Statistika

Opisna statistika (deskriptivna statistika)

Zanima nas vzorec, ki ga imamo

Porazdelitev opišemo
- grafično
- z merami središčne vrednosti in variabilnosti

Statistično sklepanje (inferenčna statistika)

Zanima nas populacija, sklepamo na podlagi vzorca

Sklepamo s pomočjo verjetnosti
- ocena povprečja
- povezanost
- bistvena velikost vzorca in rezprezentativnost

Verjetnost

- determinističnost $\Leftrightarrow$ variabilnost
- populacija $\Leftrightarrow$ vzorec
- odločanje na podlagi verjetnosti

Verjetnost in statistika

Verjetnost

Poznamo populacijo
(teorijo)

Kakšna je verjetnost
dogodka na vzorcu?

Statistika

Kakšna je verjetnost neke
povezave v populaciji

Imamo vzorec (podatke)

Verjetnost in statistika

Verjetnost

Imamo kocko, verjetnost vsakega izida $1/6$

Kakšna je verjetnost, da v 5 metih pade 5 šestic?

Statistika

Kakšna je verjetnost, da je kocka poštena?

V 5 metih je padlo 5 šestic

Verjetnost in statistika

Verjetnost

Imamo zdravilo, ki ozdravi bolezen pri 70% bolnikov

Kakšna je verjetnost, da od 30 bolnikov ozdravi le 17?

Statistika

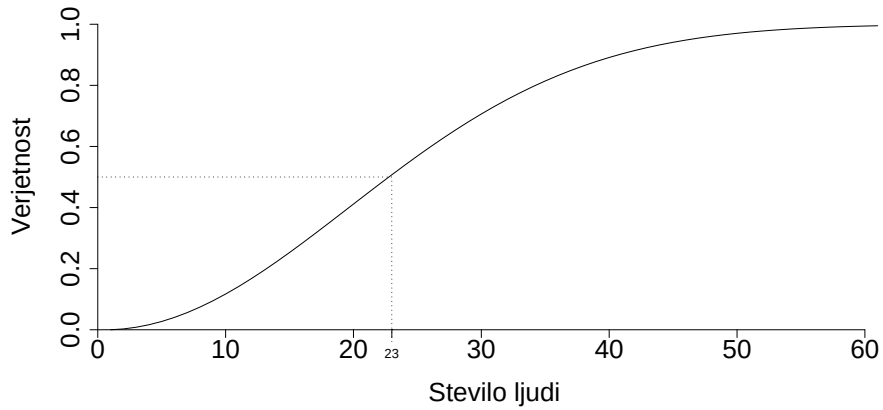
Kakšna je verjetnost, da je v populaciji učinkovitost zdravila 70%?

Od 30 bolnikov jih po prejemu zdravila ozdravi 17

Primeri iz verjetnosti

Kakšna je verjetnost, da imata v skupini 60 ljudi dva posameznika rojstni dan na isti dan v letu?

$$P(n) = 1 - \left(1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365 - (n-1)}{365}\right)$$



Pri 57 ljudeh je verjetnost že večja od 99 %.

Opomba: pri izračunu je bilo narejenih več poenostavitev - izpustili smo 29. februar, predpostavili, da med ljudmi ni dojenčkov, predpostavili, da so vsi dnevi rojstva v letu enako verjetni

Primeri iz verjetnosti

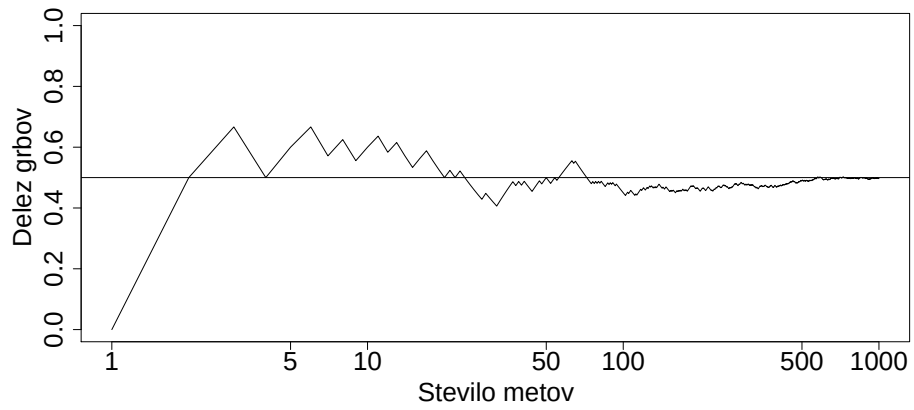
Kakšna je verjetnost, da ima izmed 60 ljudi eden rojstni dan na isti dan kot ti?

$$P(n) = 1 - \left(\frac{364}{365}\right)^n$$

Za verjetnost nad 0,5 je potrebnih 253 ljudi.

Verjetnost

Vrednosti h kateri konvergira delež na vzorcu



Verjetnost

Tabela prikazuje število porodov v Sloveniji v letih 1991 do 1995.
Vidimo, da se relativne frekvence dečkov gibljejo okrog 0,51, kar bi bilo nekako smiselno vzeti za verjetnost rojstva dečka.

Leto	1991	1992	1993	1994	1995
Porodov skupaj	21583	19982	19793	19463	18980
Dečki	11116	10333	10188	9899	9741
Deklice	10467	9649	9605	9564	9239
Relativna frekvenca	0,515	0,517	0,514	0,508	0,513

Osnove verjetnosti

- Osnovna pojma verjetnosti sta **poskus** in **dogodek**
- Poskus: neka množica dejstev, ki se vedno pojavi hkrati
- Dogodek: pojav, ki se v poskusu lahko zgodi, a to ni nujno

Primeri:

Poskus	Dogodek
met kocke	pade šestica
met dveh kock	vsota pik je večja od 10
porod	rodi se deček
porod	rodi se otrok z manj kot 3000 g porodne teže

Notacija

Dogodke bomo označevali z velikimi tiskanimi črkami, na primer

Pri metu kocke	$A = \{\text{izid je sodo število}\}$
Ob rojstvu	$B = \{\text{novorojenček je deček}\}$
Gestacijska starost	$C = \{\text{tednov nosečnosti} \geq 37\}$

Verjetnost dogodka A bomo označevali s $P(A)$. Seveda za vsak dogodek A velja

$$0 \leq P(A) \leq 1.$$

Osnove verjetnosti

- Dogodka sta neodvisna, kadar nam poznavanje enega ne da nikakršne informacije o drugem
- Dogodka sta nezdružljiva, kadar se ne moreta zgoditi hkrati
- Dogodka sta nasprotna, če sta nezdružljiva, skupaj pa predstavljata gotov dogodek

Osnove verjetnosti

Produkt dogodkov

- dogodek, ki se zgodi, kadar se zgodita A in B
- Oznaka: AB oz. $A \cap B$ (A krat B)
- $P(AB) = P(A)P(B)$ če sta dogodka neodvisna
- $P(AB) = 0$ če se dogodka ne moreta zgoditi hkrati

Unija dogodkov

- dogodek, ki se zgodi, kadar se zgodi ali A ali B
- Oznaka: $A \cup B$ (A unija B)
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Verjetnosti dogodkov torej lahko seštevamo le kadar se dogodki ne morejo zgoditi hkrati (nezdružljivi dogodki)

Pogojna verjetnost

Bolnik z rakom prostate, zanima nas verjetnost, da preživi 2 leti

- Verjetnost odvisna od tega, koliko je bila bolezen razširjena v času diagnoze
- Štirje stadiji bolezni, označimo jih z I, II, III in IV.

Relativna frekvenca bolnikov, ki so živi po 2 letih in so bili v stadiju I **ni**

$$\frac{\text{število bolnikov, ki preživijo 2 leti}}{\text{število vseh bolnikov}}$$

ampak

$$\frac{\text{število bolnikov } \textbf{v stadiju I}\text{, ki preživijo 2 leti}}{\text{število vseh bolnikov } \textbf{v stadiju I}}$$

oziroma

$$\frac{\text{število bolnikov } \textbf{v stadiju I}\text{, ki preživijo 2 leti / število vseh bolnikov}}{\text{število vseh bolnikov } \textbf{v stadiju I} \text{ / število vseh bolnikov}}$$

Pogojna verjetnost

Z oznakami:

- A dogodek, da bolnik preživi 2 leti
- B dogodek, da je bolnik v stadiju I

$P(A|B)$ pogojna verjetnost dogodka A glede na dogodek B (verjetnost, da se zgodi A, če se zgodi B)

Relativne frekvence nadomestimo z verjetnostmi in dobimo:

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

Primer: Delež dečkov med novorojenčki

Relativne frekvence dečkov glede na gestacijsko starost

teden	št. dečkov	št. deklic	% dečkov
≤ 35	148	130	53,2
36	64	70	47,8
37	170	173	49,6
38	398	372	51,7
39	838	791	51,4
40	1163	1175	49,7
41	431	393	52,3
42	20	20	50,0
skupaj	3232	3124	50,8

Pogojna verjetnost in neodvisnost

A je neodvisen od B, kadar je $P(A) = P(A|B)$.

Primer: met kocke

Definirajmo dogodke

$$A = \{2,4,6\} \quad B = \{1,2,3,4\} \quad C = \{1,2,3\}$$

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{2/6}{4/6} = \frac{1}{2}$$

$P(A|B) = P(A)$ A in B sta torej neodvisna.
Podobno bi ugotovili, da $P(A|C) \neq P(A)$ in s tem, da A in C nista neodvisna.

Pri vrstah spremenljivk lahko

Slučajne spremenljivke

- Definicija slučajne spremenljivke: merjena količina, katere vrednosti naključno variirajo
- Največkrat rečemo kar spremenljivka
- Označujemo jih ponavadi z velikimi tiskanimi črkami, njihove vrednosti pa z malimi tiskanimi črkami

Nekaj primerov

Poskus	Slučajna spremenljivka	
Met kocke	Število pik	$P(X = x) = \frac{1}{6},\ x = 1, \ldots, 6$
Met dveh kock	Vsota pik	
Rojstvo	Teža novorojenčka	

Neodvisnost slučajnih spremenljivk

Intuitivno

Dve slučajni spremenljivki sta neodvisni, če poznavanje vrednosti ene od njiju ne pove ničesar o vrednostih druge

Nekaj primerov

- Izida dveh zaporednih metov kocke sta neodvisna
- Vrednosti krvnih pritiskov dveh različnih oseb so (ponavadi) neodvisne
- Višina in teža osebe nista neodvisni

Formalno

Slučajni spremenljivki X in Y sta neodvisni, če sta dogodka $\{X \leq x\}$ in $\{Y \leq y\}$ neodvisna za vsak x in y

Pogojna neodvisnost

Spremenljivki sta neodvisni pri dani vrednosti tretje spremenljivke

Primer - umetna oploditev in Downov sindrom

- Med umetno oplojenimi ženskami je več otrok z Downovim sindromom kot sicer
- Starost je povezana z umetno oploditvijo - ženske so v povprečju starejše.
- Starost je povezana z Downovim sindromom - starejša kot je ženska, večje je tveganje za kromosomske nepravilnosti.
- Če primerjamo enako stare ženske, ni razlik v deležu otrok z Downovim sindromom

Spremenljivki ‘pojavitev Downovega sindroma’ in ‘umetna oploditev’ sta neodvisni pogojno glede na starost.

Neodvisnost

Pomemben koncept v statistiki

Idealizacija ali poenostavitev procesov v naravi, ki jo s pridom izkoriščamo pri statističnem modeliranju. Zanimalo nas bo:

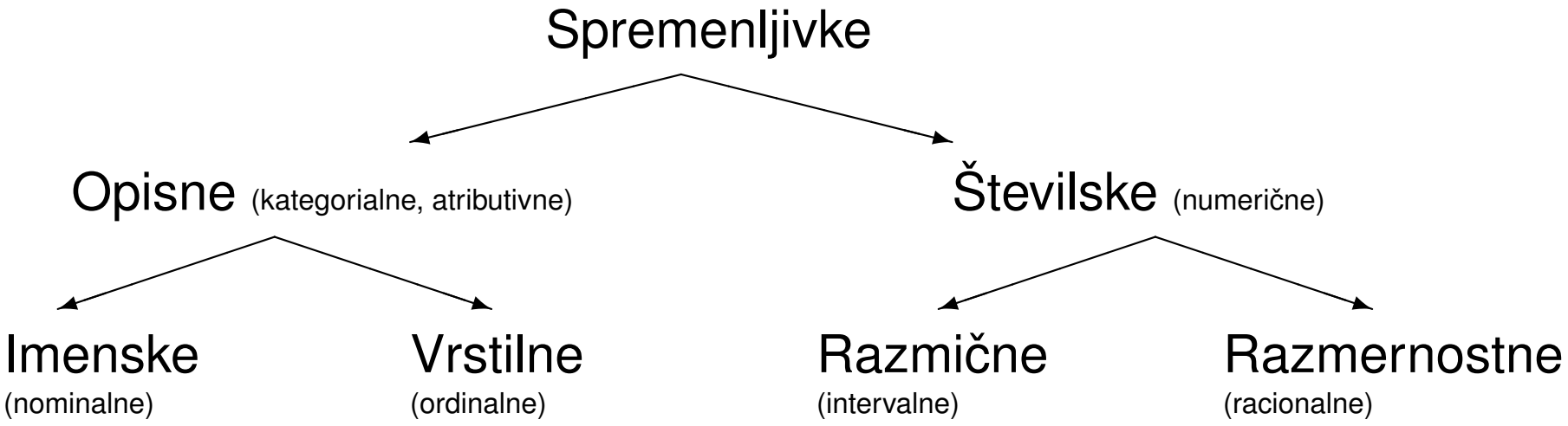
- Ali obstaja povezanost med dvema spremenljivkama (npr. doza zdravila in rezultat zdravljenja)
- Ali lahko s pomočjo lažje merljivih (morda na cenejši ali pa manj invaziven način) spremenljivk sklepamo o neki težje merljivi spremenljivki (npr. kožna guba za napovedovanje deleža telesne maščobe)
- Ali lahko s pomočjo danes dostopnih spremenljivk, napovemo nek rezultat (napovedujemo preživetje na podlagi stadija bolezni, starosti bolnika, vrste zdravljenja)

Neodvisnost - primer

- Če je prognoza za bolnika z rakom **neodvisna** od histološke klasifikacije tumorja, lahko napovedujemo, ne da bi se ozirali na takšno klasifikacijo.
- Če je prognoza za bolnika z rakom **pogojno neodvisna** od histološke klasifikacije tumorja pri dani starosti bolnika, potem nam ni potrebno poznati histološke klasifikacije, če poznamo starost bolnika.

