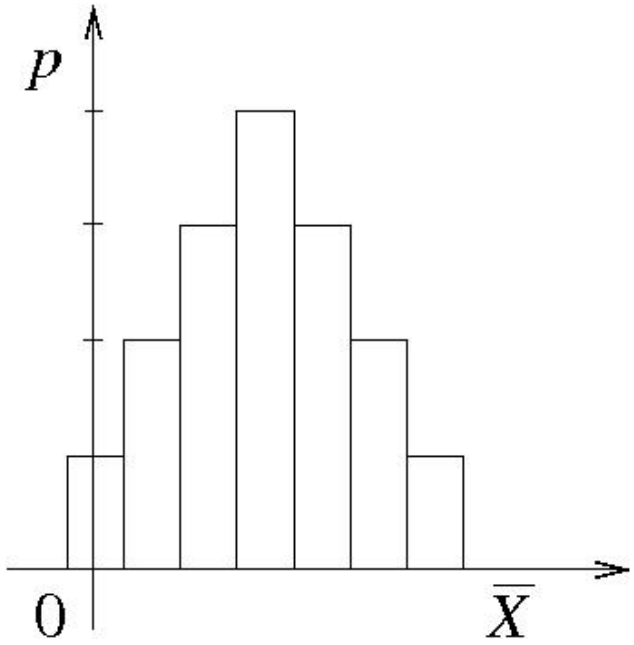
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{2}(X_1 + X_2).$$

vzorci	$\overline{X}$	vzorci	$\overline{X}$
$0,0$	0	$2,0$	1
$0,1$	$0,5$	$2,1$	$1,5$
$0,2$	1	$2,2$	2
$0,3$	$1,5$	$2,3$	$2,5$
$1,0$	$0,5$	$3,0$	$1,5$
$1,1$	1	$3,1$	2
$1,2$	$1,5$	$3,2$	$2,5$
$1,3$	2	$3,3$	3

Zapišimo verjetnostno shemo slučajne spremenljivke vzorčno povprečje $\overline{X}$:

$$\overline{X} : \begin{pmatrix} 0 & 0,5 & 1 & 1,5 & 2 & 2,5 & 3 \\ 1/16 & 2/16 & 3/16 & 4/16 & 3/16 & 2/16 & 1/16 \end{pmatrix}$$

Grafično jo predstavimo s histogramom:



... in izračunajmo matematično upanje ter disperzijo vzorčnega povprečja:

$$E(\overline{X}) = \sum_{i=1}^{m} \overline{X}_i \, p_i = \frac{0+1+3+6+6+5+3}{16} = \frac{3}{2},$$

$$D(\overline{X}) = \sum_{i=1}^{m} \left(\overline{X}_i - E(\overline{X}) \right)^2 p_i = \frac{5}{8}.$$

S tem primerom smo pokazali, da je statistika ‘vzorčna aritmetična sredina’ slučajna spremenljivka s svojo porazdelitvijo. Poglejmo, kaj lahko rečemo v splošnem o porazdelitvi vzorčnih aritmetičnih sredin.

Vzorčna porazdelitev povprečja

Centralni limitni izrek

Če je naključni vzorec velikosti n izbran iz populacije s končnim povprečjem μ in varianco σ^2, potem je lahko, če je n dovolj velik, vzorčna porazdelitev povprečja $\overline{y}$ aproksimirana z gostoto normalne porazdelitve.

Naj bo $y_1, y_2, \ldots, y_n$ naključni vzorec, ki je sestavljen iz n meritev populacije s končnim povprečjem μ in končnim standardnim odklonom σ. Potem sta povprečje in standardni odklon vzorčne porazdelitve $\overline{y}$ enaka

$$\mu_{\overline{Y}} = \mu, \qquad \text{in} \qquad \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

Hitrost centralne tendence pri CLI

Dokaz CLI je precej tehničen, kljub temu pa nam ne da občutka kako velik mora biti n, da se porazdelitev slučajne spremenljivke

$$X_1 + \cdots + X_n$$

približa normalni porazdelitvi.

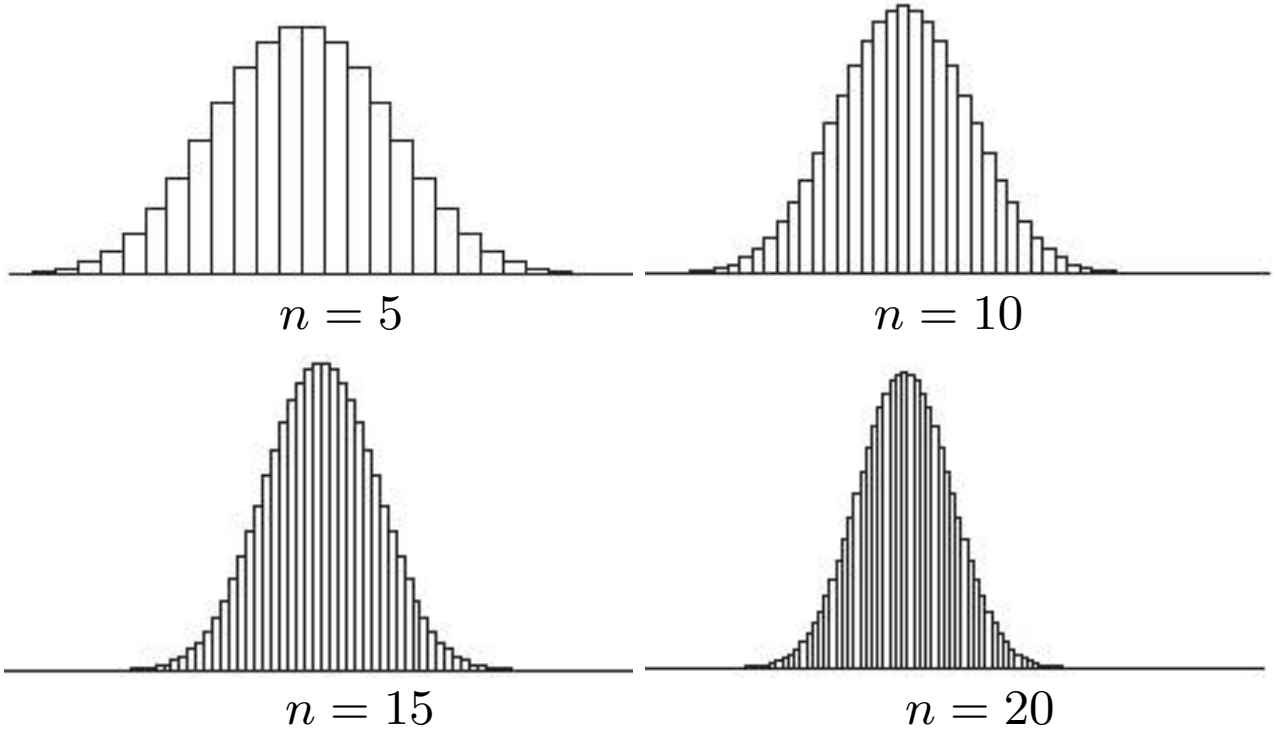
Hitrost približevanja k normalni porazdelitvi je odvisna od tega kako simetrična je porazdelitev.

To lahko potrdimo z eksperimentom: mečemo (ne)pošteno kocko, X_k naj bo vrednost, ki jo kocka pokaže pri k-tem metu.

Centralna tendenca za pošteno kocko

$p_1 = 1/6, \quad p_2 = 1/6, \quad p_3 = 1/6, \quad p_4 = 1/6, \quad p_5 = 1/6, \quad p_6 = 1/6.$

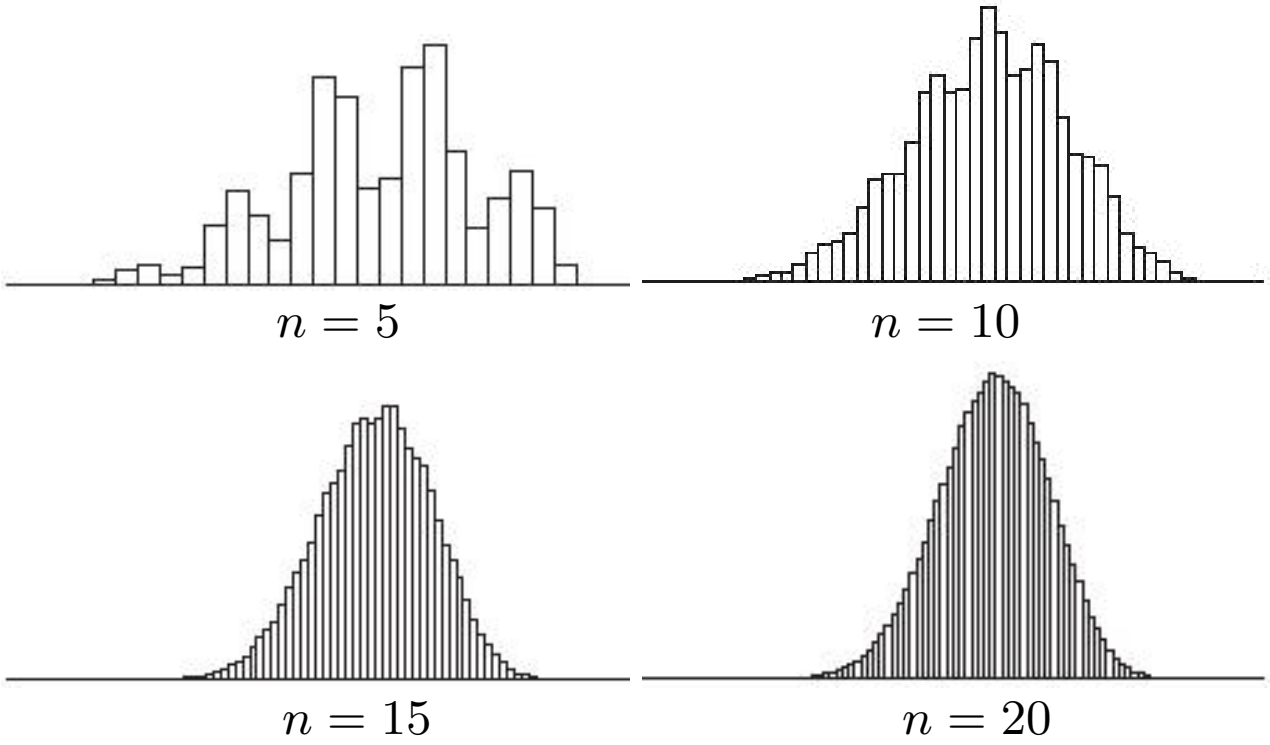
in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



Centralna tendenca za goljufivo kocko

$$p_1 = 0,2, \quad p_2 = 0,1, \quad p_3 = 0, \quad p_4 = 0, \quad p_5 = 0,3, \quad p_6 = 0,4.$$

in slučajno spremenljivko $X_1 + X_2 + \cdots + X_n$:



II.3. Cenilke

Vzorčna statistika

Vzorčna statistika je poljubna simetrična funkcija (tj. njena vrednost je neodvisna od permutacije argumentov) vzorca

$$Y = g(X_1, X_2, \ldots, X_n)$$

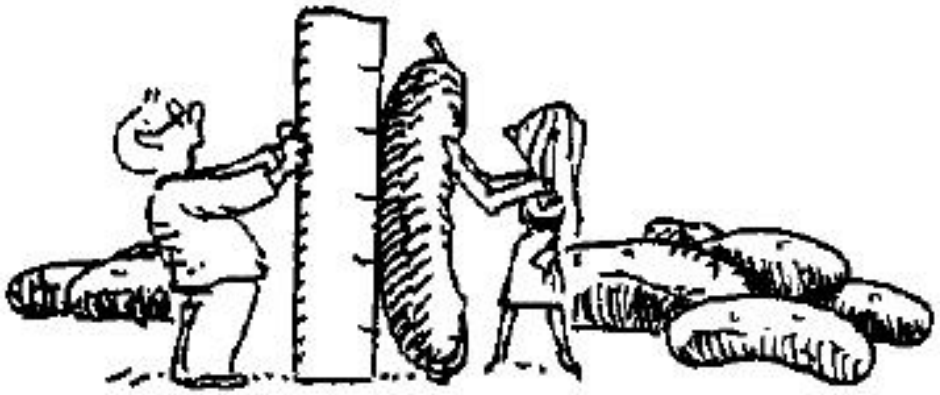
Tudi vzorčna statistika je slučajna spremenljivka, za katero lahko določimo porazdelitev iz porazdelitve vzorca.

Najzanimivejši sta značilni vrednosti

- njeno matematično upanje $\mathsf{E}Y$,
- standardni odklon σY, ki mu pravimo tudi *standardna napaka* statistike Y (angl. standard error – zato oznaka $\mathrm{SE}(Y)$).

(A) Vzorčno povprečje

Proizvajalec embalaže za kumare bi rad ugotovil **povprečno dolžino** kumarice (da se odloči za velikost embalaže), ne da bi izmeril dolžino čisto vsake.



Zato naključno izbere n kumar in izmeri njihove dolžine $X_1, \ldots, X_n$. Sedaj nam je že blizu ideja, da je vsaka dolžina X_i **slučajna spremenljivka** (numerični rezultat naključnega eksperimenta).

Če je μ (iskano/neznano) povprečje dolžin, in je σ standardni odklon porazdelitve dolžin kumar, potem velja

$$\mathsf{E}X_i = \mu, \qquad \text{in} \qquad \mathsf{D}X_i = \sigma^2,$$

za vsak i, ker bi X_i bila lahko dolžina katerekoli kumare.

. . . Vzorčno povprečje

Oglejmo si *vzorčno povprečje*, določeno z zvezo

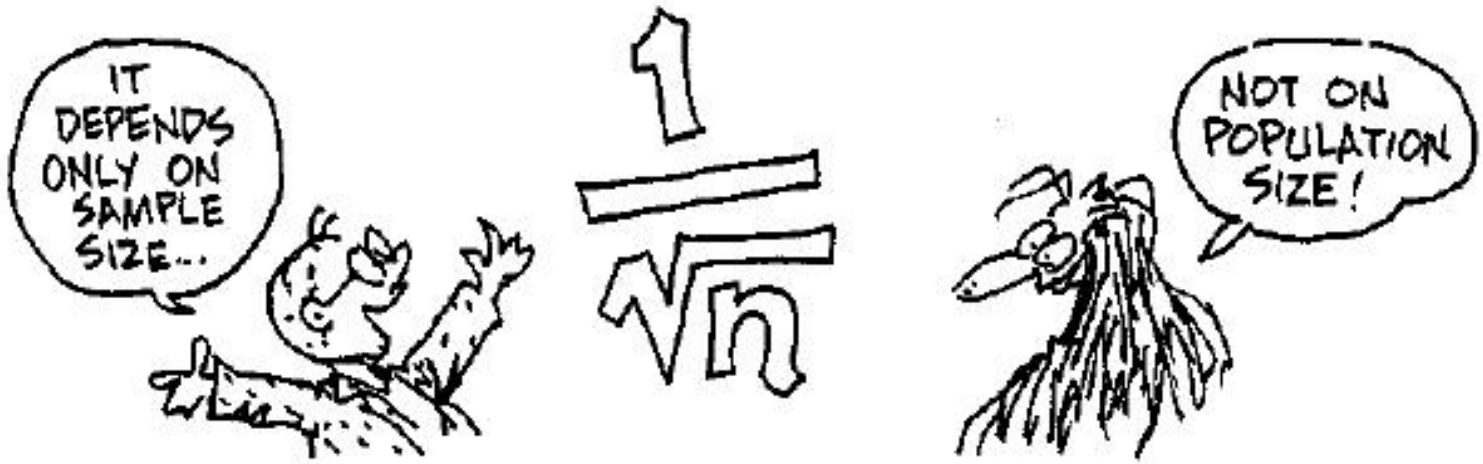
$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

ki je tudi slučajna spremenljivka. Tedaj je

$$\mathsf{E}\overline{X} = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}X_i = \mu$$

$$\mathsf{D}\overline{X} = \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{D}X_i = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Iz druge zveze vidimo, da standardna napaka $\sigma\overline{X} = \dfrac{\sigma}{\sqrt{n}}$ statistike $\overline{X}$ pada z naraščanjem velikosti vzorca, tj. $\overline{X} \to \mu$;
(enako nam zagotavlja tudi krepki zakon velikih števil).


IT DEPENDS ONLY ON SAMPLE SIZE...
$\frac{1}{\sqrt{n}}$
NOT ON POPULATION SIZE!

Denimo, da se spremenljivka X na populaciji porazdeljuje normalno $N(\mu, \sigma)$. Na vsakem vzorcu (s ponavljanjem) izračunamo vzorčno aritmetično sredino $\overline{X}$. Dokazati se da, da je **porazdelitev vzorčnih aritmetičnih sredin** normalna, kjer je

- matematično upanje vzorčnih aritmetičnih sredin enako aritmetični sredini spremenljivke na populaciji

$$E(\overline{X}) = \mu,$$

- standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Če tvorimo vzorce iz končne populacije brez vračanja, je standardni odklon vzorčnih aritmetičnih sredin

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Za dovolj velike vzorce ($n > 30$) je porazdelitev vzorčnih aritmetičnih sredin približno normalna, tudi če spremenljivka X ni normalno porazdeljena. Če se statistika X porazdeljuje vsaj približno normalno s standardno napako $\mathrm{SE}(X)$, potem se

$$Z = \frac{X - E(X)}{\mathrm{SE}(X)}$$

porazdeljuje standardizirano normalno.

Vzorčno povprečje in normalna porazdelitev

Naj bo $X : N(\mu, \sigma)$. Tedaj je $\sum_{i=1}^{n} X_i : N(n\mu, \sigma\sqrt{n})$ in dalje $\overline{X} : N(\mu, \sigma/\sqrt{n})$. Tedaj je vzorčna statistika

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n} : N(0, 1)$$

Kaj pa če porazdelitev X ni normalna? Izračun porazdelitve se lahko zelo zaplete. Toda pri večjih vzorcih ($n > 30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka Z porazdeljena skoraj standardizirano normalno. Vzorčno povprečje

$$\overline{X} = \frac{\sigma}{\sqrt{n}} Z + \mu$$

ima tedaj porazdelitev približno $N(\mu, \sigma/\sqrt{n})$.

Zgled

Odgovorimo na vprašanje: Kolikšna je verjetnost, da bo pri 36 metih igralne kocke povprečno število pik večje ali enako 4 ?

X je slučajna spremenljivka z vrednostmi 1,2,3,4,5,6 in verjetnostmi $1/6$. Zanjo je $\mu = 3,5$ in standardni odklon $\sigma = 1,7$. Vseh 36 ponovitev meta lahko obravnavamo kot slučajni vzorec velikost 36. Tedaj je

$$P(\overline{X} \geq 4) = P(Z \geq (4-\mu)\sqrt{n}/\sigma) = P(Z \geq 1,75) \approx 0,04.$$

```
> x <- 1:6
> m <- mean(x)
> s <- sd(x)*sqrt(5/6)
> z <- (4-m)*6/s
> p <- 1-pnorm(z)
> cbind(m,s,z,p)
         m        s        z          p
[1,] 3.5 1.707825 1.75662 0.03949129
```

(B) Vzorčna disperzija

Imejmo normalno populacijo $N(\mu, \sigma)$.

Kako bi določili porazdelitev za vzorčno disperzijo $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

ali popravljeno vzorčno disperzijo $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$?

Raje izračunamo porazdelitev za statistiko

$$\chi^2 = \frac{nS_0^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

. . . Vzorčna disperzija

Preoblikujemo jo lahko takole:

$$\begin{aligned}\chi^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu+\mu-\overline{X})^2 = \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{\sigma^2}(\mu-\overline{X})\sum_{i=1}^{n}(X_i-\mu)+\frac{n}{\sigma^2}(\mu-\overline{X})^2\end{aligned}$$

in, ker je $\sum_{i=1}^{n}(X_i-\mu)=n(\overline{X}-\mu)=-n(\mu-\overline{X})$, dalje

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\frac{1}{n}\left(\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}\right)^2=\sum_{i=1}^{n}Y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}Y_i\right)^2,$$

kjer so $Y_1, Y_2, \ldots, Y_n$ paroma neodvisne standardizirano normalno porazdeljene slučajne spremenljivke, $Y_i=\dfrac{X_i-\mu}{\sigma}$.

. . . Vzorčna disperzija

Porazdelitvena funkcija za χ^2 je

$$F_{\chi^2} = P(\chi^2 < z) = \int\int \ldots \int_{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 < z} e^{-(y_1^2+y_2^2+\ldots+y_n^2)/2} dy_n \ldots dy_1$$

z ustrezno ortogonalno transformacijo v nove spremenljivke $z_1, z_2, \ldots, z_n$ dobimo po nekaj računanja

$$F_{\chi^2} = \frac{1}{(2\pi)^{(n-1)/2}} \int\int \ldots \int_{\sum_{i=1}^{n-1} z_i^2 < z} e^{-(z_1^2+z_2^2+\ldots+z_{n-1}^2)/2} dz_{n-1} \ldots dz_1$$

Pod integralom je gostota vektorja $(Z_1, Z_2, \ldots, Z_{n-1})$ z neodvisnimi standardizirano normalnimi členi. Integral sam pa ustreza porazdelitveni funkciji vsote kvadratov $Z_1^2 + Z_2^2 + \ldots + Z_{n-1}^2$.
Tako je porazdeljena tudi statistika χ^2.

. . . Vzorčna disperzija

Kakšna pa je ta porazdelitev? Ker so tudi kvadrati $Z_1^2, Z_2^2, \ldots, Z_{n-1}^2$ med seboj neodvisni in porazdeljeni po zakonu $\chi^2(1)$,
je njihova vsota porazdeljena po zakonu $\chi^2(n-1)$.
Tako je torej porazdeljena tudi statistika χ^2.

Ker vemo, da je $\mathsf{E}\chi^2(n) = n$ in $\mathsf{D}\chi^2(n) = 2n$, lahko takoj izračunamo

$$\mathsf{E}S_0^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n} = \frac{(n-1)\sigma^2}{n} \qquad \mathsf{E}S^2 = \mathsf{E}\frac{\sigma^2\chi^2}{n-1} = \sigma^2$$

in

$$\mathsf{D}S_0^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n} = \frac{2(n-1)\sigma^4}{n^2} \qquad \mathsf{D}S^2 = \mathsf{D}\frac{\sigma^2\chi^2}{n-1} = \frac{2\sigma^4}{n-1}$$

. . . Vzorčna disperzija

Če je n zelo velik, je po centralnem limitnem izreku statistika χ^2 porazdeljena približno normalno in sicer po zakonu

$$N\big(n-1, \sqrt{2(n-1)}\big),$$

vzorčna disperzija S_0^2 približno po

$$N\left(\frac{(n-1)\sigma^2}{n}, \frac{\sqrt{2(n-1)}\,\sigma^2}{n}\right)$$

in popravljena vzorčna disperzija S^2 približno po

$$N\left(\sigma^2, \sqrt{\frac{2}{n-1}}\,\sigma^2\right).$$

Studentova porazdelitev

Pri normalno porazdeljeni slučajni spremenljivki X

je tudi porazdelitev $\overline{X}$ normalna, in sicer $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Statistika

$$Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$$

je potem porazdeljena standardizirano normalno.

Pri ocenjevanju parametra μ z vzorčnim povprečjem $\overline{X}$
to lahko uporabimo le, če poznamo σ;
sicer ne moremo oceniti standardne napake
– ne vemo, kako dobra je ocena za μ.

Kaj lahko naredimo, če σ ne poznamo?

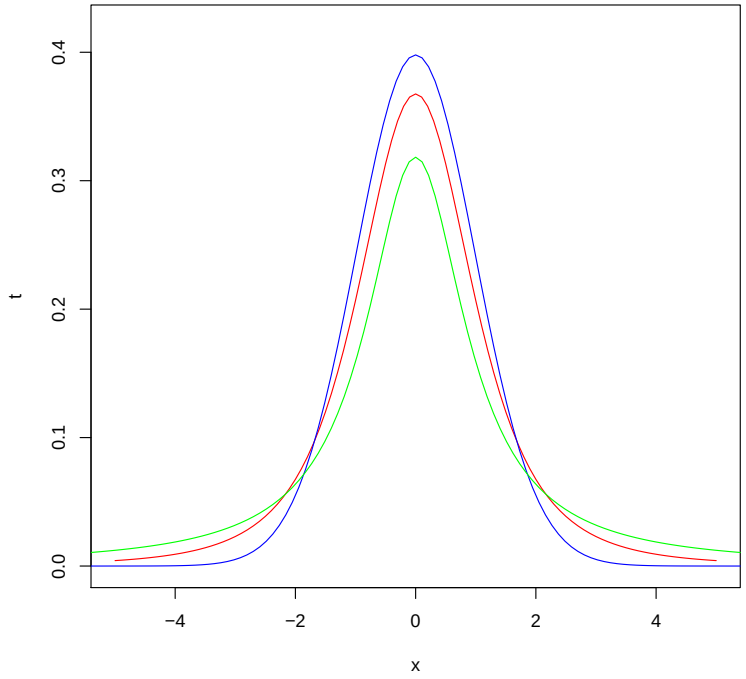
Parameter σ lahko ocenimo s S_0 ali S.
Toda S je slučajna spremenljivka in
porazdelitev statistike $\dfrac{\overline{X}-\mu}{S}\sqrt{n}$
ni več normalna $N(0,1)$
(razen, če je n zelo velik in S skoraj enak σ).

Kakšna je porazdelitev nove vzorčne statistike

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

'Student' in 1908

. . . Studentova porazdelitev



Leta 1908 je W.S. Gosset (1876-1937) pod psevdonimom 'Student' objavil članek, v katerem je pokazal, da ima statistika T porazdelitev $\boldsymbol{S(n-1)}$ z gostoto

$$p(t) = \frac{\left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}}{\sqrt{n-1}\, B\left(\frac{n-1}{2}, \frac{1}{2}\right)}$$

Tej porazdelitvi pravimo

Studentova porazdelitev

z $n-1$ *prostostnimi stopnjami*.

```
> plot(function(x) dt(x,df=3),-5,5,ylim=c(0,0.42),ylab="t",
  col="red")
> curve(dt(x,df=100),col="blue",add=T)
> curve(dt(x,df=1),col="green",add=T)
```

. . . Studentova porazdelitev

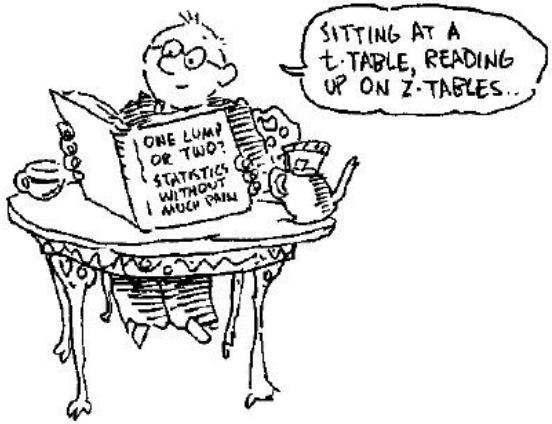
Za $S(1)$ dobimo Cauchyevo porazdelitev z gostoto

$$p(t) = \frac{1}{\pi(1+t^2)}$$

Za $n \to \infty$ pa gre $\frac{1}{\sqrt{n-1}\,B(\frac{n-1}{2},\frac{1}{2})} \to \sqrt{2\pi}$ in $(1+\frac{t^2}{n-1})^{-\frac{n}{2}} \to e^{-\frac{t^2}{2}}$.

Torej ima limitna porazdelitev gostoto

$$p(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$



standardizirane normalne porazdelitve.

Če zadnji sliki dodamo

```
> curve(dnorm(x),col="magenta",add=T)
```

ta pokrije modro krivuljo.

Fisherjeva ali Snedecorjeva porazdelitev

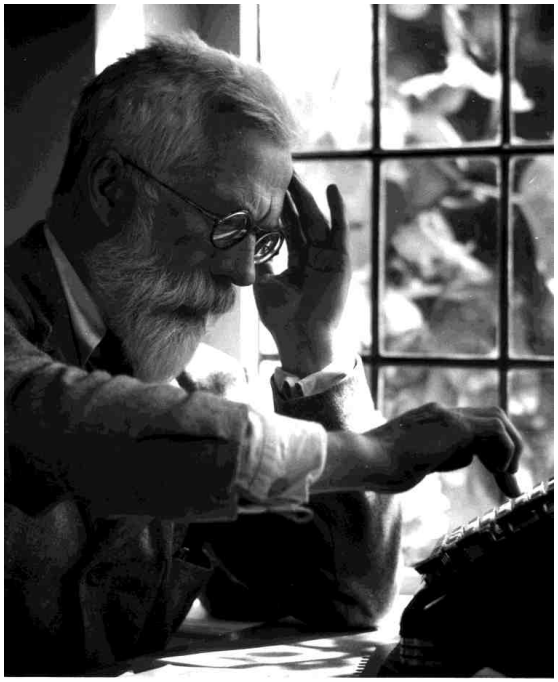
Poskusimo najti še porazdelitev kvocienta $Z = \frac{U}{V}$,

kjer sta $U : \chi^2(m)$ in $V : \chi^2(n)$ ter sta U in V neodvisni.

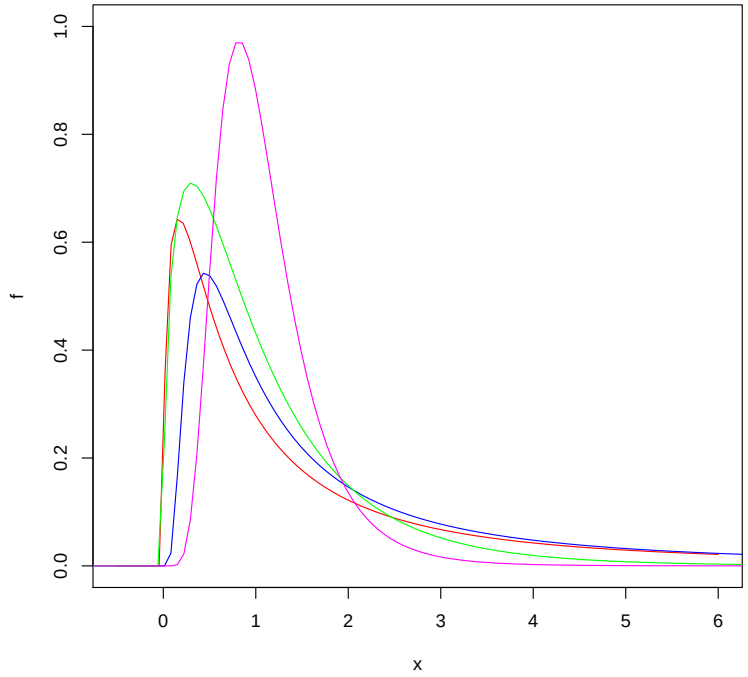
Z nekaj računanja (glej Hladnik) je mogoče pokazati, da je za $x > 0$ gostota ustrezne porazdelitve $F(m, n)$ enaka

$$p(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{\frac{m}{2}-1}}{(n + mx)^{\frac{m+n}{2}}}$$

in je enaka 0 drugje.



. . . Fisherjeva porazdelitev



Porazdelitvi $F(m,n)$ pravimo **Fisherjeva** ali tudi **Snedecorjeva porazdelitev** ***F*** *z* (m,n) *prostostnimi stopnjami*.

```
> plot(function(x) df(x,df1=3,df2=2),-0.5,6,ylim=c(0,1),ylab="f",
  col="red")
> curve(df(x,df1=20,df2=2),col="blue",add=T)
> curve(df(x,df1=3,df2=20),col="green",add=T)
> curve(df(x,df1=20,df2=20),col="magenta",add=T)
```

. . . Fisherjeva porazdelitev

Po zakonu $F(m-1, n-1)$ je na primer porazdeljena statistika

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

saj vemo, da sta spremenljivki

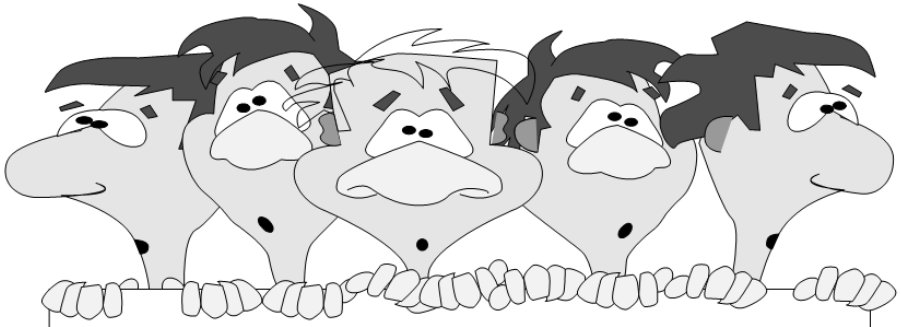
$$U = (m-1)S_X^2/\sigma_X^2 \quad \text{in} \quad V = (n-1)S_Y^2/\sigma_Y^2$$

porazdeljeni po χ^2 z $m-1$ oziroma $n-1$ prostostnimi stopnjami in sta neodvisni.

Velja še:

če je $U : F(m,n)$, je $1/U : F(n,m)$,

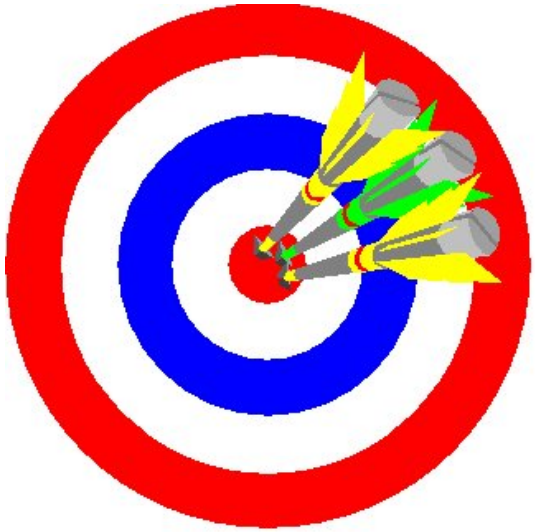
če je $U : S(n)$, je $U^2 : F(1,n)$.



Intervalno ocenjevanje in cenilke

Točkovne cenilke

Točkovna cenilka je pravilo ali formula,
ki nam pove, kako izračunati numerično oceno
parametra populacije na osnovi merjenj vzorca.



Število, ki je rezultat izračuna,
se imenuje **točkovna ocena**
(in mu ne moremo zaupati –
v smislu verjetnosti).

Cenilke

Cenilka parametra ζ je vzorčna statistika $C = C(X_1, X_2, \ldots, X_n)$,
katere porazdelitveni zakon *je odvisen* od parametra ζ,
njene vrednosti pa ležijo v prostoru parametrov.

Cenilka je simetrična funkcija:
– njena vrednost je enaka za vse permutacije argumentov.
Seveda je odvisna tudi od velikosti vzorca n .

Primeri: vzorčna mediana $\tilde{X}$ in vzorčno povprečje $\overline{X}$
sta cenilki za populacijsko povprečje μ;

popravljena vzorčna disperzija S^2 pa je cenilka
za populacijsko disperzijo σ^2 .

Doslednost

Cenilka C parametra ζ je **dosledna**, če z rastočim n zaporedje C_n verjetnostno konvergira k ζ, to je, za vsak $\varepsilon > 0$ velja

$$\lim_{n\to\infty} P(|C_n - \zeta| < \varepsilon) = 1.$$

Primeri: vzorčno povprečje $\overline{X}$ je dosledna cenilka za populacijsko povprečje μ. Tudi vsi **vzorčni začetni momenti**

$$Z_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

so dosledne cenilke ustreznih začetnih populacijskih momentov $z_k = \mathsf{E}X^k$, če le-ti obstajajo.

Vzorčna mediana $\tilde{X}$ je dosledna cenilka za populacijsko mediano.

. . . Doslednost

Če za $n \to \infty$ velja $\mathsf{E}C_n \to \zeta$ in $\mathsf{D}C_n \to 0$, je C_n dosledna cenilka parametra ζ.

To sprevidimo takole:

$$1-P(|C_n-\zeta| < \varepsilon) = P(|C_n-\zeta| \geq \varepsilon) \leq P(|C_n-\mathsf{E}C_n|+|\mathsf{E}C_n-\zeta| \geq \varepsilon).$$

Upoštevajmo še, da za dovolj velike n velja $|\mathsf{E}C_n - \zeta| < \varepsilon/2$, in uporabimo neenakost Čebiševa

$$P(|C_n - \mathsf{E}C_n| \geq \epsilon/2) \leq \frac{4\mathsf{D}C_n}{\epsilon^2} \to 0.$$

Primeri: Naj bo $X : N(\mu, \sigma)$. Ker za $n \to \infty$ velja $\mathsf{E}S_0^2 = \frac{(n-1)\sigma^2}{n} \to \sigma^2$ in $\mathsf{D}S_0^2 = \frac{2(n-1)\sigma^4}{n^2} \to 0$, je vzorčna disperzija S_0^2 dosledna cenilka za σ^2.

Nepristrana cenilka z najmanjšo varianco

Cenilka C_n parametra ζ je **nepristranska**, če je $\mathsf{E}C_n = \zeta$ (za vsak n);
in je **asimptotično nepristranska**, če je $\lim_{n\to\infty} \mathsf{E}C_n = \zeta$.

Količino $B(C_n) = \mathsf{E}C_n - \zeta$ imenujemo
pristranost (angl. *bias*) cenilke C_n.

Primeri: vzorčno povprečje $\overline{X}$ je nepristranska cenilka
za populacijsko povprečje μ ;

vzorčna disperzija S_0^2 je samo asimptotično nepristranska cenilka za σ^2,
popravljena vzorčna disperzija S^2 pa je nepristranska cenilka za σ^2.

Disperzija nepristranskih cenilk

Izmed nepristranskih cenilk istega parametra ζ je boljša tista, ki ima manjšo disperzijo – v povprečju daje bolj točne ocene.

Če je razred cenilk parametra ζ *konveksen* (vsebuje tudi njihove konveksne kombinacije), obstaja v bistvu ena sama cenilka z najmanjšo disperzijo:

Naj bo razred nepristranskih cenilk parametra ζ konveksen. Če sta C in C' nepristranski cenilki, obe z najmanjšo disperzijo σ^2, je $C = C'$ z verjetnostjo 1.