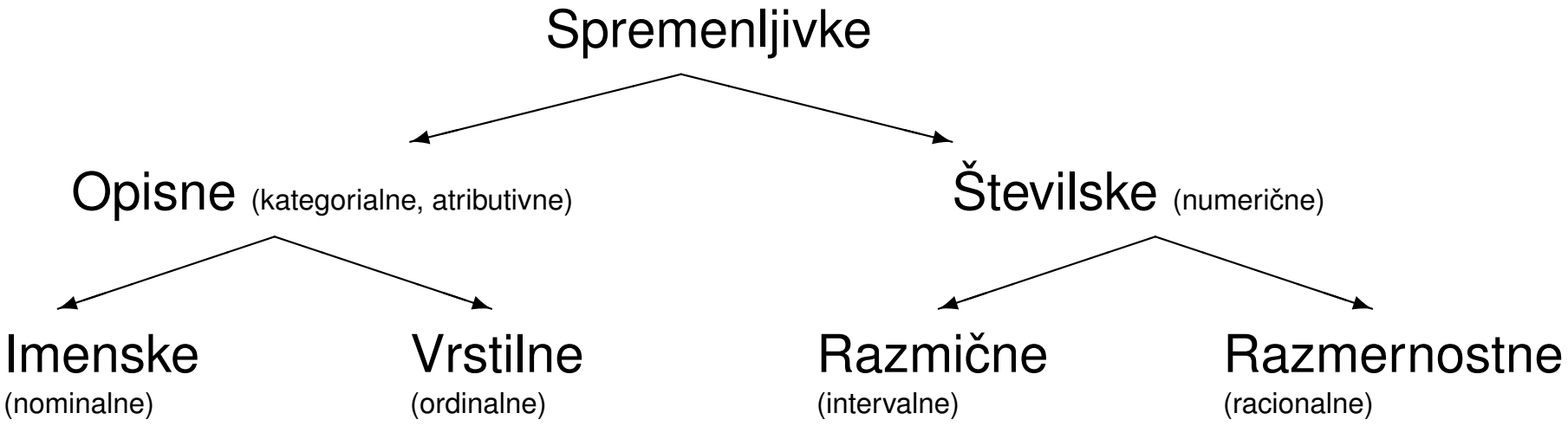
Kot bomo videli, bosta neodvisnost oz. pogojna neodvisnost pogosto predpostavki (hipotezi), ki ju bomo preverjali s statističnimi testi. Kadar bomo predpostavko o neodvisnosti zavrnili, bo pomembno opisati naravo odvisnosti.

Vrste spremenljivk



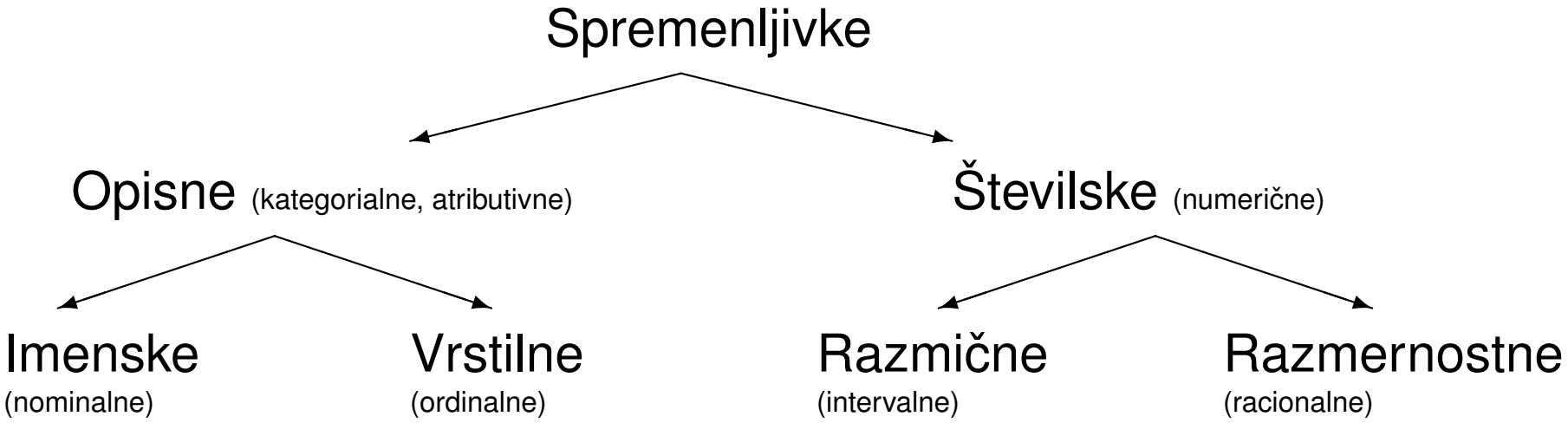
1. Opisne: vrednosti spremenljivke le opišemo.
   - Imenske: vrednosti ne moremo urediti (vsako kodiranje smiselno)
   - Vrstilne: vrednosti lahko uredimo po velikosti, razmiki med kategorijami niso nujno enaki
2. Številske: vrednosti so števila, s katerimi lahko računamo
   - Razmične: lahko odštevamo, kvocient nima pravega smisla, ne obstaja absolutna ničla
   - Razmernostne: imajo še absolutno ničlo, kvocient je smiseln

Vrste spremenljivk



1. Opisne (opišemo z besedo, pri kodiranju potrebna legenda):
   - Imenske: barva las, spol, krvna skupina, kajenje (kadilec / nekadilec /bivši kadilec
   - Vrstilne: stadij raka (I-IV), jemanje zdravil (redno, pogosto, redko, ne jemlje), kajenje (ne kadi, kadi občasno, kadi redno)
2. Številske (meritve, število, ipd., številka je naraven opis):
   - Razmične: temperatura, letnica rojstva
   - Razmernostne: starost, teža, število pokajenih cigaret na dan

Vrste spremenljivk



Glede na vrsto spremenljivke se bomo odločali za ustrezne povzetke podatkov (povprečen spol je nesmisel), grafične prikaze in statistične teste.

Vrste spremenljivk - diskretne in zvezne spremenljivke

Diskretne spremenljivke

- Obstaja nek vnaprej podan seznam možnih vrednosti
- Vse opisne spremenljivke
- Vse številske, ki predstavljajo število nečesa (število bolnikov v bolnišnici, število pozitivnih bezgavk, število enot zaužitega sadja na dan, število pokajenih cigaret)
- Zavzamejo lahko končno ali števno neskončno vrednosti (števno neskončno je vrednosti, ki jih lahko preštejemo: 0,1,2,3, . . .)

Vrste spremenljivk - diskretne in zvezne spremenljivke

Zvezne spremenljivke

- Zavzame poljubno vrednost na nekem intervalu
- Primeri so rezultati meritev (višina, krvni pritisk, laboratorijske vrednosti, delež maščob v živilu, porodna teža otroka)
- Vrednosti je neštevno neskončno

Vrste spremenljivk - nadaljevanje

Še nekaj opomb

- Spremenljivkam s samo dvema možnima vrednostima pravimo dihotomke
- Kategoriziranje številskih spremenljivk (npr. starost v starostne skupine) je pogost običaj, a le redko smiselno
- Predpogoj dobre analize so dobro definirane spremenljivke, primer vprašljive spremenljivke je merjenje bolečine (VAS)

Verjetnostna porazdelitev

Porazdelitev:

Če za neko slučajno spremenljivko poznamo vse možne izide in vemo, kako pogosto jih lahko zavzame, pravimo, da poznamo njeno verjetnostno porazdelitev

Porazdelitev diskretne spremenljivke

Porazdelitev:

Verjetnost, s katero slučajna spremenljivka zavzame vsako izmed možnih vrednosti

Primer: 105 dečkov na 100 deklic

Možna izida:	deček	deklica
Verjetnost:	0,512	0,488

Skupna vsota vseh verjetnosti je seveda vedno enaka 1

Še nekaj primerov

Met kocke

Možni izidi so 1, 2, 3, 4, 5 in 6, njihove verjetnosti pa

$$P(1) = P(2) = \cdots = P(6) = 1/6$$

Krvna skupina

Če imata starša oba krvno skupino AB, so verjetnosti krvne skupine pri otroku naslednje

$$P(A) = 1/4, \quad P(AB) = 1/2, \quad P(B) = 1/4$$

Še nekaj primerov

Primer: *Met dveh kock*

Možnih je 36 izidov, vsak ima verjetnost $1/36$

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Slučajna spremenljivka $Y =$ "vsota pik" lahko zavzame vrednosti $2, 3, 4, \ldots, 12$.

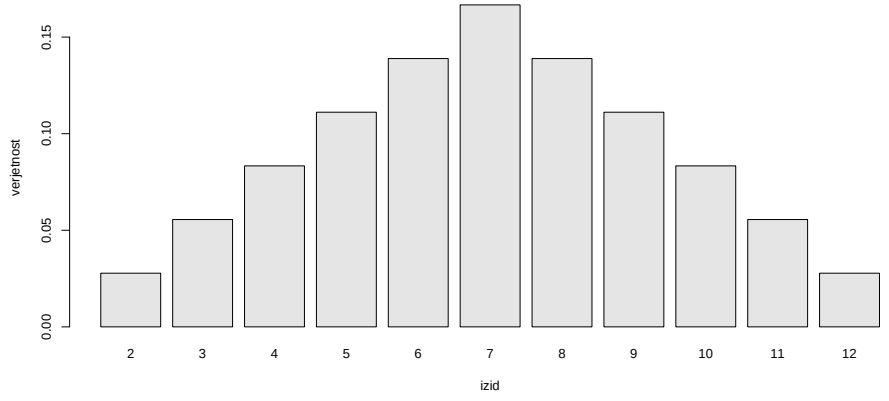
Podajanje porazdelitve diskretne spremenljivke

Primer - vsota pik na dveh kockah

Tabela

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

Stolpični diagram:



Verjetnostna porazdelitev

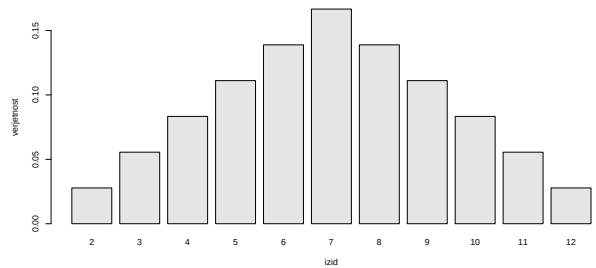
Za diskretne spremenljivke lahko navedemo verjetnosti pojava posameznih vrednosti, torej

$$p(x) = P(X = x)$$

Temu rečemo verjetnostna porazdelitev

Lastnosti

1. $p(x) > 0$
2. $\sum_{\text{vsi x}} p(x) = 1$
3. $P(a < X \leq b) = \sum_{a<x\leq b} p(x)$



Verjetnosti so pozitivne

Verjetnostna porazdelitev

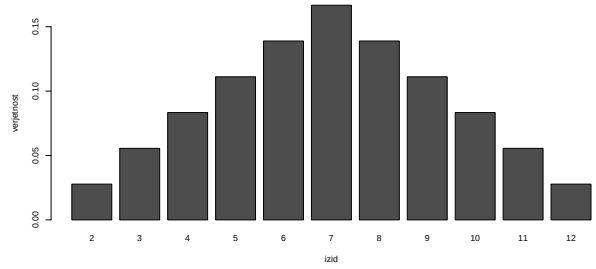
Za diskretne spremenljivke lahko navedemo verjetnosti pojava posameznih vrednosti, torej

$$p(x) = P(X = x)$$

Temu rečemo verjetnostna porazdelitev

Lastnosti

1. $p(x) > 0$
2. $\sum_{\text{vsi x}} p(x) = 1$
3. $P(a < X \leq b) = \sum_{a<x\leq b} p(x)$



Verjetnosti se seštejejo v 1

Verjetnostna porazdelitev

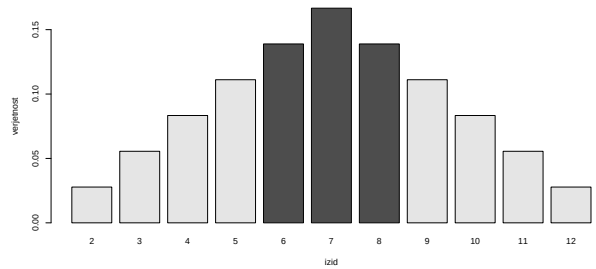
Za diskretne spremenljivke lahko navedemo verjetnosti pojava posameznih vrednosti, torej

$$p(x) = P(X = x)$$

Temu rečemo verjetnostna porazdelitev

Lastnosti

1. $p(x) > 0$
2. $\sum_{\text{vsi x}} p(x) = 1$
3. $P(a < X \leq b) = \sum_{a<x\leq b} p(x)$



Verjetnost, da je vsota med 6 in 8 je enaka 0,44

Porazdelitvena funkcija

$$F(x) = P(X \leq x)$$

Kumulativna porazdelitvena funkcija nam poda verjetnost, da je vrednost spremenljivke X manjša ali enaka neki vrednosti x.

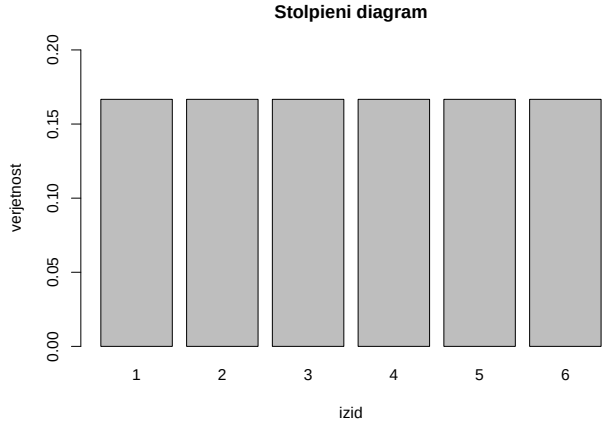
$F(x)$ je delež vrednosti X, ki so manjše od x.

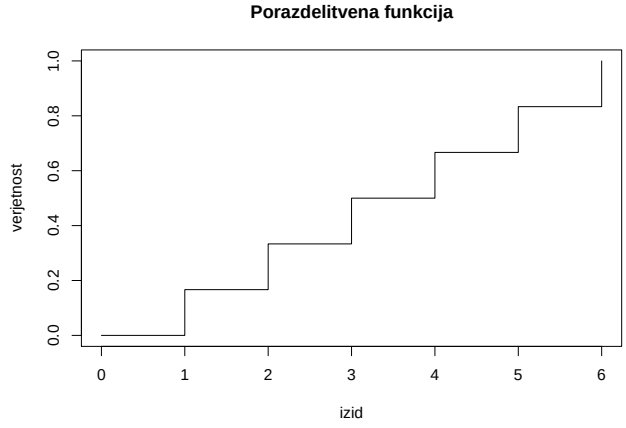
Porazdelitvena funkcija diskretne spremenljivke

Porazdelitvena funkcija pri x je definirana kot vsota verjetnosti vseh izidov, manjših ali enakih x

$$F(x) = P(X \leq x) = \sum_{y \leq x} p(x)$$

Primer - met kocke



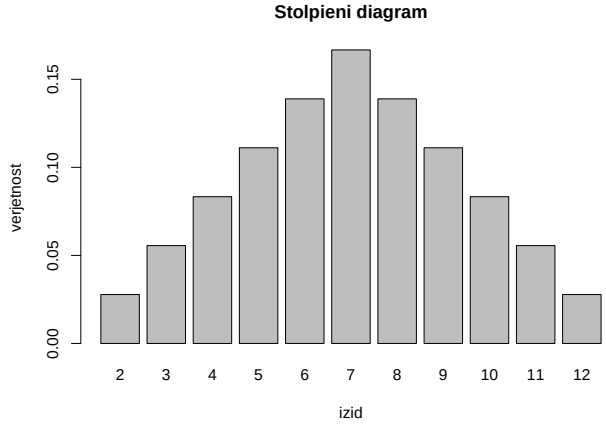


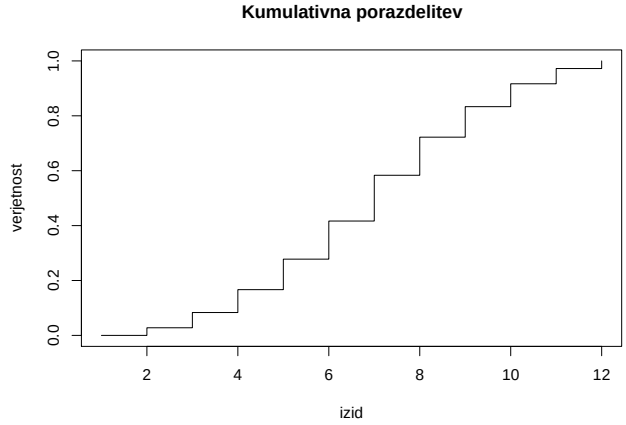
Porazdelitvena funkcija diskretne spremenljivke

Porazdelitvena funkcija pri x je definirana kot vsota verjetnosti vseh izidov, manjših ali enakih x

$$F(x) = P(X \leq x) = \sum_{y \leq x} p(x)$$

Primer - met kocke





Porazdelitev zvezne spremenljivke

Porazdelitev:

Verjetnost, s katero slučajna spremenljivka zavzame poljuben interval vrednosti

Primer: Indeks telesne teže, $\mathrm{ITT} = \frac{\text{teza}}{\text{visina}^2}$

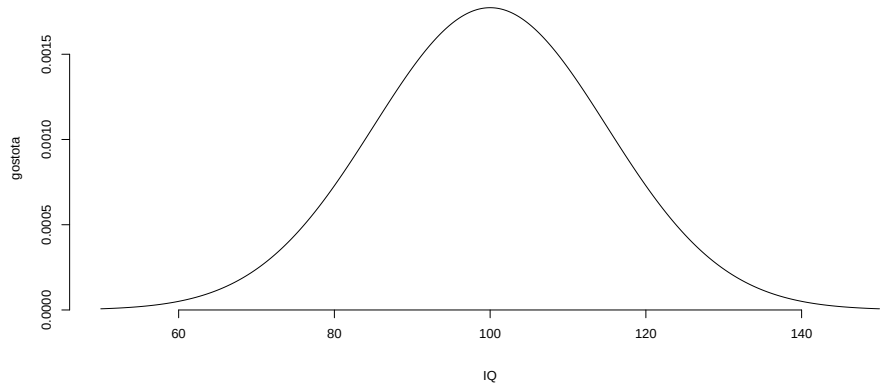
Pri odraslih je nekje med 16 in 45.
Največja je verjetnost med 20 in 40, izven gornjega intervala so vrednosti možne a malo verjetne.

Še nekaj primerov

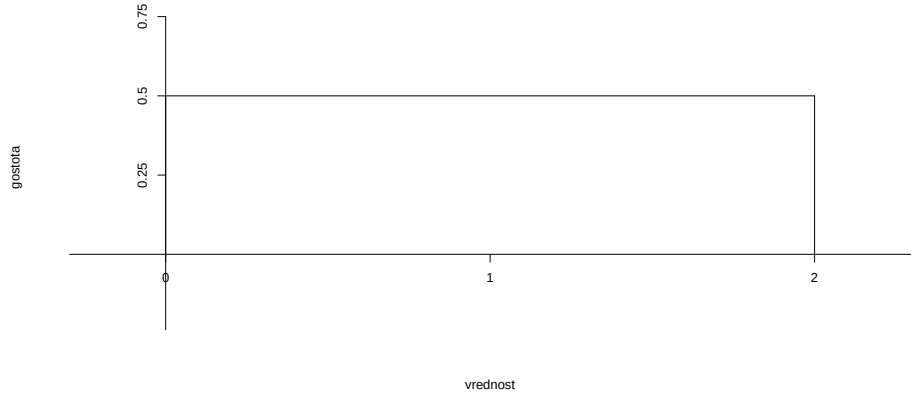
višina, krvni pritisk, laboratorijske vrednosti, delež maščob v živilu, porodna teža otroka

Podajanje porazdelitve zvezne spremenljivke

Primer - Inteligenčni količnik



Primer - Naključno realno število med 0 in 2



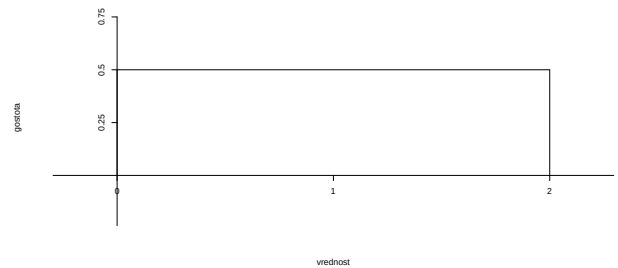
Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$



Vedno pozitivna

Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$

Celotna ploščina pod krivuljo je enaka 1

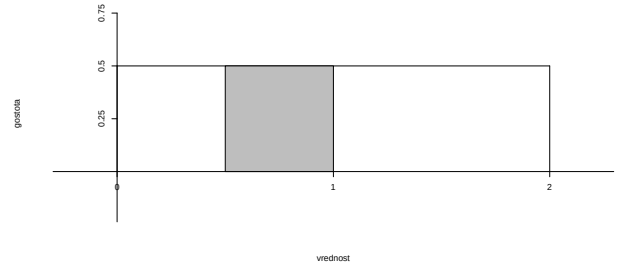
Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$



Verjetnost, da je vrednost med 0.5 in 1 je enaka $1/4$

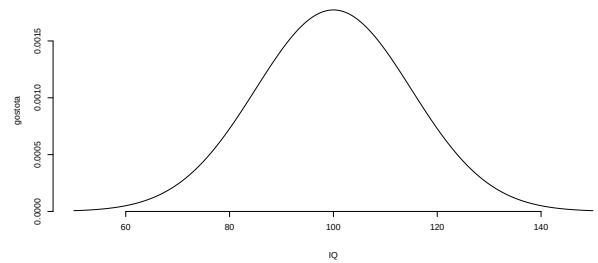
Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$



Vedno pozitivna

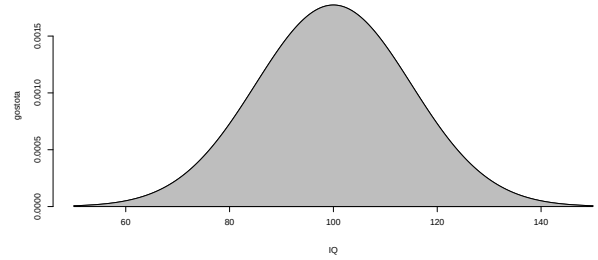
Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$



Celotna ploščina pod krivuljo je enaka 1

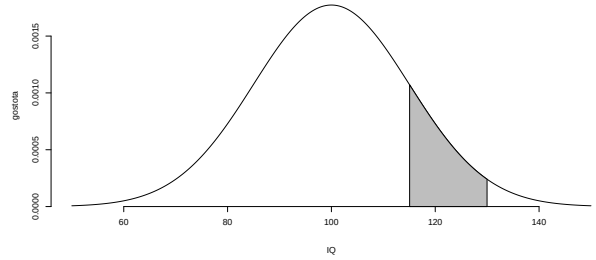
Gostota

Gostota - $f(x)$

Pove, kako goste so vrednosti okrog danega x. Verjetnost, da je vrednost spremenljivke na intervalčku širine Δx okoli vrednosti x je približno $f(x)\Delta x$

Lastnosti

1. $f(x) > 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X \leq b) = \int_{a}^{b} f(x)dx$



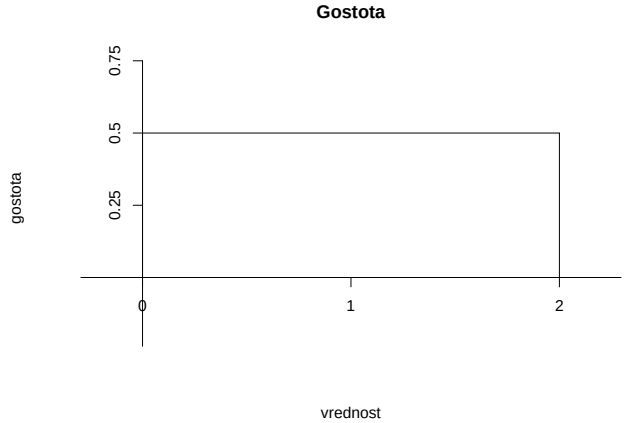
Delež ljudi z vrednostjo med 115 in 130 je 0,14

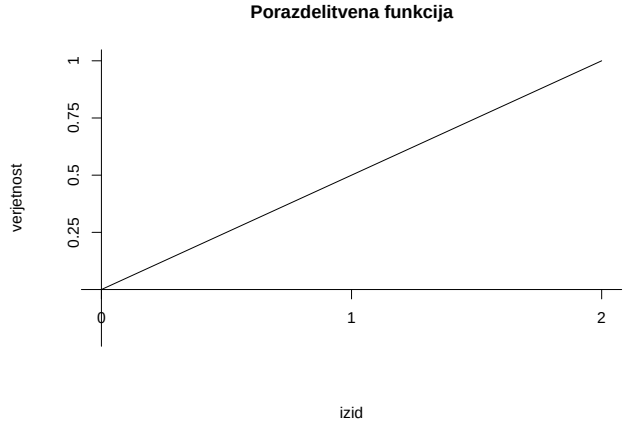
Porazdelitvena funkcija

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

Kumulativna porazdelitvena funkcija nam poda verjetnost, da je vrednost spremenljivke X manjša ali enaka neki vrednosti x
$F(x)$ je delež vrednosti X, ki so manjše od x