Za to poglejmo

$$\mathsf{D}(\frac{1}{2}(C+C')) = \frac{1}{4}(\mathsf{D}C+\mathsf{D}C'+2\mathsf{Cov}(C,C')) \leq (\frac{1}{2}(\sqrt{\mathsf{D}C}+\sqrt{\mathsf{D}C'}))^2 = \sigma^2$$

Ker sta cenilki minimalni, mora biti tudi $\mathsf{D}(\frac{1}{2}(C+C')) = \sigma^2$ in dalje $\mathsf{Cov}(C,C') = \sigma^2$ oziroma $r(C,C') = 1$. Torej je $C' = aC + b$, $a > 0$ z verjetnostjo 1. Iz $\mathsf{D}C = \mathsf{D}C'$ izhaja $a = 1$, iz $\mathsf{E}C = \mathsf{E}C'$ pa še $b = 0$.

Srednja kvadratična napaka

Včasih je celo bolje vzeti pristransko cenilko z manjšo disperzijo, kot jo ima druga, sicer nepristranska, cenilka z veliko disperzijo.

Mera *učinkovitosti* cenilk parametra ζ je *srednja kvadratična napaka*

$$q(C) = \mathsf{E}(C - \zeta)^2$$

Ker velja

$$q(C) = \mathsf{E}(C - \mathsf{E}C + \mathsf{E}C - \zeta)^2 = \mathsf{E}(C - \mathsf{E}C)^2 + (\mathsf{E}C - \zeta)^2$$

jo lahko zapišemo tudi v obliki

$$q(C) = \mathsf{D}C + B(C)^2$$

Za nepristranske cenilke je $B(C) = 0$ in zato $q(C) = \mathsf{D}C$.

Če pa je disperzija cenilke skoraj 0, je $q(C) \approx B(C)^2$.

Rao-Cramérjeva ocena

Naj bo f gostotna ali verjetnostna funkcija slučajne spremenljivke X in naj bo odvisna še od parametra ζ, tako da je $f(x;\zeta)$ njena vrednost v točki x. Združeno gostotno ali verjetnostno funkcijo slučajnega vzorca $(X_1, X_2, X_3, \ldots, X_n)$ označimo z L in ji pravimo *funkcija verjetja* (tudi *zanesljivosti*, angl. *likelihood*)

$$L(x_1, x_2, x_3, \ldots, x_n; \zeta) = f(x_1;\zeta) f(x_2;\zeta) f(x_3;\zeta) \cdots f(x_n;\zeta)$$

Velja (*): $\int \int \ldots \int L(x_1, x_2, \ldots, x_n; \zeta) dx_1 dx_2 \ldots dx_n = 1.$

$L(X_1, X_2, X_3, \ldots, X_n)$ je funkcija vzorca – torej slučajna spremenljivka.

Privzemimo, da je funkcija L vsaj dvakrat zvezno odvedljiva po ζ na nekem intervalu I in naj na tem intervalu tudi integral odvoda L po ζ enakomerno konvergira.

. . . Rao-Cramérjeva ocena

Odvajajmo enakost (*) po ζ in upoštevajmo $\frac{\partial \ln L}{\partial \zeta} = \frac{1}{L}\frac{\partial L}{\partial \zeta}$ pa dobimo

$$\int \int \ldots \int \frac{\partial \ln L}{\partial \zeta} L \, dx_1 dx_2 \ldots dx_n = 0$$

kar lahko tolmačimo kot $\mathsf{E}\dfrac{\partial \ln L}{\partial \zeta} = 0.$

Naj bo sedaj C nepristranska cenilka parametra ζ, torej $\mathsf{E}C = \zeta$, oziroma zapisano z integrali $\int \int \ldots \int CL \, dx_1 dx_2 \ldots dx_n = \zeta.$

Ker C ni odvisna od ζ, dobimo z odvajanjem po ζ:

$$\int \int \ldots \int C \frac{\partial \ln L}{\partial \zeta} L \, dx_1 dx_2 \ldots dx_n = 1$$

kar pomeni $\mathsf{E}(C\dfrac{\partial \ln L}{\partial \zeta}) = 1.$

. . . Rao-Cramérjeva ocena

Če to enakost združimo s prejšnjo (pomnoženo s ζ), dobimo:

$$\mathsf{E}\left((C-\zeta)\,\frac{\partial \ln L}{\partial \zeta}\right) = 1$$

Od tu po $(\mathsf{E}XY)^2 \le \mathsf{E}X^2\mathsf{E}Y^2$ izhaja naprej

$$1 = \left(\mathsf{E}\left((C-\zeta)\,\frac{\partial \ln L}{\partial \zeta}\right)\right)^2 \le \mathsf{E}(C-\zeta)^2\,\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2 = \mathsf{D}C\;\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2$$

kar da *Rao-Cramérjevo oceno*

$$\mathsf{D}C \ge \left(\mathsf{E}\left(\frac{\partial \ln L}{\partial \zeta}\right)^2\right)^{-1} = \left(-\mathsf{E}\frac{\partial^2 \ln L}{\partial \zeta^2}\right)^{-1} = \left(n\mathsf{E}\left(\frac{\partial \ln f}{\partial \zeta}\right)^2\right)^{-1}$$

Učinkovitost cenilk

Rao-Cramérjeva ocena da absolutno spodnjo mejo disperzije za vse nepristranske cenilke parametra ζ (v dovolj gladkih porazdelitvah).
Ta meja ni nujno dosežena. Cenilka, ki jo doseže, se imenuje *najučinkivitejša cenilka* parametra ζ in je ena sama (z verjetnostjo 1).

Kdaj pa je ta spodnja meja dosežena?

V neenakosti $(\mathsf{E}XY)^2 \leq \mathsf{E}X^2\mathsf{E}Y^2$, ki je uporabljena v izpeljavi Rao-Cramérjeve ocene, velja enakost natanko takrat, ko je $Y = cX$ z verjetnostjo 1.

. . . Učinkovitost cenilk

Torej velja v Rao-Cramérjevi oceni enakost natanko takrat, ko je

$$\frac{\partial \ln L}{\partial \zeta} = A(\zeta)(C - \zeta)$$

kjer je $A(\zeta)$ konstanta, odvisna od ζ in neodvisna od vzorca.

Zato je tudi

$$(\mathsf{D}C)^{-1} = \mathsf{E}(\frac{\partial \ln L}{\partial \zeta})^2 = A(\zeta)^2 \mathsf{E}(C - \zeta)^2 = A(\zeta)^2 \mathsf{D}C$$

oziroma končno

$$\mathsf{D}C = |A(\zeta)|^{-1}$$

Najučinkovitejše cenilke za parametre normalne porazdelitve

Naj bo $X : N(\mu, \sigma)$. Tedaj je

$$L = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-((\frac{X_1-\mu}{\sigma})^2+\cdots+(\frac{X_n-\mu}{\sigma})^2)/2}$$

in

$$\ln L = \ln \frac{1}{(2\pi)^{n/2}\sigma^n} - ((\frac{X_1-\mu}{\sigma})^2 + \cdots + (\frac{X_n-\mu}{\sigma})^2)/2$$

ter dalje

$$\frac{\partial \ln L}{\partial \mu} = \frac{X_1-\mu}{\sigma^2} + \cdots + \frac{X_n-\mu}{\sigma^2} = \frac{n}{\sigma^2}(\overline{X}-\mu)$$

Torej je vzorčno povprečje $\overline{X}$ najučinkovitejša cenilka za μ z disperzijo $\mathsf{D}\overline{X} = \frac{\sigma^2}{n}$.

. . . normalna porazdelitev

Prvi člen v izrazu za $\ln L$ lahko zapišemo tudi $-\frac{n}{2}(\ln 2\pi + \ln \sigma^2)$. Tedaj je, če privzamemo, da je μ znano število

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}((X_1 - \mu)^2 + \cdots + (X_n - \mu)^2) = \frac{n}{2\sigma^4}(S_\mu^2 - \sigma^2)$$

To pomeni, da je $S_\mu^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ najučinkovitejša cenilka za parameter σ^2 z disperzijo $\mathsf{D}S_\mu^2 = \frac{2\sigma^4}{n}$.

Poissonova porazdelitev

Za Poissonovo porazdelitev $P(\lambda)$ s parametrom λ, $p_k = \lambda^k \frac{e^{-\lambda}}{k!}$ je

$$L = e^{-n\lambda} \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}$$

in dalje

$$\ln L = -n\lambda + (x_1 + \cdots + x_n) \ln \lambda - \ln(x_1! \cdots x_n!)$$

ter končno

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1 + \cdots + x_n}{\lambda} = \frac{n}{\lambda}(\overline{X} - \lambda)$$

Najučinkovitejša cenilka za parameter λ je $\overline{X}$ z disperzijo $\mathsf{D}\overline{X} = \frac{\lambda}{n}$.

Učinkovitost cenilke

Naj bo C_0 najučinkovitejša cenilka parametra ζ in C kaka druga nepristranska cenilka. Tedaj je *učinkovitost* cenilke C določena s predpisom

$$e(C) = \frac{\mathsf{D}C_0}{\mathsf{D}C}$$

Učinkovitost najučinkovitejše cenilke je $e(C_0) = 1$.

Če najučinkovitejša cenilka ne obstaja, vzamemo za vrednost $\mathsf{D}C_0$ desno stran v Rao-Cramérjevi oceni.

Primer: Naj bo $X : N(\mu, \sigma)$. Pri velikih n-jih je vzorčna mediana $\tilde{X}$ – ocena za μ, porazdeljena približno po $N(\mu, \sigma\sqrt{\frac{\pi}{2n}})$. Torej je

$$e(\tilde{X}) = \frac{\mathsf{D}\overline{X}}{\mathsf{D}\tilde{X}} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} \approx 0.64$$

. . . Učinkovitost cenilke

Primer: Naj bo $X : N(\mu, \sigma)$. Če poznamo μ, je najučinkovitejša cenilka za σ^2 statistika S^2_μ z disperzijo $\mathsf{D}S^2_\mu = \frac{2\sigma^4}{n}$. Popravljena vzorčna disperzija S^2 pa je nepristranska cenilka istega parametra z disperzijo $\mathsf{D}S^2 = \frac{2\sigma^4}{n-1}$. Torej je učinkovitost S^2

$$e(S^2) = \frac{\mathsf{D}S^2_\mu}{\mathsf{D}S^2} = \frac{\frac{2\sigma^4}{n}}{\frac{2\sigma^4}{n-1}} = \frac{n-1}{n}$$

Iz tega vidimo, da $e(S^2) \to 1$, ko $n \to \infty$. Pravimo, da je cenilka S^2 *asimptotično najučinkovitejša cenilka* za σ^2.

Metoda momentov

Recimo, da je za zvezno slučajno spremenljivko X njena gostota f odvisna od m parametrov $f(x;\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_m)$ in naj obstajajo momenti

$$z_k = z_k(\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_m) = \int_{-\infty}^{\infty} x^k f(x;\zeta_1,\zeta_2,\zeta_3,\ldots,\zeta_m)dx$$

za $k = 1, 2, 3, \ldots, m$. Če se dajo iz teh enačb enolično izračunati parametri ζ_1, ζ_2, ζ_3, …,ζ_m kot funkcije momentov z_1, z_2, z_3, …, z_m

$$\zeta_k = \varphi_k(z_1, z_2, z_3, \ldots, z_m)$$

potem so

$$C_k = \varphi_k(Z_1, Z_2, Z_3, \ldots, Z_m)$$

cenilke parametrov ζ_k po *metodi momentov*. k-ti vzorčni začetni moment $Z_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ je cenilka za ustrezni populacijski moment z_k.

Cenilke, ki jih dobimo po metodi momentov so dosledne.

. . . Metoda momentov

Naj bo $X : N(\mu, \sigma)$. Tedaj je $z_1 = \mu$ in $z_2 = \sigma^2 + \mu^2$.

Od tu dobimo $\mu = z_1$ in $\sigma^2 = z_2 - z_1^2$.

Ustrezni cenilki sta $Z_1 = \overline{X}$ za μ in

$$Z_2 - Z_1^2 = \overline{X^2} - \overline{X}^2 = S_0^2$$

za σ^2 – torej vzorčno povprečje in disperzija.

Metoda največjega verjetja

Funkcija verjetja

$$L(x_1, x_2, x_3, \ldots, x_n; \zeta) = f(x_1; \zeta) f(x_2; \zeta) f(x_3; \zeta) \cdots f(x_n; \zeta)$$

je pri danih $x_1, x_2, x_3, \ldots, x_n$ odvisna še od parametra ζ. Izberemo tak ζ, da bo funkcija L dosegla največjo vrednost. Če je L vsaj dvakrat zvezno odvedljiva, mora veljati $\frac{\partial L}{\partial \zeta} = 0$ in $\frac{\partial^2 L}{\partial \zeta^2} < 0$. Največja vrednost parametra je še odvisna od $x_1, x_2, x_3, \ldots, x_n$:
$\zeta_{max} = \varphi(x_1, x_2, x_3, \ldots, x_n)$. Tedaj je cenilka za parameter ζ enaka

$$C = \varphi(X_1, X_2, X_3, \ldots, X_n)$$

Metodo lahko posplošimo na večje število parametrov.

Pogosto raje iščemo maksimum funkcije $\ln L$.

Če najučinkovitejša cenilka obstaja, jo dobimo s to metodo.

. . . Metoda največjega verjetja - binomska

Naj bo $X : B(1, p)$. tedaj je $f(x; p) = p^x(1-p)^{1-x}$, kjer je $x = 0$ ali $x = 1$. Ocenjujemo parameter p. Funkcija verjetja ima obliko $L = p^x(1-p)^{n-x}$, kjer je sedaj $x \in \{0, 1, 2, \ldots, n\}$. Ker je $\ln L = x \ln p + (n-x)\ln(1-p)$, dobimo

$$\frac{\partial \ln L}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p},$$

ki je enak 0 pri $p = \frac{x}{n}$. Ker je v tem primeru $\frac{\partial^2 \ln L}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0$, je v tej točki maksimum. Cenilka po metodi največjega verjetja je torej $P = \frac{X}{n}$, kjer je X binomsko porazdeljena spremenljivka – frekvenca v n ponovitvah. Cenilka P je nepristranska, saj je $\mathsf{E}P = \frac{\mathsf{E}X}{n} = p$. Ker za $n \to \infty$ gre $\mathsf{D}P = \frac{\mathsf{D}X}{n^2} = \frac{p(1-p)}{n} \to 0$, je P dosledna cenilka. P je tudi najučinkovitejša $\frac{\partial \ln L}{\partial p} = \frac{X}{p} - \frac{n-X}{1-p} = \frac{n}{p(1-p)}(\frac{X}{n} - p) = \frac{n}{p(1-p)}(P - p)$.

. . . Metoda največjega verjetja - Poissonova

Za Poissonovo porazdelitev $P(\lambda)$ s parametrom λ, $p_x = \lambda^x \frac{e^{-\lambda}}{x!}$ je

$$L = e^{-n\lambda} \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}$$

in dalje

$$\ln L = -n\lambda + (x_1 + \cdots + x_n) \ln \lambda - \ln(x_1! \cdots x_n!)$$

ter končno

$$\frac{\partial \ln L}{\partial \lambda} = -n + \frac{x_1 + \cdots + x_n}{\lambda} = \frac{n}{\lambda}(\overline{X} - \lambda)$$

Odvod je enak 0 za $\lambda = \overline{X}$. Drugi odvod v tej točki je $\frac{\partial^2 \ln L}{\partial \lambda^2} = -\frac{x_1+\cdots+x_n}{\lambda^2} < 0$. V točki je maksimum.

Cenilka za λ po metodi največjega verjetja je vzorčno povprečje $\overline{X}$.

Je tudi najučinkovitejša cenilka za λ z disperzijo $\mathsf{D}\overline{X} = \frac{\lambda}{n}$.

Porazdelitev vzorčnih aritmetičnih sredin

Primer: Denimo, da se spremenljivka inteligenčni kvocient na populaciji porazdeljuje normalno z aritmetično sredino $\mu = 100$ in standardnim odklonom $\sigma = 15$.

$$X : N(100, 15)$$

Denimo, da imamo vzorec velikosti $n = 225$. Tedaj se vzorčne aritmetične sredine porazdeljujejo normalno

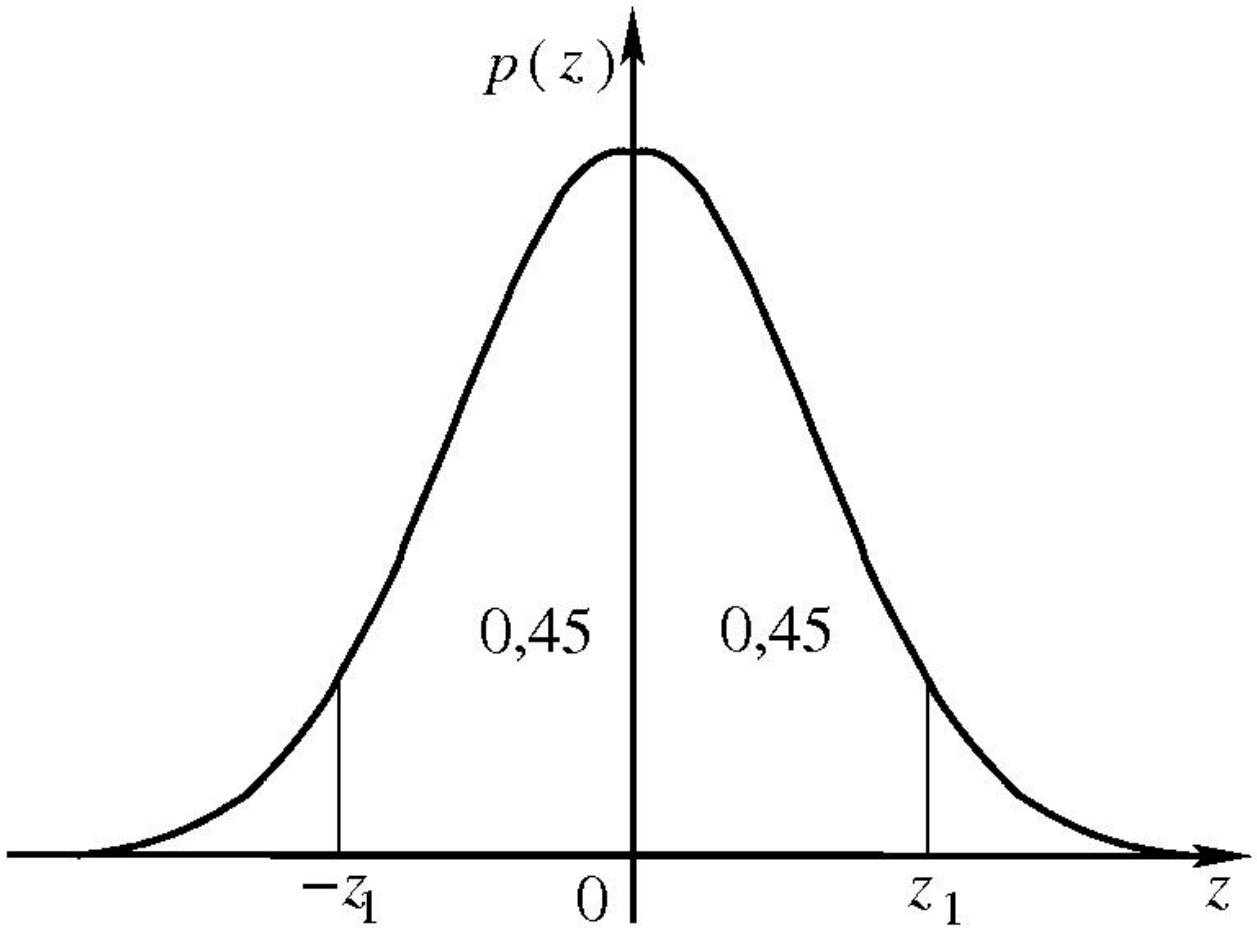
$$\overline{X} : N(100, \frac{15}{\sqrt{225}}) = N(100, 1)$$

Izračunajmo, kolikšne vzorčne aritmetične sredine ima 90% vzorcev (simetrično na povprečje). 90% vzorčnih aritmetičnih sredin se nahaja na intervalu:

$$P(\overline{X}_1 < \overline{X} < \overline{X}_2) = 0,90$$

$$P(-z_1 < z < z_1) = 0,90 \quad \Longrightarrow \quad 2\Phi(z_1) = 0,90$$

$$\Phi(z_1) = 0,45 \quad \Longrightarrow \quad z_1 = 1,65$$


p (z)
0,45
0,45
−z1
0
z1
z

Potem se vzorčne aritmetične sredine nahajajo v intervalu

$$P\left(\mu - z_1 \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_1 \frac{\sigma}{\sqrt{n}}\right) = 0,90$$

oziroma konkretno

$$P\left(100 - 1,65 < \overline{X} < 100 + 1,65\right) = 0,90$$

90% vseh slučajnih vzorcev velikosti 225 enot bo imelo povprečja za inteligenčni kvocient na intervalu

$$(98,35\,;\,101,65).$$

Lahko preverimo, da bi bil ta interval v primeru večjega vzorca ožji.
Npr. v primeru vzorcev velikosti $n = 2500$ je ta interval

$$P\left(100 - 1,65\frac{15}{\sqrt{2500}} < \overline{X} < 100 + 1,65\frac{15}{\sqrt{2500}}\right) = 0,90$$

oziroma

$$(99,5\,;\,100,5).$$

Porazdelitev vzorčnih deležev

Denimo, da želimo na populaciji oceniti delež enot π z določeno lastnostjo.

Zato na vsakem vzorcu poiščemo vzorčni delež p.

Pokazati se da, da se za dovolj velike slučajne vzorce s ponavljanjem (za deleže okoli 0,5 je dovolj 20 enot ali več) vzorčni deleži porazdeljujejo približno normalno z

- aritmetično sredino vzorčnih deležev, ki je enaka deležu na populaciji

$$\mathsf{E}p = \pi,$$

- standardnim odklonom vzorčnih deležev

$$\mathsf{SE}(p) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Za manjše vzorce se vzorčni deleži porazdeljujejo binomsko.

Cenilka populacijskega deleža je nepristranska cenilka, ker velja

$$\mathsf{E}p = \pi.$$

Primer: V izbrani populaciji prebivalcev je polovica žensk $\pi = 0,5$. Če tvorimo vzorce po $n = 25$ enot, nas zanima, kolikšna je verjetnost, da je v vzorcu več kot 55 % žensk? To pomeni, da iščemo verjetnost $P(p > 0,55)$.

Vzorčni deleži p se porazdeljujejo približno normalno

$$p \ : \ N\left(0,5\ ,\ \sqrt{\frac{\pi(1-\pi)}{n}}\right) = N\left(0,5\ ,\ \sqrt{\frac{0,5 \cdot 0,5}{25}}\right) = N(0,5\ ,\ 0,1).$$

$$P(p > 0,55) = P\left(Z > \frac{0,55 - 0,5}{0,1}\right) = P(Z > 0,5) =$$

$$= 0,5 - \Phi(0,5) = 0,5 - 0,1915 = 0,3085.$$

Rezultat pomeni, da lahko pričakujemo, da bo pri približno 31% vzorcev delež žensk večji od 0,55.

Poglejmo, kolikšna je ta verjetnost, če bi tvorili vzorce velikosti $n = 2500$ enot:

$$P(p > 0,55) = P\left(Z > \frac{0,55 - 0,5}{\sqrt{\frac{0,5(1-0,5)}{2500}}}\right)$$

$$= P(Z > 5) = 0,5 - \Phi(5) = 0,5 - 0,5 = 0.$$

V 10-krat večjih vzorcih kot prej ne moremo pričakovati več kot 55% žensk.

Porazdelitev razlik vzorčnih aritmetičnih sredin

Denimo, da imamo dve populaciji velikosti N_1 in N_2 in se spremenljivka X na prvi populaciji porazdeljuje normalno $N(\mu_1, \sigma)$, na drugi populaciji pa $N(\mu_2, \sigma)$ (standardna odklona sta na obeh populacijah enaka!).

V vsaki od obeh populacij tvorimo neodvisno slučajne vzorce velikosti n_1 in n_2. Na vsakem vzorcu (s ponavljanjem) prve populacije izračunamo vzorčno aritmetično sredino $\overline{X}_1$ in podobno na vsakem vzorcu druge populacije $\overline{X}_2$.

Dokazati se da, da je porazdelitev razlik vzorčnih aritmetičnih sredin normalna, kjer je

- matematično upanje razlik vzorčnih aritmetičnih sredin enako

$$\mathsf{E}(\overline{X}_1 - \overline{X}_2) = \mathsf{E}\overline{X}_1 - \mathsf{E}\overline{X}_2 = \mu_1 - \mu_2,$$

- disperzija razlik vzorčnih aritmetičnih sredin enaka

$$\mathsf{D}(\overline{X}_1 - \overline{X}_2) = \mathsf{D}\overline{X}_1 + \mathsf{D}\overline{X}_2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \cdot \frac{n_1 + n_2}{n_1 n_2}.$$

Primer: Dvema populacijama študentov na neki univerzi (tehnikom in družboslovcem) so izmerili neko sposobnost s povprečjema $\mu_t = 70$ in $\mu_d = 80$ točk in standardnim odklonom, ki je na obeh populacijah enak, $\sigma = 7$ točk.

Kolikšna je verjetnost, da je aritmetična sredina slučajnega vzorca družboslovcev ($n_d = 36$) večja za več kot 12 točk od aritmetične sredine vzorca tehnikov ($n_t = 64$)? Zanima nas torej verjetnost:

$$P(\overline{X}_d - \overline{X}_t > 12) = P\left(Z > \frac{12-10}{7\sqrt{\frac{36+64}{36\cdot 64}}}\right)$$

$$= P(Z > 1,37) = 0,5 - \Phi(1,37) =$$

$$= 0,5 - 0,4147 = 0,0853.$$

Torej, približno 8,5% parov vzorcev je takih, da je povprečje družboslovcev večje od povprečja tehnikov za 12 točk.

Porazdelitev razlik vzorčnih deležev

Podobno kot pri porazdelitvi razlik vzorčnih aritmetičnih sredin naj bosta dani dve populaciji velikosti N_1 in N_2 z deležema enot z neko lastnostjo π_1 in π_2. Iz prve populacije tvorimo slučajne vzorce velikosti n_1 in na vsakem izračunamo delež enot s to lastnostjo p_1.

Podobno naredimo tudi na drugi populaciji: tvorimo slučajne vzorce velikosti n_2 in na njih dolocimo deleže p_2. Pokazati se da, da se za dovolj velike vzorce razlike vzorčnih deležev porazdeljujejo približno normalno z

- matematičnim upanjem razlik vzorčnih deležev

$$\mathsf{E}(p_1 - p_2) = \mathsf{E}p_1 - \mathsf{E}p_2 = \pi_1 - \pi_2,$$

- disperzijo razlik vzorčnih deležev

$$\mathsf{D}(p_1 - p_2) = \mathsf{D}p_1 + \mathsf{D}p_2 = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}.$$

II.4. Intervali zaupanja

Denimo, da s slučajnim vzorcem ocenjujemo parameter γ.
Poskušamo najti statistiko g, ki je nepristranska, tj. $\mathsf{E}g = \gamma$
in se na vseh možnih vzorcih vsaj približno normalno porazdeljuje s standardno napako $\mathsf{SE}(g)$. Nato poskušamo najti interval, v katerem se bo z dano gotovostjo $(1 - \alpha)$ nahajal ocenjevani parameter:

$$P(a < \gamma < b) = 1 - \alpha$$

a je spodnja meja zaupanja, b je zgornja meja zaupanja,
α verjetnost tveganja oziroma $1 - \alpha$ verjetnost gotovosti.

Ta interval imenujemo **interval zaupanja** in ga interpretiramo takole:
z verjetnostjo tveganja α se parameter γ nahaja v tem intervalu.

Konstruirajmo interval zaupanja.

Na osnovi omenjenih predpostavk o porazdelitvi statistike g lahko zapišemo, da se statistika

$$Z = \frac{g - \mathsf{E}g}{\mathsf{SE}(g)} = \frac{g - \gamma}{\mathsf{SE}(g)}$$

porazdeljuje standardizirano normalno $N(0, 1)$.

Če tveganje α porazdelimo polovico na levo in polovico na desno na konce normalne porazdelitve, lahko zapišemo

$$P\left(-z_{\alpha/2} < \frac{g-\gamma}{\mathsf{SE}(g)} < z_{\alpha/2}\right) = 1-\alpha.$$

Po ustrezni preureditvi lahko izpeljemo naslednji interval zaupanja za parameter γ

$$P\left(g - z_{\alpha/2}\,\mathsf{SE}(g) < \ \gamma\ < g + z_{\alpha/2}\,\mathsf{SE}(g)\right) = 1-\alpha$$

$z_{\alpha/2}$ je določen le s stopnjo tveganja α.

Vrednosti $z_{\alpha/2}$ lahko razberemo iz tabele za verjetnosti za standardizirano normalno porazdelitev, ker velja

$$\Phi(z_{\alpha/2}) = 0,5 - \frac{\alpha}{2}$$

Podajmo vrednost $z_{\alpha/2}$ za nekaj najbolj standardnih tveganj:

- $\alpha = 0,10, \quad z_{\alpha/2} = 1,65$
- $\alpha = 0,05, \quad z_{\alpha/2} = 1,96$
- $\alpha = 0,01, \quad z_{\alpha/2} = 2,58$

Pomen stopnje tveganja pri intervalih zaupanja

Za vsak slučajni vzorec lahko ob omenjenih predpostavkah izračunamo ob izbrani stopnji tveganja α interval zaupanja za parameter γ.

Ker se podatki vzorcev razlikujejo, se razlikujejo vzorčne ocene parametrov in zato tudi izračunani intervali zaupanja za parameter γ.
To pomeni, da se intervali zaupanja od vzorca do vzorca razlikujejo.
Meji intervala sta slučajni spremenljivki.

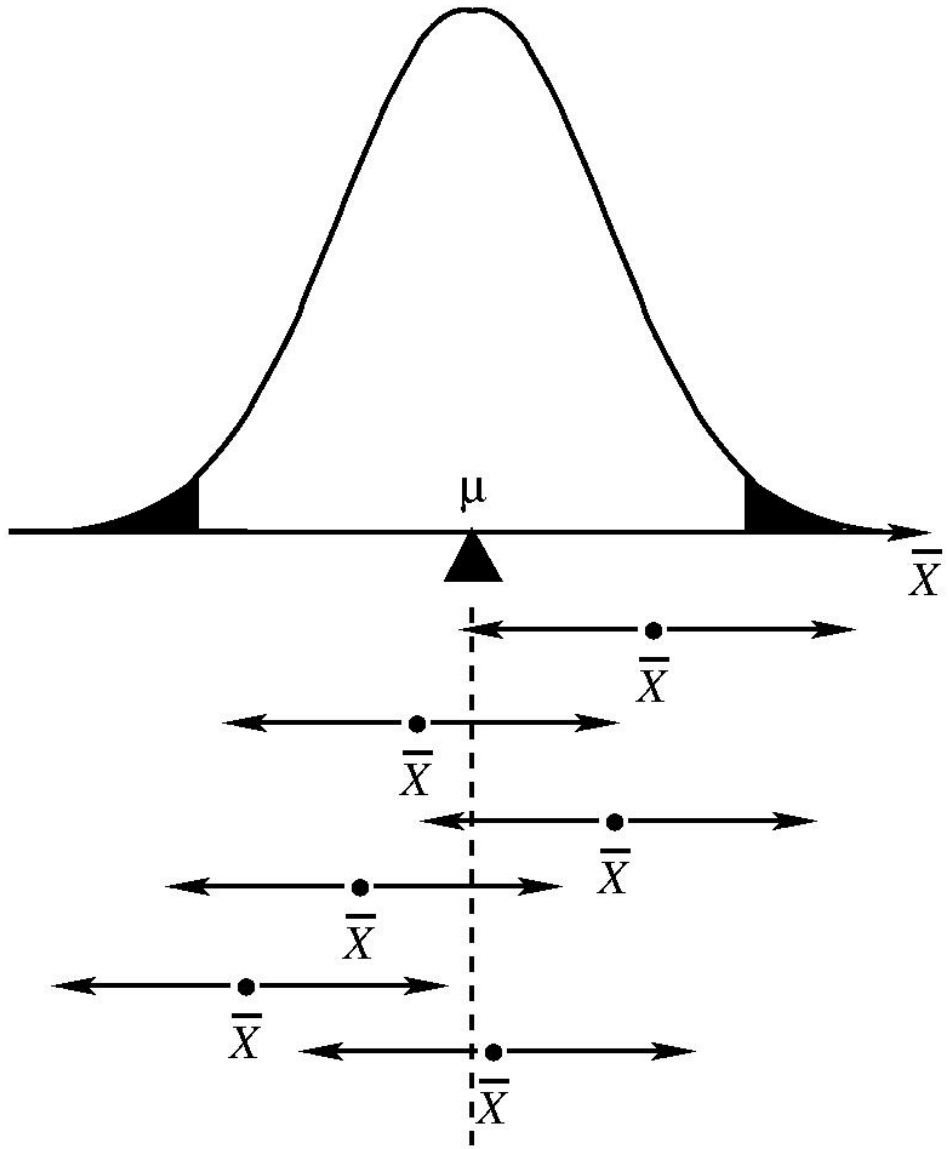
Vzemimo stopnjo tveganja $\alpha = 0,05$. Denimo, da smo izbrali 100 slučajnih vzorcev in za vsakega izračunali interval zaupanja za parameter γ.

Tedaj lahko pričakujemo, da 5 intervalov zaupanja od 100 ne bo pokrilo iskanega parametra γ.

Povedano je lepo grafično predstavljeno na naslednji strani.
V tem primeru ocenjujemo parameter aritmetično sredino inteligenčnega kvocienta. Kot vemo, se vzorčne aritmetične sredine $\overline{X}$ za dovolj velike vzorce porazdeljujejo nomalno.

Denimo, da v tem primeru poznamo vrednost parametra ($\mu = 100$).
Za več slučajnih vzorcev smo izračunali in prikazali interval zaupanja za μ ob stopnji tveganja $\alpha = 0,05$.

Predstavitev več intervalov zaupanja za aritmetično sredino μ pri 5% stopnji tveganja: približno 95% intervalov pokrije parameter μ.


μ
$\overline{X}$
$\overline{X}$
$\overline{X}$
$\overline{X}$
$\overline{X}$
$\overline{X}$
$\overline{X}$

Intervalsko ocenjevanje parametrov

Naj bo X slučajna spremenljivka na populaciji G z gostoto verjetnosti odvisno od parametra ζ.

Slučajna množica $M \subset \mathbb{R}$, ki je odvisna le od slučajnega vzorca, ne pa od parametra ζ, se imenuje *množica zaupanja* za parameter ζ, če obstaja tako število α, $0 < \alpha < 1$, da velja $P(\zeta \in M) = 1 - \alpha$. Število $1 - \alpha$ imenujemo tedaj *stopnja zaupanja*; število α pa *stopnja tveganja*.

Stopnja zaupanja je običajno 95% ali 99% – $\alpha = 0,05$ ali $\alpha = 0,01$.
Pove nam, kakšna je verjetnost, da M vsebuje vrednost parametra ζ ne glede na to, kakšna je njegova dejanska vrednost.

Če je množica M interval $M = [A, B]$, ji rečemo *interval zaupanja* (za parameter ζ).
Njegovi krajišči sta funkciji slučajnega vzorca – torej statistiki.

. . . Intervalsko ocenjevanje parametrov

Naj bo $X : N(\mu, \sigma)$ in recimo, da poznamo parameter σ in ocenjujemo parameter μ. Izberimo konstanti a in b, $b > a$, tako da bo $P(a \leq Z \leq b) = 1 - \alpha$, kjer je $Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$. Tedaj je

$$P\left(\overline{X} - \frac{b\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} - \frac{a\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Označimo

$$A = \overline{X} - \frac{b\sigma}{\sqrt{n}} \quad \text{in} \quad B = \overline{X} - \frac{a\sigma}{\sqrt{n}}.$$

Za katera a in b je interval $[A, B]$ najkrajši?

Pokazati je mogoče (Lagrangeova funkcija),
da mora biti $a = -b$ in $\Phi(b) = (1-\alpha)/2$;
oziroma, če označimo $b = z_{\alpha/2}$, velja $P(Z > z_{\alpha/2}) = \alpha/2$.

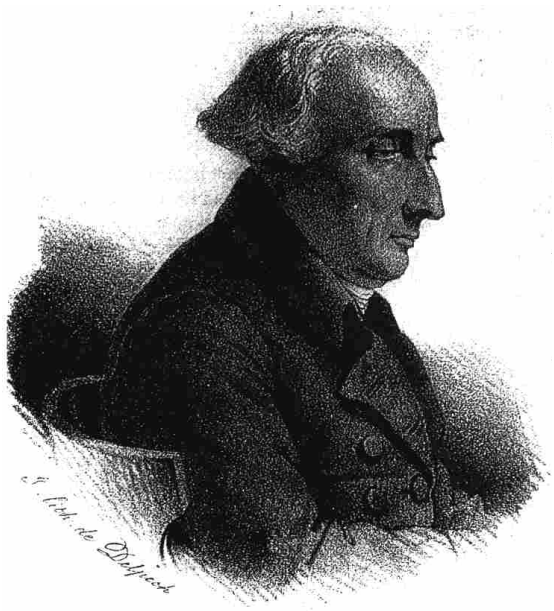
Iskani interval je torej

$$A = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ in } B = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

tj., z verjetnostjo $1-\alpha$ je $|\overline{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Od tu dobimo, da *mora za to,*
da bo napaka manjša od ε z verjetnostjo $1-\alpha$,

veljati $n > (\frac{z_{\alpha/2}\sigma}{\varepsilon})^2$.



. . . Intervalsko ocenjevanje parametrov

Če pri porazdelitvi $X : N(\mu, \sigma)$ tudi parameter σ ni znan, ga nadomestimo s cenilko S in moramo zato uporabiti Studentovo statistiko $T = \frac{\overline{X}-\mu}{S}\sqrt{n}$. Ustrezni interval je tedaj

$$A = \overline{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \qquad B = \overline{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

kjer je $P(T > t_{\alpha/2}) = \alpha/2$.

Če pa bi ocenjevali parameter σ^2, uporabimo statistiko $\chi^2 = (n-1)\frac{S^2}{\sigma^2}$, ki je porazdeljena po $\chi^2(n-1)$. Tedaj sta

$$A = \frac{(n-1)S^2}{b}, \qquad B = \frac{(n-1)S^2}{a}$$

Konstanti a in b včasih določimo iz pogojev $P(\chi^2 < a) = \alpha/2$ in $P(\chi^2 > b) = \alpha/2$; najkrajši interval pa dobimo, ko velja zveza $a^2 p(a) = b^2 p(b)$ in seveda $\int_a^b p(t)dt = 1 - \alpha$.