Primer



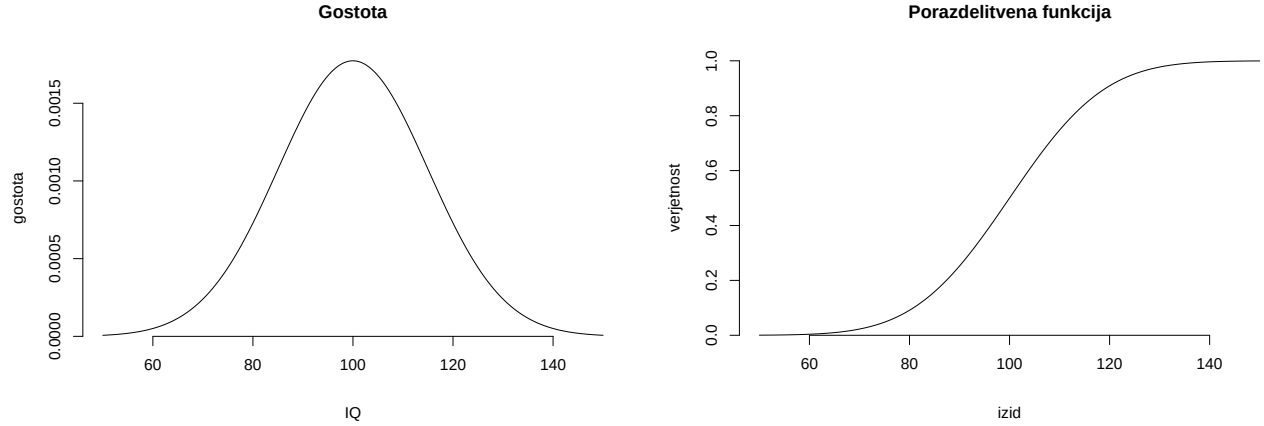


Porazdelitvena funkcija

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

Kumulativna porazdelitvena funkcija nam poda verjetnost, da je vrednost spremenljivke X manjša ali enaka neki vrednosti x
$F(x)$ je delež vrednosti X, ki so manjše od x

Primer



Opisovanje empirične porazdelitve

Teoretična porazdelitev - opisuje variabilnost populacije
Empirična porazdelitev - opisuje variabilnost vzorca

Opisovanje

Porazdelitvena funkcija, verjetnostna funkcija in gostota so teoretične funkcije, ki opisujejo variabilnost v populaciji.
Za opisovanje variabilnosti na vzorcih uporabljamo analogno definirane empirične funkcije.

Opisovanje empirične porazdelitve - diskretna spremenljivka

Primer - porazdelitev spola pri novorojenčkih

Možna izida:	deček	deklica
Verjetnost:	0,512	0,488
Leto 2006:	9762	9170
Delež:	0,515	0,485

Opisovanje

Porazdelitvena funkcija, verjetnostna funkcija in gostota so teoretične funkcije, ki opisujejo variabilnost v populaciji.
Za opisovanje variabilnosti na vzorcih uporabljamo analogno definirane empirične funkcije.

Opisovanje empirične porazdelitve - diskretna spremenljivka

Tabela vrednosti (frekvenčna tabela)

Tabela, v kateri so podane možne vrednosti in njihova pogostost

Stolpični diagram

Graf frekvenc, ali relativnih frekvenc za vsako vrednost spremenljivke

Empirična porazdelitvena funkcija

Graf kumulativne relativne frekvence

Opisovanje empirične porazdelitve - diskretna spremenljivka - Primer: met kocke

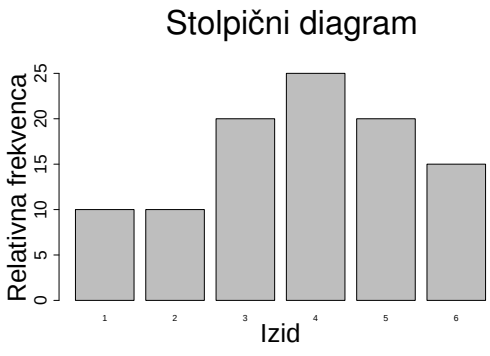
Kocko vržemo 20-krat

Met	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Izid	1	3	3	6	5	5	4	1	2	3	4	5	4	3	4	6	6	2	4	5

Tabela vrednosti (frekvenčna tabela)

Izid	1	2	3	4	5	6
Relativna frekvenca	10%	10%	20%	25%	20%	15%

Grafični prikazi



Opisovanje empirične porazdelitve - zvezna spremenljivka

Tabela vrednosti (frekvenčna tabela)

Različnih vrednosti je hitro preveč, potrebno je združevanje v razrede

Histogram

V nasprotju s stolpičnim diagramom ne prikazujemo pogostosti posameznih vrednosti temveč celoten razpon razdelimo na enako široke intervale.

Empirična porazdelitvena funkcija

Graf kumulativne relativne frekvence - to ostaja smiseln prikaz, je pa nekoliko težje berljiv

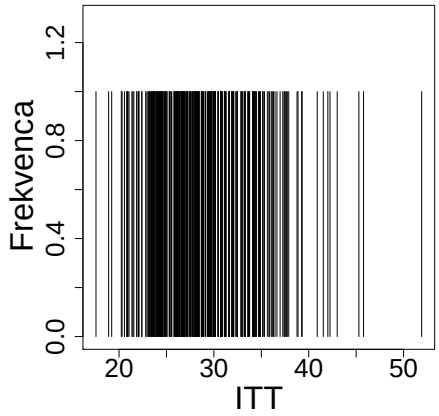
Opisovanje empirične porazdelitve - zvezna spremenljivka

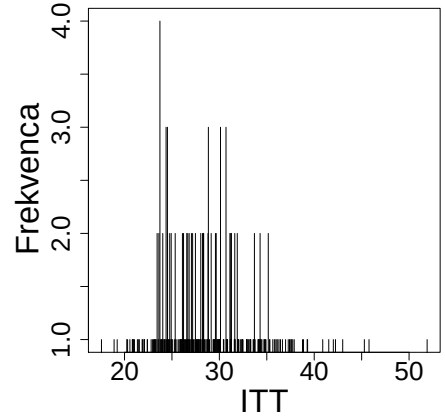
Primer: ITT za ženske med 40 in 80 letom

Vzorec 617 žensk

Stolpični diagram

Vsaka vrednost svoj stolpec





Slika odvisna od natančnosti merjenja, združevanje je nujno!

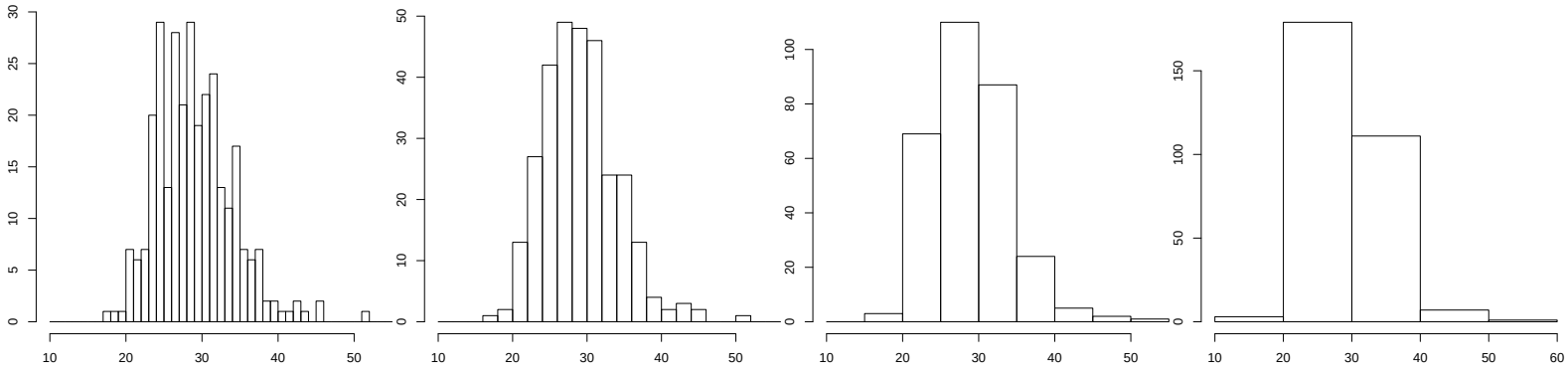
Opisovanje empirične porazdelitve - zvezna spremenljivka

Primer: ITT za ženske med 40 in 80 letom

Vzorec 617 žensk

Histogram

Ploščina pravokotnika nad intervalom sorazmerna (relativni) frekvenci.



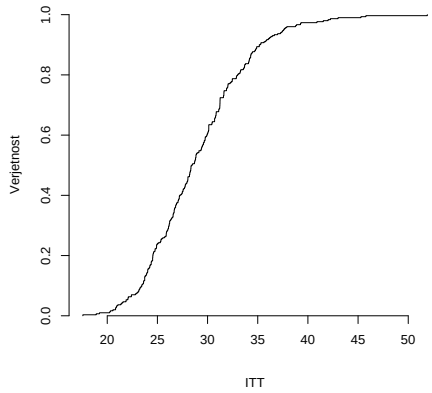
Intervali široki 1, 2, 5, 10 - širina intervala vpliva na izgled.

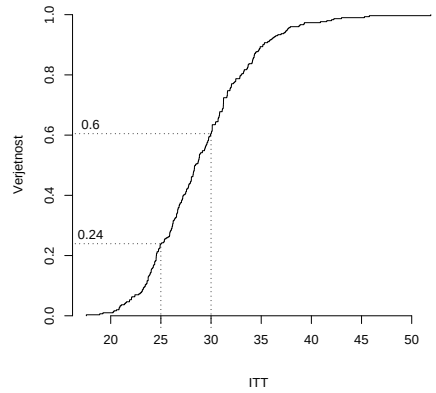
Opisovanje empirične porazdelitve - zvezna spremenljivka

Primer: ITT za ženske med 40 in 80 letom

Vzorec 617 žensk

Empirična porazdelitvena funkcija





Nekoliko težje berljivo, odčitamo lahko poljuben interval

Mere središčnosti in variabilnosti

Porazdelitvena funkcija in gostota dajeta popolno informacijo, a pogosto jo želimo strniti

Mere središčnosti
- povprečje
- mediana
- modus

Mere razpršenosti
- varianca (std. odklon)
- interkvartilni razmik
- razpon

Mere središčnosti - primer

Dobite vabilo za službo v uspešnem slovenskem podjetju, kjer je povprečna plača 2000 EUR. Sprejmete?

Delavec	650 EUR
Strokovni sodelavec	825 EUR
Vodja oddelka	1000 EUR
Upravni odbor	12000 EUR
Direktor	20000 EUR

Mere središčnosti - vzorčno povprečje

Aritmetična sredina - oznaka $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n}[x_i \cdot \frac{1}{n}]$$

V kakšnem smislu je $\bar{x}$ sredina?

Ker je $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, sta vsoti levih in desnih odmikov enaki!

Mere središčnosti - populacijsko povprečje

Pričakovana vrednost - oznaka $E(X)$ ali μ

Diskretna porazdelitev:

$$E(X) = \mu = \sum_x [x \cdot p(x)]$$

Primer - met kocke:

$$p(1) = p(2) = \cdots = p(6) = \frac{1}{6}$$

$$\begin{aligned} E(X) &= \sum_{x=1}^{6} [x \cdot p(x)] \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3{,}5 \end{aligned}$$

Zvezna porazdelitev

$$E(X) = \mu = \int x f(x) dx$$

Mere središčnosti - mediana (Me)

Na vzorcu:

srednja vrednost glede na range (pri sodem n aritmetična sredina srednjega para).
Primeri:

$1,\ 7,\ 2,\ 4,\ 9 \quad \rightarrow \quad 1,\ 2,\ \ 4\ ,\ 7,\ 9$

$1,\ 7,\ 2,\ 4,\ 900 \quad \rightarrow \quad 1,\ 2,\ \ 4\ ,\ 7,\ 900$

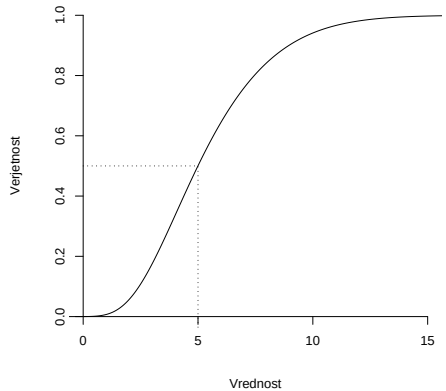
$1,\ 7,\ 2,\ 4,\ 9,\ 10 \rightarrow \quad 1,\ 2,\ 4,\ 7,\ 9,\ 10$

$5{,}5$

Drugi kvantili: podobno definiramo: 25. percentil (= 1. kvartil) je tista vrednost, od katere je 25% vrednosti manjših, 75% pa večjih.

Mere središčnosti - populacijska mediana (Me)

- Tista vrednost, za katero velja $F(x) = 0{,}5$



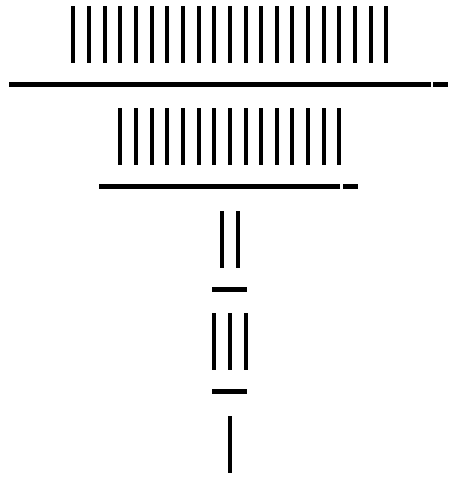
- Podobno definiramo druge kvantile
- Za diskretne porazdelitve je stvar nekoliko nerodna

Mere središčnosti - modus (Mo)

- Najpogostejša vrednost
- Smiselno definirati le za diskretne spremenljivke z malo različnimi vrednostmi

Mere središčnosti - primer

Dobite vabilo za službo v uspešnem slovenskem podjetju, kjer je povprečna plača 2000 EUR. Sprejmete?



Delavec	19	650 EUR
Strokovni sodelavec	14	825 EUR
Vodja oddelka	2	1000 EUR
Upravni odbor	3	12000 EUR
Direktor	1	20000 EUR

Povprečje = 2000 EUR mediana = 737,5 EUR modus= 650 EUR

Mere razpršenosti - vzorčna varianca in std. odklon

Varianca - oznaka s^2

povprečen kvadriran odmik od aritmetične sredine

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

Uporabljamo v teoriji

Primer: šolska ocena:

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	3	2	4
3	3	0	0
1	3	-2	4
4	3	1	1
4	3	1	1
1	3	-2	4
		$(\sum_i = 0)$	$\sum_i = 14$

$$s^2 = \tfrac{1}{5}14 = 2{,}8$$

Standardni odklon - oznaka s (včasih tudi $\hat{\sigma}$)

$$s = \sqrt{s^2}$$

Uporabljamo pri interpretaciji (v enakih enotah kot spremenljivka).

Primer:

$$s = \sqrt{2{,}8} = 1{,}67$$

Mere variabilnosti - populac. varianca in std. odklon

Varianca - oznaka σ^2

Pričakovana vrednost kvadriranega odmika od populacijskega povprečja

$$\text{Diskretna porazdelitev} \quad \text{Var(X)} = \sigma^2 = \sum_x [x - E(X)]^2 \cdot p(x)$$

$$\text{Zvezna porazdelitev} \quad \text{Var(X)} = \sigma^2 = \int [x - E(X)]^2 f(x)$$

Primer - met kocke:

$$Var(X) = (1 - 3{,}5)^2 \cdot \frac{1}{6} + (2 - 3{,}5)^2 \cdot \frac{1}{6} + \cdots + (6 - 3{,}5)^2 \cdot \frac{1}{6} = 2{,}9167.$$

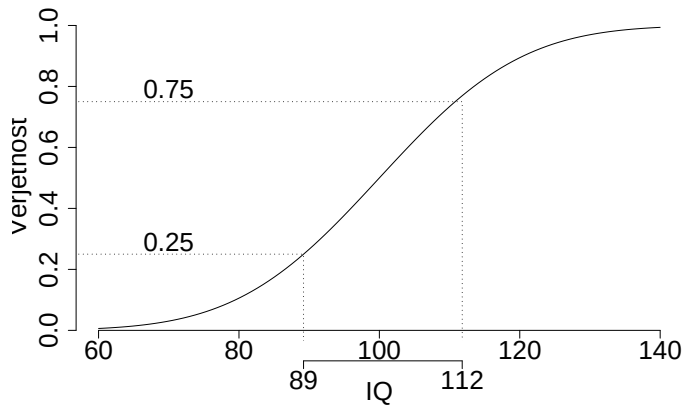
Standardni odklon - $\sigma = \sqrt{\sigma^2}$

Primer: $s = \sqrt{2{,}9167} = 1{,}7078$

Mere variabilnosti - interkvartilni razmik

Razlika med 3. in 1. kvartilom porazdelitve.

Primer - Porazdelitev IQ



Interkvartilni razmik je 23

Primer - plače

1. kvartil = 650 EUR
3. kvartil = 825 EUR

Srednjih 50 % plač se razlikuje za 175 EUR

Za primerjavo

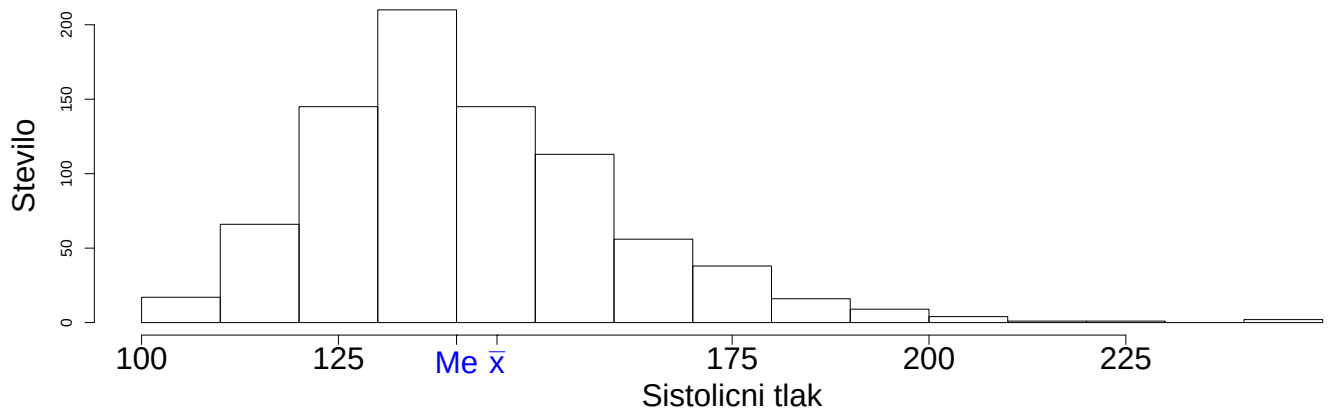
std. odklon = 4087 EUR

Interkvartilni razmik je mera razpršenosti, na katero ekstremne vrednosti ne vplivajo

Povprečje vs. mediana

- Mediana razdeli vse podatke na dva dela, v vsakem je 50% podatkov.
- Aritmetična sredina je teoretično enaka mediani, kadar je porazdelitev simetrična.
- Kaj lahko rečemo v primeru asimetrične porazdelitve?
- Bi raje doživeli povprečno pričakovano starost ali mediano pričakovane starosti?
- Kaj lahko sklepamo, če je standardni odklon večji od povprečja, negativne vrednosti spremenljivke pa niso možne?

Sistolični krvni tlak bolnikov s periferno arterijsko boleznijo



$$\bar{x} = 145$$

$$Me = 140$$

Povprečje vs. mediana

- Razlike lahko povzročijo tudi tujki, to so nenavadne vrednosti daleč od ‘sredine’. Te vplivajo na aritmetično sredino, ne pa na mediano.
- Mediano bomo rajši uporabljali pri nesimetričnih porazdelitvah, če ne bomo posebej želeli, da mera središčnosti upošteva dejanske vrednosti spremenljivke.
- Za izračun mediane moramo poznati le prvih 50% vrednosti (analiza preživetja)
- Mediana ne upošteva dejanskih vrednosti spremenljivke (razen pri rangiranju)
- Mediana je nerodna za računanje. (npr. mediana dveh združenih vzorcev se ne da izraziti z medianama posameznih vzorcev)

Lastnosti pričakovane vrednosti in variance

Prištevanje/odštevanje konstante

Primer - merjenje temperature:

Pod pazduho: $\bar{x} = 37{,}1\,℃$, $s = 0{,}5\,℃$, $s^2 = 0{,}25\,℃^2$

V ustih - vsaka meritev je za 0,3 stopinje višja: $\bar{x}, s, s^2$?

Pravilo:

$$E(X + c) = E(X) + c,\ Var(X + c) = Var(X)$$

$$E(X) = \sum_x [x \cdot p(x)] \qquad Var(X) = \sum_x [x - E(X)]^2 \cdot p(x)$$

$$E(X) = \int x f(x) dx \qquad Var(X) = \int [x - E(X)]^2 f(x) dx$$

Lastnosti pričakovane vrednosti in variance

Množenje/deljenje s konstanto

Primer - merjenje obsega pasu:

$\bar{x} = 0{,}995m$, $s = 0{,}111m$, $s^2 = 0{,}0123m^2$

Kakšne vrednosti bi dobili, če bi meritve izrazili v cm?

Pravilo:

$E(cX) = cE(X)$, $Var(cX) = c^2 Var(X)$

$E(X) = \sum_x [x \cdot p(x)]$ $\qquad$ $Var(X) = \sum_x [x - E(X)]^2 \cdot p(x)$

$E(X) = \int xf(x)dx$ $\qquad$ $Var(X) = \int [x - E(X)]^2 f(x)dx$

Lastnosti pričakovane vrednosti in variance

Seštevanje neodvisnih slučajnih spremenljvk

Primer - met kocke:

$\bar{x} = 3{,}5$, $s = 1{,}7$, $s^2 = 2{,}9$

Kakšne vrednosti dobimo pri metu dveh kock?

Pravilo:

$$E(X_1 + X_2) = E(X_1) + E(X_2),$$
$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$$

$E(X) = \sum_x [x \cdot p(x)]$ $\qquad$ $Var(X) = \sum_x [x - E(X)]^2 \cdot p(x)$

$E(X) = \int x f(x) dx$ $\qquad$ $Var(X) = \int [x - E(X)]^2 f(x) dx$

Lastnosti pričakovane vrednosti in variance

Odštevanje neodvisnih slučajnih spremenljvk

Primer - met kocke:

$\bar{x} = 3{,}5$, $s = 1{,}7$, $s^2 = 2{,}9$

Kakšne vrednosti dobimo, če nas zanima razlika dveh metov?

Pravilo:

$$E(X_1 - X_2) = E(X_1) - E(X_2),$$
$$Var(X_1 - X_2) = Var(X_1) + Var(X_2)$$

Opomba - to bi lahko izračunali z uporabo že znanih formul

$$E(X) = \sum_x [x \cdot p(x)] \qquad Var(X) = \sum_x [x - E(X)]^2 \cdot p(x)$$

$$E(X) = \int x f(x) dx \qquad Var(X) = \int [x - E(X)]^2 f(x) dx$$

Lastnosti pričakovane vrednosti in variance

Seštevanje/odštevanje odvisnih slučajnih spremenljvk

Primer - X=pulz pred obremenitvijo, Y=pulz po obremenitvi:

$\bar{x} = 70$, $s_x = 15$, $\bar{y} = 140$, $s_x = 20$

Kaj lahko rečemo o razliki $Y - X$?

Pravilo:

$E(X_1 \pm X_2) = E(X_1) \pm E(X_2)$
varianca je odvisna od povezanosti spremenljivk

$E(X) = \sum_x [x \cdot p(x)]$ $\quad$ $Var(X) = \sum_x [x - E(X)]^2 \cdot p(x)$

$E(X) = \int xf(x)dx$ $\quad$ $Var(X) = \int [x - E(X)]^2 f(x)dx$

Lastnosti pričakovane vrednosti in variance

Slučajni vzorec - teorija

- Če so slučajne spremenljivke $X_1, X_2, \ldots, X_n$ paroma neodvisne in identično porazdeljene (iid), govorimo o **slučajnem vzorcu**.
- **Vzorčno povprečje** $\bar{X} = (\sum X_i)/n$ je seveda spet slučajna spremenljivka.

Slučajni vzorec - primer

- Populacija moških med 40 in 80 letom ima v povprečju krvni pritisk enak 139 mmHg.
- Vzorec, $n = 20$, izračunamo $\bar{x}$.
- Bo povprečje na vzorcu enako populacijskemu?
- Vzorčno povprečje je slučajna spremenljivka, ima neko pričakovano vrednost in varianco.

Lastnosti pričakovane vrednosti in variance

Pričakovana vrednost vzorčnega povprečja

$$
\begin{aligned}
E(\bar{X}) &= E(\frac{1}{n} \cdot [X_1 + X_2 + \cdots + X_n]) \\
&= \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n} \cdot [E(X_1) + E(X_2) + \cdots + E(X_n)] \\
&= \frac{1}{n} \cdot [\mu + \mu + \cdots + \mu] = \mu
\end{aligned}
$$

Lastnosti pričakovane vrednosti in variance

Varianca vzorčnega povprečja

$$
\begin{aligned}
Var(\bar{X}) &= Var(\frac{1}{n} \cdot [X_1 + X_2 + \cdots + X_n]) \\
&= \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n^2} \cdot [Var(X_1) + Var(X_2) + \cdots + Var(X_n)] \\
&= \frac{1}{n^2} \cdot [\sigma^2 + \sigma^2 + \cdots + \sigma^2] = \frac{\sigma^2}{n}
\end{aligned}
$$

$$
sd(\bar{X}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}
$$

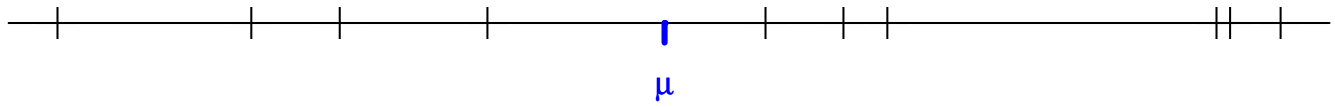
Standardnemu odklonu vzorčnega povprečja rečemo **standardna napaka**.

Ocena variance na vzorcu

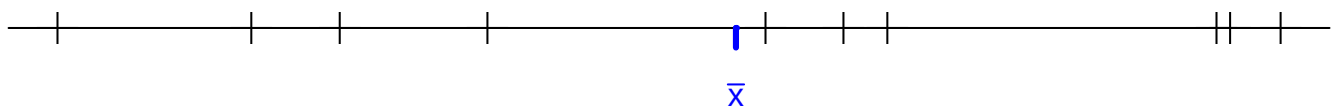
Populacijsko varianco na vzorcu ocenimo z

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

Zakaj ne n ?
Oceno bi izboljšali, če bi namesto $\bar{x}$ uporabili 'pravo'vrednost, torej μ.



Odmiki od $\bar{x}$ so nekoliko manjši od odmikov od μ.