Teoretična interpretacija koeficienta zaupanja $(1-\alpha)$

Če zaporedoma izbiramo vzorce velikosti n
iz dane populacije in konstruiramo
$[(1-\alpha)100]\%$ interval zaupanja
za vsak vzorec, potem lahko pričakujemo,
da bo $[(1-\alpha)100]\%$ intervalov
vsebovalo pravo vrednost parametra.

stopnja tveganja $=1-$ stopnja zaupanja

I. $(1-\alpha)\%$-ni interval zaupanja za povprečje μ populacije, kadar poznamo standardni odklon σ:

točki

$$\overline{y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

prestavljata krajišči intervala zaupanja, pri čemer je

$z_{\alpha/2}$ vrednost spremenljivke, ki zavzame površino $\alpha/2$ na svoji desni;

σ je standardni odklon za populacijo;

n je velikost vzorca;

$\overline{y}$ je vrednost vzorčnega povprečja.

II. Veliki vzorec za $(1-\alpha)\%$-ni interval zaupanja za povprečje μ populacije:

$$\overline{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right),$$

kjer je s standardni odklon vzorca.

Primer: Na vzorcu velikosti $n = 151$ podjetnikov v majhnih podjetjih v Sloveniji, ki je bil izveden v okviru ankete 'Drobno gospodarstvo v Sloveniji' (Prašnikar, 1993), so izračunali, da je povprečna starost anketiranih podjetnikov $\overline{X} = 40,4$ let in standardni odklon $s = 10,2$ let. Pri 5 % tveganju želimo z intervalom zaupauja oceniti povprečno starost podjetnikov v majhnih podjetjih v Sloveniji.

$$P\left(\overline{y} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{y} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$

$$40,4 - \frac{1,96 \times 10,2}{\sqrt{151}} < \mu < 40,4 + \frac{1,96 \times 10,2}{\sqrt{151}}$$

$$40,4 - 1.6 < \mu < 40,4 + 1,6$$

95 % interval zaupanja za povprečno starost podjetnikov v majhnih podjetjih v Sloveniji je med $38,8$ in $42,0$ leti.

III. Majhen vzorec za $(1-\alpha)\%$-ni interval zaupanja za povprečje μ populacije:

$$\overline{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right),$$

kjer je porazdelitev spremenljivke y vzeta na osnovi $(n-1)$ prostostnih stopenj.

Privzeli smo: populacija, iz katere smo izbrali vzorec,
ima **približno normalno porazdelitev**.

Primer: Vzemimo, da se spremenljivka X - število ur branja dnevnih časopisov na teden - porazdeljuje normalno $N(\mu, \sigma)$.

Na osnovi podatkov za 7 slučajno izbranih oseb ocenimo interval zaupanja za aritmetično sredino pri 10% tveganju.

Podatki in ustrezni izračuni so:

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
5	-2	4
7	0	0
9	2	4
7	0	0
6	-1	1
10	3	9
5	-2	4
49	0	22

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{49}{7} = 7,$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2}{n-1} = \frac{22}{6} = 3,67.$$

Iz tabele za t–porazdelitev preberemo, da je

$$t_{\alpha/2}(n-1) = t_{0,05}(6) = 1,943$$

in interval zaupanja je

$$7 - 1,943 \cdot \frac{1,9}{\sqrt{7}} < \mu < 7 + 1,943 \cdot \frac{1,9}{\sqrt{7}}$$

$$7 - 1,4 < \mu < 7 + 1,4.$$

IV. $(1-\alpha)$%-ni interval zaupanja za razliko $\mu_1 - \mu_2$, če poznamo odklona σ_1 in σ_2 in sta vzorca izbrana neodvisno:

$$\overline{y}_1 - \overline{y}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

V. Veliki vzorec za $(1-\alpha)\%$-ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2, vzorce pa izbramo neodvisno:

$$\overline{y}_1 - \overline{y}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

VI. Majhen vzorec za $(1-\alpha)\%$-ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2, ki pa sta si enaka, vzorci pa so majhni in izbrani neodvisno:

$$\overline{y}_1 - \overline{y}_2 \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$\text{kjer je} \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

Privzeli smo:

– obe populaciji sta **približno normalni**,
– varianci sta enaki,
– naključni vzorci so izbrani **neodvisno**.

VII. Majhen vzorec za $(1-\alpha)\%$-ni interval zaupanja za razliko $\mu_1 - \mu_2$, kadar ne poznamo odklonov σ_1 in σ_2, vzorci pa so majhni in izbrani neodvisno:

$$\overline{y}_1 - \overline{y}_2 \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

kjer je

$$\nu = \left\lfloor \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)} \right\rfloor$$

Primer:

Naslednji podatki predstavljajo dolžine filmov,
ki sta jih naredila dva filmska studija.

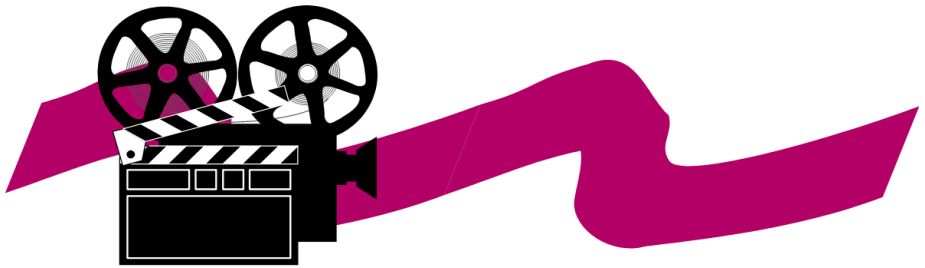
Izračunaj 90%-ni interval zaupanja
za razliko med povprečnim časom filmov,
ki sta jih producirala ta dva studija.

Predpostavimo, da so dolžine filmov
porazdeljene **približno normalno**.

Čas (v minutah)

Studio 1: 103 94 110 87 98

Studio 2: 97 82 123 92 175 88 118



Podatke vnesemo v Minitab

Film.MTV

Studio 1:	Studio 2:
103	97
94	82
110	123
87	92
98	175
	88
	118

Dva vzorca T-Test in interval zaupanja

```
Dva vzorca T za C1 : C2
      N    povpr.  St.odk.  SE povpr.
C1    5    98.40    8.73       3.9
C2    7    110.7    32.2       12

90%-ni interval zaupanja za mu C1-mu C2:

T-TEST mu C1=mu C2 (vs ni=):
T = -0.96  P = 0.37  DF = 7
```

VIII. $(1-\alpha)$%-ni interval zaupanja za razliko $\mu_d = \mu_1 - \mu_2$ ujemajočih se parov v velikih vzorcih:

$$\overline{d} \pm z_{\alpha/2}\left(\frac{s_d}{\sqrt{n}}\right), \quad \text{kjer je } n \text{ število parov.}$$

IX. $(1-\alpha)\%$-ni interval zaupanja za razliko $\mu_d = \mu_1 - \mu_2$ ujemajočih se parov v majhnih vzorcih:

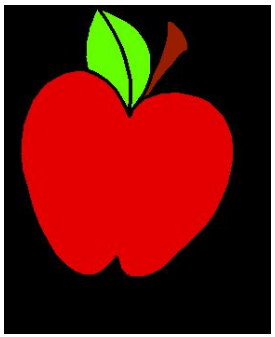
$$\overline{d} \pm t_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right), \quad \text{kjer je } n \text{ število parov.}$$

Privzeli smo: populacija razlik parov je normalno porazdeljena.

Nal. 8-39. Špricanje jabolk lahko pozroči kontaminacijo zraka. Zato so v času najbolj intenzivnega špricanja zbrali in analizirali vzorce zraka za vsak od 11ih dni.

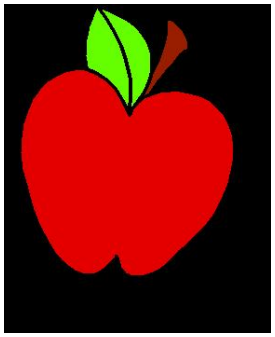
Raziskovalci želijo vedeti ali se povprečje ostankov škropiv (diazinon) razlikuje med dnevom in nočjo.

Analiziraj podatke za 90% interval zaupanja.

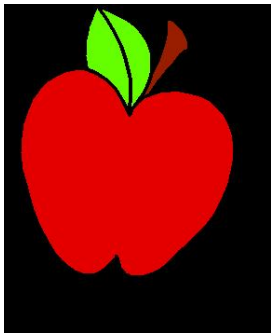


Nal. 8-39

Diazinon Residue		
Datum	dan	no"c
Jan. 11	5,4	24,3
12	2,7	16,5
13	34,2	47,2
14	19,9	12,4
15	2,4	24,0
16	7,0	21,6
17	6,1	104,3
18	7,7	96,9
19	18,4	105,3
20	27,1	78,7
21	16,9	44,6



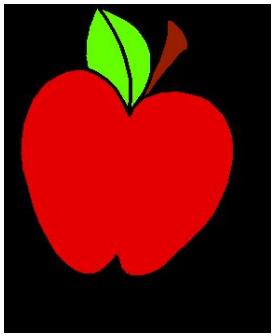
Datum	Diazinon dan	Residue no"c	razlika dan-no"c
Jan. 11	5,4	24,3	-18,9
12	2,7	16,5	-13,8
13	34,2	47,2	-13,0
14	19,9	12,4	7,5
15	2,4	24,0	-21,6
16	7,0	21,6	-14,6
17	6,1	104,3	-98,2
18	7,7	96,9	-89,2
19	18,4	105,3	-86,9
20	27,1	78,7	-51,6
21	16,9	44,6	-27,7



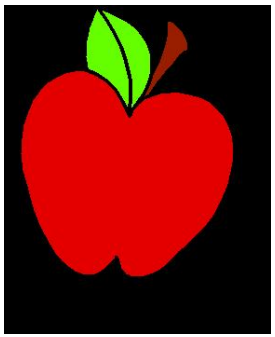
Podatke vnesemo v Minitab:

Ex8-39.MTW

C1	C2
5.4	24.3
2.7	16.5
34.2	47.2
19.2	12.4
2.4	24.0
7.0	21.6
6.1	104.3
7.7	96.9
18.4	105.3
27.1	78.7
16.9	44.6



```
MTB > Let C3=C1-C2.

T interval zaupanja

Spremen. N   povpr. Stdev SEpovpr.
C3       11   -38.9   36.6    11.0
```



90,0 % interval zaupanja je $(58,9\,;\,18,9)$.

Za deleže

$p =$ delež populacije

$\hat{p} =$ delež vzorca,

kjer je $\hat{p} = y/n$ in je y število uspehov v n poskusih.

X. $(1-\alpha)$%-ni interval zaupanja za delež populacije p, kadar poznamo $\sigma_{\hat{p}}$:

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}}$$

XI. Veliki vzorec za $(1-\alpha)\%$-ni interval zaupanja za delež populacije:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Privzeli smo: velikost vzorca n je dovolj velika, da je aproksimacija veljavna.

Dobro pravilo (angl. rule of thumb) za izpolnitev pogoja

"dovolj velik vzorec" je:

$$n\hat{p} \geq 4 \quad \text{in} \quad n\hat{q} \geq 4.$$

Primer: Na vzorcu ($n = 151$), ki je bil izveden v okviru ankete 'Drobno gospodarstvo v Sloveniji', so izračunali. da je delež obrtnih podjetij $\hat{p} = 0,50$. Pri 5% tveganju želimo z intervalom zaupanja oceniti delež obrtnih majhnih podjetij v Sloveniji.

$$0,50 - 1.96\,\frac{0,50 \times 0,50}{\sqrt{151}} \;<\; p \;<\; 0,50 + 1.96\,\frac{0,50 \times 0,50}{\sqrt{151}}$$

$$0,50 - 0,08 \;<\; p \;<\; 0,50 + 0,08$$

S 5% stopnjo tveganja trdimo, da je delež obrtnih majhnih podjetij v Sloveniji glede na vsa majhna podjetja med 0,42 in 0,58.

XII. $(1-\alpha)$%-ni interval zaupanja za razliko deležev $p_1 - p_2$, kadar poznamo $\sigma_{\hat{p}_1-\hat{p}_2}$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{\hat{p}_1-\hat{p}_2}.$$

XIII. Veliki vzorec za $(1-\alpha)\%$-ni interval zaupanja za razliko deležev $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Privzeli smo: velikost vzorca n je dovolj velika, da je aproksimacija veljavna.

Kot splošno pravilo za dovolj velika vzorca privzamemo naslednje:

$$n_1\hat{p}_1 \geq 4 \qquad n_1\hat{q}_1 \geq 4.$$
$$n_2\hat{p}_2 \geq 4 \quad \text{in} \quad n_2\hat{q}_2 \geq 4.$$

XIV. Veliki vzorec za $(1-\alpha)\%$-ni interval zaupanja za varianco populacije σ^2:

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}$$

Privzeli smo: populacija iz katere izbiramo vzorce, ima **približno normalno porazdelitev**.

Primer: Vzemimo prejšnji primer spremenljivke o številu ur branja dnevnih časopisov na teden. Za omenjene podatke iz vzorca ocenimo z intervalom zaupanja varianco pri 10% tveganju.

Iz tabele za χ^2–porazdelitev preberemo, da je

$$\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0,95}(6) = 12,6,$$

$$\chi^2_{\alpha/2}(n-1) = \chi^2_{0,05}(6) = 1,64.$$

90 % interval zaupanja za varianco je tedaj

$$\frac{6 \cdot 3,67}{12,6} < \sigma^2 < \frac{6 \cdot 3,67}{1,64},$$

$$1,75 < \sigma^2 < 13,43.$$

XV. $(1-\alpha)\%$-ni interval zaupanja za kvocient varianc dveh populacij σ_1^2/σ_2^2:

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_2-1,n_1-1}}$$

Privzeli smo:

– obe populaciji iz katerih izbiramo vzorce, imata **približno normalni porazdelitvi** relativnih frekvenc.

– naključni vzorci so izbrani **neodvisno** iz obeh populacij.

XVI. Izbira velikosti vzorca za oceno populacijskega povprečja μ znotraj E enot z verjetnostjo $(1-\alpha)$:

Spomnimo se (glej str. 386), da z verjetnostjo $1-\alpha$ velja

$$|\overline{X}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Od tu dobimo, da mora za to, da bo napaka manjša od ε z verjetnostjo $1-\alpha$, veljati

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$$

Populacijski odklon mora biti običajno aproksimiran.

Primer: Denimo, da želimo oceniti povprečno starost podjetnikov majhnih podjetij, tako da bo razlika med populacijskim povprečjem in ocenjenim povprečjem manjša od enega leta.

Če vemo, da je standardni odklon $\sigma = 10$ let in izberemo 5% tveganje, lahko ocenimo, kako velik vzorec potrebujemo:

$$n > \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left(\frac{1,96 \times 10}{1}\right)^2 = 384,2$$

Če želimo doseči postavljeno natančnost ocenjevanja,
potrebujemo vsaj 385 enot v slučajnem vzorcu.

XVII. Izbira velikosti vzorca za oceno razlike $\mu_1 - \mu_2$ med parom populacijskih povprečij, ki je pravilna znotraj E enot z verjetnostjo $(1 - \alpha)$:

$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2)$$

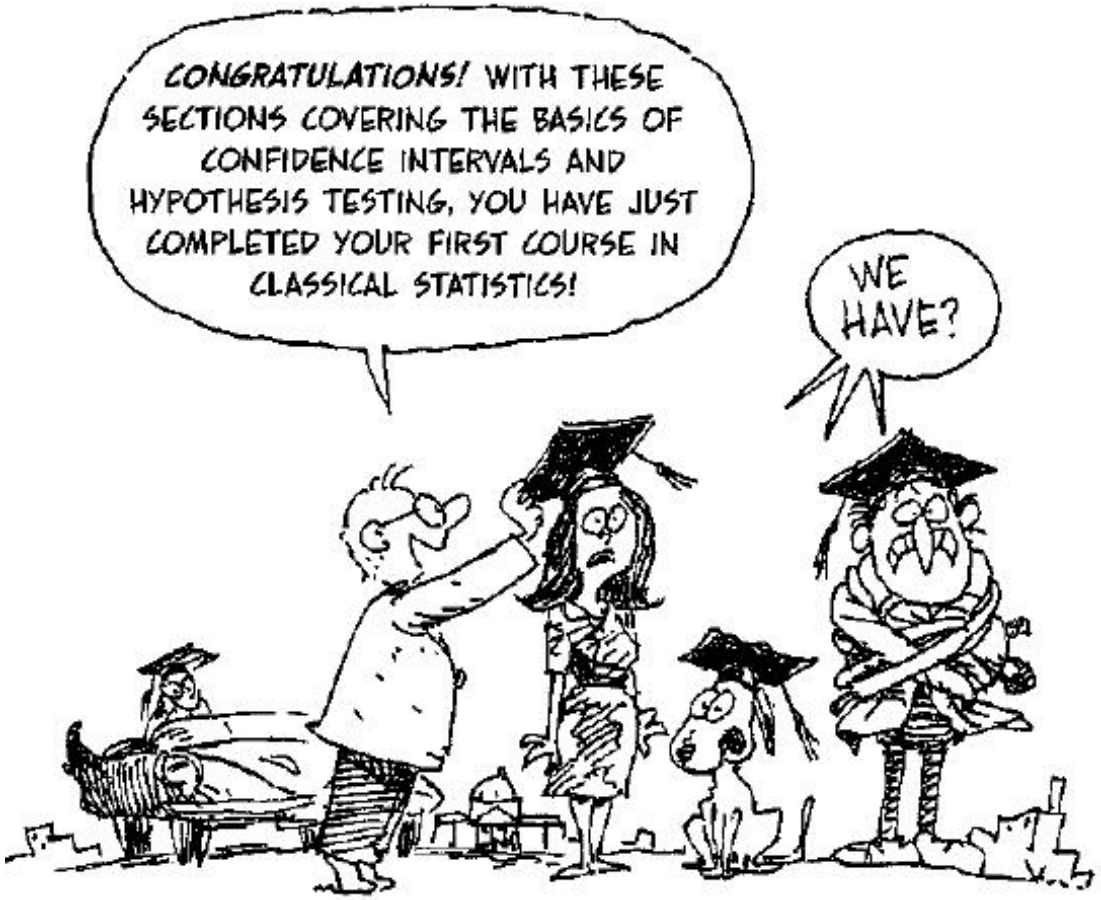
XVIII. Izbira velikosti vzorca za oceno deleža populacije p, ki je pravilna znotraj E enot z verjetnostjo $(1-\alpha)$:

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 pq$$

Opozorilo: v tem primeru potrebujemo oceni za p in q.
Če nimamo nobene na voljo, potem uporabimo $p = q = 0,5$
za konzervativno izbiro števila n.

XIX. Izbira velikosti vzorca za cenilko razlike $p_1 - p_2$ med dvema deležema populacije, ki je pravilna znotraj E enot z verjetnostjo $(1 - \alpha)$:

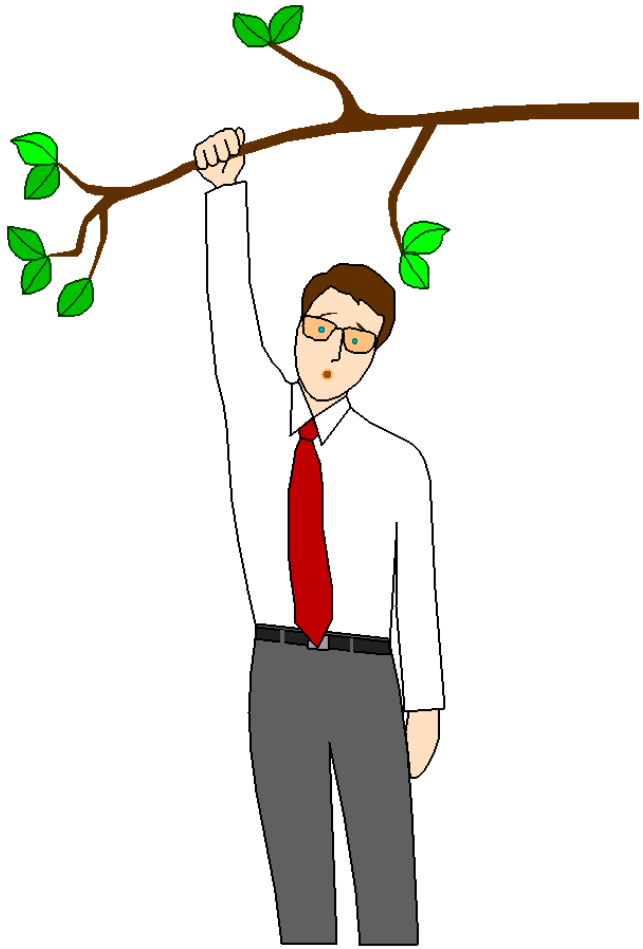
$$n_1 = n_2 = \left(\frac{z_{\alpha/2}}{E}\right)^2 (p_1 q_1 + p_2 q_2)$$


CONGRATULATIONS! WITH THESE SECTIONS COVERING THE BASICS OF CONFIDENCE INTERVALS AND HYPOTHESIS TESTING, YOU HAVE JUST COMPLETED YOUR FIRST COURSE IN CLASSICAL STATISTICS!
WE HAVE?

II.5. Preizkušanje statističnih domnev

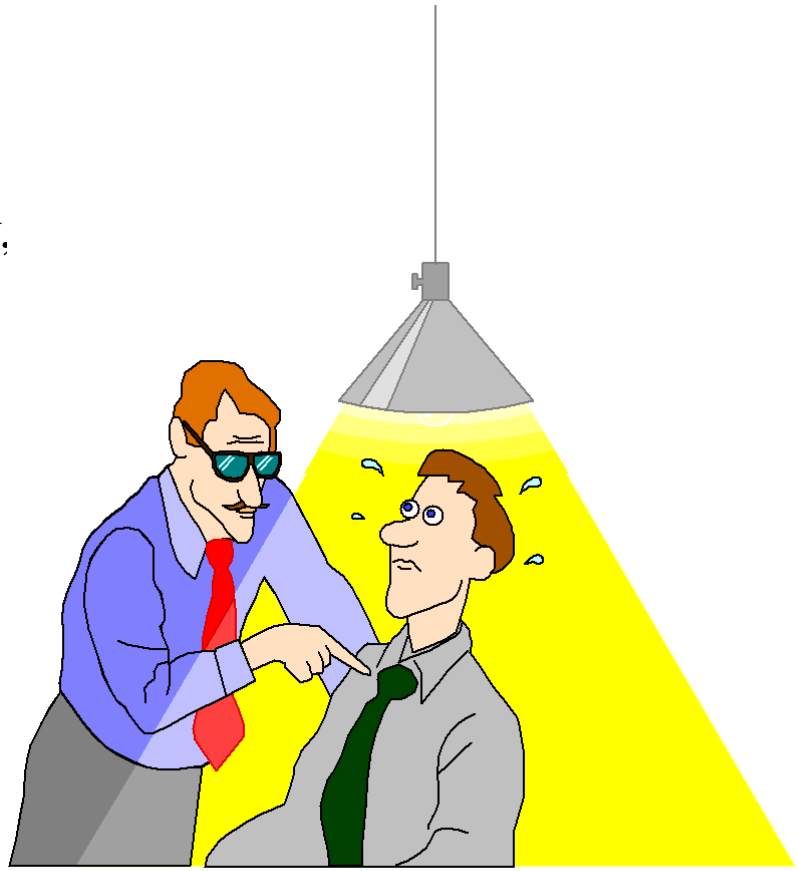
Načrt

- postopek
- elementi
  - napake 1. in 2. vrste
  - značilno razlikovanje
  - moč statističnega testa
- testi
  - centralna tendenca
  - delež
  - varianca



Uvod

- postavimo trditev o populaciji,
- izberemo vzorec,
  s katerim bomo preverili trditev,
- zavrni ali sprejmi trditev.

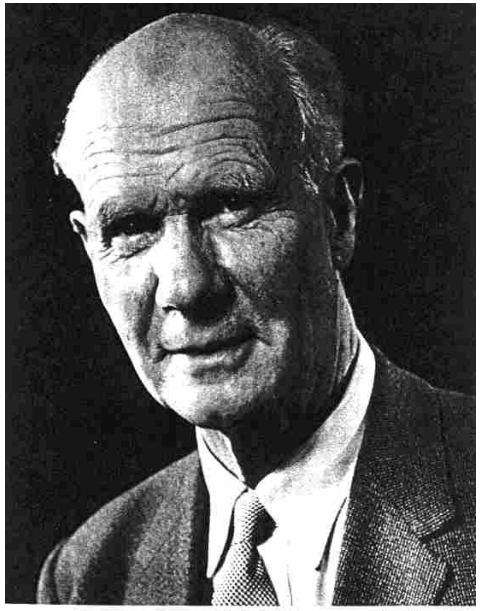


Hipoteza je testirana
z določanjem verjetja,
da dobimo določen rezultat,
kadar jemljemo vzorce iz
populacije s predpostavljenimi
vrednostimi parametrov.

Zgodovina

Teorijo preizkušanja domnev sta v
20. in 30. letih prejšnjega stoletja
razvila J. Neyman in E.S. Pearson.

Statistična domneva (ali hipoteza)
je vsaka domneva o porazdelitvi
slučajne spremenljivke X na populaciji.



EGON SHARPE PEARSON

Če poznamo vrsto (obliko) porazdelitve $f(x;\zeta)$ in postavljamo/raziskujemo domnevo o parametru ζ, govorimo o *parametrični domnevi*.

Če pa je vprašljiva tudi sama vrsta porazdelitve,
je domneva *neparametrična*.

Domneva je *enostavna*, če natačno določa porazdelitev
(njeno vrsto in točno vrednost parametra); sicer je *sestavljena*.

Primer: Naj bo $X : N(\mu, \sigma)$.

Če poznamo σ, je domneva $H : \mu = 0$ enostavna;

Če pa parametra σ ne poznamo, je sestavljena.

Primer sestavljene domneve je tudi $H : \mu > 0$.

Statistična domneva je lahko pravilna ali napačna.
Želimo seveda sprejeti pravilno domnevo in zavrniti napačno.
Težava je v tem, da o pravilnosti/napačnosti domneve
ne moremo biti gotovi, če jo ne preverimo na celotni populaciji.
Ponavadi se odločamo le na podlagi vzorca.

Če vzorčni podatki preveč odstopajo od domneve, rečemo,
da niso *skladni* z domnevo, oziroma,
da so *razlike značilne*, in domnevo zavrnemo.

Če pa podatki domnevo podpirajo, jo ne zavrnemo
– včasih jo celo sprejmemo.
To ne pomeni, da je domneva pravilna, temveč
da ni zadostnega razloga za zavrnitev.

Postopek testiranja hipoteze

- postavi ničelno in alternativno hipotezo,
- izberi testno statistiko,
- določi zavrnitveni kriterij,
- izberi naključni vzorec,
- izračunaj vrednost na osnovi testne statistike,
- sprejmi odločitev,
- naredi ustrezen zaključek.

Hipoteza

- Ničelna hipoteza (H_0)
  - je trditev o lastnosti populacije za katero predpostavimo, da drži (oziroma za katero verjamemo, da je resnična),
  - je trditev, ki jo test skuša ovreči.

- Alternativna (nasprotna) hipoteza (H_a)
  - je trditev nasprotna ničelni hipotezi,
  - je trditev, ki jo s testiranjem skušamo dokazati.

... Hipoteza

- ničelna hipoteza (H_0)
  - obtoženec je nedolžen,
- alternativna hipoteza (H_a)
  - obtoženec je kriv.

Odločitev in zaključek

- Porota je spoznala obtoženca za **krivega**.
  Zaključimo, da je bilo dovolj dokazov, ki nas prepričajo,
  da je obtoženec storil kaznivo dejanje.
- Porota je spoznala obtoženca za **nedolžnega**.
  Zaključimo, da je ni bilo dovolj dokazov, ki bi nas prepričali,
  da je obtoženec storil kaznivo dejanje.

Elementi testiranja hipoteze

dejansko stanje	*odločitev*	
	nedolžen	**kriv**
nedolžen	pravilna odločitev	napaka 1. vrste (α)
kriv	napaka 2. vrste (β)	moč $(1-\beta)$

... Elementi testiranja hipoteze

- verjetnost napake 1. vrste (α)
  verjetnost za obtožbo nedolžnega obtoženca.
- značilno razlikovanje (signifikantno) oziroma **stopnja značilnosti**
- količina dvoma (α), ki ga bo porota še sprejela.
  - Kriminalna tožba: Beyond a reasonable doubt...
  - Civilna tožba: The preponderance of evidence must suggest...

... Elementi testiranja hipoteze

- verjetnost napake 2. vrste: (β)
  - verjetnost, da spoznamo krivega obtoženca za nedolžnega,
- moč testa: ($1 - \beta$)
  - verjetnost, da obtožimo krivega obtoženca.

Sodba

- breme dokazov,
- potrebno je prepričati poroto, da je obtoženi kriv (alternativna hipoteza)
  preko določene stopnje značilnosti.
  - Criminal: Reasonable Doubt
  - Civil: Preponderance of evidence

Obramba

- Ni bremena dokazovanja.
- Povzročiti morajo dovolj dvoma pri poroti, če je obtoženi resnično kriv.

Statistična ničelna hipoteza

$H_0 \ : \ \mu = 9mm$

(Premer 9 milimetrskega kroga),

$H_0 \ : \ \mu = 600$ km

(Proizvalajec trdi, da je to doseg novih vozil),

$H_0 \ : \ \mu = 3$ dnevi

Neusmerjena alternativna hipoteza

$H_0 \ : \ \mu = 9mm$

$H_a \ : \ \mu \neq 9mm$

Premer 9 milimetrskega kroga.

"Manj kot" alternativna hipoteza

$H_0 \ : \ \mu = 600$ km

$H_a \ : \ \mu < 600$ km

Proizvalajec trdi, da je to doseg novih vozil.

"Več kot" alternativna hipoteza

H_0 : $\mu = 3$ dnevi

H_a : $\mu > 3$ dnevi

Čas odsotnosti določenega artikla
pri neposredni podpori.

Definicije

1. Zavrnitev ničelne hipoteze, če je le-ta pravilna, je **napaka 1. vrste.**

   Verjetnost, da naredimo napako 1. vrste, označimo s simbolom α in ji pravimo **stopnja tveganja**, $(1-\alpha)$ pa je **stopnja zaupanja**.

2. Če ne zavrnemo ničelno hipotezo, v primeru, da je napačna, pravimo, da gre za **napako 2. vrste**.

   Verjetnost, da naredimo napako 2. vrste, označimo s simbolom β.

3. **Moč statističnega testa**, $(1-\beta)$
   je verjetnost zavrnitve ničelne hipoteze v primeru, ko je le-ta v resnici napačna.

Statistična hipoteza

- ničelna hipoteza
  - $H_0 \ : \ q = q_0$
- alternativna hipoteza
  - $H_a \ : \ q \neq q_0$
  - $H_a \ : \ q > q_0$
  - $H_a \ : \ q < q_0$

Primer testiranja hipoteze

Predpostavimo, da je dejanska mediana (τ) pH iz določene regije 6,0.

Da bi preverili to trditev, bomo izbrali 10 vzorcev zemlje iz te regije,
da ugotovimo, če empirični vzorci močno podpirajo,
da je dejanska mediana manšja ali enaka 6,0?

Predpostavke

- naključni vzorec
  - neodvisen
  - enako porazdeljen (kot celotna populacija),
- vzorčenje iz zvezne porazdelitve,
- verjetnostna porazdelitev ima mediano.

Postavitev statistične hipoteze in izbira testne statistike

- ničelna hipoteza
  - $H_0 \ : \ \tau = 6,0$ (mediana populacije τ_0)
- alternativna hipoteza
  - $H_a \ : \ \tau < 6,0$

Testna statistika (TS)

- $S_+ =$ število vzorcev, ki so večji od mediane τ_0 iz hipoteze,
- $S_- =$ število vzorcev, ki so manšji od mediane τ_0 iz hipoteze.

Porazdelitev testne statistike

- vsak poskus je bodisi uspeh ali neuspeh,
- fiksen vzorec, velikosti n,
- naključni vzorci
  - neodvisni poskusi,
  - konstantna verjetnost uspeha.

Torej gre za

- binomsko porazdelitev: $S_+ \approx B(n, p)$,
- s parameteri $n = 10$ in $p = 0,5$,
- in pričakovano vrednostjo (matematičnim upanjem): $E(X) = np = 5$.

Testiranje hipoteze

		odločitev	
		FTR H_0	**zavrni** H_0
dejansko stanje	H_0 **je pravilna**	pravilna odločitev	
	H_0 **je napačna**		pravilna odločitev

Napaka 1. vrste

		odločitev	
		FTR H_0	**zavrni** H_0
dejansko stanje	H_0 **je pravilna**		napaka 1. vrste
	H_0 **je napačna**		

Napaka 2. vrste

		odločitev	
		FTR H_0	**zavrni** H_0
dejansko stanje	H_0 **je pravilna**		
	H_0 **je napačna**	napaka 2. vrste	

Moč testa

dejansko stanje	*odločitev*	
	FTR H_0	**zavrni** H_0
H_0 **je pravilna**		
H_0 **je napačna**		$(1-\beta)$

Elementi testiranja hipoteze

		odločitev	
		FTR H_0	**zavrni** H_0
dejansko stanje	H_0 **je pravilna**		(α)
	H_0 **je napačna**	(β)	$(1-\beta)$

... Elementi testiranja hipoteze

- verjetnost napake 1. vrste (α)
  - Če hipoteza H_0 drži, kakšna je možnost, da jo zavržemo.
- stopnja značilnosti testa (signifikantnosti)
  - Največji α, ki ga je vodja eksperimenta pripravljen sprejeti (zgornja meja za napako 1. vrste).
- verjetnost napake 2. vrste (β)
  - Če hipoteza H_0 ne drži, kakšna je možnost, da je **ne** zavržemo.
- moč statističnega testa: $(1 - \beta)$
  - Če hipoteza H_0 ne drži, kakšna je možnost, da jo zavržemo.

... Elementi testiranja hipoteze

velikost vzorca	napaka 1.vrste	napaka 2.vrste	moč
n	α	β	$1-\beta$
konst.	$\uparrow$	$\downarrow$	$\uparrow$
konst.	$\downarrow$	$\uparrow$	$\downarrow$
povečanje	$\downarrow$	$\downarrow$	$\uparrow$
zamnjšanje	$\uparrow$	$\uparrow$	$\downarrow$

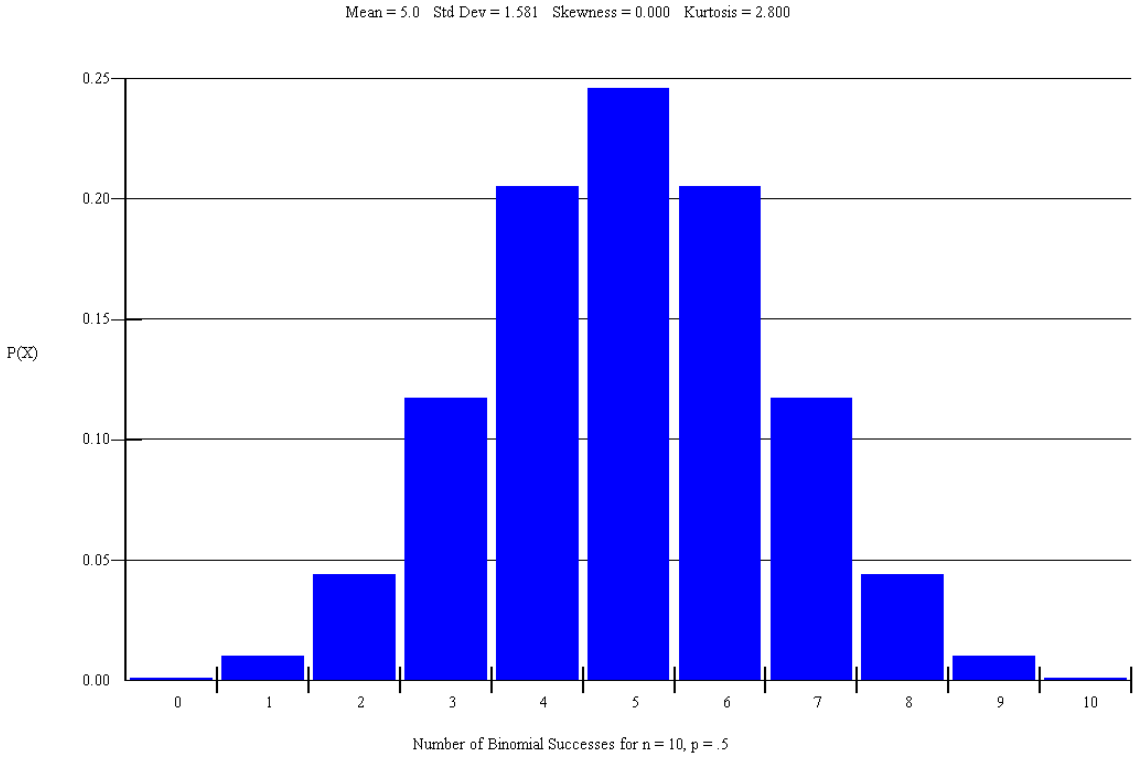
Primer (A) testiranja hipoteze

Predpostavimo, da je dejanska mediana (τ) pH iz določene regije 6,0.

Da bi preverili to trditev, bomo izbrali 10 vzorcev zemlje iz te regije, da ugotovimo, če empirični vzorci močno podpirajo, da je dejanska mediana manjša ali enaka 6,0?

- Hipotezi
  - $H_0 \ : \ \tau = 6,0$
  - $H_a \ : \ \tau < 6,0$
- Testna statistika
  - $S_+ =$ število vzorcev večjih od predpostavljene mediane τ_0,
  - $S_+ \approx B(n,p) = B(10; 0,5)$,
  - $E(S_+) = 5$.

Porazdelitev testne statistike



Določimo zavrnitveni kriterij

x	$P(X = x)$	$F(x)$
0	$0,000977$	$0,00098$
1	$0,009766$	$0,01074$
2	$0,043945$	$0,05469$
3	$0,117188$	$0,17188$
4	$0,205078$	$0,37695$
5	$0,246094$	$0,62305$
6	$0,205078$	$0,82813$
7	$0,117188$	$0,94531$
8	$0,043945$	$0,98926$
9	$0,009766$	$0,99902$
10	$0,000977$	$1,00000$

... Določimo zavrnitveni kriterij

- Stopnja značilnosti testa (α) = 0,01074,
- Kritična vrednost
  - $S_+ = 1$,
- Območje zavrnitve
  - $S_+ \leq 1$,

Izberemo naključni vzorec

Predpostavimo, da je dejanska mediana (τ) pH iz določene regije 6,0.

Da bi preverili to trditev, smo izbrali 10 vzorcev zemlje iz te regije in jih podvrgli kemični analizi in na ta način določili pH vrednost za vsak vzorec.

Ali empirični podatki podpirajo trditev,
da je dejanska mediana manšja ali enaka 6,0?

$$5,93;\ 6,08;\ 5,86;\ 5,91;\ 6,12;\ 5,90;\ 5,95;\ 5,89;\ 5,98;\ 5,96.$$

Izračunaj vrednost iz testne statistike

pH	predznak
$5,93$	$-$
$6,08$	$+$
$5,86$	$-$
$5,91$	$-$
$6,12$	$+$
$5,90$	$-$
$5,95$	$-$
$5,89$	$-$
$5,98$	$-$
$5,96$	$-$

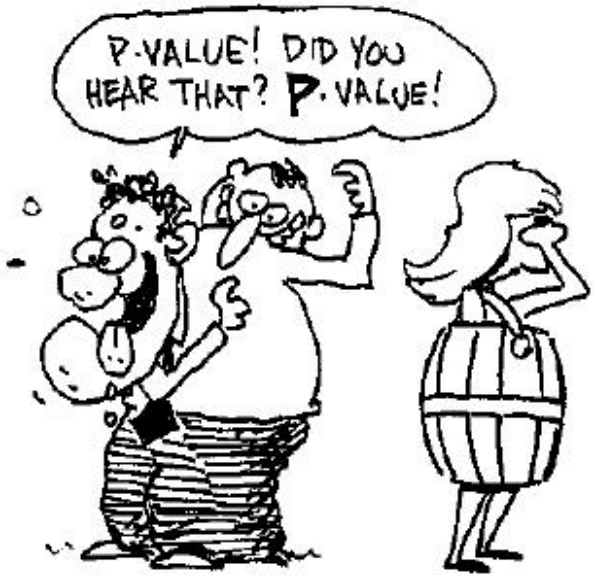
$S_+ = 2$

Naredimo odločitev

- Izračunana vrednost S_+ leži zunaj zavrnitvenega območja.
- Ni osnove za zavrnitev hipoteze H_0.

P-vrednost

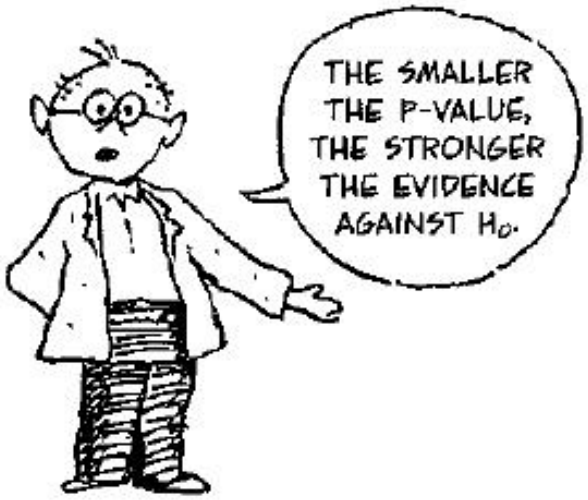
***P*-vrednost** (ali ugotovljena bistvena stopnja
za določen statistični test) je verjetnost
(ob predpostavki, da drži H_0),
da ugotovimo vrednost testne statistike,
ki je vsaj toliko v protislovju s H_0
in podpira H_a kot tisto,
ki je izračunana iz vzorčnih podatkov.



P-vrednost

- Sprejemljivost hipoteze H_0 na osnovi vzorca
  - Verjetnost, da je opazovani vzorec (ali podatki) bolj ekstremni, če je hipoteza H_0 pravilna.

- Najmanjši α pri katerem zavrnemo hipotezo H_0.
  - če je P-vrednost $> \alpha$, potem FTR H_0,
  - če je P-vrednost $< \alpha$, potem zavrni H_0.



... P-vrednost

x	$P(X = x)$	$F(x)$
0	$0,000977$	$0,00098$
1	$0,009766$	$0,01074$
2	$0,043945$	$0,05469$
3	$0,117188$	$0,17188$
4	$0,205078$	$0,37695$
5	$0,246094$	$0,62305$
6	$0,205078$	$0,82813$
7	$0,117188$	$0,94531$
8	$0,043945$	$0,98926$
9	$0,009766$	$0,99902$
10	$0,000977$	$1,00000$

P-vrednost $= P(S_+ \geq 2 \,|\, \tau = 6,0) = 0,05469.$

Izračunaj vrednost iz testne statistike

pH	predznak
$5,93$	$-$
$6,08$	$+$
$5,86$	$-$
$5,91$	$-$
$6,12$	$+$
$5,90$	$-$
$5,95$	$-$
$5,89$	$-$
$5,98$	$-$
$5,96$	$-$

$S_- = 8$

... P-vrednost

x	$P(X = x)$	$F(x)$
0	$0,000977$	$0,00098$
1	$0,009766$	$0,01074$
2	$0,043945$	$0,05469$
3	$0,117188$	$0,17188$
4	$0,205078$	$0,37695$
5	$0,246094$	$0,62305$
6	$0,205078$	$0,82813$
7	$0,117188$	$0,94531$
8	$0,043945$	$0,98926$
9	$0,009766$	$0,99902$
10	$0,000977$	$1,00000$

P-vrednost $= P(S_- \geq 8 \,|\, \tau = 6,0) = 0,05469.$

Odločitev in zaključek

- P-vrednost $> \alpha = 0,01074$.
- Ni osnove za zavrnitev hipoteze H_0.
- Zavrni ničelno hipotezo.
  - Zaključimo, da empirični podatki sugerirajo, da velja alternativna trditev.
- Ni osnove za zavrnitev ničelne hipoteze (angl. fail to reject, kratica FTR).
  - Zaključimo, da nimamo dovolj osnov, da bi dokazali, da velja alternativna trditev.
- Premalo podatkov, da bi pokazali, da je dejanska mediana pH manjša od 6,0.
- Privzemimo, da je pH enaka 6,0 v tej konkretni regiji.

Naloga 9.4 na strani 429

Pascal je visoko-nivojski programski jezik,
ki smo ga nekoč pogosto uporabljali na
miniračunalnikih in microprocesorjih.

Narejen je bil eksperiment, da bi ugotovili
delež Pascalovih spremenljivk, ki so tabelarične spremenljivke
(v kontrast skalarim spremenljivkam,
ki so manj učinkovite, glede na čas izvajanja).

20 spremenljivk je bilo naključno izbranih iz
množice Pascalovih programov,
pri tem pa je bilo zabeleženo število
tabelaričnih spremenljivk Y.

Predpostavimo, da želimo testirati hipotezo,
da je Pascal bolj učinkovit jezik kot Agol,
pri katerem je 20% spremenljivk tabelaričnih.

To pomeni, da bomo testirali $H_0 \ : \ p = 0,20$,
proti $H_a \ : \ p > 0,20$, kjer je p verjetnost,
da imamo tabelarično spremenljivko na vsakem poskusu.

Predpostavimo, da je 20 poskusov neodvisnih.

(a) Določi α za območje zavrnitve $y > 8$.

Izračunati želimo verjetnost, da se bo zgodila napaka 1. vrste, torej da bomo zavrnili pravilno hipotezo.

Predpostavimo, da je hipoteza H_0 pravilna, tj. $Y : \mathrm{B}(20; 0{,}2)$. Če se bo zgodilo, da bo Y pri izbranem vzorcu večji ali enak 8, bom hipotezo zavrnili, čeprav je pravilna. Torej velja:

$$\begin{aligned}
\alpha &= \mathrm{P}(Y \geq 8) = 1 - \mathrm{P}(Y \leq 7) \\
&= 1 - \sum_{i=0}^{7} \mathrm{P}(Y = i) \\
&= 1 - \sum_{i=0}^{7} \binom{20}{i} 0{,}2^i \, 0{,}2^{20-i} \\
&= 1 - 0{,}9679 = 0{,}0321 = 3{,}21\%.
\end{aligned}$$

(b) Določi α za območje zavrnitve $y \geq 5$.

Do rezultata pridemo na enak način kot v prejšnji točki:

$$\begin{aligned}
\alpha &= \mathrm{P}(Y \geq 5) = 1 - \mathrm{P}(Y \leq 4) \\
&= 1 - \sum_{i=0}^{4} \mathrm{P}(Y = i) \\
&= 1 - \sum_{i=0}^{4} \binom{20}{i} 0{,}2^{i}\ 0{,}2^{20-i} \\
&= 1 - 0{,}6296 = 0{,}3704 = 37{,}04\%.
\end{aligned}$$

(c) Določi β za območje zavrnitve $Y \geq 8$, če je $p = 0,5$.

Izračunati želimo verjetnost, da se bo zgodila napaka 2. vrste,
torej da bomo sprejeli napačno hipotezo.
Ker vemo, da je $p = 0{,}5$, velja $Y \sim \mathrm{B}(20; 0{,}5)$.

Napačno hipotezo bomo sprejeli,
če bo y pri izbranem vzorcu manjši od 8.

$$\beta = \mathrm{P}(y \leq 7) = \sum_{i=0}^{7} \binom{20}{i} 0{,}5^i \, 0{,}5^i = 0{,}1316 = 13{,}16\%.$$

(d) Določi β za območje zavrnitve $y \geq 5$, če je $p = 0{,}5$.

Do rezultata pridemo na enak način kot v prejšnji točki:

$$\beta = \mathrm{P}(y \leq 4) = \sum_{i=0}^{4} \binom{20}{i} 0{,}5^i \; 0{,}5^i = 0{,}0059 = 0{,}59\%.$$

(e) Katero območje zavrnitve $y \geq 8$ ali $y \geq 5$ je bolj zaželjeno,
če želimo minimizirati verjetnost napake 1. stopnje
oziroma če želimo minimizirati verjetnost napake 2. stopnje.

Napako 1. stopnje minimiziramo z izbiro območja $y \geq 8$,
napako 2. stopnje pa z izbiro območja $y \geq 5$.

(f) Določi območje zavrnitve $y \geq a$ tako, da je α približno 0,01.

Na osnovi točke (e) zaključimo, da se z večanjem števila a manjša verjetnost α in s poskušanjem (ki ga pričnemo na osnovi izkušenje iz točke (a) pri 9) pridemo do $a = 9$.

(g) Za območje zavrnitve določeno v točki (f) določi moč testa, če je v resnici $p = 0,4$.

Moč testa je $1 - \beta$. Verjetnost β izračunamo enako kot v točkah (c) in (d). Velja $Y \sim \mathrm{B}(20; 0{,}4)$ in

$$\beta = \mathrm{P}(y \leq 8) = \sum_{i=0}^{8} \binom{20}{i} 0{,}4^i \; 0{,}6^i = 0{,}5956 = 59{,}56\%.$$

Moč testa znaša $0{,}4044$.

(h) Za območje zavrnitve določeno v točki (f) določi moč testa, če je v resnici $p = 0{,}7$.

Tokrat velja $Y \sim \mathrm{B}(20; 0{,}4)$ in

$$\beta = \mathrm{P}(y \leq 8) = \sum_{i=0}^{8} \binom{20}{i} 0{,}7^i \; 0{,}3^i = 0{,}0051 = 0{,}51\%.$$

Moč testa znaša $0{,}995$.

Formalen postopek za testiranje hipotez

1. Postavi hipotezi:
   - ničelna,
   - alternativna.
2. Za parameter poiščemo kar se da dobro cenilko
   (npr. nepristransko) in njeno porazdelitev ali
   porazdelitev ustrezne statistike (izraz, v katerem nastopa cenilka).
3. Določi odločitveno pravilo.
   Izberemo stopnjo značilnosti (α).
   Na osnovi stopnje značilnosti in porazdelitve statistike
   določimo kritično območje;
4. Zberi/manipuliraj podatke ter na vzorčnih podatkih
   izračunaj (eksperimentalno) vrednost testne statistike.

5. Primerjaj in naredi zaključek.

- če eksperimentalna vrednost pade v kritično območje, ničelno domnevo zavrni in sprejmi osnovno domnevo ob stopnji značilnosti α.
- če eksperimentalna vrednost ne pade v kritično območje, pa pravimo da vzorčni podatki kažejo na statistično neznačilne razilike med parametrom in vzorčno oceno.

I. $H_0 : \mu = \mu_0$

Če poznamo odklon σ, potem

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$ **sledi z-porazdelitev.**

Primer (B)

Proizvajalec omake za špagete da v vsako posodo 28 unče omake za špagete. Količina omake, ki je v vsaki posodi, je porazdeljena normalno s standardnim odklonom $0,005$ unče.

Podjetje ustavi proizvodni trak in popravi napravo za polnenje, če so posode bodisi

– premalo napolnjene (to razjezi kupce),

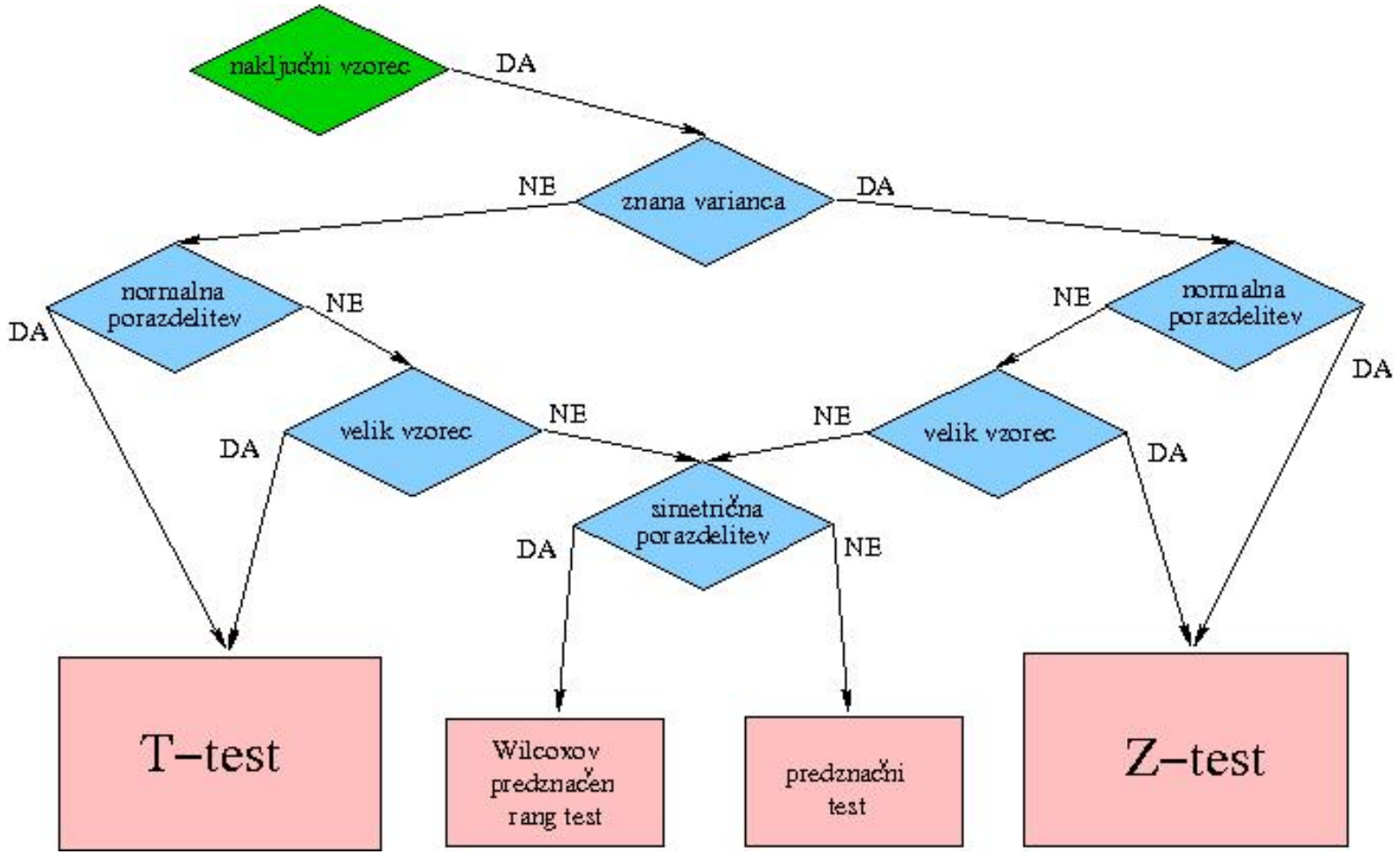
– ali preveč napolnjene (kar seveda pomeni manjši profit).

Ali naj na osnovi vzorca iz 15ih posod ustavijo proizvodno linijo? Uporabi stopnjo značilnosti $0,05$.

Postavimo hipotezo

- ničelna hipoteza
  - $H_0 \ : \ \mu = 28$
- alternativna hipoteza
  - $H_a \ : \ \mu \neq 28$

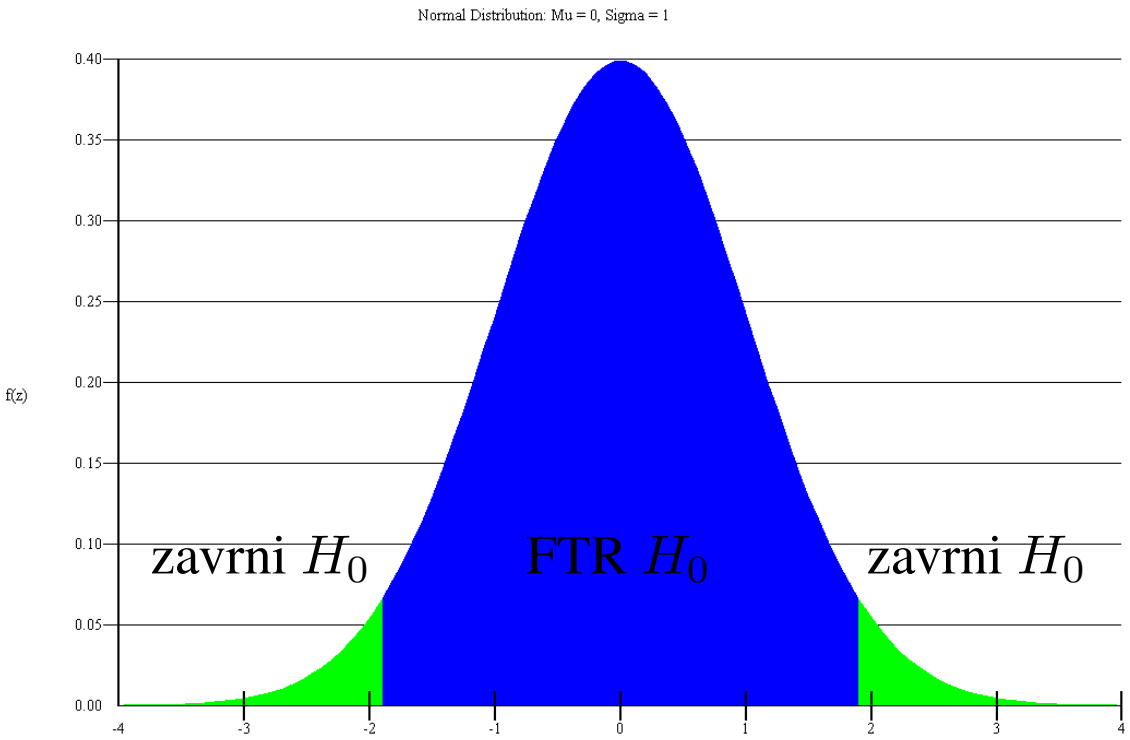
Izberimo testno statistiko



Z-Test

- test
  - $H_0 \ : \ \mu = \mu_0$ (povprečje populacije)
- predpostavke
  - naključno vzorčenje
  - poznamo varianco populacije
  - izbiramo vzorce iz normalne porazdelitve in/ali imamo vzorec pri katerem je n velik.

Določimo zavrnitveni kriterij



Rezultati testiranja

- naredi naključni vzorec
  - vzorčno povprečje: $28,0165$
- izračunaj vrednost testne statistike

$$Z = (28,0165 - 28)/0,0129 = 1,278$$

- naredi odločitev
  - FTR H_0
- zaključek
  - privzemi $\mu = 28$

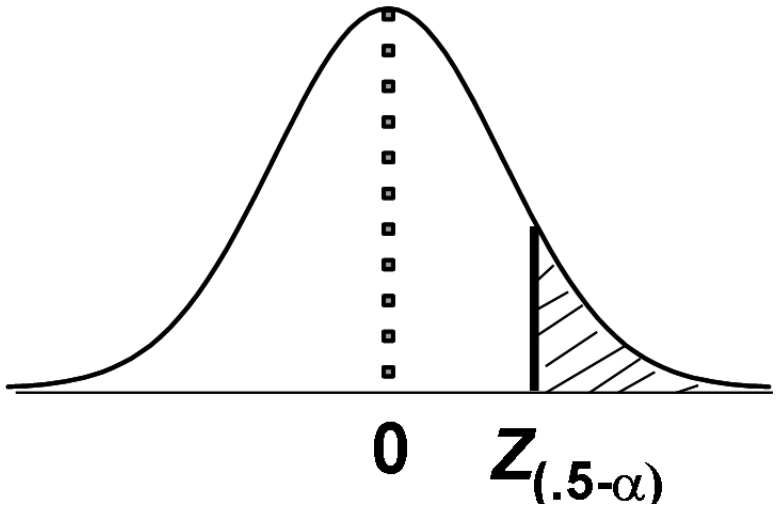
P-vrednost

- Sprejemljivost hipoteze H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je hipoteza H_0 pravilna
  - P-vrednost $= (2)P(Z > 1,278) = (2)(0,1003) = 0,2006$
- Najmanjši α pri katerem zavrnemo hipotezo H_0
  - P-vrednost $> \alpha$, zato FTR H_0

Za $H_a : \mu > \mu_0$

odločitveno pravilo: zavrni H_0, če je

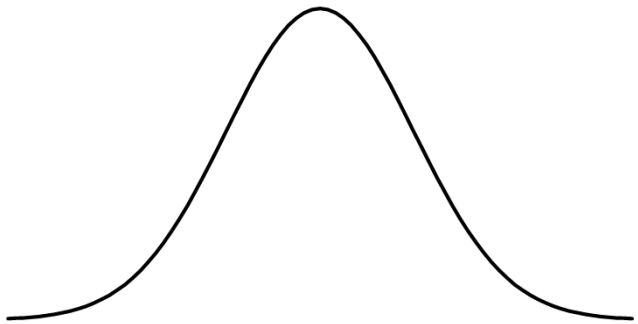
$$\text{T.S.} \geq z_{(0,5-\alpha)}$$



Za $H_a : \mu < \mu_0$

odločitveno pravilo: zavrni H_0, če je

$$\mathbf{T.S.} \leq z_{(0,5-\alpha)}$$

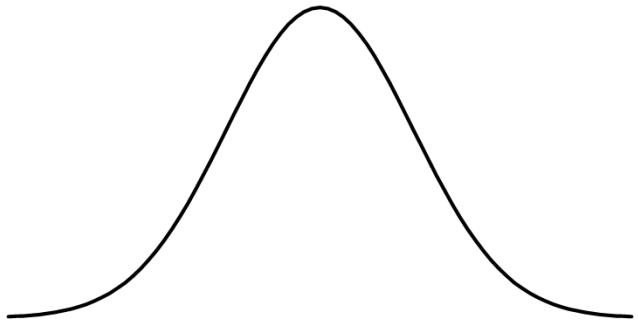


Za $H_a : \mu \neq \mu_0$

odločitveno pravilo: zavrni H_0

če je T.S. $\leq -z_{(0,5-\alpha)}$

ali če je T.S. $\geq z_{(0,5-\alpha)}$



II. $H_0 : \mu = \mu_0$

Če ne poznamo odklona σ in je $n \geq 30$, potem

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{\frac{s}{\sqrt{n}}} \quad \textbf{sledi}$$

t-porazdelitev z $n-1$ prostostnimi stopnjami.

(Velja omeniti še, da se pri tako velikem n z- in t-porazdelitev tako ne razlikujeta kaj dosti.)

III. $H_0 : \mu = \mu_0$

Če ne poznamo odklona σ, populacija je normalna in je $n < 30$, potem

$$\text{T.S.} = \frac{\overline{y} - \mu_0}{\frac{s}{\sqrt{n}}} \quad \textbf{sledi}$$

t-porazdelitev z $n-1$ prostostnimi stopnjami.

Primer (C2)

Za slučajni vzorec: 16-ih odraslih Slovencev smo izračunali povprečno število in variance priznanih let šolanja: $\overline{X} = 9$ in $s^2 = 9$.
Predpostavljamo, da se spremenljivka na populaciji porazdeljuje normalno.

Ali lahko sprejmemo domnevo, da imajo odrasli Slovenci v povprečju več kot osemletko pri 5% stopnji značilnosti?

Postavimo najprej ničelno in osnovno domnevo:

$$H_0 \,:\, \mu = 8 \quad \text{in} \quad H_1 \,:\, \mu > 8.$$

Ustrezna statistike je

$$t = \frac{\overline{X} - \mu_H}{s}\sqrt{n},$$

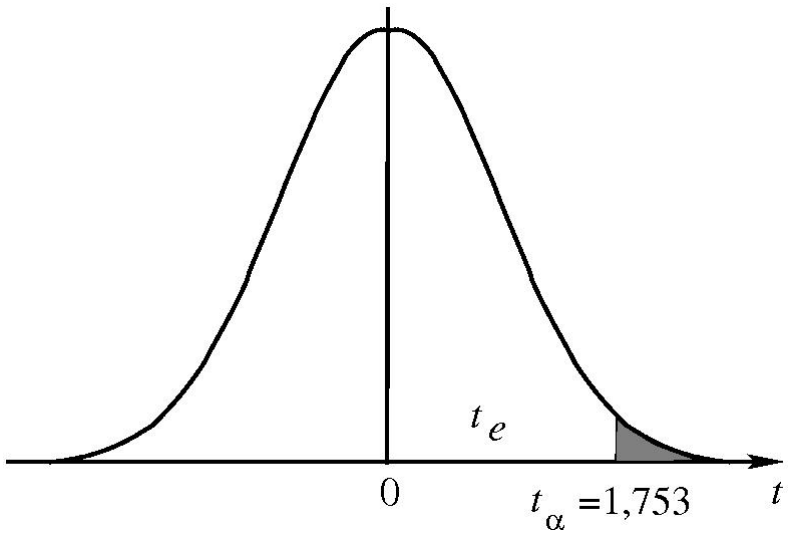
ki se porazdeljnje po t-porazdelitvi s 15 prostostnimi stopnjami.

Ker gre za enostranski test,
je glede na osnovno domnevo
kritično območje na desni strani
porazdelitve in kritična vrednost

$$t_{0,05}(15) = 1,753.$$

Izračunajmo eksperimentalno vrednost
statistike:

$$t_e = \frac{9-8}{3}\sqrt{16} = 1,3$$



Eksperimentalna vrednost (T.S.) ne pade v kritično območje.
Zato ničelne domneve ne moremo zavrniti
in sprejeti osnovne domneve,
da imajo odrasli Slovenci več kot osemletko.

Primer (C)

Ravnatelj bežigrajske gimnazije trdi, da imajo najboljši PT program v Sloveniji s povprečjem APFT 240.

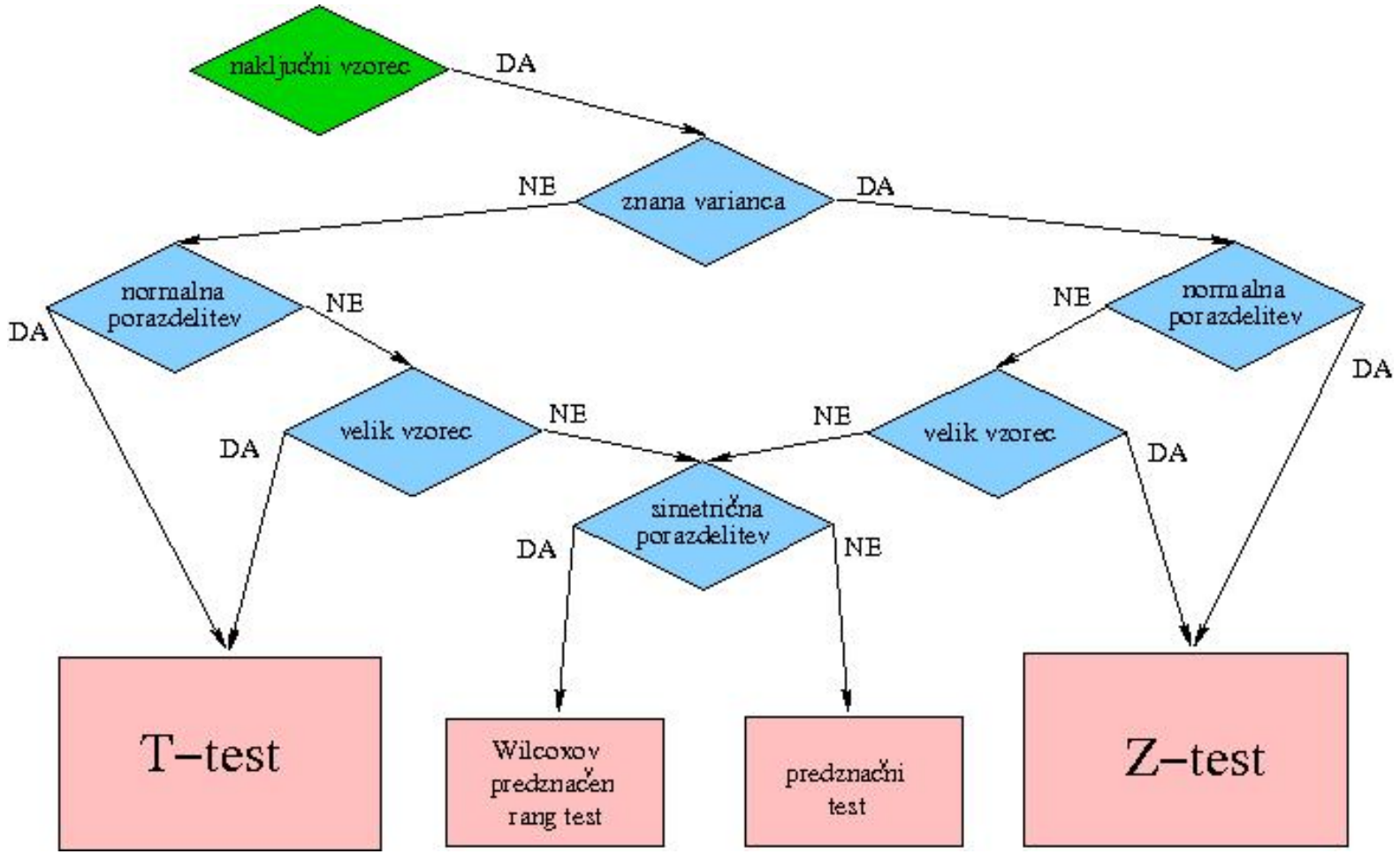
Predpostavi, da je porazdelitev rezultatov testov približno normalna.

Uporabi $\alpha = 0,05$ za določitev ali je povprečje APFT rezultatov šestih naključno izbranih dijakov iz bežigrajske gimnazije statistično večje od 240?

Postavimo hipotezi:

- $H_0 \ : \ \mu = 240$
- $H_a \ : \ \mu > 240$

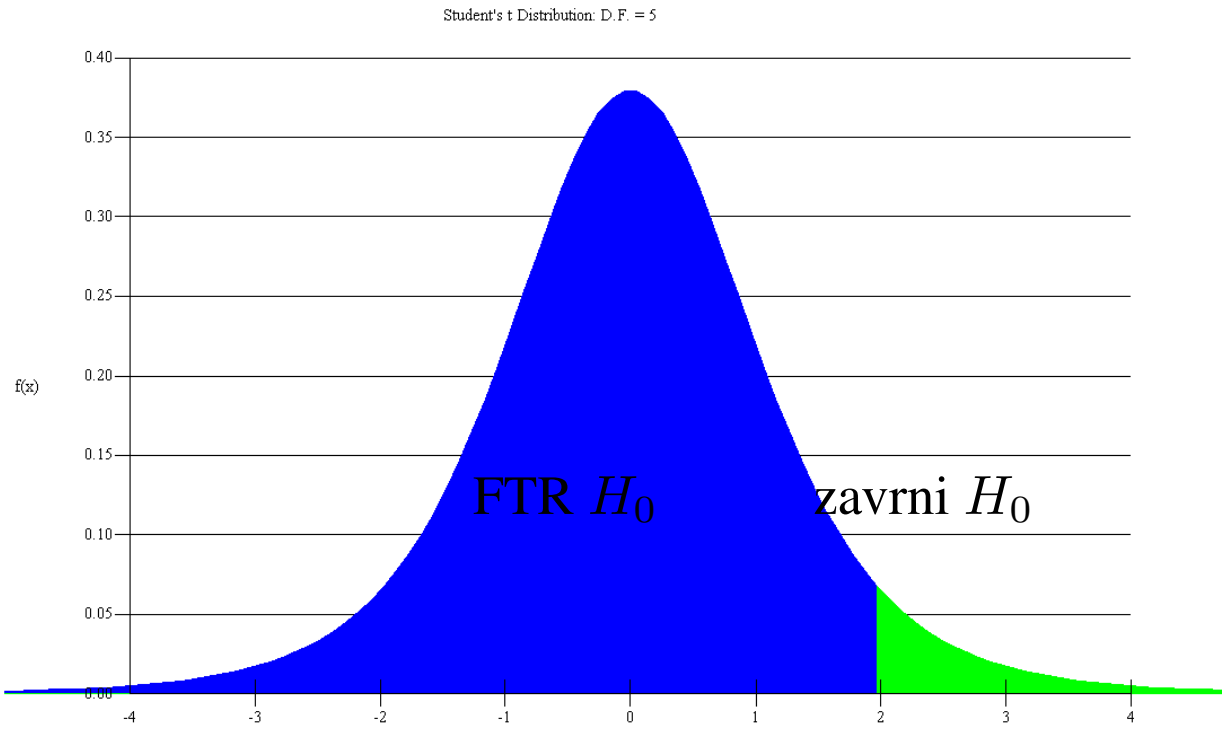
Izberimo testno statistiko



T-test

- test
  - $H_0 \ : \ \mu = \mu_0$ (povprečje populacije)
- predpostavke
  - naključno vzorčenje
  - ne poznamo varianco populacije
  - izbiramo vzorce iz normalne porazdelitve
    in/ali imamo vzorec pri katerem je n velik.

Določimo zavrnitveni kriterij



Rezultati testov

- naredi naključni vzorec
  - vzorčno povprečje: 255,4
  - vzorčni standardni odklon: 40,07
- izračunaj vrednost testne statistike

$$T = (255,4 - 240)/16,36 = 0,9413$$

- sprejmi odločitev
  - FTR H_0
- zaključek
  - Bežigrajska gimnazija ne more pokazati, da imajo višje povprečje APFT rezultatov, kot slovensko povprečje.

P-vrednost

- Sprejemljivost hipoteze H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je hipoteza H_0 pravilna
  - P-vrednost $= P(T > 0,9413) = 0,1949.$
- Najmanjši α pri katerem zavrnemo hipotezo H_0
  - P-vrednost $> \alpha$, zato FTR H_0 .

Vstavimo podatke v Minitab
(Ex9-23.MTV)

C1:
2610
2750
2420
2510
2540
2490
2680

T-test povprečja

```
Test of mu = 2500.0 vs mu > 2500.0

   N       MEAN    STDEV     SE MEAN
C1 7     2571.4    115.1        43.5

      T      p-VALUE
   1.64        0.076
```

Razlaga P-vrednosti

1. Izberi največjo vrednost za α,
   ki smo jo pripravljeni tolerirati.

2. Če je P-vrednost testa manjša kot
   maksimalna vrednost parametra α,
   potem zavrni ničelno hipotezo.

Razlika povprečij dveh populaciji

IV. $H_0 : \mu_1 - \mu_2 = D_0$

Če poznamo σ_1 in σ_2

ter jemljemo vzorce neodvisno, potem

$$\text{T.S.} = \frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ **sledi z-porazdelitev.**

Primer 1: Preveriti želimo domnevo, da so dekleta na izpitu boljša od fantov. To domnevo preverimo tako, da izberemo slučajni vzorec 36 deklet in slučajni vzorec 36 fantov, za katere imamo izpitne rezultate, na katerih izračunamo naslednje statistične karakteristike:

$$\overline{X}_F = 7,0, \quad s_F = 1$$

$$\overline{X}_D = 7,2, \quad s_D = 1$$

Domnevo preverimo pri 5% stopnji značilnosti.

Postavimo ničelno in osnovno domnevo:

$$H_0 : \mu_D = \mu_F \quad \text{oziroma} \quad \mu_D - \mu_F = 0,$$

$$H_1 : \mu_D > \mu_F \quad \text{oziroma} \quad \mu_D - \mu_F > 0.$$

Za popularijsko razliko aritmetičnih sredin na vzorcih računamo vzorčno razliko aritmetičnih sredin, ki se za dovolj velike vzorce porazdeljuje normalno

$$\overline{X}_D - \overline{X}_F \ : \ N\left(\mu_D - \mu_F, \sqrt{\frac{s_D^2}{n_D} + \frac{s_F^2}{n_F}}\right).$$

oziroma statistika

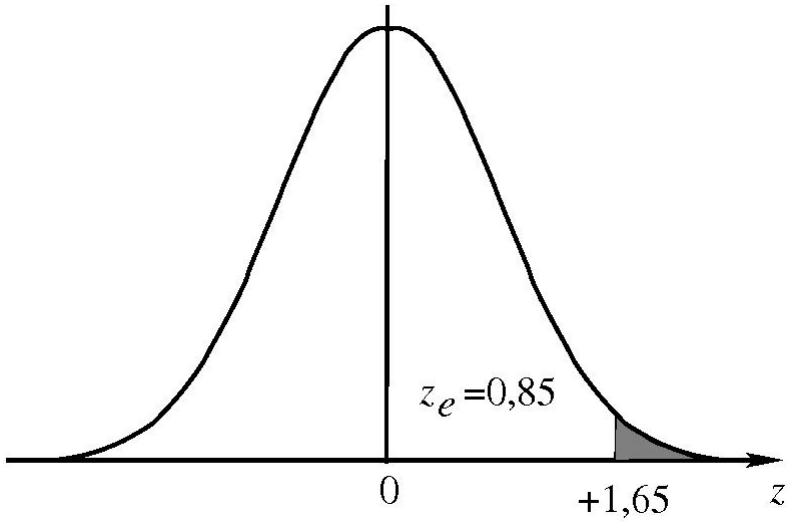
$$z = \frac{\overline{X}_D - \overline{X}_F - (\mu_D - \mu_F)_H}{\sqrt{\frac{s_D^2}{n_D} + \frac{s_F^2}{n_F}}}$$

standardizirano normalno $N(0,1)$.