Torej bo varianca izračunana s pomočjo $\bar{x}$ nekoliko manjša.
Delimo z manj $\rightarrow$ povečamo

Risanje grafov

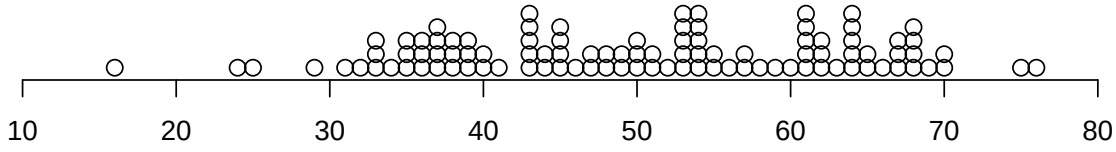
“... drawing graphs, like motor-car driving and love-making, is one of those activities which almost every researcher thinks he or she can do well without instruction.”

Wainer & Thissen, 1991 Annual Review of Psychology

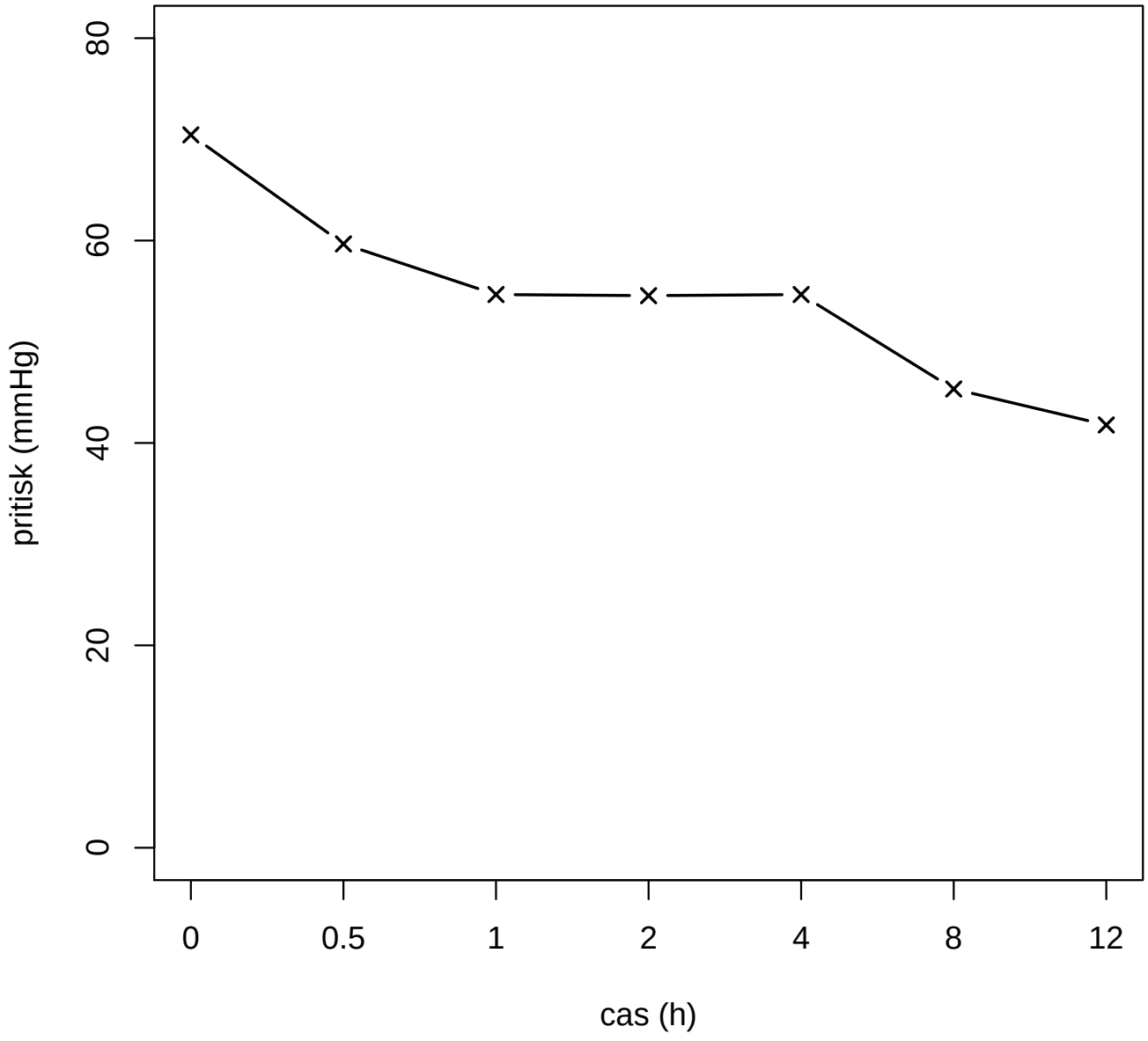
Nekateri principi konstrukcije grafov

1. Izločite nepotrebne dimenzije
2. Oznake merilne lestvice (“tick marks”) naj kažejo navzven
3. Razmislite o vključevanju ničle v graf (včasih dobro, včasih ne)
4. Zaznavanje relativnih razdalj je najbolj natančno - ploščine so težje
5. Izogibajte se odvečnosti - mislite na razmerje črnilo : informacija
6. Koristni grafi so lahko zahtevni - ne nujno preprosti in takoj dojemljivi

Točkovni diagram



Linijski diagram

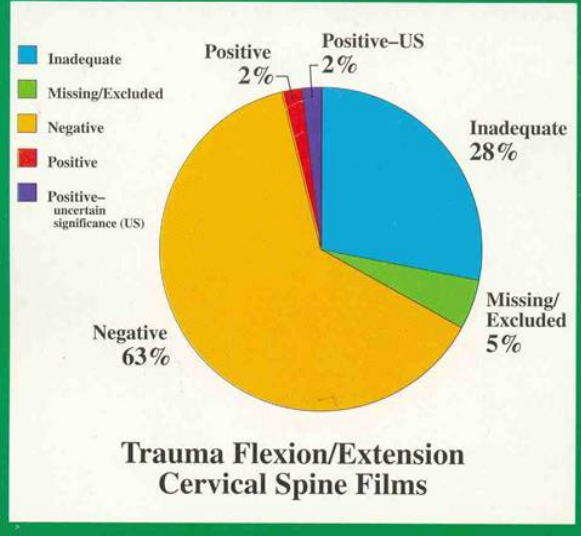


Volume 52 ▪ Number 1 ▪ January 2002

LIPPINCOTT
WILLIAMS
& WILKINS

The Journal of TRAUMA®

Injury, Infection, and Critical Care



Trauma Flexion/Extension Cervical Spine Films

American Association for the Surgery of Trauma
Eastern Association for the Surgery of Trauma
Trauma Association of Canada/L'Association Canadienne de Traumatologie
Western Trauma Association

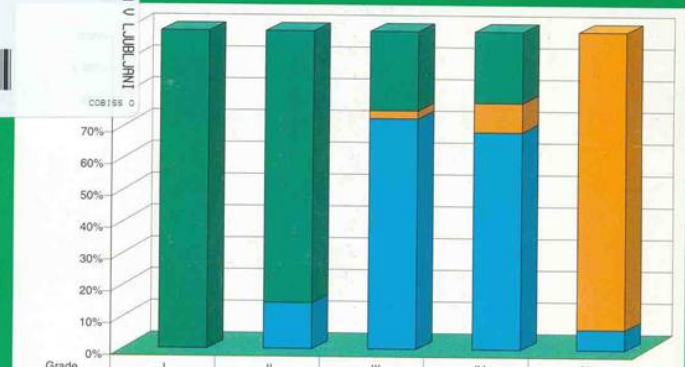
www.jtrauma.com

Volume 50 ▪ Number 2 ▪ February 2001

LIPPINCOTT WILLIAMS & WILKINS

The Journal of TRAUMA®

Injury, Infection, and Critical Care



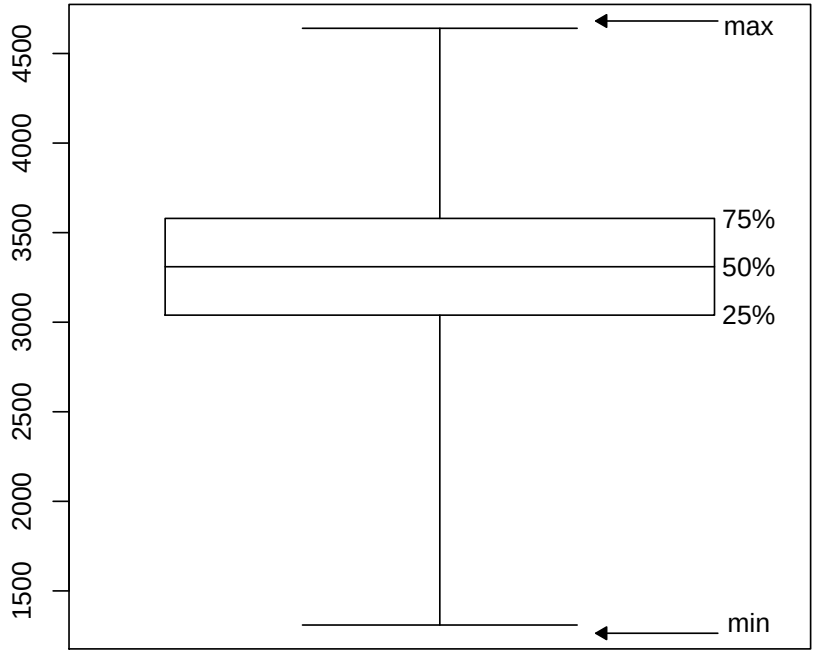
Grade	I	II	III	IV	V
Bedrest	100%	85%	24%	22%	0%
Nephrectomy	0%	0%	3%	9%	86%
Repair	0%	15%	73%	69%	7%

American Association for the Surgery of Trauma
Eastern Association for the Surgery of Trauma
Trauma Association of Canada/L'Association Canadienne de Traumatologie
Western Trauma Association

www.jtrauma.com

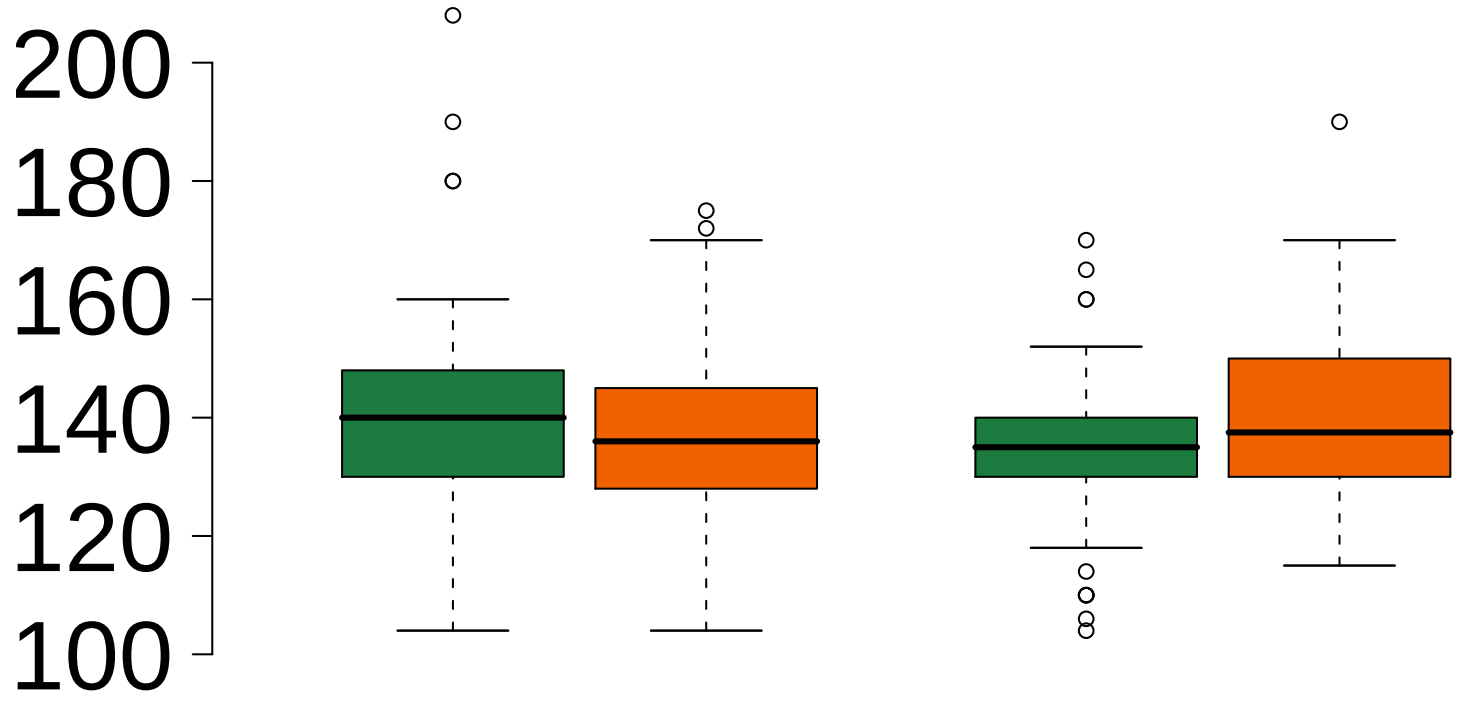
Graf kvantilov (box and whiskers plot)

- srednja črta je mediana
- pravokotnik prikazuje razpon od prvega do tretjega kvantila - zajema srednjih 50 % podatkov
- daljice segajo do minimuma oz. maximuma (točke, ki so oddaljene za več kot 1,5x IQR so pogosto izvzete)