Osnovna domneva kaže enostranski test: možnost napake 1. vrste je le na desni strani normalne porazdelitve, kjer zavračamo ničelno domnevo. Zato je kritično območje določeno z vrednostmi večjimi od 1,65.

Eksperimentalna vrednost statistike je

$$z_e = \frac{7,2 - 7 - 0}{\sqrt{\frac{1}{36} + \frac{1}{36}}} = 0,852.$$



Eksperimentalna vrednost ne pade v kritično območje.

Ničelne domneve ne moremo zavrniti.

Povprečna uspešnost deklet in fantov ni statistično značilno različna.

V. $H_0 : \mu_1 - \mu_2 = D_0$

Če ne poznamo σ_1 in/ali σ_2,

vzorce jemljemo neodvisno,

$n_1 \geq 30$ in/ali $n_2 \geq 30$, potem

$$\text{T.S.} = \frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

sledi z-porazdelitev.

VI. $H_0 : \mu_1 - \mu_2 = D_0$

Če ne poznamo σ_1 in/ali σ_2,
vzorce jemljemo neodvisno,
populaciji sta normalno porazdeljeni,
varianci obeh populacij sta enaki,
$n_1 < 30$ ali $n_2 < 30$, potem

T.S. $=$

$$\frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$$

sledi t-porazdelitev z $n_1 + n_2 - 2$ prostostnimi stopnjami.

Privzeli smo:

1. Populaciji iz katerih jemljemo vzorce imata obe približno **normalno** relativno porazdelitev frekvenc.
2. Varianci obeh populacij sta **enaki**.
3. Naključni vzorci so izbrani **neodvisno** iz obeh populacij.

VII. $H_0 : \mu_1 - \mu_2 = D_0$

Če ne poznamo σ_1 in/ali σ_2,

vzorce jemljemo neodvisno,

spremenljivki sta vsaka na svoji populacij

normalno porazdeljeni, njuni varianci nista enaki,

$n_1 < 30$ ali $n_2 < 30$, potem

$$\mathbf{T.S.} = \frac{(\overline{y}_1 - \overline{y}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \textbf{sledi}$$

t-porazdelitev z ν prostostnimi stopnjami,

kjer je

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Če ν ni naravno število, zaokroži ν navzdol do najbližjega naravnega števila za uporabo t-tabele.

VIII. $H_0 : \mu_d = D_0$

Če vzorce jemljemo neodvisno

in če je $n \geq 30$, potem

$$\text{T.S.} = \frac{\overline{d} - D_0}{\frac{s_d}{\sqrt{n}}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

IX. $H_0 : \mu_d = D_0$

Če vzorce ne jemljemo neodvisno,

če je populacija razlik normalno porazdeljena

in če je $n \leq 30$, potem

$$\text{T.S.} = \frac{\overline{d} - D_0}{\frac{s_d}{\sqrt{n}}} \quad \textbf{sledi } t\textbf{-porazdelitev}$$

z $n - 1$ prostostnimi stopnjami.

naloga	clovek. urnik	avtomatizirana metoda
1	185,4	180,4
2	146,3	248,5
3	174,4	185,5
4	184,9	216,4
5	240,0	269,3
6	253,8	249,6
7	238,8	282,0
8	263,5	315,9

naloga	clovek. urnik	avtomatizirana metoda	razlika
1	185,4	180,4	5,0
2	146,3	248,5	-102,2
3	174,4	185,5	-11,1
4	184,9	216,4	-31,5
5	240,0	269,3	-29,3
6	253,8	249,6	-4,2
7	238,8	282,0	-43,2
8	263,5	315,9	-52,4

Vstavimo podatke v Minitab (Ex9-40.MTV)

C1: 185,4 146,3 174,4 184,9 240,0 253,8 238,8 263,5

C2: 180,4 248,5 185,5 216,4 269,3 249,6 282,0 315,9

Test za parjenje in interval zaupanja

Parjen T za C1-C2

	N	povpr.	StDev	SE povpr.
C1	8	210,9	43,2	15.3
C2	8	243,4	47,1	16,7
Razlika	8	032,6	35,0	12,4

95% interval zaupanja za razliko povprečja: $(-61,9; -3,3)$

T-test za razliko povpr. = 0 (proti $\neq 0$):

T-vrednost=-2,63

P-vrednost=0,034.

X. $H_0 : p = p_0$

Če je n dovolj velik, potem

$$\text{T.S.} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \quad \textbf{sledi } z\textbf{-porazdelitev.}$$

Kot splošno pravilo bomo zahtevali, da velja

$$n\hat{p} \geq 4 \quad \text{in} \quad n\hat{q} \geq 4.$$

Primer (C0)

Postavimo domnevo o vrednosti parametra, npr. π – delež enot z določeno lastnostjo na populaciji. Denimo, da je domneva

$$H \ : \ \pi_H = 0,36$$

Tvorimo slučajne vzorce npr. velikosti $n = 900$ in na vsakem vzorcu določimo vzorčni delež p (delež enot z določeno lastnostjo na vzorcu). Ob predpostavki, da je domneva pravilna, vemo, da se vzorčni deleži porazdeljujejo približno normalno

$$N\left(\pi_H, \sqrt{\frac{\pi_H(1-\pi_H)}{n}}\right)$$

Vzemimo en slučajni vzorec z vzorčnim deležem p. Ta se lahko bolj ali manj razlikuje od π_H. Če se zelo razlikuje, lahko podvomimo o resničnosti domneve π_H. Zato okoli π_H naredimo območje sprejemanja domneve in izven tega območja območje zavračanja domneve.

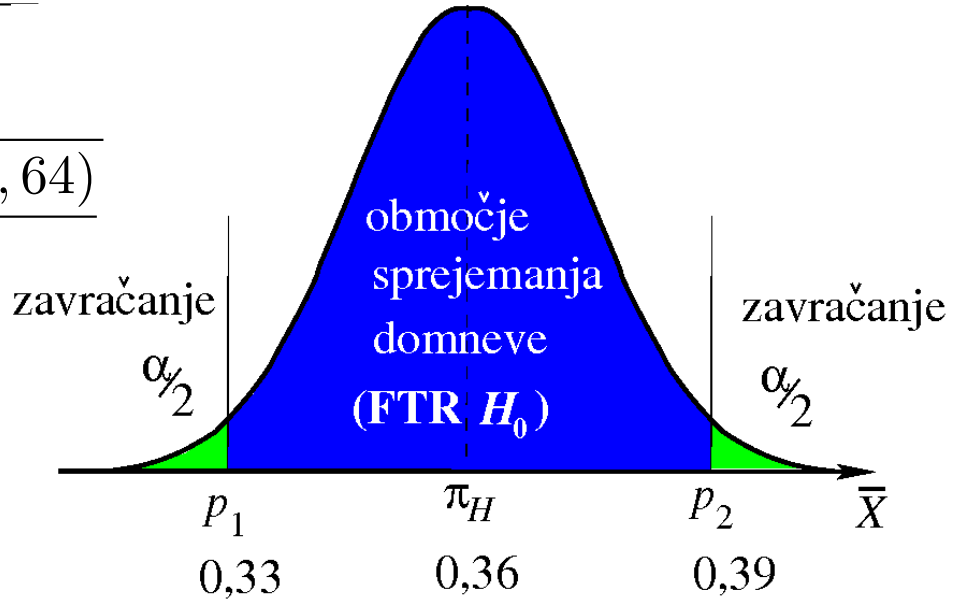
Denimo, da je območje zavračanja določeno s 5% vzorcev,
ki imajo ekstremne vrednosti deležev (2,5% levo in 2,5% desno).

Deleža, ki ločita območje sprejemanja
od območja zavračanja lahko izračunamo takole:

$$p_{1,2} = \pi_H \pm z_{\alpha/2}\sqrt{\frac{\pi_H(1-\pi_H)}{n}}$$

$$p_{1,2} = 0,36 \pm 1,96\sqrt{\frac{0,36 \times 0,64)}{900}}$$

$$= 0,36 \pm 0,03$$



Kot smo že omenili, je sprejemanje ali zavračanje domnev po opisanem postopku lahko napačno v dveh smislih:

Napaka 1. vrste (α):
Če vzorčna vrednost deleža pade v območje zavračanja, domnevo π_H zavrnemo. Pri tem pa vemo, da ob resnični domnevi π_H obstajajo vzorci, ki imajo vrednosti v območju zavračanja.
Število α je verjetnost, da vzorčna vrednost pade v območje zavračanja ob prodpostavki, da je domneva resnična.
Zato je α verjetnost, da zavrnemo pravilno domnevo – **napaka 1. vrste**.
Ta napaka je merljiva in jo lahko poljubno manjšamo.

Napaka 2. vrste (β):

Vzorčna vrednost lahko pade v območje sprejemanja,
čeprav je domnevna vrednost parametra napačna.

V primeru, ki ga obravnavamo,
naj bo prava vrednost deleža na populaciji $\pi = 0,40$.

Tedaj je porazdelitev vzorčnih deležev

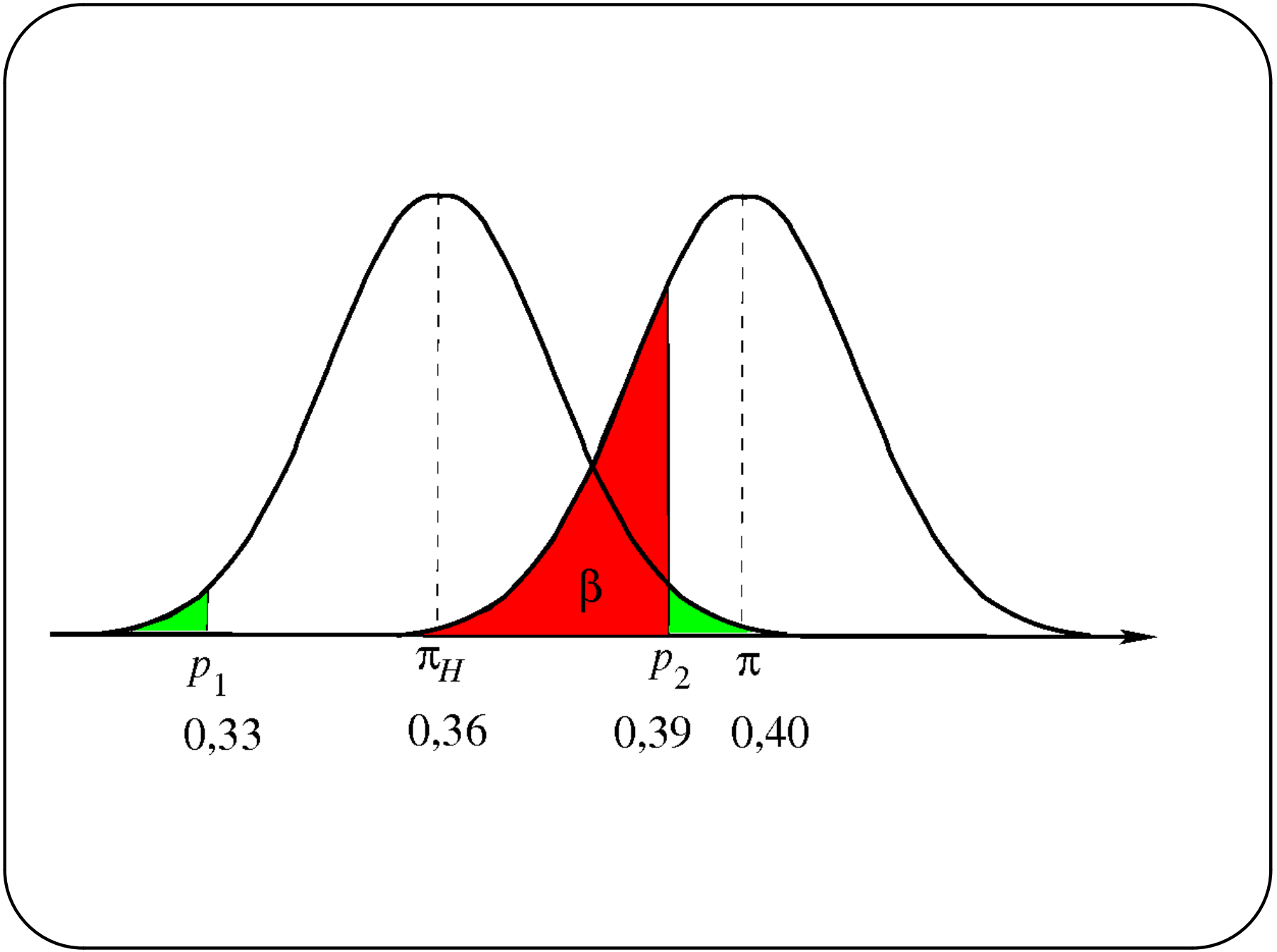
$$N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right) = N(0,40; 0,0163)$$

Ker je območje sprejemanja, domneve v intervalu $0,33 < p < 0,39$,
lahko izračunamo verjetnost, da bomo sprejeli napačno domnevo takole:

$$\beta = P(0,33 < p < 0,39) = 0,27$$

Napako 2. vrste lahko izračunamo le,
če imamo znano resnično vrednost parametra π.

Ker ga ponavadi ne poznamo, tudi ne poznamo napake 2. vrste.
Zato ne moremo sprejemati domnev.


β
p_1
π_H
p_2
π
0,33
0,36
0,39
0,40

Primer (D)

Državni zapisi indicirajo, da je od vseh vozil, ki gredo skozi testiranje izpušnih plinov v preteklem letu, 70% uspešno opravilo testiranje v prvem poskusu.

Naključni vzorec 200ih avtomobilov testiranih v določeni pokrajni v tekočem letu je pokazalo, da jih je 156 šlo čez prvi test.

Ali to nakazuje, da je dejanski delež populacije za to pokrajno v tekočem letu različno od preteklega državnega deleža?

Pri testiranju hipoteze uporabi $\alpha = 0,05$.

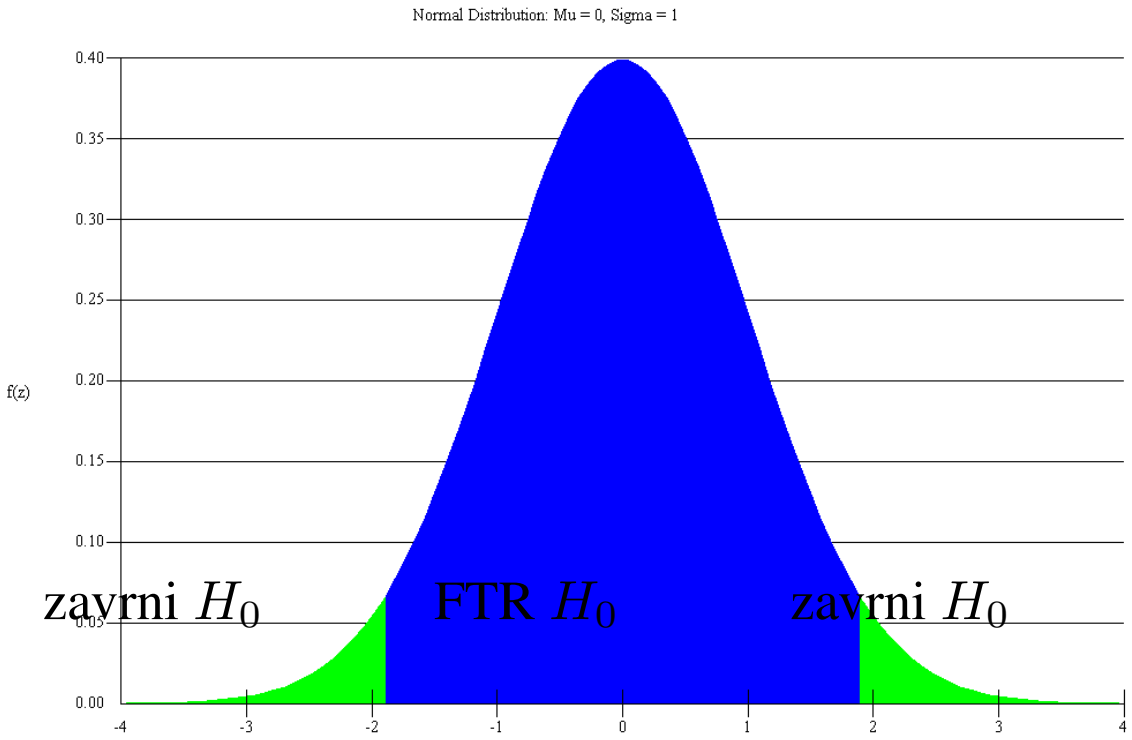
Testiranje hipoteze za delež

- Ničelna hipoteza $H_0 \ : \ p = 0,7$
- Alternativna hipoteza $H_a \ : \ p \neq 0,7$
- Test
  - $H_0 \ : \ p = p_0$ (delež populacije)
- Predpostavke
  - naključni vzorec
  - izbiranje vzorca iz binomske porazdelitve
- Testna statistika

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$$

$$n\hat{p} \geq 4 \quad \text{in} \quad n\hat{q} \geq 4.$$

Določimo zavrnitveni kriterij



Rezultati testiranja

- Naredi naključni vzorec
  - delež vzorca: $156/200 = 0,78$
- Izračunaj vrednost testne statistike

$$Z = (0,78 - 0,7)/0,0324 = 2,4688$$

- Naredi odločitev
  - zavrni hipotezo H_0
- Zaključek
  - Pokrajna ima drugačen kriterij.

P-vrednost

- Sprejemljivost hipoteze H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je hipoteza H_0 pravilna
  - P-vrednost $= (2) * P(Z > 2,469) = (2) * (0,0068) = 0,0136$

- Najmanjši α pri katerem zavrnemo hipotezo H_0
  - P-vrednost $< \alpha$, zato zavrni hipotezo H_0

Razlika deležev dveh populaciji

Velik vzorec za testiranje hipoteze o $p_1 - p_2$

Kot splošno pravilo bomo zahtevali, da velja

$$n_1\hat{p}_1 \geq 4, \qquad n_1\hat{q}_1 \geq 4,$$

$$n_2\hat{p}_2 \geq 4 \quad \text{in} \quad n_2\hat{q}_2 \geq 4.$$

XI. Velik vzorec za testiranje hipoteze o $p_1 - p_2$, kadar je $D_0 = 0$.

$$\text{T.S.} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ **sledi z-porazdelitev.**

kjer je $\hat{p} = \dfrac{y_1 + y_2}{n_1 + n_2}$.

Primer (D3)

Želimo preveriti, ali je predsedniški kandidat različno priljubljen med mestnimi in vaškimi prebivalci. Zato smo slučajni vzorec mestnih prebivalcev povprašali, ali bi glasovali za predsedniškega kandidata.

Od 300 vprašanih (n_1) jih je 90 glasovalo za kandidata (k_1).
Od 200 slučajno izbranih vaških prebivalcev (n_2) pa je za kandidata glasovalo 50 prebivalcev (k_2).

Domnevo, da je kandidat različno priljubljen v teh dveh območjih preverimo pri 10% stopnji značinosti.

$$H_0 \ : \ \pi_1 = \pi_2 \quad \text{oziroma} \quad \pi_1 - \pi_2 = 0,$$

$$H_1 \ : \ \pi_1 \neq \pi_2 \quad \text{oziroma} \quad \pi_1 - \pi_2 \neq 0.$$

Vemo, da se razlika vzorčnih deležev porazdeljuje približno normalno:

$$p_1 - p_2 \ : \ N\left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}\right).$$

Seveda π_1 in π_2 nista znana. Ob predpostavki, da je ničelna domneva pravilna, je matematično upanje razlike vzorčnih deležev hipotetična vrednost razlike deležev, ki je v našem primeru enaka 0. Problem pa je, kako oceniti standardni odklon. Ker velja domneva $\pi_1 = \pi_2 = \pi$, je disperzija razlike vzorčnih deležev

$$\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} = \frac{\pi(1-\pi)}{n_1} + \frac{\pi(1-\pi)}{n_2} = \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Populacijski delež π ocenimo z obteženim povprečjem vzorčnih deležev p_1 in p_2

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{k_1 + k_2}{n_1 + n_2}.$$

Vrnimo se na primer. Vzorčna deleža sta:

$$p_1 = \frac{90}{300} = 0,30, \qquad p_2 = \frac{50}{200} = 0,25.$$

Ocena populacijskega deleža je

$$P = \frac{50 + 90}{200 + 300} = 0,28.$$

Kot smo že omenili, se statistika

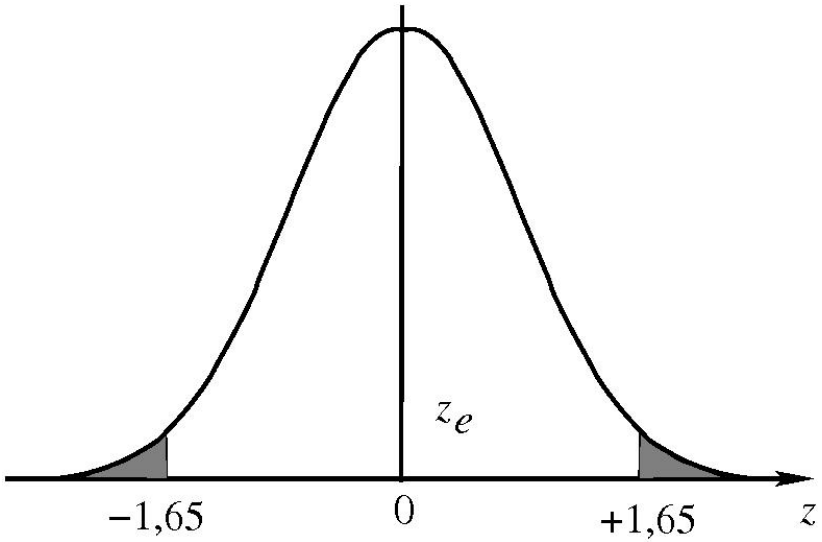
$$z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)_H}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

porazdeljuje približno standardizirano normalno $N(0,1)$.

Ker gre za dvostranski test, sta kritični vrednosti $\pm z_{\alpha/2} = \pm 1,65$.

Eksperimentalna vrednost statistike pa je

$$z_e = \frac{0,30 - 0,025 - 0}{\sqrt{0,28(1-0,28)\left(\frac{1}{300} + \frac{1}{200}\right)}} = 1,22.$$



Eksperimentalna vrednost ne pade v kritično območje. Zato ničelne domneve ne moremo zavrniti. Priljubljenost predsedniškega kandidata ni statistično značilno različna med mestnimi in vaškimi prebivalci.

XII. Velik vzorec za testiranje hipoteze o $p_1 - p_2$, kadar je $D_0 \neq 0$.

$$\text{T.S.} = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

sledi z-porazdelitev.

Primer

Neka tovarna cigaret proizvaja dve znamki cigaret. Ugotovljeno je, da ima 56 od 200 kadilcev raje znamko A in da ima 29 od 150 kadilcev raje znamko B.

Testiraj hipotezo pri $0,06$ stopnji zaupanja, da bo prodaja znamke A boljša od prodaje znamke B za 10% proti alternativni hipotezi, da bo razlika manj kot 10%.

Analiza variance

Če opravljamo isti poskus v nespremenjenih pogojih, kljub temu v rezultatu poskusa opažamo spremembe (variacije) ali odstopanja.

Ker vzrokov ne poznamo in jih ne moremo kontrolirati, spremembe pripisujemo *slučajnim vplivom* in jih imenujemo *slučajna odstopanja*.

Če pa enega ali več pogojev v poskusu spreminjamo, seveda dobimo dodatna odstopanja od povprečja. Analiza tega, ali so odstopanja zaradi sprememb različnih faktorjev ali pa zgolj slučajna, in kateri faktorji vplivajo na variacijo, se imenuje *analiza variance*.

Zgleda:

(a) Namesto dveh zdravil proti nespečnosti kot v Studentovem primeru lahko preskušamo učinkovitost več različnih zdravil A, B, C, D,... in s preskušanjem hipoteze $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ raziskujemo, ali katero od zdravil sploh vpliva na rezultat. Torej je to posplošitev testa za $H_0 : \mu_1 = \mu_2$

(b) Raziskujemo hektarski donos pšenice. Nanj vplivajo različni faktorji: različne sorte pšenice, različni načini gnojenja, obdelave zemlje itd., nadalje klima, čas sejanja itd .

Analiza variance je nastala prav v zvezi z raziskovanjem v kmetijstvu. Glede na število faktorjev, ki jih spreminjamo, ločimo t.i. *enojno klasifikacijo* ali *enofaktorski eksperiment*, *dvojno klasifikacijo* ali *dvofaktorski eksperiment*, itd.

. . . Analiza variance

V izrazu

$$Q_v^2 = \sum_{i=1}^{r}\sum_{k=1}^{n_i}(X_{ik} - \overline{X}_i)^2 = \sum_{i=1}^{r}(n_i - 1)S_i^2$$

je

$$S_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(X_{ik} - \overline{X}_i)^2$$

nepristranska cenilka za disperzijo v i-ti skupini; neodvisna od S_j^2, za $i \neq j$.
Zato ima

$$\frac{Q_v^2}{\sigma^2} = \sum_{i=1}^{r}(n_i - 1)\frac{S_i^2}{\sigma^2}$$

porazdelitev $\chi^2(n-r)$, saj je ravno $\sum_{i=1}^{r}(n_i - 1) = n - r$ prostostnih stopenj.

Ker je $\mathsf{E}\dfrac{Q_v^2}{\sigma^2} = n - r$, je tudi $S_v^2 = \dfrac{1}{n-r}Q_v^2$ nepristranska cenilka za σ^2.

. . . Analiza variance

Izračunajmo še Q_m^2 pri predpostavki o veljavnosti osnovne domneve H_0. Dobimo

$$Q_m^2 = \sum_{i=1}^{r} n_i(\overline{X}_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

Torej je

$$\frac{Q_m^2}{\sigma^2} = \sum_{i=1}^{r} n_i\Big(\frac{\overline{X}_i - \mu}{\sigma/\sqrt{n_i}}\Big)^2 - n\Big(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\Big)^2$$

od tu sprevidimo, da je statistika $\frac{Q_m^2}{\sigma^2}$ porazdeljena po $\chi^2(r-1)$.

Poleg tega je $S_m^2 = \dfrac{Q_m^2}{r-1}$ nepristranska cenilka za σ^2, neodvisna od S_v^2.

. . . Analiza variance

Ker sta obe cenilki za varianco σ^2, pri domnevi H_0, njuno razmerje $F = \frac{S_m^2}{S_v^2}$ ne more biti zelo veliko. Iz

$$F = \frac{S_m^2}{S_v^2} = \frac{Q_m^2/(r-1)}{Q_v^2/(n-r)} = \frac{\frac{Q_m^2}{\sigma^2}/(r-1)}{\frac{Q_v^2}{\sigma^2}/(n-r)}$$

vidimo da gre za Fisherjevo (Snedecorjevo) porazdelitev $F(r-1, n-r)$.

Podatke zapišemo v *tabelo analize variance*

VV	VK	PS	PK	F
faktor	Q_m^2	$r-1$	S_m^2	F
slučaj	Q_v^2	$n-r$	S_v^2	
	Q^2	$n-1$		

Analiza variance v R-ju

Zgled: Petnajst enako velikih njiv je bilo posejanih z isto vrsto pšenice, vendar gnojeno na tri različne načine – z vsakim po pet njiv.

```
> ena <- c(47,47,40,32,40)
> dva <- c(76,68,71,46,54)
> tri <- c(49,40,34,36,44)
> d <- stack(list(e=ena,d=dva,t=tri))
> names(d)
[1] "values" "ind"
> oneway.test(values ~ ind, data=d, var.equal=TRUE)

        One-way analysis of means

data:  values and ind
F = 10.5092, num df = 2, denom df = 12, p-value = 0.002304
> av <- aov(values ~ ind, data=d)
> summary(av)
            Df  Sum Sq Mean Sq F value   Pr(>F)
ind          2 1628.93  814.47  10.509 0.002304 **
Residuals   12  930.00   77.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Domnevo H_0 zavrnemo.

XIII. Testiranje hipoteze o varianci populacije $H_0 : \sigma^2 = \sigma_0^2$

$$\text{T.S.} = \frac{(n-1)s^2}{\sigma_0^2} \text{ sledi } \chi^2\text{−porazd.}$$

Če je

- $H_a : \sigma^2 > \sigma_0^2$, potem je **odločitveno pravilo**:
  zavrni ničelno hipotezo, če je test statistike večji ali enak $\chi^2_{(\alpha,n-1)}$.
- $H_a : \sigma^2 < \sigma_0^2$ potem je **odločitveno pravilo**:
  zavrni ničelno hipotezo, če je test statistike manjši ali enak $\chi^2_{(\alpha,n-1)}$.
- $H_a : \sigma^2 \neq \sigma_0^2$, potem je **odločitveno pravilo**:
  zavrni ničelno hipotezo, če je test statistike manjši ali enak $\chi^2_{(\alpha,n-1)}$
  ali če je test statistike večji ali enak $\chi^2_{(\alpha,n-1)}$.

Primer (E)

Količina pijače, ki jo naprava za mrzle napitke zavrže je normalno porazdeljena s povprečjem 12 unčev in standardnim odklonom 0,1 unče.

Vsakič, ko servisirajo napravo, si izberejo 10 vzorcev
in izmerijo zavrženo tekočino.

Če je razpršenost zavržene količine prevelika, potem mora naprava na servis.

Ali naj jo odpeljejo na servis?

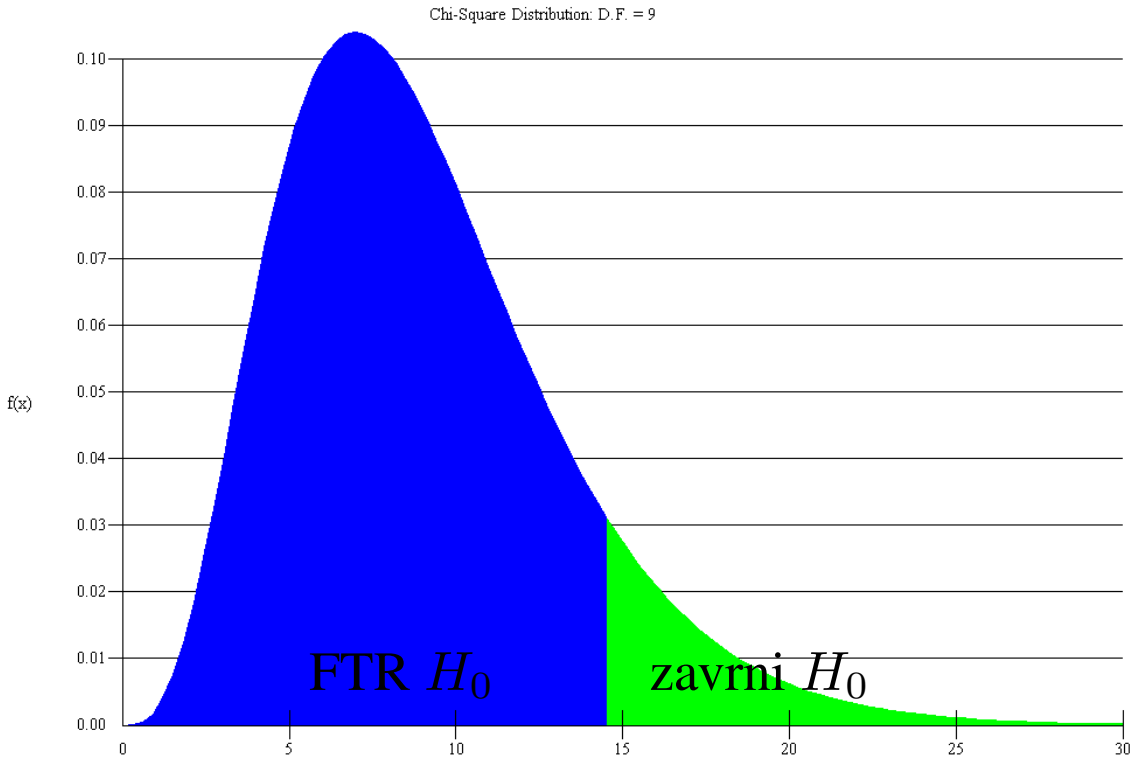
Uporabi $\alpha = 0,1$.

Testiranje hipoteze za varianco

- Ničelna hipoteza $H_0 \ : \ \sigma^2 = 0,01$,
- Alternativna hipoteza $H_a \ : \ \sigma^2 > 0,01$,
- Predpostavke
  - naključni vzorec
  - vzorčenje iz normalne porazdelitve.
- Testna statistika

$$\chi^2_{\nu=n-1} = \frac{S^2(n-1)}{\sigma_0^2}.$$

Določimo zavrnitveni kriterij



14,684

Rezultati testiranja

- naredi naključni vzorec
  - varianca vzorca: $0,02041$
- izračunaj vrednost testne statistike

$$\chi^2 = (0,02041)(9)/(0,01) = 18,369$$

- naredi odločitev
  - zavrni H_0
- zaključek
  - popravi napravo

P-vrednost

- Sprejemljivost hipoteze H_0 na osnovi vzorca
  - možnost za opazovanje vzorca (ali bolj ekstremno podatkov), če je hipoteza H_0 pravilna
  - P-vrednost $= P(\chi^2 > 18,369) = 0,0311$
- Najmanjši α pri katerem zavrnemo hipotezo H_0
  - P-vrednost $< \alpha$, zato zavrni hipotezo H_0.

XIV. Testiranje hipoteze o kvocientu varianc neodvisnih vzorcev

$$H_0 : \sigma_1^2/\sigma_2^2 = 1$$

Če velja

$$H_a : \sigma_1^2/\sigma_2^2 > 1,$$

potem je **test statistike** enak s_1^2/s_2^2,

odločitveno pravilo pa je:

zavrni ničelno hipotezo, če velja

$$\text{T.S.} \geq F_{\alpha,n_1-1,n_2-1}.$$

Če velja $H_a : \sigma_1^2/\sigma_2^2 < 1$, potem je **test statistike** enak

$$\frac{\text{varianca večjega vzorca}}{\text{varianca manjšega vzorca}}.$$

odločitveno pravilo pa je:

zavrni ničelno hipotezo, če velja $s_1^2 > s_2^2$ in

$$\text{T.S.} \geq F_{\alpha,n_1-1,n_2-1}$$

oziroma zavrni ničelno hipotezo, če velja $s_1^2 < s_2^2$ in

$$\text{T.S.} \geq F_{\alpha,n_2-1,n_1-1}.$$

Preverjanje domnev o porazdelitvi spremenljivke

Do sedaj smo ocenjevali in preverjali domnevo o parametrih populacije kot μ, σ in π.

Sedaj pa bomo preverjali, če se spremenljivka porazdeljuje po določeni porazdelitvi.

Test je zasnovan na dejstvu, kako dobro se prilegajo empirične (eksperimentalne) frekvence vrednosti spremenljivke hipotetičnim (teoretičnim) frekvencam, ki so določene s predpostavljeno porazdelitvijo.

Preverjanje domneve o enakomerni porazdelitvi

Za primer vzemimo met kocke in za spremenljivko število pik pri metu kocke. Preizkusimo domnevo, da je kocka poštena, kar je enakovredno domnevi, da je porazdelitev spremenljivke enakomerna. Tedaj sta ničelna in osnovna domneva

H_0 : spremenljivka se porazdeljuje enakomerno,

H_1 : spremenljivka se ne porazdeljuje enakomerno.

Denimo, da smo 120-krat vrgli kocko ($n = 120$)
in štejemo kolikokrat smo vrgli posamezno število pik.

To so empirične ali opazovane frekvence, ki jih označimo s f_i.
Teoretično, če je kocka poštena, pričakujemo, da bomo dobili vsako vrednost z verjetnostjo 1/6 oziroma 20 krat.

To so teoretične ali pričakovane frekvence, ki jih označimo s f_i'.

Podatke zapišimo v naslednji tabeli

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6
f_i'	20	20	20	20	20	20
f_i	20	22	17	18	19	24

S primerjavo empiričnih frekvenc z ustreznimi teoretičnim frekvencami se moramo odločiti, če so razlike posledica le vzorčnih učinkov in je kocka poštena ali pa sp razlike prevelike, kar kaže, da je kocka nepoštena. Statistika, ki meri prilagojenost empiričnih frekvenc teoretičnim je

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

ki se porazdeljuje po χ^2 porazdelitvi z $m = k-1$ prostostnimi stopnjami, ki so enake številu vrednosti spremenijivke ali celic (k) minus število količin dobljenih iz podatkov, ki so uporabljene za izračun teoretičnih frekvenc.

V našem primeru smo uporabili le eno količino in sicer skupno število metov kocke ($n = 120$). Torej število prostostnih stopenj je $m = k-1 = 6-1 = 5$. Ničelna in osnovna domneva sta tedaj

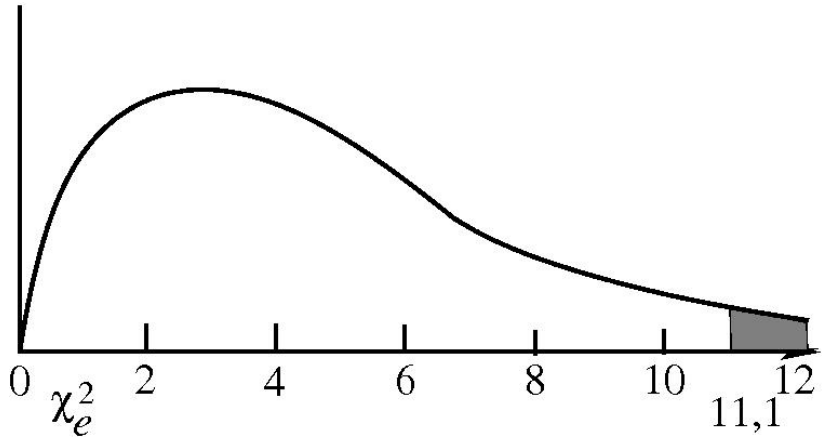
$$H_0 \,:\, \chi^2 = 0 \quad \text{in} \quad H_1 \,:\, \chi^2 > 0.$$

Doomnevo preverimo pri stopnji značilnosti $\alpha = 5\%$.
Ker gre za enostranski test, je kritična vrednost enaka

$$\chi^2_{1-\alpha}(k-1) = \chi^2_{0,95}(5) = 11,1.$$

Eksperimentalna vrednost statistike pa je

$$\begin{aligned}\chi^2_e &= \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} \\ &+ \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} \\ &= \frac{4+9+4+1+16}{20} = \frac{34}{20} = 1,7.\end{aligned}$$



Ker ekperimentalna vrednost statistike ne pade v kritično območje, ničelne domneve ne moremo zavrniti. Empirične in teoretične frekvence niso statistično značilno različne med seboj.

Preverjanje domneve o normalni porazdelitvi

Omenjeni test najpogosteje uporabljamo za preverjanje ali se spremenjivka porazdeljuje normalno.

V tem primeru je izračun teoretičnih frekvenc potrebno vložiti malo več truda.

Primer: Preizkusimo domnevo, da se spremenljivka telesna višina porazdeljuje normalno $N(177, 10)$. Domnevo preverimo pri 5% stopnji značilnosti.

Podatki za 100 slučajno izbranih oseb so urejeni v frekvenčni porazdelitvi takole:

	f_i
nad 150-160	2
nad 160-170	20
nad 170-180	40
nad 180-190	30
nad 190-200	8
	100

Ničelna in osnovna domneva sta tedaj

$$H_0 : \chi^2 = 0 \quad \text{in} \quad H_1 : \chi^2 \neq 0.$$

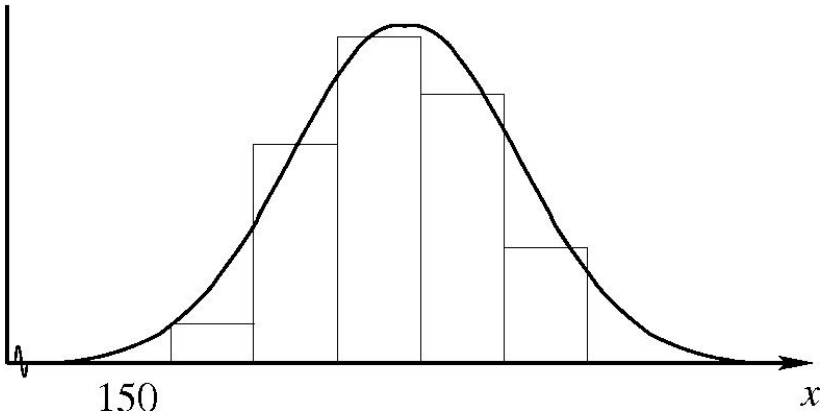
Za test uporabimo statistiko

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

ki se porazdeljuje po χ^2 porazdelitvi z $m = 5 - 1$ prostostnimi stopnjami.
Kritična vrednost je

$$\chi^2_{0,95}(4) = 9,49.$$

V naslednjem koraku je potrebno izračunati teoretične frekvence.
Najprej je potrebno za vsak razred izračunati verjetnost p_i, da spremenljivka zavzame vrednosti določenega intervala, če se porazdeljuje normalno.
To lahko prikažemo na sliki:



Tako je na primer verjetnost, da je višina med 150 in 160 cm:

$$P(150 < X < 160) = P\left(\frac{150-177}{10} < Z < \frac{160-177}{10}\right)$$

$$= P(-2,7 < Z < -1,7) = H(2,7) - H(1,7) = 0,4965 - 0,4554 = 0,0411.$$

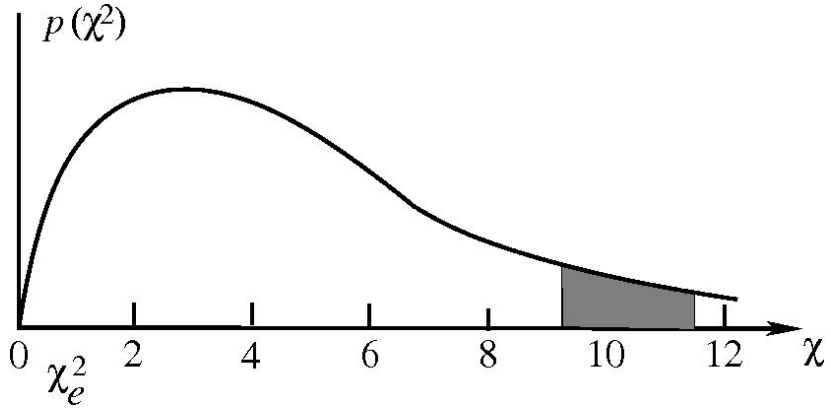
Podobno lahko izračunamo ostale verjetnosti. Teoretične frekvence so $f_i' = n \times p_i$. Izračunane verjetnosti p_i in teoretične frekvence f_i' so

	f_i'	p_i	f_i'
nad 150-160	2	0,0411	4,11
nad 160-170	20	0,1974	19,74
nad 170-180	40	0,3759	37,59
nad 180-190	30	0,2853	28,53
nad 190-200	8	0,0861	8,61
	100		98,58

Eksperimentalna vrednost statistike je tedaj

$$\chi_e^2 = \frac{(2-4.11)^2}{4,11} + \frac{(20-19,74)^2}{19,74} + \frac{(40-37,59)^2}{37,59}$$

$$+\frac{(30-28,53)^2}{28,53} + \frac{(8-8,61)^2}{8,61} \approx 1$$



Ker ekperimentalna vrednost ne pade v kritično območje, ne moremo zavrniti ničelne domneve, da je spremenljivka normalno porazdeljena.

Obstajajo tudi drugi testi za preverjanje porazdelitve spremenljivke, npr. Kolmogorov-Smirnov test.

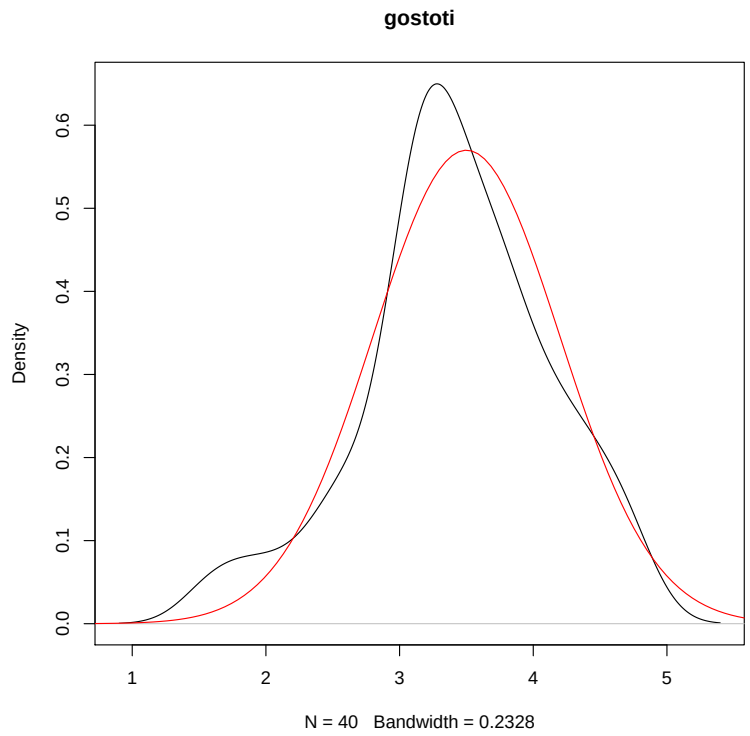
Pri objavi anketiranih rezultatov je potrebno navesti:

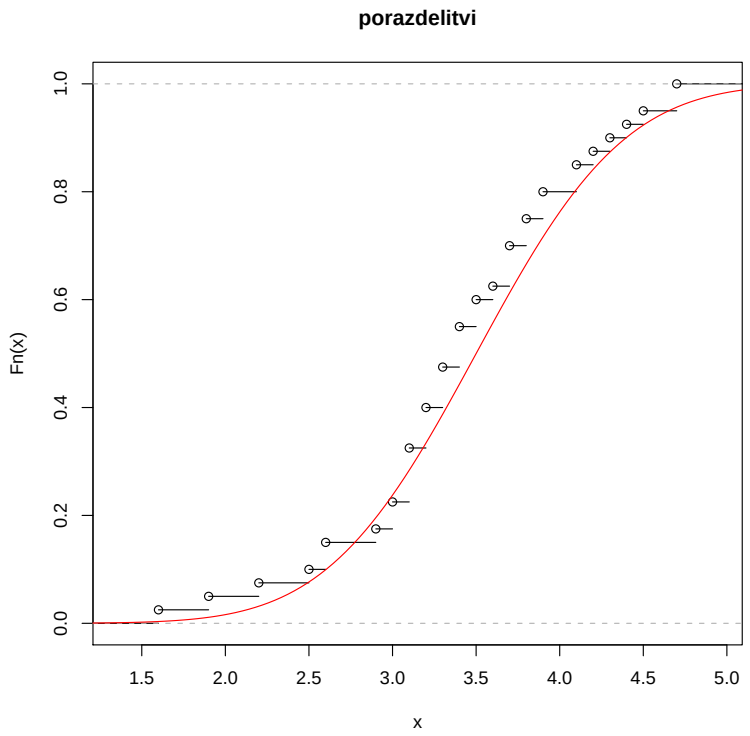
1. naročnika in izvajalca,
2. populacijo in vzorčni okvir,
3. opis vzorca,
4. velikost vzorca in velikost realiziranega vzorca (stopnja odgovorov)
5. čas, kraj in način anketiranja,
6. anketno vprašanje,
7. vzorčno napako.

Preizkus Kolmogorov-Smirnova v R-ju