Graf kvantilov

Graf primeren za primerjavo večin skupin:
Primer: primerjava dveh skupin (bolniki, kontrola) ob dveh obiskih



Porazdelitev dihotomne spremenljivke 0/1 (Bernoullijeva porazdelitev)

Primer - met kocke

- $X =$ število šestic - spremenljivka z dvema vrednostima
- $P(X = 1) = \frac{1}{6} = \pi$

Pričakovana vrednost

- $E(X) = \sum_x [x \cdot p(x)] = 0 \cdot \frac{5}{6} + 1 \cdot \frac{1}{6} = \frac{1}{6}$
- V splošnem $E(X) = P(X = 1) = \pi$

Varianca

- $Var(X) = E\left(X - E(X)\right)^2 = E\left(X - \frac{1}{6}\right)^2 =$
  $(0 - \frac{1}{6})^2 \cdot \frac{5}{6} + (1 - \frac{1}{6})^2 \cdot \frac{1}{6} = \frac{5}{6} \cdot \frac{1}{6}$
- V splošnem $Var(X) = P(X = 1)P(X = 0) = \pi(1 - \pi)$

Spremenljivko opišemo z enim samim parametrom: $X \sim B(\pi)$

Binomska porazdelitev

- zanima nas opisna spremenljivka z dvema vrednostima (dihotomka)
- opazujemo jo na n med seboj neodvisnih enotah
- verjetnost dogodka A je pri vsaki enoti enaka $P(A) = \pi$
- Zanima nas slučajna spremenljivka X: število dogodkov A pri n enotah

Primer: met 5 kock

- $A = \{\text{pade šestica}\}$
- $n = 5$
- $P(A) = \pi = \frac{1}{6}$
- $X =$število šestic v 5 metih

Primer - met 5 kock

Kolikšna je verjetnost, da padeta dve šestici

- $P(A) = \frac{1}{6}$, $P(B) = \frac{5}{6}$
- Izid ABBAB ima verjetnost

A	B	B	A	B	ABBAB
$\frac{1}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{6}$	$\frac{5}{6}$	$(\frac{1}{6})^2(\frac{5}{6})^3$

- V n metih bo prvih k izidov šestica z verjetnostjo: $(\frac{1}{6})^k(\frac{5}{6})^{n-k}$
- Na n enotah bo prvih k izidov dogodek A: $\pi^k(1-\pi)^{n-k}$

Primer - met 5 kock, nadaljevanje

Kolikšna je verjetnost, da padeta dve šestici

- Upoštevati je potrebno še različne vrstne rede - 2 šestici lahko padeta na 10 načinov:

  AABBB, ABABB, ABBAB, ABBBA, BAABB, BABAB, BABBA, BBAAB, BBABA, BBBAA

  Verjetnost dveh šestic je torej

$$P(X = 2) = 10(\frac{1}{6})^2(\frac{5}{6})^3 = 0{,}16$$

- Število vrstnih redov izračunamo z binomskim simbolom:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

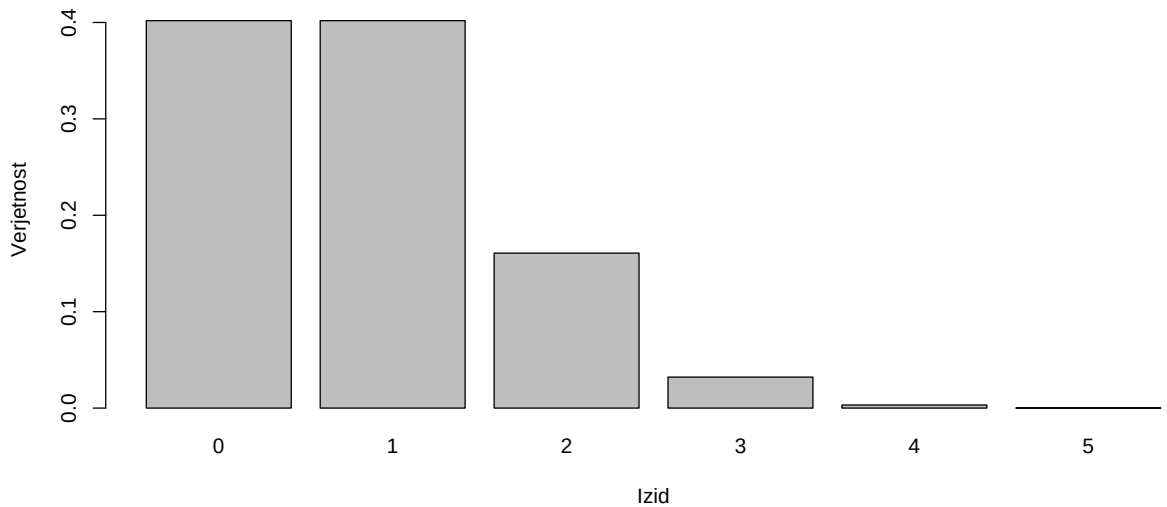
- V našem primeru torej

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2} = 10$$

Primer - met 5 kock, nadaljevanje

Verjetnosti

k	0	1	2	3	4	5
$P(X = k)$	0,402	0,402	0,161	0,032	0,003	0,0001



Binomska porazdelitev

Slučajna spremenljivka X je binomsko porazdeljena, če:

- zanima nas opisna spremenljivka z dvema vrednostima (dihotomka)
- opazujemo jo na n med seboj neodvisnih enotah
- verjetnost dogodka A je pri vsaki enoti enaka $P(A) = \pi$
- Zanima nas slučajna spremenljivka X: število dogodkov A pri n enotah

Verjetnost poljubnega izida izračunamo kot

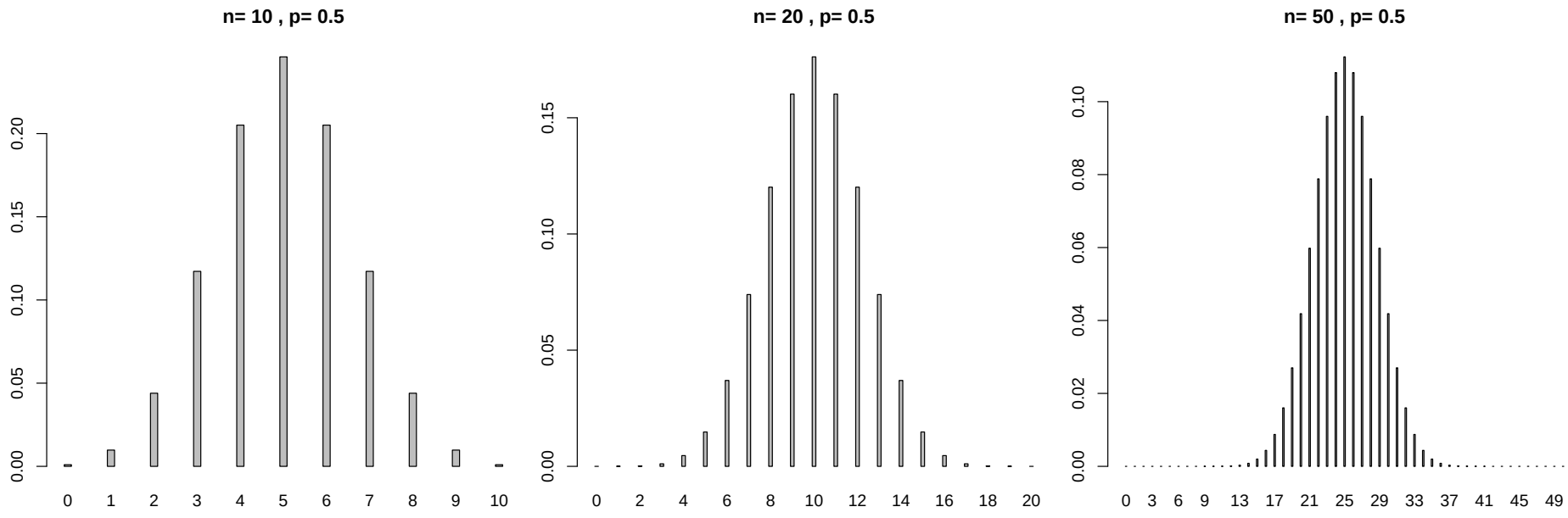
$$P(X = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \quad \text{za k} = 0,1,2,\ldots,\text{n}$$

Porazdelitev je določena z 2 parametroma $X \sim Bin(n,\pi)$

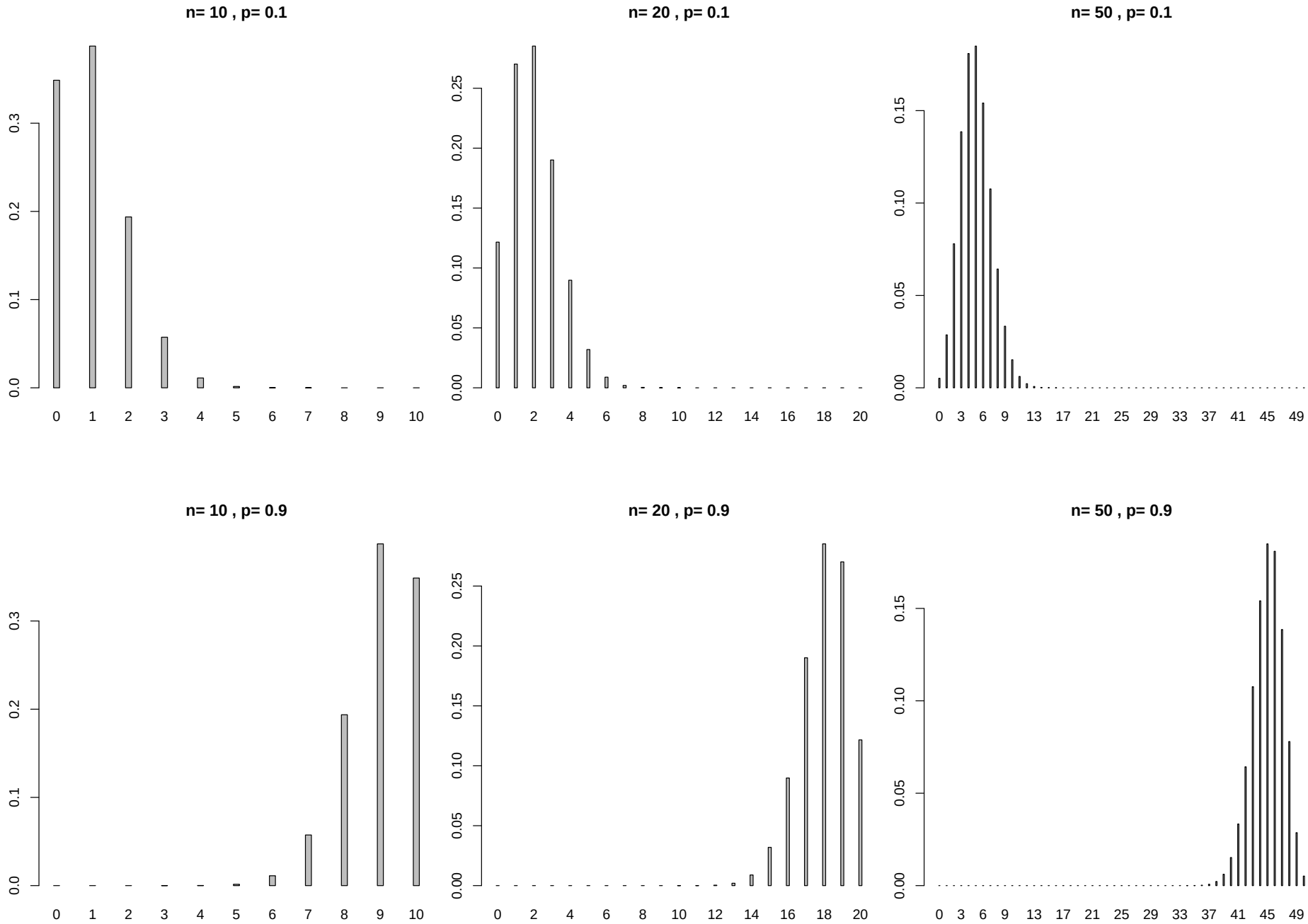
Če poznamo vrednost n in π, lahko izračunamo verjetnost poljubnega dogodka

Primeri binomsko porazdeljenih spremenljivk

Binomska porazdelitev je primer diskretne porazdelitev - prikažemo jo npr. s stolpičnim diagramom.