```
> t <- c(1.6,1.9,2.2,2.5,2.6,2.6,2.9,3.0,3.0,3.1,3.1,3.1,
+ 3.1,3.2,3.2,3.2,3.3,3.3,3.3,3.4,3.4,3.4,3.5,3.5,3.6,3.7,
+ 3.7,3.7,3.8,3.8,3.9,3.9,4.1,4.1,4.2,4.3,4.4,4.5,4.7,4.7)
> z <- sample(t)
> z
 [1] 4.1 2.2 4.7 3.8 4.7 3.5 3.3 3.6 4.4 2.9 4.3 4.2 3.9 3.1
[15] 3.1 2.5 4.5 3.0 2.6 3.4 1.9 3.5 3.2 3.1 3.7 2.6 3.1 3.2
[29] 3.4 3.7 3.4 3.3 4.1 1.6 3.9 3.3 3.0 3.7 3.2 3.8
> "lot(density(z),main="gostoti")
> "urve(dnorm(x,mean=3.5,sd=0.7),add=TRUE,col="red")
> "lot(ecdf(z),main="porazdelitvi")
> curve(pnorm(x,mean=3.5,sd=0.7),add=TRUE,col="red")
> ks.test(z,"pnorm",mean=3.5,sd=0.7)

        One-sample Kolmogorov-Smirnov test

data:  z
D = 0.1068, p-value = 0.7516
alternative hypothesis: two.sided

Warning message:
cannot compute correct p-values with ties in:
  ks.test(z, "pnorm", mean = 3.5, sd = 0.7)
```

Preizkus Kolmogorov-Smirnova v R-ju





V R-ju so pri preizkusih izpisane vrednosti $\texttt{p-value} = \Pi$
(preizkusna statistika ima pri veljavnosti osnovne domneve vrednost vsaj tako ekstremno, kot je zračunana).

$[0, 0.001]$ – izjemno značilno (***);
$(0.001, 0.01]$ – zelo značilno (**);
$(0.01, 0.05]$ – statistično značilno (*);
$(0.05, 0.1]$ – morda značilno;
$(0.1, 1]$ – neznačilno.

Osnovno domnevo zavrnemo, če je `p-value` pod izbrano stopnjo značilnosti.

Preizkus Kolmogorov-Smirnova v R-ju

```
> p <- pnorm(t,mean=3.5,sd=0.7)
> s <- (1:40)/40; d <- abs(s-p)
> options(digits=3)
> cbind(t,p,s,d)
        t       p     s       d          t       p     s       d
 [1,] 1.6 0.00332 0.025 0.02168  [21,] 3.4 0.44320 0.525 0.08180
 [2,] 1.9 0.01114 0.050 0.03886  [22,] 3.4 0.44320 0.550 0.10680
 [3,] 2.2 0.03165 0.075 0.04335  [23,] 3.5 0.50000 0.575 0.07500
 [4,] 2.5 0.07656 0.100 0.02344  [24,] 3.5 0.50000 0.600 0.10000
 [5,] 2.6 0.09927 0.125 0.02573  [25,] 3.6 0.55680 0.625 0.06820
 [6,] 2.6 0.09927 0.150 0.05073  [26,] 3.7 0.61245 0.650 0.03755
 [7,] 2.9 0.19568 0.175 0.02068  [27,] 3.7 0.61245 0.675 0.06255
 [8,] 3.0 0.23753 0.200 0.03753  [28,] 3.7 0.61245 0.700 0.08755
 [9,] 3.0 0.23753 0.225 0.01253  [29,] 3.8 0.66588 0.725 0.05912
[10,] 3.1 0.28385 0.250 0.03385  [30,] 3.8 0.66588 0.750 0.08412
[11,] 3.1 0.28385 0.275 0.00885  [31,] 3.9 0.71615 0.775 0.05885
[12,] 3.1 0.28385 0.300 0.01615  [32,] 3.9 0.71615 0.800 0.08385
[13,] 3.1 0.28385 0.325 0.04115  [33,] 4.1 0.80432 0.825 0.02068
[14,] 3.2 0.33412 0.350 0.01588  [34,] 4.1 0.80432 0.850 0.04568
[15,] 3.2 0.33412 0.375 0.04088  [35,] 4.2 0.84134 0.875 0.03366
[16,] 3.2 0.33412 0.400 0.06588  [36,] 4.3 0.87345 0.900 0.02655
[17,] 3.3 0.38755 0.425 0.03745  [37,] 4.4 0.90073 0.925 0.02427
[18,] 3.3 0.38755 0.450 0.06245  [38,] 4.5 0.92344 0.950 0.02656
[19,] 3.3 0.38755 0.475 0.08745  [39,] 4.7 0.95676 0.975 0.01824
[20,] 3.4 0.44320 0.500 0.05680  [40,] 4.7 0.95676 1.000 0.04324
> options(digits=7)
> max(d)
[1] 0.1067985
```


Preizkus χ^2 v R-ju

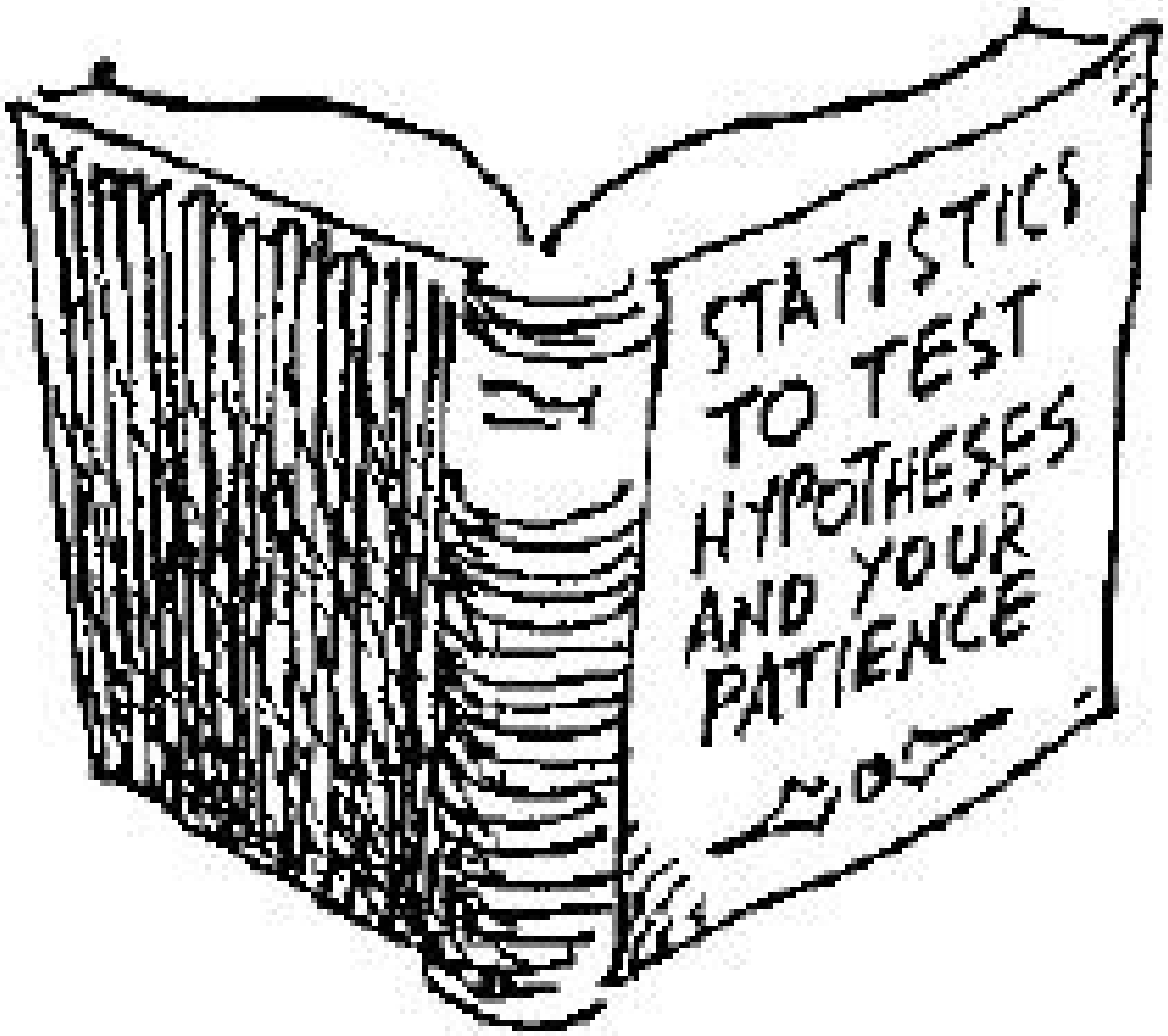
```
> a <- rbind(c(80,5,15),c(40,20,20),c(20,30,20))
> rownames(a) <- c("za","proti","neodlocen")
> colnames(a) <- c("do 25","25-50","nad 50")
> a
          do 25 25-50 nad 50
za           80     5     15
proti        40    20     20
neodlocen    20    30     20
> chisq.test(a)

        Pearson's Chi-squared test

data:  a
X-squared = 51.4378, df = 4, p-value = 1.808e-10
```

Poleg `ks.test` in `chisq.test` obstaja v R-ju še več drugih preizkusov: `prop.test`, `binom.test`, `t.test`, `wilcox.test`, `var.test`, `shapiro.test`, `cor.test`, `fisher.test`, `kruskal.test`.

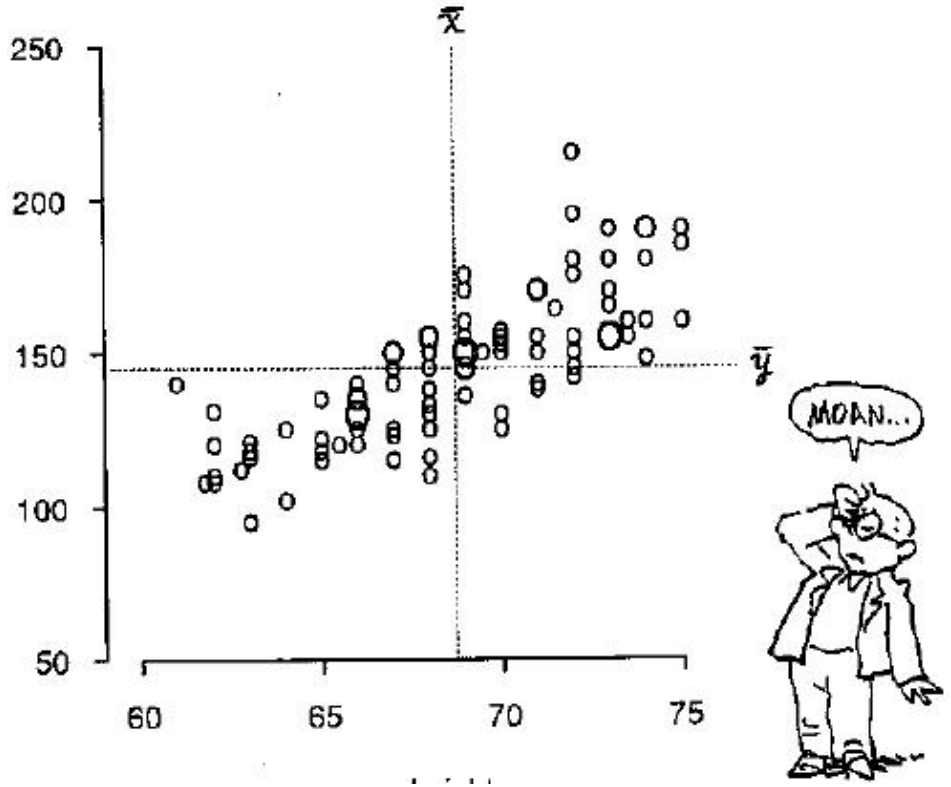
Opise posameznega preizkusa dobimo z zahtevo `help(`*preizkus*`)` .

Spearmanov preizkus povezanosti v R-ju

```
> slo <- c(5,3,1,2,4)
> mat <- c(5,2,3,1,4)
> slo
[1] 5 3 1 2 4
> mat
[1] 5 2 3 1 4
> cor.test(slo,mat,method="spearm")

        Spearman's rank correlation rho

data:  slo and mat
S = 6, p-value = 0.2333
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7
```



STATISTICS
TO TEST
HYPOTHESES
AND YOUR
PATIENCE

II.6. Bivariatna analiza in regresija



Bivariatna analiza

$X \longleftrightarrow Y$ povezanost

$X \longrightarrow Y$ odvisnost

Mere povezanosti ločimo glede na tip spremenljivk:

1. NOMINALNI tip para spremenljivk (ena od spremenljivk je nominalna): χ^2, kontingenčni koeficienti, koeficienti asociacije;
2. ORDINALNI tip para spremenljivk (ena spremenljivka je ordinalna druga ordinalna ali boljša) koeficient korelacije rangov;
3. ŠTEVILSKI tip para spremenljivk (obe spremenljivki sta številski): koeficient korelacije.

Preverjanje domneve o povezanosti dveh nominalnih spremenljivk

Vzemimo primer:

- ENOTA: dodiplomski študent neke fakultete v letu 1993/94;
- VZOREC: slučajni vzorec 200 študentov;
- 1. SPREMENLJIVKA: spol;
- 2. SPREMENLJIVKA: stanovanje v času študija.

Zanima nas ali študentke drugače stanujejo kot študentje oziroma: ali sta spol in stanovanje v času študija povezana. V ta namen podatke študentov po obeh spremenljivkah uredimo v dvodimenzionalno frekvenčno porazdelitev. To tabelo imenujemo kontingenčna tabela.

Denimo, da so podatki za vzorec urejeni v naslednji kontingenčni tabeli:

	starši	št. dom	zasebno	skupaj
moški	16	40	24	80
ženske	48	36	36	120
skupaj	64	76	60	200

Ker nas zanima ali študentke drugače stanujejo v času študija kot študentje, moramo porazdelitev stanovanja študentk primerjati s porazdelitvijo študentov.

Ker je število študentk različno od števila študentov, moramo zaradi primerjave izračunati relativne frekvence:

	starši	št. dom	zasebno	skupaj
moški	20	50	30	100
ženske	40	30	30	100
skupaj	32	38	30	100

Če med spoloma ne bi bilo razlik, bi bili obe porazedelitvi (za moške in ženske) enaki porazdelitvi pod "skupaj". Naš primer kaže, da se odstotki razlikujejo: npr. le 20% študentov in kar 40% študentk živi med študijem pri starših. Odstotki v študentskih domovih pa so ravno obratni.
Zasebno pa stanuje enak odstotek deklet in fantov. Že pregled relativnih frekvenc (po vrsticah) kaže, da sta spremenljivki povezani med seboj.

Relativne frekvence lahko računamo tudi po stolpcih:

	starši	št. dom	zasebno	skupaj
moški	25	56, 6	40	40
ženske	75	43, 4	60	60
skupaj	100	100	100	100

Relativno frekvenco lahko prikažemo s stolpci ali krogi.

Kontingenčna tabela kaže podatke za slučajni vzorec.

Zato nas zanima, ali so razlike v porazdelitvi tipa stanovanja v času študija po spolu statistčno značilne in ne le učinek vzorca.

H_0: spremenljivki nista povezani
H_1: spremenljivki sta povezani

Za preverjanje domneve o povezanosti med dvema nominalnima spremenljivkama na osnovi vzorčnih podatkov, podanih v dvo-razsežni frekvenčni porazdelitvi, lahko uporabimo χ^2 test.

Ta test sloni na primerjavi empiričnih (dejanskih) frekvenc s teoretičnimi frekvencami, ki so v tem primeru frekvence, ki bi bile v kontingenčni tabeli, če spremenljivki ne bi bili povezani med seboj.

To pomeni, da bi bili porazdelitvi stanovanja v času študija deklet in fantov enaki.

Če spremenljivki nista povezani med seboj, so verjetnosti hkratne zgoditve posameznih vrednosti prve in druge spremenljivke enake produktu verjetnosti posameznih vrednosti. Npr., če označimo moške z M in stanovanje pri starših s S, je:

$$P(M) = \frac{80}{200} = 0,40;$$

$$P(S) = \frac{64}{200} = 0,32;$$

$$P(M \cap S) = P(M) \cdot P(S) = \frac{80}{200} \cdot \frac{64}{200} = 0,128.$$

Teoretična frekvenca je verjetnost $P(M \cap S)$ pomnožena s številom enot v vzorcu:

$$f'(M \cap S) = n \cdot P(M \cap S) = 200 \cdot \frac{80}{200} \cdot \frac{64}{200} = 25,6.$$

Podobno izračunamo teoretične frekvence tudi za druge celice kontingenčne tabele.

Če teoretične frekvence zaokrožimo na cela števila, je tabela izračunanih teoretičnih frekvenc f_i' naslednja:

	starši	št. dom	zasebno	skupaj
moški	26	30	24	80
ženske	38	46	36	120
skupaj	64	76	60	200

Spomnimo se tabel empiričnih (dejanskih) frekvenc f_i:

χ^2 statistika, ki primerja dejanske in teoretične frekvence je

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'},$$

kjer je k število celic v kontingenčni tabeli. Statistika χ^2 se porazdeljuje po χ^2 porazdelitvi s $(s-1)(v-1)$ prostostnimi stopnjami, kjer je s število vrstic v kontingenčni tabeli in s število stolpcev.

Ničelna in osnovna domneva sta v primeru tega testa

H_0: $\chi^2 = 0$ (spremenljivki nista povezani)
H_1: $\chi^2 > 0$ (spremenljivki sta povezani)

Iz tabele za porazdelitev χ^2 lahko razberemo kritične vrednost te statistike pri 5% stopnji značilnosti:

$$\chi^2_{1-\alpha}[(s-1)(v-1)] = \chi^2_{0,95}(2) = 5,99.$$

Eksperimentalna vrednost statistike χ^2 pa je:

$$\chi^2_e = \frac{(16-26)^2}{26} + \frac{(40-30)^2}{30} + \frac{(24-24)^2}{24}$$

$$+\frac{(48-38)^2}{38} + \frac{(36-46)^2}{46} + \frac{(36-36)^2}{36} = 12.$$

Ker je ekperimentalna vrednost večja od kritične vrednosti, pomeni, da pade v kritično območje.

To pomeni, da ničelno domnevo zavrnemo.

Pri 5% stopnji značilnosti lahko sprejmemo osnovno domnevo, da sta spremenljivki statistično značilno povezani med seboj.

Statistika χ^2 je lahko le pozitivna. Zavzame lahko vrednosti v intervalu $[0, \chi^2_{\max}]$, kjer je $\chi^2_{\max} = n(k-1)$, če je $k = \min(v, s)$.

χ^2 statistika v splošnem ni primerljiva. Zato je definiranih več **kontingenčnih koeficientov**, ki so bolj ali manj primerni. Omenimo naslednje:

1. **Pearsonov koeficient**:

$$\Phi = \frac{\chi^2}{n},$$

ki ima zgornjo mejo $\Phi^2_{\max} = k - 1$.

2. **Cramerjev koeficient**:

$$\alpha = \sqrt{\frac{\Phi^2}{k-1}} = \sqrt{\frac{\chi^2}{n(k-1)}},$$

ki je definiran na intervalu $[0, 1]$.

3. **Kontingenčni koeficient**:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

ki je definiran na intervalu $[0, C_{\max}]$, kjer je $C_{\max} = \sqrt{k/(k-1)}$.

Koeficienti asociacije

Denimo, da imamo dve nominalni spremenljivki, ki imata le po dve vrednosti (sta dihotomni). Povezanost med njima lahko računamo poleg kontingenčnih koeficientov s **koeficienti asociacije** na osnovi frekvenc iz kontingenčne tabele 2×2:

$Y \backslash X$	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
	$a+c$	$b+d$	N

kjer je $N = a + b + c + d$. Na osnovi štirih frekvenc v tabeli je definirnih več koeficientov asociacije:

- **Yulov koeficient asociacije**:

$$Q = \frac{ad - bc}{ad + bc} \in [-1, 1].$$

- **Sokal Michenerjev koeficient**:

$$S = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \in [0,1].$$

- **Pearsonov koeficient**:

$$\phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1,1].$$

Velja

$$\chi^2 = N \cdot \phi^2.$$

- **Jaccardov koeficient**:

$$J = \frac{a}{a+b+c} \in [0,1],$$

- in še več drugih.

Primer:

Vzemimo primer, ki kaže povezanost mod kaznivimi dejanji in alkoholizmom. Tabela kaže podatke za $N = 10.750$ ljudi

alk. \ kaz. d.	DA	NE	skupaj
DA	50	500	550
NE	200	10.000	10.200
skupaj	250	10.500	10.750

Izračunajmo koeficiente asociacije:

$$Q = \frac{50 \times 10000 - 200 \times 500}{50 \times 10000 + 200 \times 500} = 0,67;$$

$$S = \frac{10050}{10750} = 0,93;$$

$$J = \frac{50}{50 + 500 + 200} = 0,066.$$

Izračunani koeficienti so precej različni. Yulov in Sokal Michenerjev koeficient kažeta na zelo močno povezanost med kaznjivimi dejanji in alkoholizmom, medtem kot Jaccardov koeficient kaže, da med spremenljivkama ni povezanosti. Pri prvih dveh koeficientih povezanost povzroča dejstvo, da večina alkoholiziranih oseb ni naredila kaznivih dejanj in niso alkoholiki (frekvenca d). Ker Jaccardov koeficient upošteva le DA DA ujemanje, je lažji za interpretacijo. V našem primeru pomeni, da oseba, ki je naredila kaznivo dejanje, sploh ni nujno alkoholik.

Preverjanje domneve o povezanosti dveh ordinalnih spremenljivk

V tem primeru gre za študij povezanosti med dvema spremenljivkama, ki sta vsaj ordinalnega značaja.

Primer:

Vzemimo slučajni vzorec šestih poklicev in ocenimo, koliko so odgovorni (O) in koliko fizično naporni (N).

V tem primeru smo poklice uredili od najmanj odgovornega do najbolj odgovornega in podobno od najmanj fizično napornega do najbolj napornega.

Poklicem smo torej priredili range po odgovornosti (R_0) in po napornosti (R_N) od 1 do 6.

Podatki so podani v tabeli:

poklic	R_0	R_N
A	1	6
D	2	4
C	3	5
D	4	2
E	5	3
F	6	1

Povezanost med spremenljivkama lahko merimo s koeficientom korelacije rangov r_s (Sperman), ki je definiran takole:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}.$$

kjer je d_i razlika med **rangoma** v i-ti enoti.

Koeficient korelacije rangov lahko zavzame vrednosti v intervalu $[-1, 1]$. Če se z večanjem rangov po prvi spremenjivki večajo rangi tudi po drugi spremenljivki, gre za pozitivno povezanost. Tedaj je koeficient pozitiven in blizu 1. Če pa se z večanjem rangov po prvi spremenljivki rangi po drugi spremenljivki manjšajo, gre za negativno povezanost. Koeficient je tedaj negativen in blizu -1. V našem preprostem primeru gre negativno povezanost. Če ne gre za pozitivno in ne za negativno povezanost, rečemo, da spremenljivki nista povezani.

Izračunajmo koeficient korelacije rangov za primer šestih poklicev:

poklic	R_0	R_N	d_i	d_i^2
A	1	6	-5	25
B	2	4	-2	4
C	3	5	-2	4
D	4	2	2	4
E	5	3	2	4
F	6	1	5	25
vsota			0	66

$$r_s = 1 - \frac{6 \cdot 66}{6 \cdot 35} = 1 - 1,88 = -0,88.$$

Res je koeficient blizu, kar kaže na močno negativno povezanost teh 6-ih poklicev.

Omenili smo, da obravnavamo 6 slučajno izbranih poklicev.

Zanima nas, ali lahko na osnovi tega vzorca posplošimo na vse poklice, da sta odgovornost in fizična napornost poklicev (negativno) povezana med seboj.

Upoštevajmo 5% stopnjo značilnosti.

Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho_s = 0$ (spremenljivki nista povezani)
H_1: $\rho_s \neq 0$ (spremenljivki sta povezani)

kjer populacijski koeficient označimo s ρ_s.

Pokaže se, da se statistika

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

porazdeljuje približno po porazdelitvi z $m = (n-2)$ prostostnimi stopnjami. Ker gre za dvostranski test, sta kritični vrednosti enaki

$$\pm t_{\alpha/2} = \pm t_{0,025}(4) = \pm 2,776.$$

Eksperimentalna vrednost statistike je za naš primer

$$t_e = \frac{-0,88 \times 2}{\sqrt{1-(-0,88)^2}} = \frac{-1,76}{0,475} = -3,71.$$

Eksperimentalna vrednost pade v kritično območje. Pri 5% stopnji značilnosti lahko rečemo, da sta odgovornost in fizična napornost (negativno) povezani med seboj.

Če je ena od obeh spremenljivk številska, moramo vrednosti pred izračunom d_i rangirati. Če so kakšne vrednosti enake, zanje izračunamo povprečne pripadajoče range.

Preverjanje domneve o povezanosti dveh številskih spremenljivk

Vzemimo primer dveh številskih spremenljivk:

X - izobrazba (število priznanih let šole)
Y - število ur branja dnevnih časopisov na teden

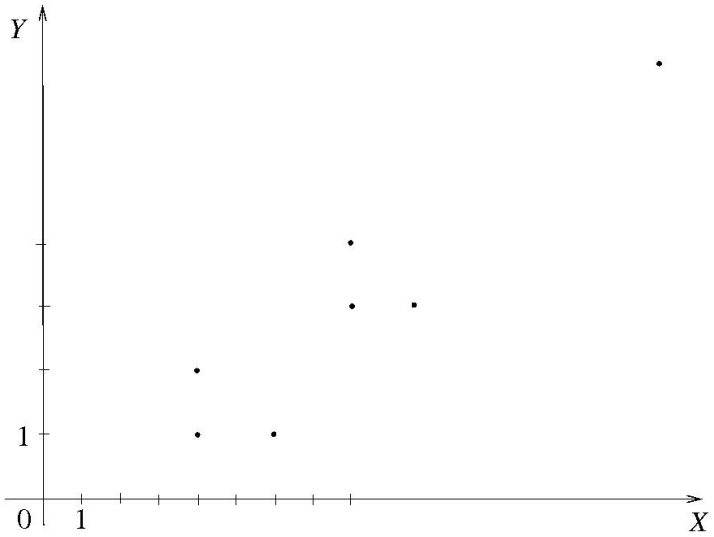
Podatki za 8 slučajno izbranih oseb so:

X	10	8	16	8	6	4	8	4
Y	3	4	7	3	1	2	3	1

Grafično lahko ponazorimo povezanost med dvema številskima spremenljivkama z razsevnim grafikonom.

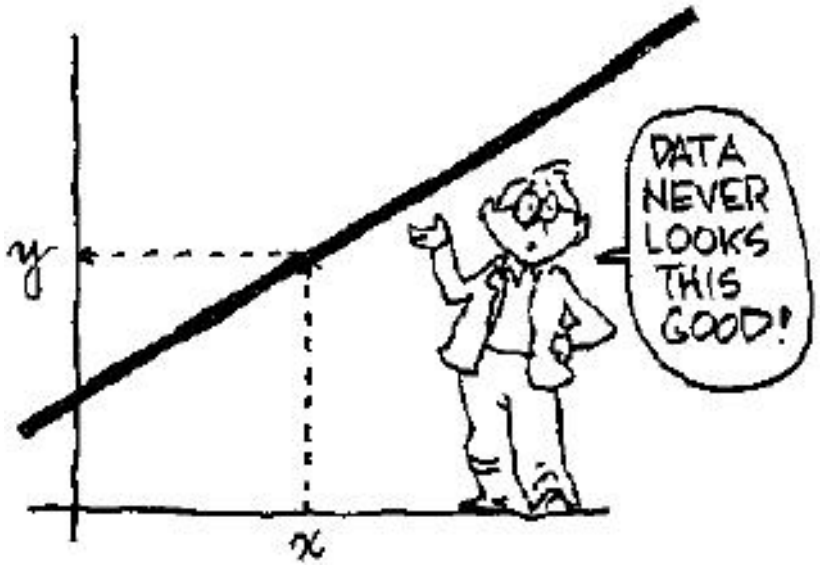
To je, da v koordinatni sistem, kjer sta koordinati obe spremenljivki, vrišemo enote s pari vrednosti.

V našem primeru je izgleda razsevni grafikon takole:

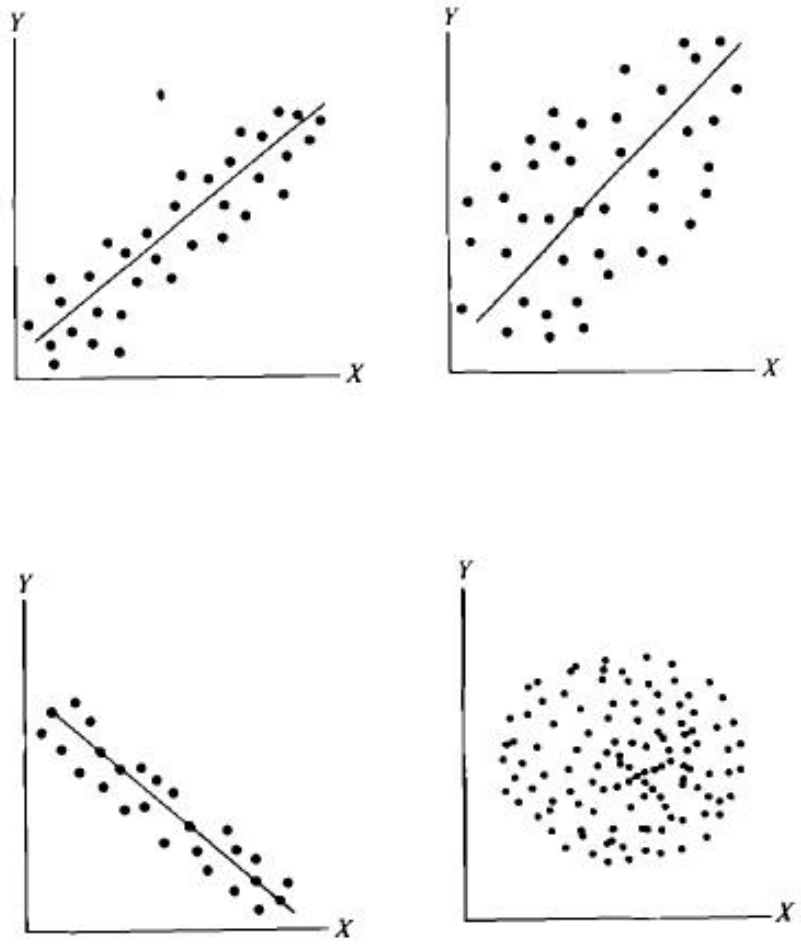


Tipi povezanosti:

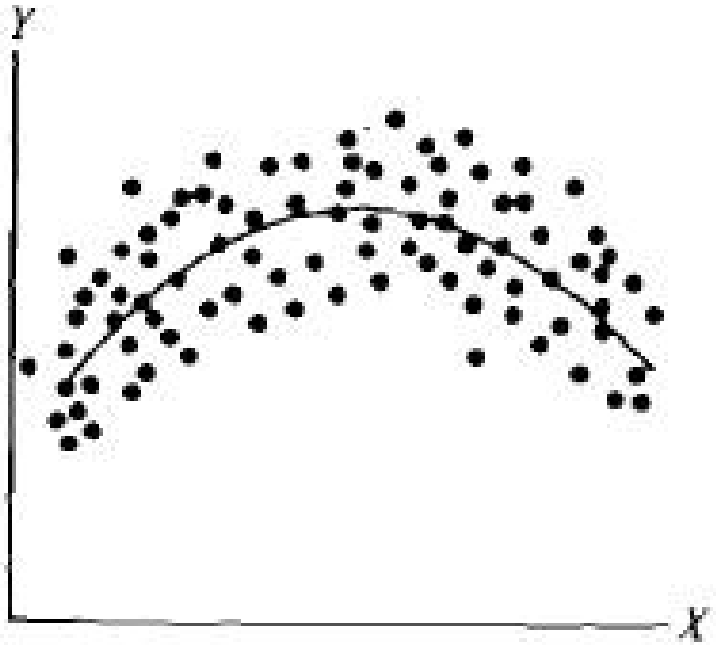
- **funkcijska** povezanost: vse točke ležijo na krivulji:
- **korelacijska** (stohastična] povezanost: točke so od krivulje bolj ali manj odklanjajo (manjša ali večja povezanost).



Tipični primeri linearne povezanosti spremenljivk:



Primer nelinearne povezanosti spremenljivk:



Kovarianca

$$\mathsf{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)\cdot(y_i - \mu_Y)$$

meri linearno povezanost med spremenljivkama.
$\mathsf{Cov}(X,Y) > 0$ pomeni pozitivno linearno povezanost,
$\mathsf{Cov}(X,Y) = 0$ pomeni da ni linearne povezanosti,
$\mathsf{Cov}(X,Y) < 0$ pomeni negativno linearno povezanost.
(Pearsonov) koeficient korelacije je

$$\rho_{XY} = \frac{\mathsf{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)\cdot(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_Y)^2}}.$$

Koeficient korelacije lahko zavzame vrednosti v intervalu $[-1, 1]$.

Če se z večanjem vrednosti prve spremenljivke
večajo tudi vrednosti druge spremenljivke,
gre za *pozitivno* povezanost.
Tedaj je koeficient povezanosti blizu 1.

Če pa se z večanjem vrednosti prve spremenljivke
vrednosti druge spremenljivke manjšajo,
gre za *negativno* povezanost.
Koeficient je tedaj negativen in blizu -1.

Če ne gre za pozitivno in ne za negativno povezanost,
rečemo da spremenljivki nista povezani in koeficient je blizu 0.

Statistično sklepanje o korelacijski povezanosti:

Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho = 0$ (spremenljivki nista linearno povezani)
H_1: $\rho \neq 0$ (spremenljivki sta linearno povezani)

Pokaže se, da se statistika

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

porazdeljuje po t porazdelitvi z $m = (n-2)$ prostostnimi stopnjami.

Z r označujemo koeficient korclacije na vzorcu in
z ρ koeficient korelacije na populaciji.

Primer: Preverimo domnevo, da sta izobrazba (število priznanih let šole) in število ur branja dnevnih časopisov na teden povezana med seboj pri 5% stopnji značilnosti. Najprej izračunajmo vzorčni koeficient korelacije:

x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x) \cdot (y_i - \mu_y)$
10	3	2	0	4	0	0
8	4	0	1	0	1	0
16	7	8	4	64	16	32
8	3	0	0	0	0	0
6	1	-2	-2	4	4	4
4	2	-4	-1	16	1	4
8	3	0	0	0	0	0
4	1	-4	-2	16	4	8
64	24	0	0	104	26	48

$$r = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2 \sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

$$= \frac{48}{\sqrt{104 \cdot 26}} = 0,92.$$

Ker gre za dvostranski test, je kritično območje določeno s kritičnima vrednostima

$$\pm t_{\alpha/2}(n-2) = \pm t_{0,025}(6) = \pm 2,447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \frac{0,92\sqrt{8-2}}{\sqrt{1-0,92^2}} = 2,66.$$

Eksperimentalna vrednost pade v kritično območje.

Zaključek: ob 5% stopnji značilnosti lahko rečemo, da je izobrazba linearno povezana z branjem dnevnih časopisov.

Parcialna korelacija

Včasih je potrebno meriti zvezo med dvema spremenljivkama in odstraniti vpliv vseh ostalih spremenljivk.

To zvezo dobimo s pomočjo koeficienta parcialne korclacije. Pri tem seveda predpostavljamo, da so vse spremenljivke med seboj linearno povezane. Če hočemo iz zveze med spremenljivkama X in Y odstraniti vpliv tretje spremenljivke Z, je **koeficient parcialne korelacije**:

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{1 - r_{XZ}^2}\,\sqrt{1 - r_{YZ}^2}}.$$

Tudi ta koeficient, ki zavzema vrednosti v intervalu $[-1, 1]$, interpretiramo podobno kot običajni koeficient korelacije.

S pomočjo tega obrazca lahko razmišljamo naprej, kako bi izločili vpliv naslednjih spremenljivk.

Primer:

V neki ameriški raziskavi, v kateri so proučevali vzroke za kriminal v mestih, so upoštevali naslednje spremenljivke:

X : % nebelih prebivalcev,

Y : % kaznivih dejanj,

Z : % revnih prebivalcev,

U : velikost mesta.

Izračunali so naslednje koeficiente korelacije:

	X	Z	U	Y
X	1	$0,51$	$0,41$	$0,36$
Z		1	$0,29$	$0,60$
U			1	$0,49$
Y				1

Zveza med nebelim prebivalstvom in kriminalom je

$$r_{XY} = 0,36.$$

Zveza je kar močna in lahko bi mislili,
da nebeli prebivalci povzročajo več kaznivih dejanj.

Vidimo pa še, da je zveza med revščino in kriminalom tudi precejšna

$$r_{YZ} = 0,60.$$

Lahko bi predpostavili, da revščina vpliva na zvezo med nebelci in kriminalom, saj je tudi zveza med revnimi in nebelimi precejšna $r_{XZ} = 0,51$.

Zato poskusim odstraniti vpliv revščine iz zveze:

“nebelo prebivalstvo : kazniva dejanja”:

$$r_{XY,Z} = \frac{0,36 - 0,51 \cdot 0,60}{\sqrt{1 - 0,51^2}\sqrt{1 - 0,60^2}}.$$

Vidimo, da se je linearna zveza zelo zmanjšala.

Če pa odstranimo še vpliv velikosti mesta,
dobimo parcialno korelacijo $-0,02$
oziroma zveze praktično ni več.

Regresijska analiza

Regresijska funkcija $Y' = f(X)$ kaže, kakšen bi bil vpliv spremenljivke X na Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y. Ker pa so ponavadi še drugi vplivi na proučevano spremenljivko Y, se točke, ki predstavljajo enote v razsevnem grafikonu, odklanjajo od idealne regresijske krivulje

$$Y = Y' + E = f(X) + E$$

kjer X imenujemo neodvisna spremenljivka, Y odvisna spremenljivka in E člen napake (ali motnja, disturbanca).

Če je regresijska fukcija linearna:

$$Y' = f(X) = a + bX$$

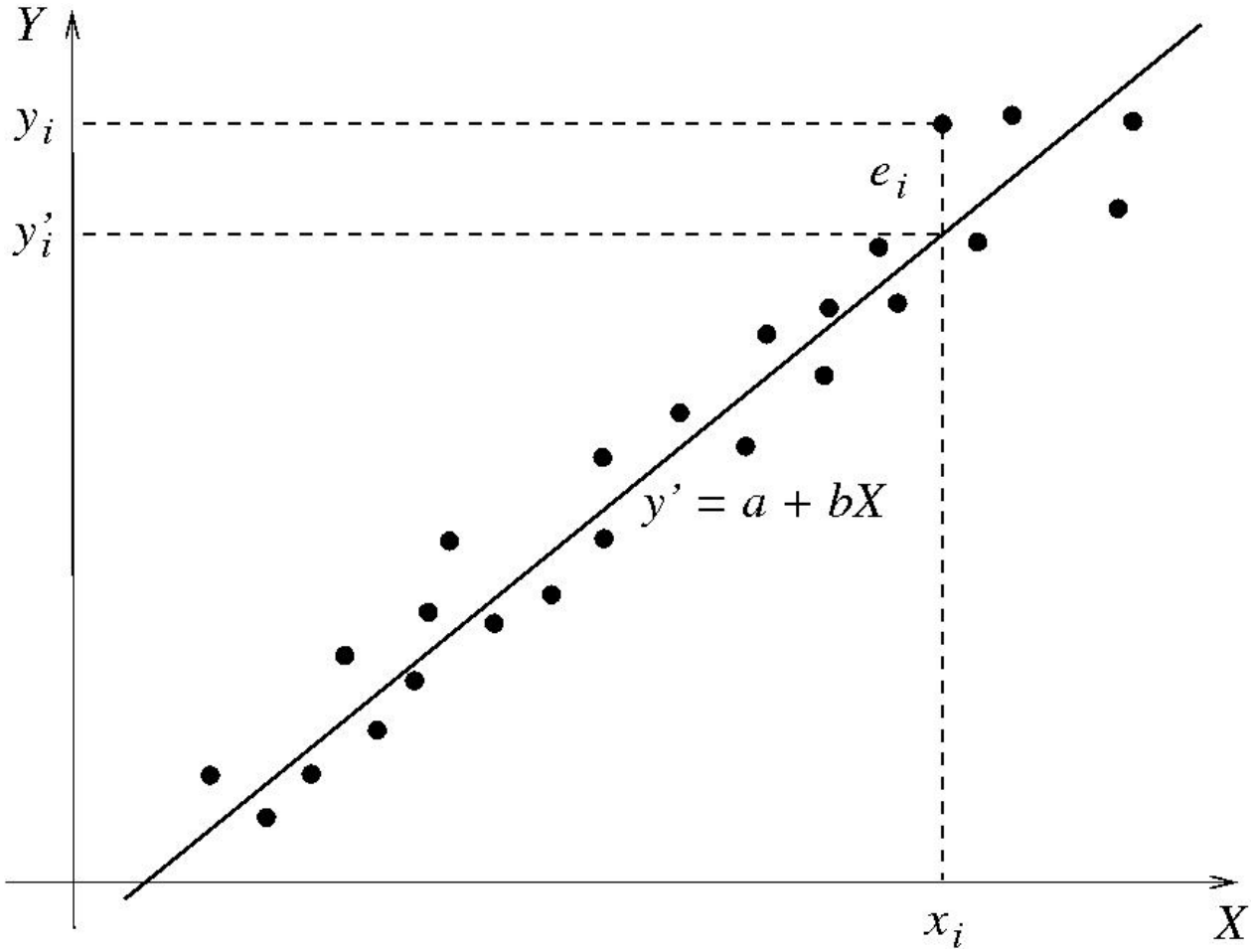
je regresijska odvisnost

$$Y = Y' + E = a + bX + E$$

oziroma za i to enoto

$$y_i = y_i' + e_i = a + bx_i + e_i$$

Regresijsko odvisnost si lahko zelo nazorno predstavimo v razsevnem grafikonu:



Regresijsko funkcijo lahko v splošnem zapišemo

$$Y' = f(X, a, b, \cdots),$$

kjer so a, b, $\ldots$ parametri funkcije.

Ponavadi se moramo na osnovi pregleda razsevnega grafikona odločiti za tip regresijske funkcije in nato oceniti parametre funkcije, tako da se regresijska krivulja kar se da dobro prilega točkam v razsevnem grafikonu.

... Regresijska analiza

Pri dvorazsežno normalno porazdeljenem slučajnem vektorju $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je, kot vemo

$$\mathsf{E}(Y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

Pogojna porazdelitev Y glede na X je tudi normalna:

$$N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y\sqrt{1 - \rho^2}).$$

Regresija je linearna in regresijska krivulja premica,
ki gre skozi točko (σ_x, σ_y).

Med Y in X ni linearne zveze, sta le 'v povprečju' linearno odvisni.

Če označimo z $\boldsymbol{\beta} = \rho\frac{\sigma_y}{\sigma_x}$ *regresijski koeficient*, $\boldsymbol{\alpha} = \mu_y - \beta\mu_x$ in $\boldsymbol{\sigma^2} = \sigma_y\sqrt{1 - \rho^2}$, lahko zapišemo zvezo v obliki

$$y = \alpha + \beta x.$$

Preizkušanje regresijskih koeficientov

Po metodi momentov dobimo cenilki za α in β:

$$B = R\frac{C_y}{C_x} = \frac{C_{xy}}{C_x^2} \qquad \text{in} \qquad A = \overline{Y} - B\overline{X},$$

kjer so $C_x^2 = \sum_{i=1}^n (X_i - \overline{X})^2$, $C_y^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2$ in
$C_{xy} = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$.

Kako sta cenilki B in A porazdeljeni?

$$B = \frac{C_{xy}}{C_x^2} = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}(Y_i - \overline{Y}) = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2} Y_i.$$

Ker proučujemo pogojno porazdelitev Y glede na X (torej so vrednost X poznane), obravnavamo spremenljivke $X_1, \ldots, X_n$ kot konstante.
Ker je B linearna funkcija spremenljivk $Y_1, \ldots, Y_n$, ki so normalno porazdeljene $Y_i : N(\alpha + \beta X_i, \sigma)$, je tudi B normalno porazdeljena.

. . . Preizkušanje regresijskih koeficientov

Določimo parametra te porazdelitve:

$$\mathsf{E}B = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{C_x^2} \mathsf{E}Y_i = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{C_x^2} (\alpha + \beta X_i) = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{C_x^2} \beta (X_i - \overline{X}) = \beta.$$

Pri tem upoštevamo, da je $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$ in da sta α ter $\overline{X}$ konstanti.

$$\mathsf{D}B = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{C_x^4} \mathsf{D}Y_i = \frac{\sigma^2}{C_x^2}.$$

Torej je $B : N\left(\beta, \frac{\sigma}{C_x}\right)$, oziroma $\frac{B - \beta}{\sigma} C_x : N(0, 1)$.

Podobno dobimo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\right).$$

. . . Preizkušanje regresijskih koeficientov

Težje se je dokopati do cenilke za parameter σ^2.
Označimo $Q^2 = \sum_{i=1}^{n}(Y_i - A - BX_i)^2$.

Po nekaj računanja se izkaže, da velja $\mathsf{E}\dfrac{Q^2}{\sigma^2} = n - 2$.

Torej je $S^2 = \dfrac{Q^2}{n-2} = \dfrac{\sigma^2}{n-2}\,\dfrac{Q^2}{\sigma^2}$ nepristranska cenilka za σ^2.

S^2 je neodvisna od A in B. Testni statistiki za A in B sta tedaj

$$T_A = \frac{A - \mathsf{E}A}{\sqrt{\mathsf{D}A}} = \frac{A-\alpha}{S}\sqrt{\frac{nC_x^2}{C_x^2 + n\overline{X}^2}} = \frac{A-\alpha}{S}C_x\sqrt{\frac{n}{\sum_{i=1}^{n}X_i^2}},$$

$$T_B = \frac{B - \mathsf{E}B}{\sqrt{\mathsf{D}B}} = \frac{B-\beta}{S}C_x,$$

ki sta obe porazdeljeni po Studentu $S(n-2)$. Statistika za σ^2 pa je spremenljivka $\dfrac{Q^2}{\sigma^2} = (n-2)\dfrac{S^2}{\sigma^2}$, ki je porazdeljena po $\chi^2(n-2)$.

. . . Preizkušanje regresijskih koeficientov

Pokazati je mogoče tudi, da velja

$$Q^2 = C_y^2 - B^2 C_x^2 = C_y^2(1 - R^2).$$

To nam omogoča S v statistikah zapisati z C_y in R.

Te statistike uporabimo tudi za določitev intervalov zaupanja za parametre α, β in σ^2.

Linearni model

Pri proučevanju pojavov pogosto teorija postavi določeno funkcijsko zvezo med obravnavanimi spremenljivkami. Oglejmo si primer *linernega modela*, ko je med spremenljivkama x in y linearna zveza

$$y = \alpha + \beta x$$

Za dejanske meritve se pogosto izkaže, da zaradi različnih vplivov, ki jih ne poznamo, razlika $u = y - \alpha - \beta x$ v splošnem ni enaka 0, čeprav je model točen. Zato je ustreznejši *verjetnostni linearni model*

$$Y = \alpha + \beta X + U,$$

kjer so X, Y in U slučajne spremenljivke in $\mathsf{E}U = 0$ – model je vsaj v povprečju linearen.

. . . Linearni model

Slučajni vzorec (meritve) $(X_1, Y_1), \ldots, (X_n, Y_n)$ je realizacija slučajnega vektorja. Vpeljimo spremenljivke

$$U_i = Y_i - \alpha - \beta X_i$$

in predpostavimo, da so spremenljivke U_i med seboj neodvisne in enako porazdeljene z matematičnim upanjem 0 in disperzijo σ^2. Torej je:

$$\mathsf{E}U_i = 0, \quad \mathsf{D}U_i = \sigma^2 \quad \text{in} \quad \mathsf{E}(U_i U_j) = 0, \quad \text{za } i \neq j.$$

Običajno privzamemo še, da lahko vrednosti X_i točno določamo – X_i ima vedno isto vrednost. Poleg tega naj bosta vsaj dve vrednosti X različni.

Težava je, da (koeficientov) premice $y = \alpha + \beta x$ ne poznamo.
Recimo, da je približek zanjo premica $y = a + bx$.

. . . Linearni model

Določimo jo po *načelu najmanjših kvadratov* z minimizacijo funkcije

$$f(a,b) = \sum_{i=1}^{n} (y_i - (bx_i + a))^2.$$

Naloga zadošča pogojem izreka. Iz pogoja $\nabla P = 0$ dobimo enačbi

$$\begin{aligned} \frac{\partial f}{\partial a} &= -\sum_{i=1}^{n} 2(y_i - (bx_i + a)) = 0, \\ \frac{\partial f}{\partial b} &= -\sum_{i=1}^{n} 2(y_i - (bx_i + a))x_i = 0, \end{aligned}$$

z rešitvijo

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}, \qquad a = \frac{1}{n}\Big(\sum y - b\sum x\Big).$$

. . . Linearni model

oziroma, če vpeljemo oznako $\overline{z} = \frac{1}{n}\sum z$:

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \qquad a = \overline{y} - b\overline{x}.$$

Poglejmo še Hessovo matriko

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial a^2} & \frac{\partial^2 f}{\partial a \partial b} \\ \frac{\partial^2 f}{\partial b \partial a} & \frac{\partial^2 f}{\partial b^2} \end{bmatrix} = 2 \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}.$$

Ker je $\Delta_1 = 2\sum x^2 > 0$ in

$$\Delta_2 = 4(n\sum x^2 - (\sum x)^2) = 2\sum\sum (x_i - x_j)^2 > 0,$$

je matrika H pozitivno definitna in zato funkcija P strogo konveksna.
Torej je *regresijska premica* enolično določena.

. . . Linearni model

Seveda sta parametra a in b odvisna od slučajnega vzorca – torej slučajni spremenljivki. Iz dobljenih zvez za a in b dobimo že znani cenilki za koeficients α in β

$$B = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \overline{Y} - B\overline{X}.$$

Iz prej omenjenih predpostavk lahko (brez poznavanja porazdelitve Y in U) pokažemo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2\Big(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\Big), \quad \mathsf{E}B = \beta \quad \text{in} \quad \mathsf{D}B = \frac{\sigma^2}{C_x^2},$$

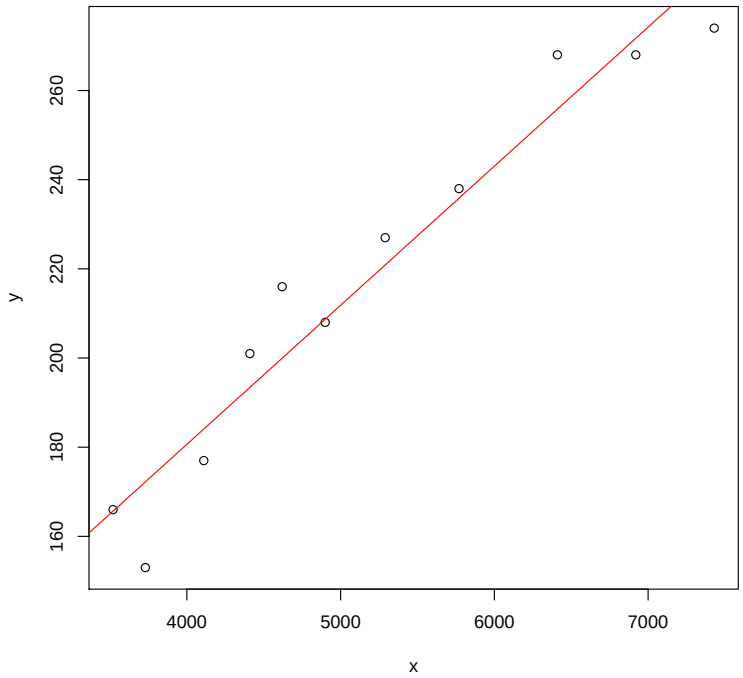
$$K(A,B) = -\sigma^2\frac{\overline{X}}{C_x^2}.$$

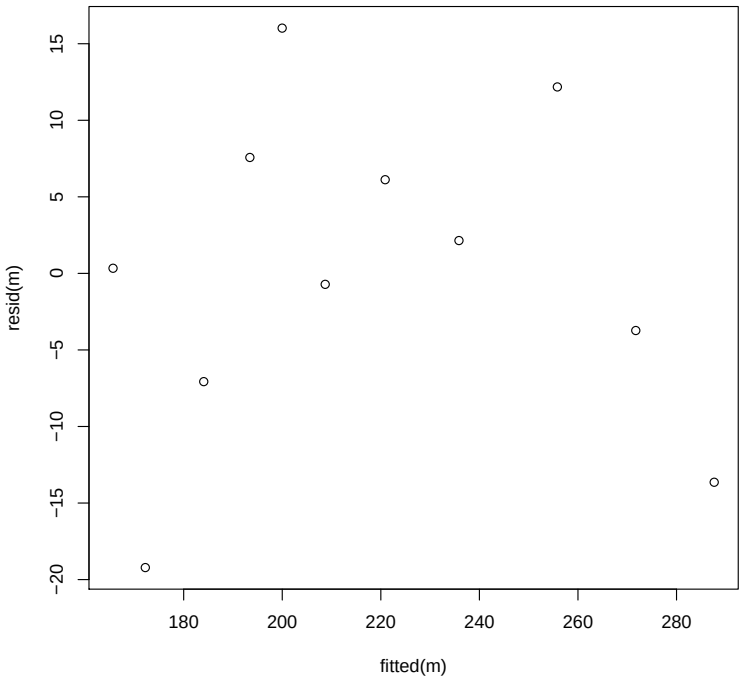
Cenilki za A in B sta najboljši linearni nepristranski cenilki za α in β.

. . . Linearni model

```
> x <- c(3520, 3730,4110,4410,4620,4900,5290,5770,6410,6920,7430)
> y <- c(166,153,177,201,216,208,227,238,268,268,274)
> l <- 1947:1957
> plot(y ~ x); abline(lm(y ~ x),col="red")
> m <- lm(y ~ x)
> summary(m)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2149  -5.4003   0.3364   6.8453  16.0204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.852675  14.491253   3.854  0.00388 **
x            0.031196   0.002715  11.492 1.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.18 on 9 degrees of freedom
Multiple R-Squared: 0.9362,     Adjusted R-squared: 0.9291
F-statistic: 132.1 on 1 and 9 DF,  p-value: 1.112e-06

> plot(fitted(m),resid(m))
```


. . . Linearni model






```
> coef(m)
(Intercept)           x
 55.8526752   0.0311963
```

To metodo ocenjevanja parametrov regresijske funkcije imenujemo **metoda najmanjših kvadratov**.

Linearni model

Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

$$Y = \mu_Y + \frac{\mathsf{Cov}(X,Y)}{\sigma_X^2}(X - \mu_X).$$

To funkcijo imenujemo tudi **prva** regresijska funkcija.

Podobno bi lahko ocenili linearno regresijsko funkcijo

$$X = a^* + b^* Y.$$

Če z metodo najmanjših kvadratov podobno ocenimo parametra a^* in b^*, dobimo:

$$X = \mu_X + \frac{\mathsf{Cov}(X,Y)}{\sigma_Y^2}(Y - \mu_Y).$$

To funkcijo imenujemo **druga** regresijska funkcija.

Primer: Vzemimo primer 8 osob, ki smo ga obravnavali v poglavju o povezanosti dveh številskih spremenljivk.

Spremenjivki sta bili:
X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.
Spomnimo se podatkov za teh 8 slučajno izbranh osob:

X	10	8	16	8	6	4	8	4
Y	3	4	7	3	1	2	3	1

Zanje izračunajmo obe regresijski premici in ju vrišimo v razsevni grafikon.

Ko smo računali koeficient korelacije smo že izračunali aritmetični sredini

$$\mu_X = \frac{64}{8} = 8, \qquad \mu_Y = \frac{24}{8} = 3,$$

vsoti kvadratov odklonov od aritmetične sredine za obe spremenljivki

$$\sum_{i=1}^{n}(x_i - \mu_X)^2 = 104, \qquad \sum_{i=1}^{n}(y_i - \mu_Y)^2 = 26$$

in vsoto produktov odklonov od obeh aritmetičnih sredin

$$\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) = 48.$$

Potem sta regresijski premici

$$Y = \mu_Y + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\sum_{i=1}^{N}(x_i-\mu_X)^2}(X-\mu_X),$$

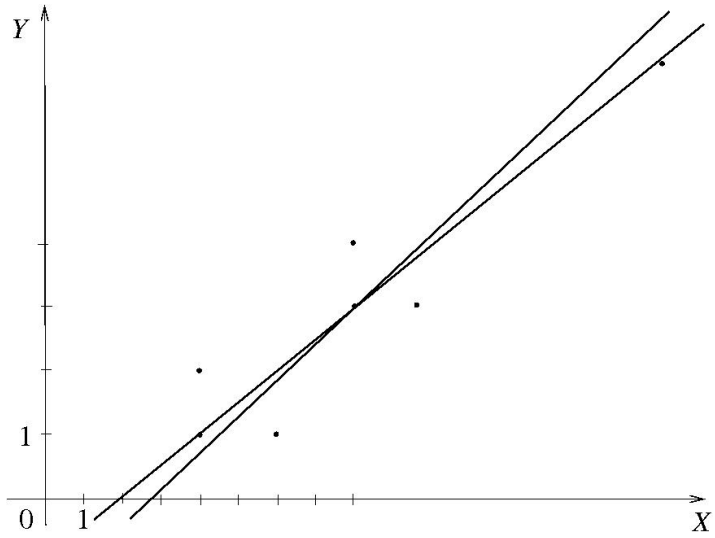
$$X = \mu_X + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\sum_{i=1}^{N}(y_i-\mu_Y)^2}(Y-\mu_Y),$$

oziroma

$$Y = 3 + \frac{48}{104}(X-8) = -0,68 + 0,46X,$$

$$X = 8 + \frac{48}{26}(Y-3) = -2,46 + 1,85Y.$$

Obe regresijski premici lahko vrišemo v razsevni grafikon in preverimo, če se res najbolje prilegata točkam v grafikonu:



Regresijski premici se sečeta v točki, določeni z aritmetičnima sredinama spremenljivk X in Y.

Dokažite, da se premici vedno sečeta v tej točki.

Statistično sklepanje o regresijskem koeficientu

Vpeljmo naslednje oznake:

$Y = \alpha + \beta X$ regresijska premica na populaciji,
$Y = a + bX$ regresijska premica na vzorcu.

Denimo, da želimo preveriti domnevo o regresijskem koeficientu β. Postavimo ničelno in osnovno domnevo takole:

H_0: $\beta = \beta_0$,
H_1: $\beta \neq \beta_0$.

Nepristranska cenilka za regresijski koeficient β je $b = \mathsf{Cov}(X, Y)/s_X^2$, ki ima matematično upanje in standardno napako:

$$\mathsf{E}b = \beta; \qquad \mathsf{SE}(b) = \frac{s_Y\sqrt{1-r^2}}{s_X\sqrt{n-2}}.$$

Testna statistika za zgornjo ničelno domnevo je:

$$t = \frac{s_Y\sqrt{n-2}}{s_X\sqrt{1-r^2}}\,(b-\beta_0),$$

ki se porazdeljuje po t-porazdelitvi z $m = (n-2)$ prostostnimi stopnjami.

Primer: Vzemimo primer, ki smo ga že obravnavali.

Spremenljivki sta

X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.

Podatke za slučajno izbrane enote ($n = 8$) najdemo na prejšnjih prosojnicah.

Preverimo domnevo, da je regresijski koeficient različen od 0 pri $\alpha = 5\%$.

Postavimo najprej ničelno in osnovno domnevo:

H_0: $\beta = 0$,
H_1: $\beta \neq 0$.

Gre za dvostranski test. Zato je ob 5% stopnji značilnosti kritično območje določeno s kritičnimia vrednostima:

$$\pm t_{\alpha/2}(n-2) = \pm t_{0,025}(6) = \pm 2,447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \sqrt{\frac{104 \cdot (8-2)}{26 \cdot (1-0,92^2)}} \cdot (0,46-0) = 5,8.$$

Regresijski koeficient je statistično značilno različen od 0.

Pojasnjena varianca (ang. ANOVA)

Vrednost odvisne spremenljivke Y_i lahko razstavimo na tri komponente:

$$y_i = \mu_Y + (y_i' - \mu_Y) + (y_i - y_i'),$$

kjer so pomeni posameznih komponent

μ_Y : rezultat splošnih vplivov,

$(y_i' - \mu_Y)$: rezultat vpliva spremenljivke X (regresija),

$(y_i - y_i')$: rezultat vpliva drugih dejavnikov (napake/motnje).

Če zgornjo enakost najprej na obeh straneh kvadriramo, nato seštejemo po vseh enotah in končno delimo s številom enot (N), dobimo:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_Y)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i' - \mu_Y)^2 + \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2.$$

To lahko zapišemo takole:

$$\sigma_Y^2 = \sigma_{Y'}^2 + \sigma_e^2,$$

kjer posamezni členi pomenijo:

σ_Y^2 : celotna varianca spremenljivke Y,

$\sigma_{Y'}^2$: pojasnjena varianca spremenljvke Y,

σ_e^2 : nepojasnjena varianca spremnenljivke Y.

Delež pojasnjene variance spremenljivke Y s spremenljivko X je

$$R = \frac{\sigma_{Y'}^2}{\sigma_Y^2}.$$

Imenujemo ga **determinacijski koeficient** in je definiran na intervalu $[0, 1]$.

Pokazati se da, da je v primeru linearne regresijske odvisnosti determinacijski koeficient enak

$$R = \rho^2,$$

kjer je ρ koeficient korelacije.

Kvadratni koren iz nepojasnjene variance σ_e imenujemo **standardna napaka regresijske ocene**, ki meri razpršenost točk okoli regresijske krivulje.

Standarna napaka ocene meri kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo.

V primeru linearne regresijske odvisnosti je standardna napaka enaka:

$$\sigma_e = \sigma_Y \sqrt{1 - \rho^2}.$$

Primer: Vzemimo spremenljivki

X - število ur gledanja televizije na teden

Y - število obiskov kino predstav na mesec

Podatki za 6 oseb so:

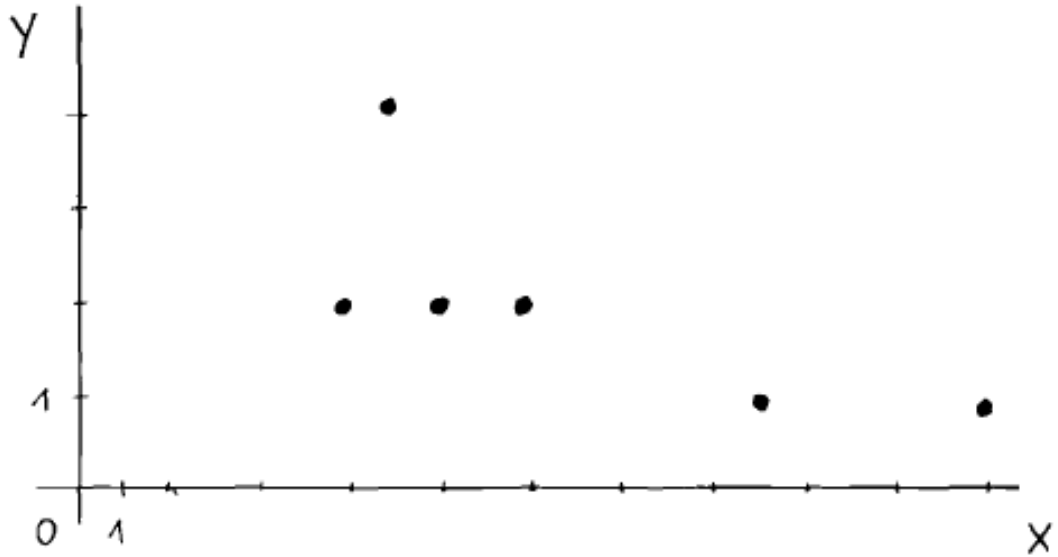
X	10	15	6	7	20	8
Y	2	1	2	4	1	2

Z linearno regresijsko funkcijo ocenimo, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden televizijo.

Kolikšna je standardna napaka?

Kolikšen delež variance obiska kinopredstav lahko pojasnimo z gledanjem televizije?

Najprej si podatke predstavimo v razsevnem grafikonu:



Za odgovore potrebujemo naslednje izračune:

x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$	$(x_i - \mu_x) \cdot (y_i - \mu_y)$
10	2	-1	0	1	0	0
15	1	4	-1	16	1	-4
6	2	-5	0	25	0	0
7	4	4	2	16	4	-8
20	1	9	-1	81	1	-9
8	2	-3	0	9	0	0
66	12	0	0	148	6	21

$$Y' = 2 - \frac{21}{148}\,(X - 11) = 3,54 - 0,14X$$

$$y'(18) = 3,54 - 0,14 \cdot 18 = 1,02$$

$$\rho = \frac{21}{146 \cdot 6} = -0,70$$

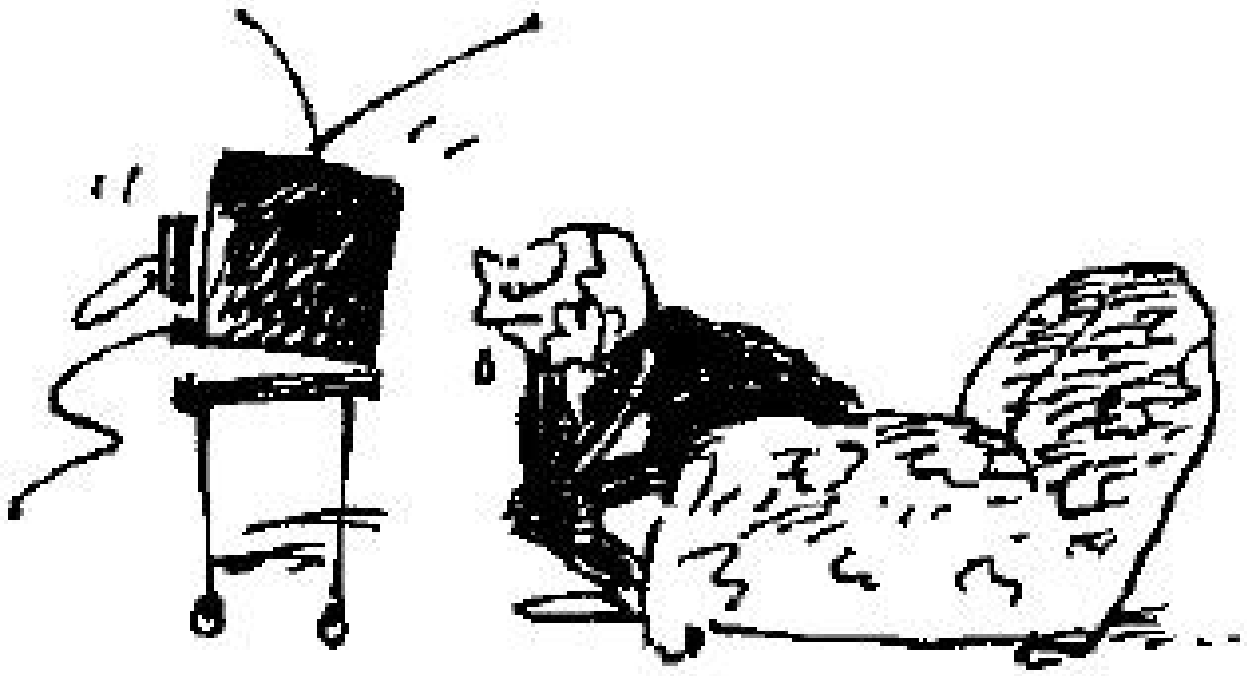
$$\sigma_e^2 = \frac{6}{6}\sqrt{1 - (-0,70)^2} = \sqrt{0,51} = 0,71$$

$$R = (0,70)^2 = 0,49$$

Če oseba gleda 18 ur na teden televizijo, lahko pričakujemo, da bo 1-krat na mesec šla v kino, pri čemer je standardna napaka $0,7$.

49% variance obiska kino predstav lahko pojasnimo z gledanjem televizije.

II.7 Časovne vrste



Družbeno-ekonomski pojavi so časovno spremenljivi. Spremembe so rezultat delovanja najrazličnejših dejavnikov, ki tako ali dugače vplivajo na pojave. Sliko dinamike pojavov dobimo s časovimi vrstami.

Časovna vrsta jo niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke.

Osnovni namen analize časovnih vrst je

- opazovati časovni razvoj pojavov,
- iskati njihove zakonitosti in
- predvidevati nadaljni razvoj.

Seveda to predvidevanje ne more biti popolnoma zanesljivo, ker je skoraj nemogoče vnaprej napovedati in upoštevati vse faktorje, ki vplivajo na proučevani pojav. Napoved bi veljala strogo le v primeru, če bi bile izpolnjene predpostavke, pod katerimi je napoved izdelana.

Časovne vrste prikazujejo individualne vrednosti neke spremenljivke v času. Čas lahko interpretiramo kot trenutek ali razdobje; skladno s tem so časovne vrste

- trenutne, npr. število zaposlenih v določenem trenutku:
- intervalne, npr. družbeni proizvod v letu 1993.

Časovne vrste analiziramo tako, da opazujemo spreminjanje vrednosti členov v časovih vrstah in iščemo zakonitosti tega spreminjanja. Naloga enostavne analize časovnih vrst je primerjava med členi v isti časovni vrsti.

Z metodami, ki so specifične za analizo časovnih vrst, analiziramo zakonitosti dinamike ene same vrste, s korelacijsko analizo pa zakonitosti odvisnosti v dinamiki več pojavov, ki so med seboj v zvezi.

Primer:

Vzemimo število nezaposlenih v Sloveniji v letih od 1981 do 1990.

V metodoloških pojasnilih v Statističnem letopisu Republike Slovenije 1991, so nezaposlni (spremenljivka X) opredeljeni takole:

"Brezposelna oseba je oseba, ki je sposobna in voljna delati ter je pripravljena sprejeti zaposlitev, ki ustreza njeni strokovni izobrazbi oz. z delom pridobljeni delovni zmožnosti, vendar brez svoje krivde nima dela in možnosti, da si z delom zagotavlja sredstva za preživetje in se zaradi zaposlitve prijavi pri območni enoti Zavoda za zaposlovanje (do leta 1989 skupnosti za zaposlovanje)."

leto	X_k
1981	12.315
1982	13.700
1983	15.781
1984	15.300
1985	11.657
1986	14.102
1987	15.184
1988	21.311
1989	28.218
1990	44.227

Primerljivost členov v časovni vrsti

Kljub temu, da so členi v isti časovni vrsti istovrstne količine, dostikrat niso med seboj neposredno primerljivi.

> Osnovni pogoj za primerljivost členov v isti časovni vrst je pravilna in nedvoumna opredelitev pojava, ki ga časovna vrsta prikazuje.
> Ta opredelitev mora biti vso dobo opazovanja enaka in se ne sme spreminjati.

Ker so spremembe pojava, ki ga časovna vrsta prikazuje bistveno odvisne od časa, je zelo koristno, če so **časovni razmiki med posameznimi členi enaki**. Na velikost pojavov dostikrat vplivajo tudi **administrativni** ukrepi, ki z vsebino proučevanja nimajo neposredne zveze.

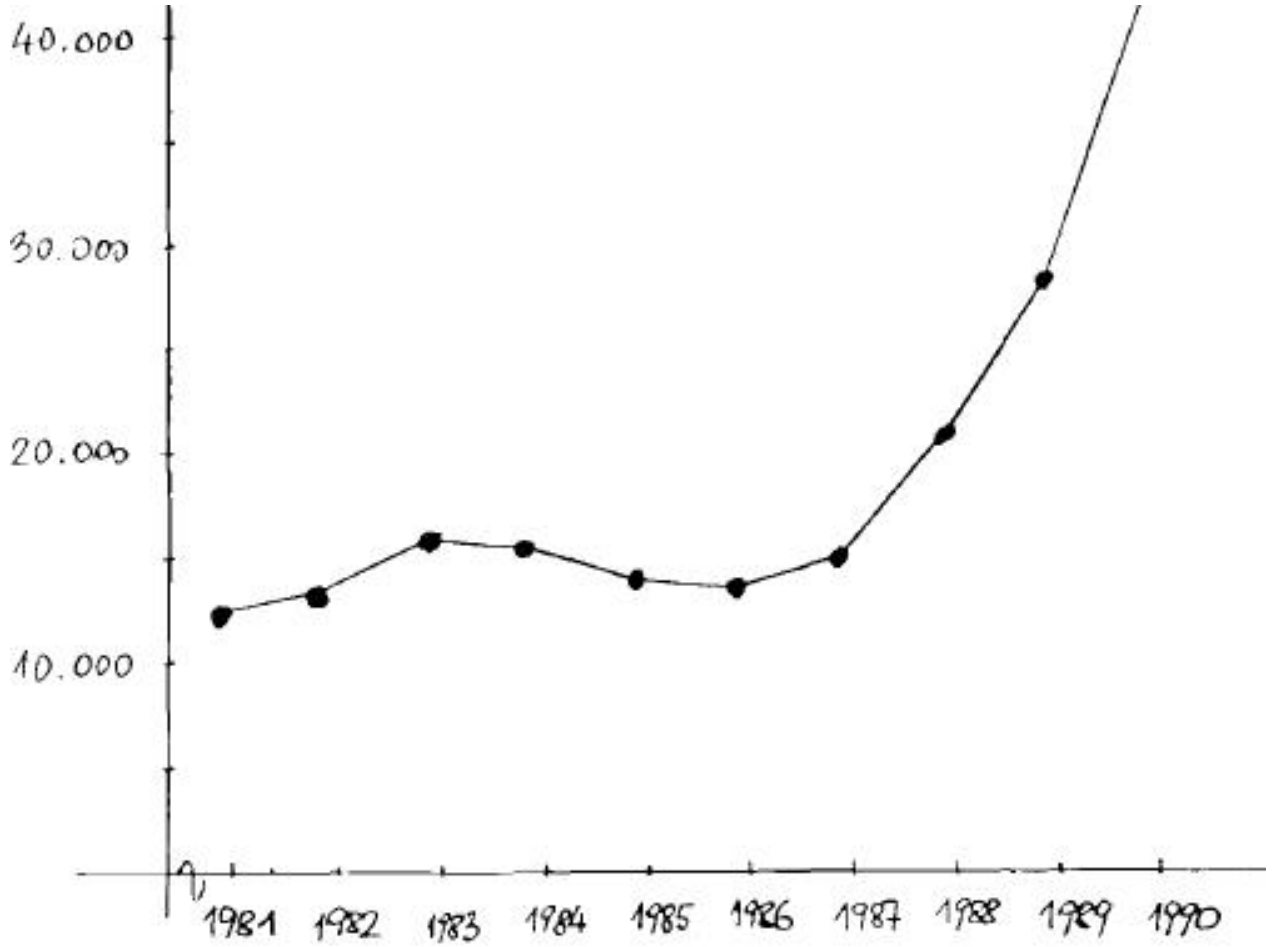
En izmed običajnih vzrokov so upravnoteritorialne spremembe, s katerimi se spremeni geografska opredelitev pojava, ki onemogoča primerljivost podatkov v časovni vrsti. V tem primeru je potrebno podatke časovne vrste za nazaj preračunati za novo območje.

Grafični prikaz časovne vrste

Kompleksen vpogled v dinamiko pojavov dobimo z grafičnim prikazom časovnih vrst v koordinatnem sistemu, kjer nanašamo na abscisno os čas in na ordinatno vrednosti dane spremenljivke. V isti koordinatni sistem smemo vnašati in primerjati le istovrstne časovne vrste.

Primer:

Grafično prikažimo število brezposelnih v Sloveniji v letih od 1981 do 1990.



Indeksi

Denimo, da je časovna vrsta dana z vrednostmi neke spremenljivke v časovnih točkah takole:

$$X_1, X_2, \ldots, X_n$$

o indeksih govorimo, kadar z relativnimi števili primerjamo istovrstne podatke.

Glede na to, kako določimo osnovo, s katero primerjamo člene v časovni vrsti, ločimo dve vrsti indeksov:

- **Indeksi s stalno osnovo**
  Člene časovnih vrst primerjamo z nekim stailnim členom v časovni vrsti, ki ga imenujemo osnova X_0

$$I_{k/0} = \frac{X_k}{X_0} \cdot 100.$$

- **Verižni indeksi**
  Za dano časovno vrsto računamo vrsto verižnih indeksov tako, da za vsak člen vzamemo za osnovo predhodni člen

$$I_k = \frac{X_k}{X_{k-1}} \cdot 100.$$

  člene časovne vrste lahko primerjamo tudi z absolutno in relativno razliko med členi:

- **Absolutna razlika**

$$D_k = X_k - X_{k-1}.$$

- **Stopnja rasti** (relativna razlika med členi)

$$T_k = \frac{X_k - X_{k-1}}{X_{k-1}} \cdot 100 = I_k - 100.$$

Interpretacija indeksov

	pojav		
indeks	raste	stagnira	pada
s stalno osnovo	$I_{k+1/0} > I_{k/0}$	$I_{k+1/0} = I_{k/0}$	$I_{k+1/0} < I_{k/0}$
verižni indeks	$I_k > 100$	$I_k = 100$	$I_k < 100$
stopnja rasti	$T_k > 0$	$T_k = 0$	$T_k < 0$

Primer: Izračunajmo omenjene indekse za primer brezposelnih v Sloveniji:

leto	X_k	$I_{k/0}$	I_k	T_k
1981	12.315	100	–	–
1982	13.700	111	111	11
1983	15.781	128	115	15
1984	15.300	124	97	−3
1985	11.657	119	96	−4
1986	14.102	115	97	−3
1987	15.184	124	107	7
1988	21.311	173	141	41
1989	28.218	229	132	32
1990	44.227	359	157	57

Rezultati kažejo, da je bila brezposenost v letu 1990 kar 3,5 krat večja kot v letu 1981 (glej indeks s stalno osnovo).

Iz leta 1989 na leto 1990 je bil prirast nezposlenih 57% (glej stopnjo rasti).

Sestavine dinamike v časovnih vrstah

Posamezne vrednosti časovnih vrst so rezultat številnih dejavnikov,
ki na pojav vplivajo.

Iz časovne vrste je moč razbrati skupen učinek dejavnikov,
ki imajo širok vpliv na pojav, ki ga proučujemo.

Na časovni vrsti opazujemo naslednje vrste sprememb:

1. **Dolgoročno gibanje ali trend - X_T**
   podaja dolgoročno smer razvoja.
   Običajno ga je mogoče izraziti s
   preprostimi rahlo ukrivljenimi krivuljami.

2. **Ciklična gibanja -** X_C,
   so oscilarijo okoli trenda.
   Poriode so ponavdi daljše od enega leta
   in so lahko različno dolge.

3. **Sezonske oscilacije -** X_S
   so posledice vzrokov, ki se pojavljajo na stalno razdobje.
   Periode so krajše od enega leta, ponavadi sezonskega značaja.

4. **Naključne spremembe -** X_E
   so spremembe, ki jih ne moremo razložiti s
   sistematičnimi gibanji (1, 2 in 3).

Časovna vrsta ne vsebuje nujno vseh sestavin. Zvezo med sestavinami je mogoče prikazati z nekaj osnovnim modeli. Npr.:

$$X = X_T + X_C + X_S + X_E$$

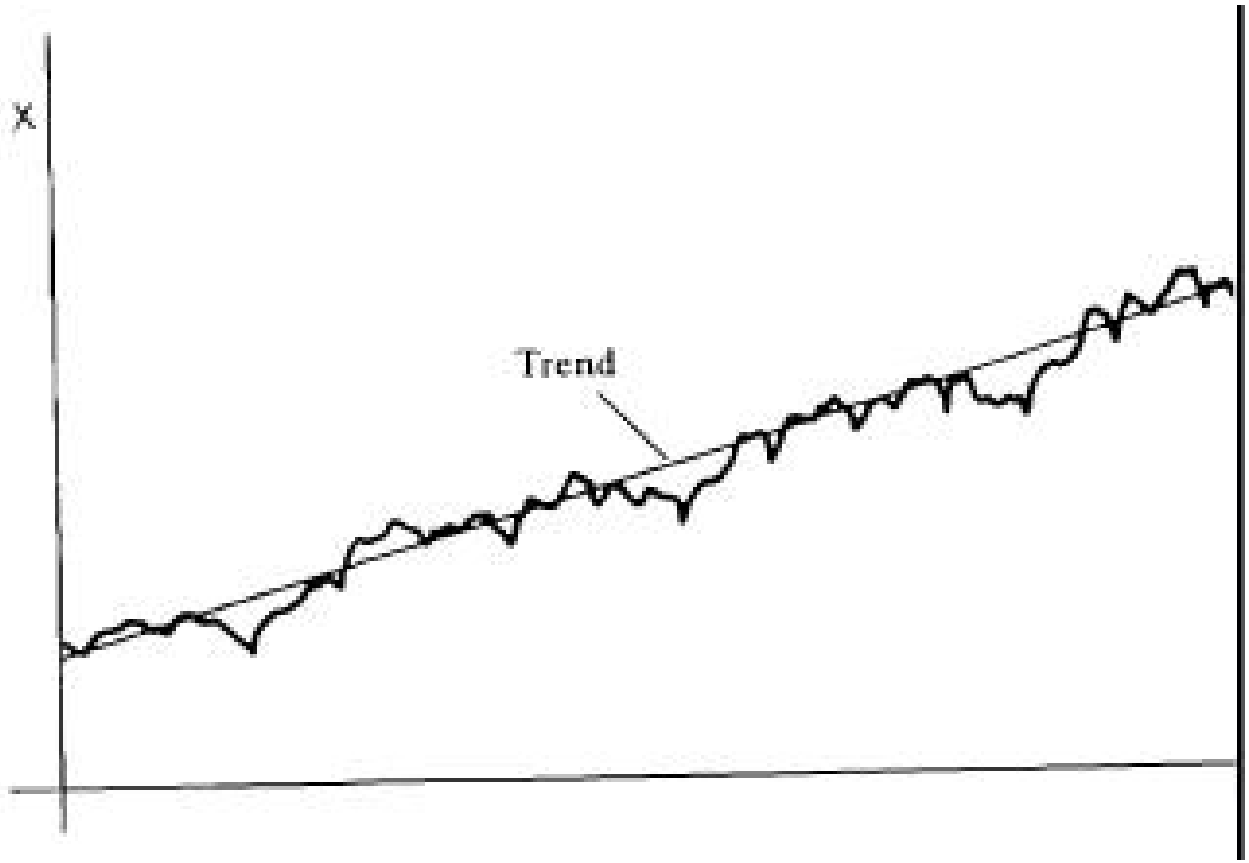
ali

$$X = X_T \cdot X_C \cdot X_S \cdot X_F;$$

ali

$$X = X_T \cdot X_C \cdot X_S + X_E.$$

Primer časovne vrste z vsemi štirimi sestavinami:



Ali je v časovni vrsti trend?

Obstaja statistični test, s katerim preverjamo ali trend obstaja v časovni vrsti. Med časom in spremenljivko izračunamo koeficient korelacije rangov

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

kjer je d_i, razlika med rangoma i tega časa in pripadajoče vrednosti spremenljivke. Ničelna in osnovna domneva sta:

H_0: $\rho_e = 0$ trend ne obstaja
H_1: $\rho_e \neq 0$ trend obstaja

Ustrezna statistika je

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

ki se porazdeluje približno po t porazdelitvi z $(n-2)$ prostostnimi stopnjami.

Metode določanja trenda

- Prostoročno
- Metoda drsečih sredin
- Metoda najmanjših kvadratov
- Druge analitične metode

Drseče sredine

Metoda drsečih sredin lahko pomaga pri določitvi ustreznega tipa krivulje trenda. V tem primeru namesto člena časovne vrste zapišemo povprečje določenega števila sosednjih članov. Če se odločimo za povprečje treh členov, govorimo o tričlenski vrsti drsečih sredin. Tedaj namesto članov v osnovni časovni vrsti X_k: tvorimo tričlenske drseče sredine X :

$$X'_k = \frac{X_{k-1} + X_k + X_{k+1}}{3}.$$

V tem primeru prvega in zadnjega člena časovne vrste moramo izračunati.

- Včasih se uporablja obtežena aritmetična sredina, včasih celo geometrijska za izračun drsečih sredin.
- Če so v časovni vrsti le naključni vplivi, dobimo po uporabi drsečih sredin ciklična gibanja (učinek Slutskega).
- Če so v časovni vrsti stalne periode, lahko drseče sredine zabrišejo oscilacije v celoti.
- V splošnem so drseče sredine lahko dober približek pravemu trendu.

Primer: Kot primer drsečih sredin vzemimo zopet brezposelne v Sloveniji. Izračunajmo tričlensko drsečo sredino:

T	X_k	tričl. drs. sred.
1981	12.315	–
1982	13.700	13.032
1983	15.781	14.030
1984	15.240	15.249
1985	15.300	14.710
1986	14.657	14.678
1987	14.102	15.184
1988	21.341	21.581
1989	28.218	31.262
1990	44.227	–

Analitično določanje trenda

Trend lahko obravnavamo kot posebni primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend

$$X_T = f(T),$$

lahko parametre trenda določimo z metoda najmanjših kvadratov

$$\sum_{i=1}^{n} (X_i - X_{iT})^2 = \min .$$

V primeru linearnega trenda

$$X_T = a + bT,$$

$$\sum_{i=1}^{n}(X_i - a - bT_i)^2 = \min .$$

dobimo naslednjo oceno trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X})(T_i - \overline{T})}{\sum_{i=1}^{n}(T_i - \overline{T})^2}\,(T - \overline{T}).$$

Ponavadi je čas T transformiran tako, da je $t = 0$. Tedaj je ocena trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \cdot t_i}{\sum_{i=1}^{n} t_i^2}\,t.$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je

$$\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{iT})^2}.$$

Primer:

Kot primer ocenimo število doktoratov znanosti v Sloveniji v razdobju od leta 1986 do 1990. Z linearnim trendom ocenimo koliko doktorjev znanosti je v letu 1991. Izračunajmo tudi standardno napako ocene.

Izračunajmo najprej trend:

T	Y_i	t_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})t_i$	t_i^2
1986	89	-2	$-19,8$	$39,6$	4
1987	100	-1	$-8,8$	$8,8$	1
1988	118	0	$9,2$	0	0
1989	116	1	$7,2$	$7,2$	1
1990	121	2	$12,2$	$24,4$	4
	544	0		80	10

$$\overline{Y} = \frac{544}{4} = 1088,$$

$$Y_T = 108,8 + \frac{80}{10}\, t = 108,8 + 8t,$$

$$Y_T(1991) = 108,8 + 8 \cdot 3 = 132,8.$$

Ocena za leto 1991 je približno 133 doktorjev znanosti.

Zdaj pa izračunajmo standardno napako ocene.

Za vsako leto je potrebno najprej izračunat

T	Y_i	Y_{iT}	$Y_i - Y_{iT}$	$(Y_i - Y_{iT})^2$
1986	89	$92,8$	$-3,8$	$14,14$
1987	100	$100,8$	$-0,8$	$0,64$
1988	118	$108,8$	$9,2$	$84,64$
1989	116	$116,8$	$-0,8$	$0,64$
1990	121	$124,8$	$-3,8$	$14,44$
	544	544	0	$114,8$

$$\sigma_e = \sqrt{\frac{114,8}{5}} = 4,8.$$

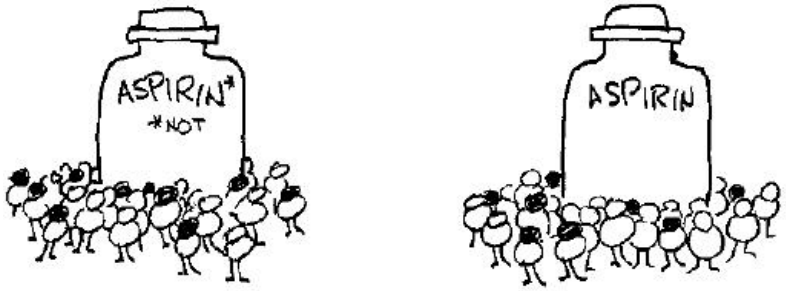
II.8. Načrtovanje eksperimentov

Načrtovanje eksperimentov se pogosto neposredno prevede v uspeh oziroma neuspeh.

V primeru parjenja lahko statistik spremeni svojo vlogo iz pasivne v aktivno.

Predstavimo samo osnovne ideje, podrobno numerično analizo pa prepustimo statistični programski opremi.

Elementi načrta so eksperimentalne enote ter terapije, ki jih želimo uporabiti na enotah.



- medicina: bolniki (enote) in zdravila (terapije),
- optimizacija porabe: taxi-ji (enote) in različne vrste goriva (terapije),
- agronomija: območja na polju in različne vrste kulture, gnojiva, špricanja,...

Danes uporabljamo ideje načrtovanja eksperimentov na številnih področjih:

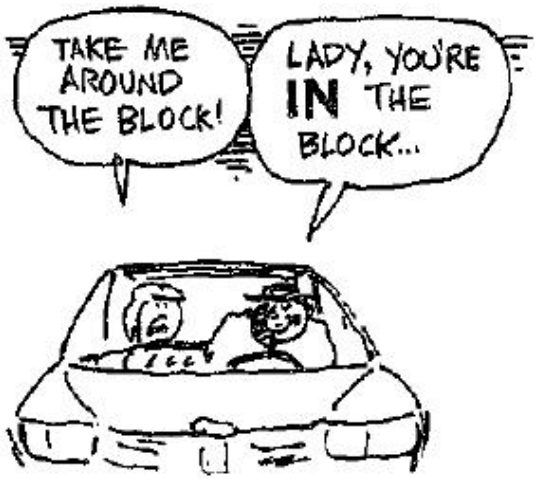
- optimizacija industrijskih procesov,
- medicina,
- sociologija.

Na primeru bomo predstavili tri osnovne principe načrtovanja eksperimentov:

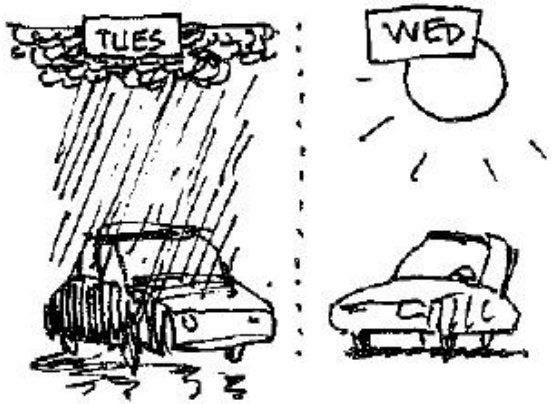
1. Ponavljanje: enake terapije pridružimo različnim enotam, saj ni mogoče oceniti naravno spremenljivost (ang. natural variability) in napake pri merjenju.

2. Lokalna kontrola pomeni vsako metodo, ki zmanjša naravno spremenljivost.

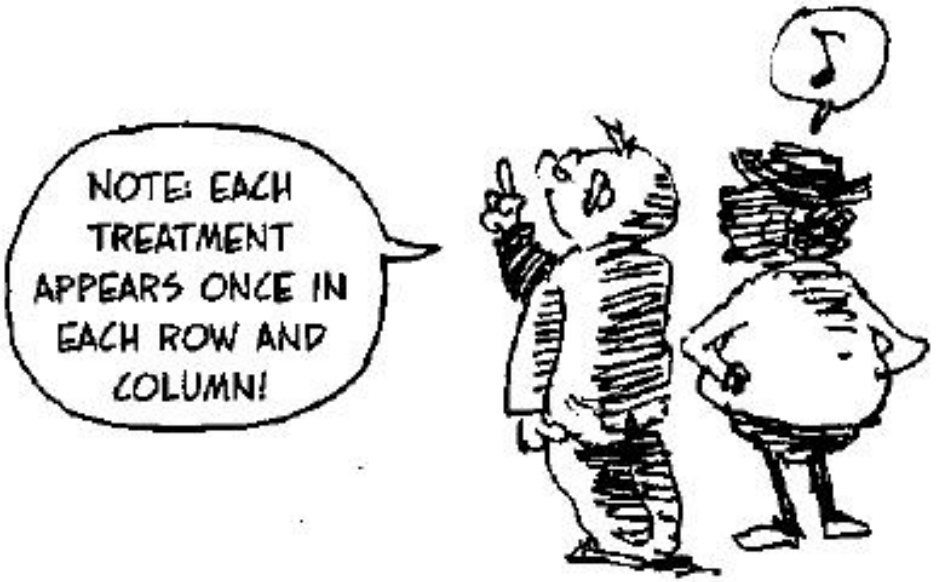
En od načinov grupira podobne enote eksperimentov v **bloke**. V primeru taxijev uporabimo obe vrsti goriva na vsakem avtomobilu in rečemo, da je avto blok.



3. Naključna izbira je bistven korak povsod v statistiki! Terapije za enote izbiramo naključno. Za vsak taksi izberemo vrsto goriva za torek oziroma sredo z metom kovanca. Če tega ne bi storili, bi lahko razlika med torkom in sredo vplivala na rezultate.



		DAY 1	2	3	4
CAB	1	a	b	c	d
	2	b	c	d	a
	3	c	d	a	b
	4	d	a	b	c

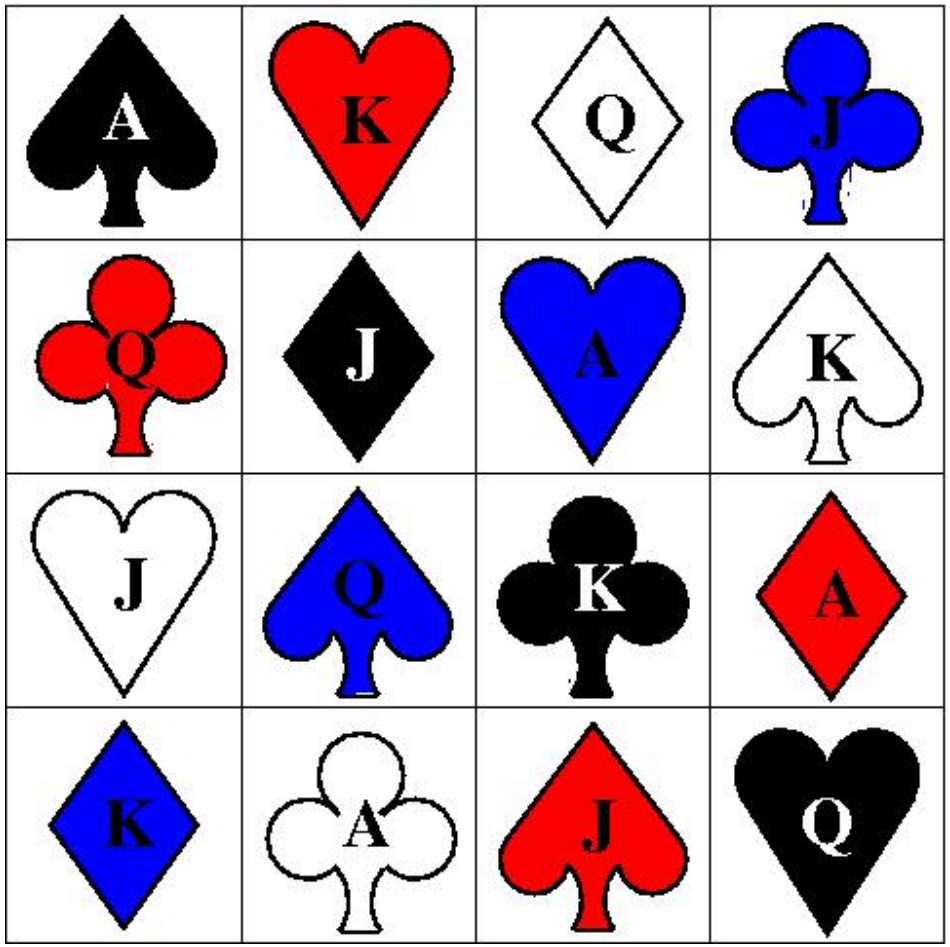


Latinski kvadrati

Latinski kvadrat reda v je $v \times v$-razsežna matrika, v kateri vsi simboli iz množice

$$\{1, \ldots, v\}$$

nastopajo v vsaki vrstici in vsakem stolpcu.



Trije paroma ortogonalni latinski kvadrati reda 4,

tj. vsak par znak-črka ali črka-barva ali barva-znak

se pojavi natanko enkrat.

III. ZAKLJUČKI

Osnovni principi in orodja,
ki smo jih spoznali pri VIS,
lahko posplošimo in razširimo
do te mere, da se dajo
z njimi rešiti tudi
bolj kompleksni problemi.

Spoznali smo kako predstaviti **eno** spremenljivko (dot-plot, histogrami,...) in **dve** spremenljivki (razsevni diagram).

Kako pa predstavimo več kot dve spremenljivki na ravnem listu papirja?

Med številnimi možnostmi moramo omeniti idejo **Hermana Chernoffa**, ki je uporabil človeški obraz, pri čemer je vsako lastnost povezal z eno spremenljivko.

Oglejmo si Chernoffov obraz:

X =naklon obrvi,

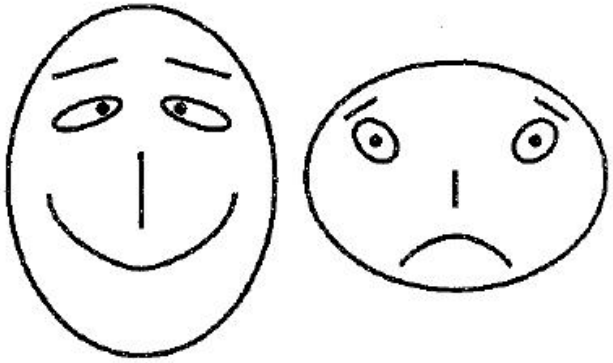
Y =velikost oči,

Z =dolžina nosu,

T =dolžina ust,

U =višino obraza,

itd.

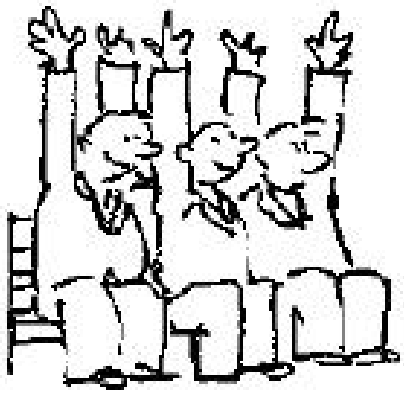


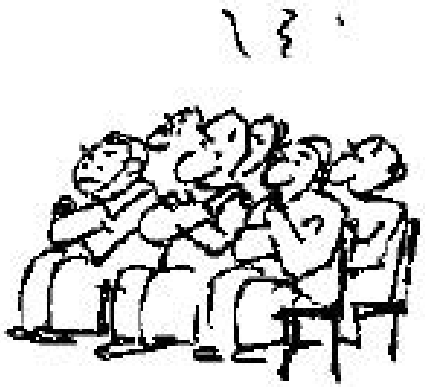
Multivariantna analiza

Širok izbor multivariantnih modelov nam omogoča analizo in ponazoritev n-razsežnih podatkov.

Združevalna/grozdna tehnika (ang. cluster technique):

Iskanje delitve populacije
na homogene podskupine,
npr. z analizo vzorcev
senatorskih glasovanj v ZDA
zaključimo, da *jug* in *zahod*
tvorita dva različna grozda.





Diskriminacijska analiza

je obraten proces. Npr. odbor/komisija za sprejem novih študentov bi rad našel podatke, ki bi že vnaprej opozorili ali bodo prijavljeni kandidati nekega dne uspešno zaključili program (in finančno pomagali šoli - npr. z dobrodelnimi prispevki) ali pa ne bodo uspešni (gre delati dobro po svetu in šola nikoli več ne sliši zanj(o).

Analiza faktorjev

išče poenostavljeno razlago večrazsežnih podatkov z manjšo skupino spremenljivk.

Npr. Psihiater lahko postavi 100 vprašanj, skrivoma pa pričakuje, da so odgovori odvisni samo od nekaterih faktorjev:

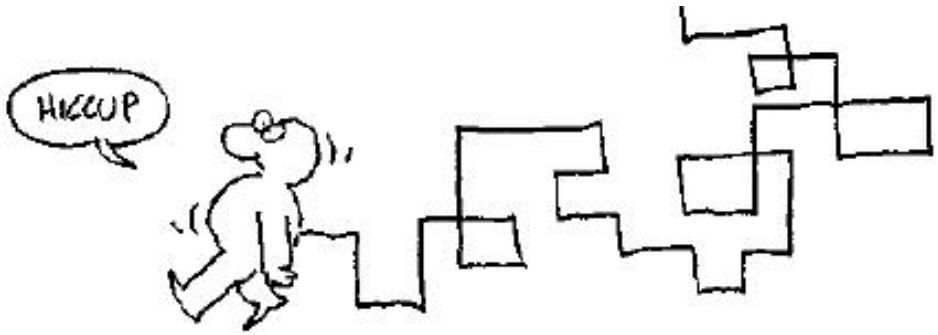
ekstravertiranost,
avtoritativnost,
alutarizem, itd.

Rezultate testa lahko potem povzamemo le z nekaterimi sestavljenimi rezultati v ustreznih dimenzijah.

Naključni sprehodi

pričnejo z metom kovanca, recimo, da se pomaknemo korak nazaj, če pade grb, in korak naprej, če pade cifra. (z dvema kovancema se lahko gibljemo v 2-razsežnemu prostoru - tj. ravnini).

Če postopek ponavljamo, pridemo do
stohastičnega procesa,
ki ga imenujemo
naključni sprehod
(ang. random walk).

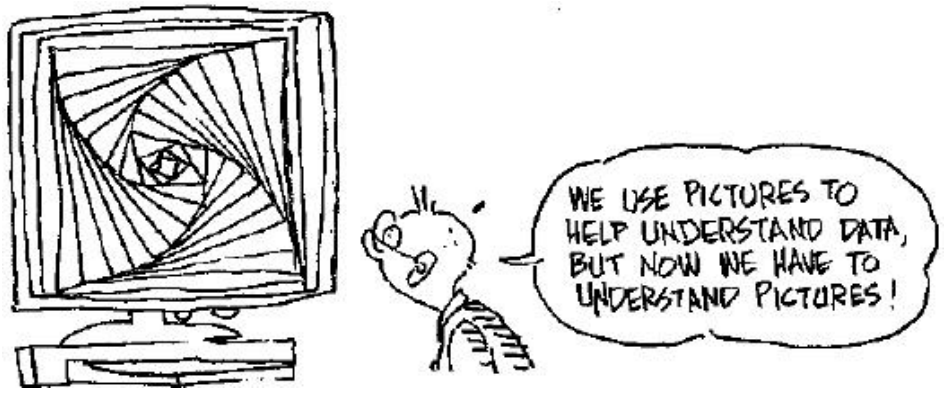


Modeli na osnovi naključnih sprehodov se uporabljajo za nakup/prodajo delnic in portfolio management.

Vizualizacija in analiza slik

Sliko lahko sestavlja 1000×1000 pikslov,
ki so predstavljeni z eno izmed 16,7 milijonov barv.

Statistična analiza slik
želi najti nek pomen iz
"informacije" kot je ta.



Ponovno vzorčenje

Pogosto ne moremo izračunati standardne napake in limite zaupanja.

Takrat uporabimo tehniko ponovnega vzorčenja,
ki tretira vzorec, kot bi bila celotna populacija.
Za takšne tehnike uporabljamo
pod imeni:

randomization
Jackknife, in
Bootstrapping.

Kvaliteta podatkov

navidezno majhne napake pri vzorčenju, merjenju, zapisovanju podatkov, lahko povzročijo katastrofalne učinke na vsako analizo.

R. A. Fisher, genetik in ustanovitelj moderne statistike ni samo načrtoval in analiziral eksperimentalno rejo, pač pa je tudi čistil kletke in pazil na živali. Zavedal se je namreč, da bi izguba živali vplivala na rezultat.

Moderni statistiki, z njihovimi računalniki in podatkovnimi bazami ter vladnimi projekti (beri denarjem) si pogosto ne umažejo rok.

Inovacija

Najboljše rešitve niso vedno v knjigah
(no vsaj najti jih ni kar tako).

Npr. Mestni odpad je najel strokovnjake,
da ocenijo kaj sestavljajo odpadki,
le-ti pa so se znašli pred zanimivimi problemi,
ki se jih ni dalo najti v standardnih učbenikih.

Komunikacija

Še tako uspešna in bistroumna analiza je zelo malo vredna,
če je ne znamo jasno predstaviti,
vključujoč stopnjo statistične značilnosti? v zaključku.

Npr. V medijih danes veliko bolj natančno poročajo o velikosti napake pri svojih anketah.

Timsko delo

V današnji kompleksni družbi.

Reševanje številnih problemov zahteva *timsko delo*.

Inženirji, statistiki in delavci sodelujejo,
da bi izboljšali kvaliteto produktov.

Biostatistiki, zdravniki, in AIDS-aktivisti združeno sestavljajo klinične poiskuse, ki bolj učinkovito ocenijo terapije.