Primeri binomsko porazdeljenih spremenljivk



Pričakovana vrednost in varianca

Primer - met kocke

- Naj bo Y dihotomna slučajna spremenljivka, $Y =$ število šestic v enem metu, $P(Y = 1) = \frac{1}{6}$
- Vemo že $E(Y) = \frac{1}{6}$, $Var(Y) = \frac{1}{6}\frac{5}{6}$
- $X = \sum_{i=1}^{5} Y_i$
- $E(X) = \sum_{i=1}^{5} E(Y_i) = 5 \cdot E(Y_i) = \frac{5}{6}$
- $Var(X) = \sum_{i=1}^{5} Var(Y_i) = 5\frac{1}{6}\frac{5}{6}$

Formule

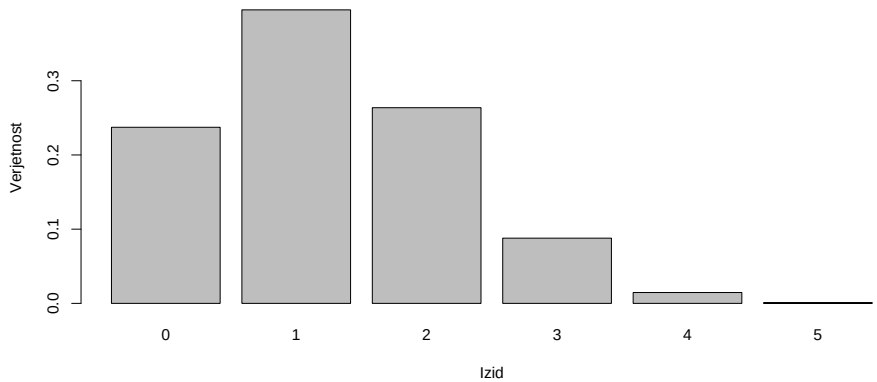
- $E(X) = n\pi$
- $Var(X) = n\pi(1 - \pi)$

Test

5 vprašanj, vsako ima 4 možne odgovore

Pravilna rešitev je ADDCA

Teoretična porazdelitev odgovorov



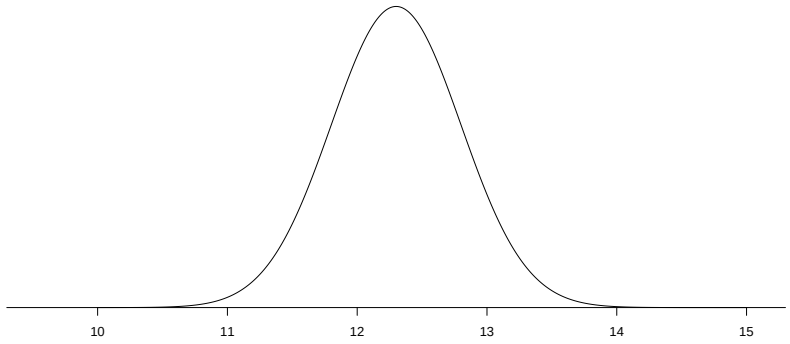
$E(X) = \frac{5}{4}$
$SD = 0{,}96$
$P(X = 5) = 0{,}00098.$

Verjetnost, da ima v skupini 50 ljudi nekdo pravilnih vseh 5 odgovorov je $1 - (1 - 0{,}00098)^{50} = 0{,}047$

primer, formula

Normalna porazdelitev

- zvezna porazdelitev
- simetrična (povprečje je enako mediani)
- vrednosti blizu povprečja so bolj verjetne
- verjetnost pada z oddaljenostjo od povprečja, možne vrednosti so vsa realna števila

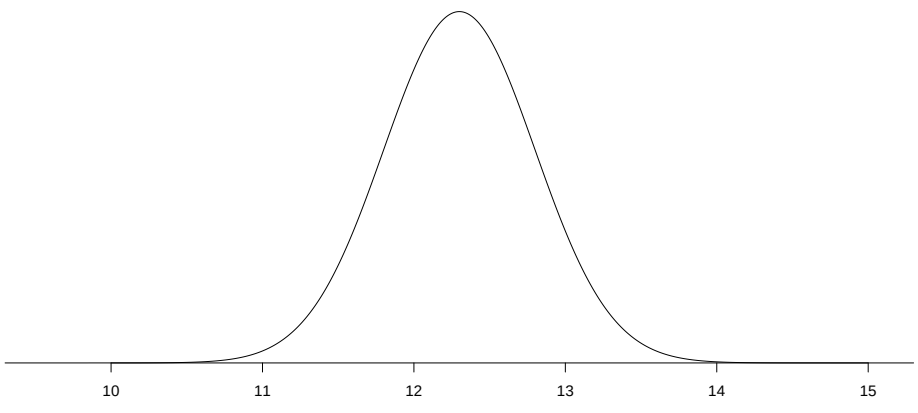


- ključna za statistično teorijo
- v praksi pogosto predpostavimo, da je porazdelitev neke spremenljivke približno normalna

Normalna porazdelitev - primer

Primer: napake pri merjenju

- Merimo dolžino hodnika, prava vrednost je 12,3 m
- Meritve bodo bolj ali manj natančne, verjetnost napake je enaka v $+$ ali $-$ smer
- Bolj verjetno je, da bomo zgrešili manj
- V povprečju bomo dobili pravo vrednost

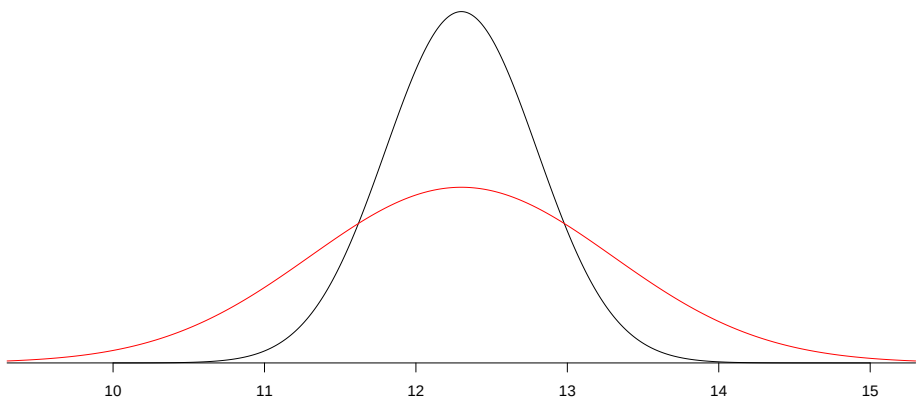


šiviljski meter
$\mu = 12{,}3m$
$\sigma = 0{,}5m$
ravnilo
$\mu = 12{,}3m$
$\sigma = 1m$

Normalna porazdelitev - primer

Primer: napake pri merjenju

- Merimo dolžino hodnika, prava vrednost je 12,3 m
- Meritve bodo bolj ali manj natančne, verjetnost napake je enaka v $+$ ali $-$ smer
- Bolj verjetno je, da bomo zgrešili manj
- V povprečju bomo dobili pravo vrednost



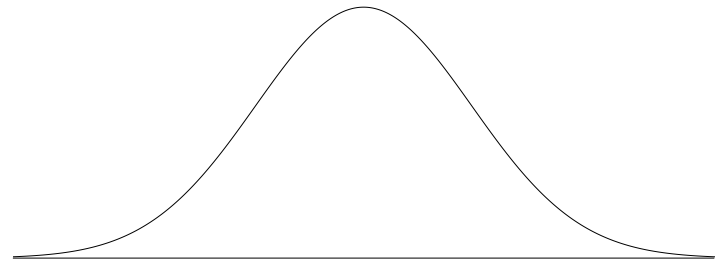
šiviljski meter
$\mu = 12{,}3m$
$\sigma = 0{,}5m$
ravnilo
$\mu = 12{,}3m$
$\sigma = 1m$

Normalna porazdelitev - gostota

Enačba funkcije

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

Verjetnost, da je dogodek na poljubnem intervalu $[a,b]$ izračunamo kot



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

Funkcija je torej popolnoma določena z dvema parametroma:

- **povprečje** μ
- **standardni odklon** σ

To bomo zapisali kot $X \sim \mathcal{N}(\mu,\sigma^2)$

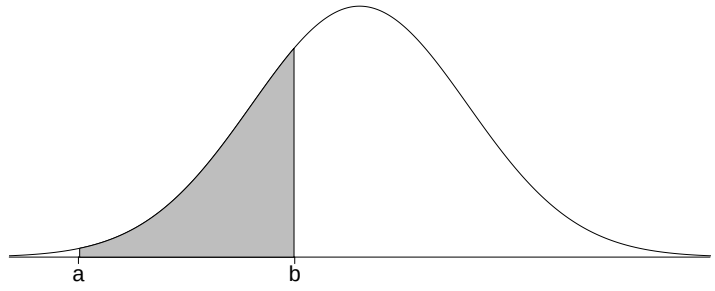
Normalna porazdelitev - gostota

Enačba funkcije

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

Verjetnost, da je dogodek na poljubnem intervalu $[a,b]$ izračunamo kot

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



Funkcija je torej popolnoma določena z dvema parametroma:

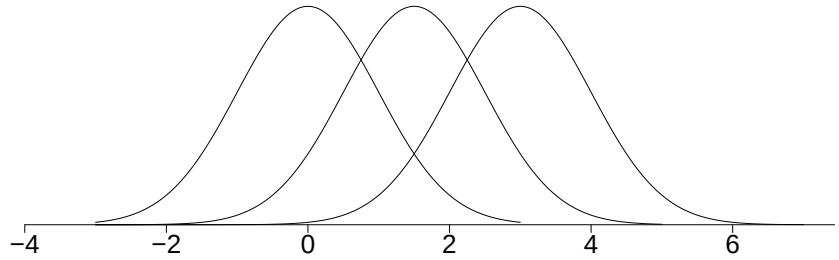
- **povprečje** μ
- **standardni odklon** σ

To bomo zapisali kot $X \sim \mathcal{N}(\mu,\sigma^2)$

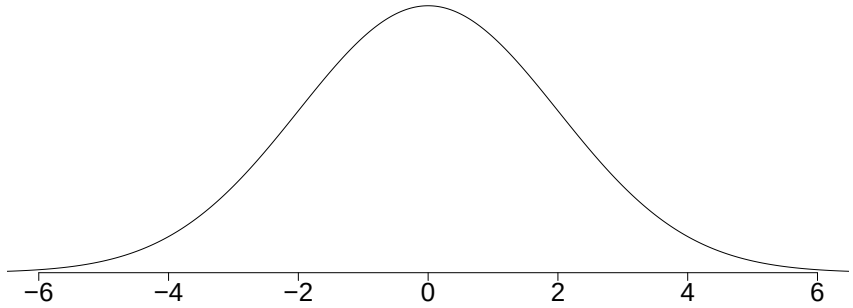
Normalna porazdelitev - lastnosti

povprečje in std. odklon

Spreminjanje povprečja



Spreminjanje variance (spreminjamo skalo na osi)



Ploščina pod krivuljo je seveda vedno enaka 1

Krivulja ohranja enako obliko ne glede na μ in σ

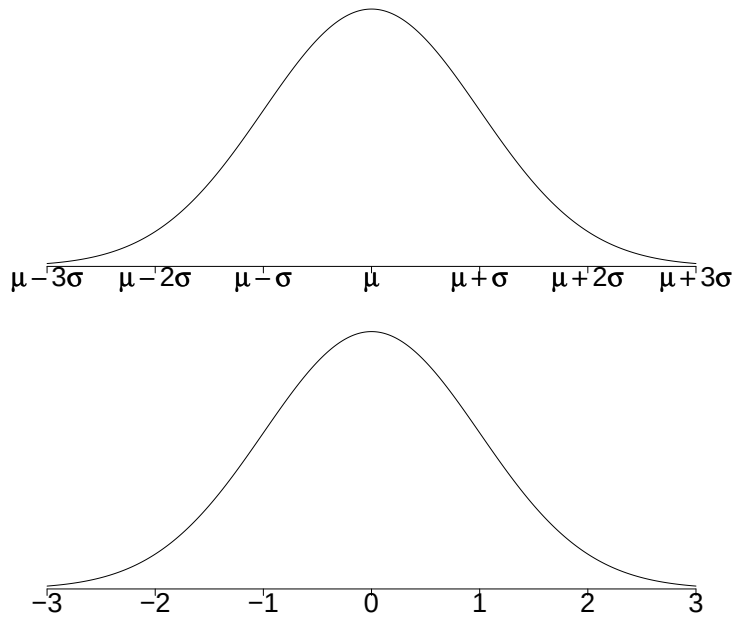
Normalna porazdelitev - lastnosti

Formalno

- Normalna porazdelitev je neobčutljiva za linearne transformacije
- Če je X normalno porazdeljena, je tudi $Y = a + bX$ normalno porazdeljena
- Če sta X in Y normalno porazdeljeni, je normalno porazdeljena tudi $Z = X + Y$

Standardizirana normalna porazdelitev

- To lastnost lahko uporabimo pri računanju verjetnosti
- Poznati bo treba le verjetnosti za eno krivuljo
- Standardizirana normalna porazdelitev ima $\mu = 0$ in $\sigma = 1$



Standardizirana normalna porazdelitev

V teoriji:

Če je X normalno porazdeljena z $E(X) = \mu$ in $Var(X) = \sigma^2$, je

$$Y = a + b \cdot X$$

tudi normalno porazdeljena z

$$E(Y) = a + b\mu \quad \text{in} \quad Var(Y) = b^2\sigma^2.$$

Standardizirana spremenljivka

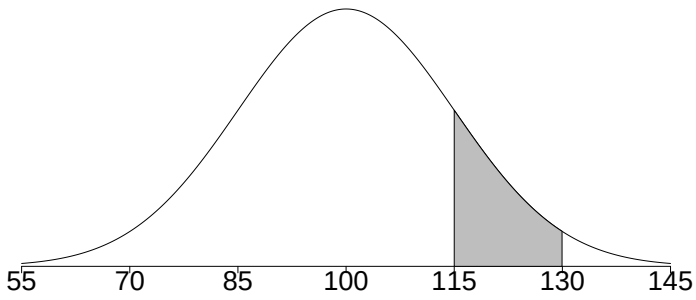
$$Z = \frac{X - \mu}{\sigma} = -\frac{\mu}{\sigma} + \frac{1}{\sigma} \cdot X$$

je potem normalno porazdeljena z

$$E(Z) = 0 \quad \text{in} \quad Var(Z) = 1.$$

Standardizirana normalna porazdelitev - primer

- Zanima nas delež ljudi z IQ med 115 in 130
- $\text{IQ} \sim \text{N}(100, 15^2)$

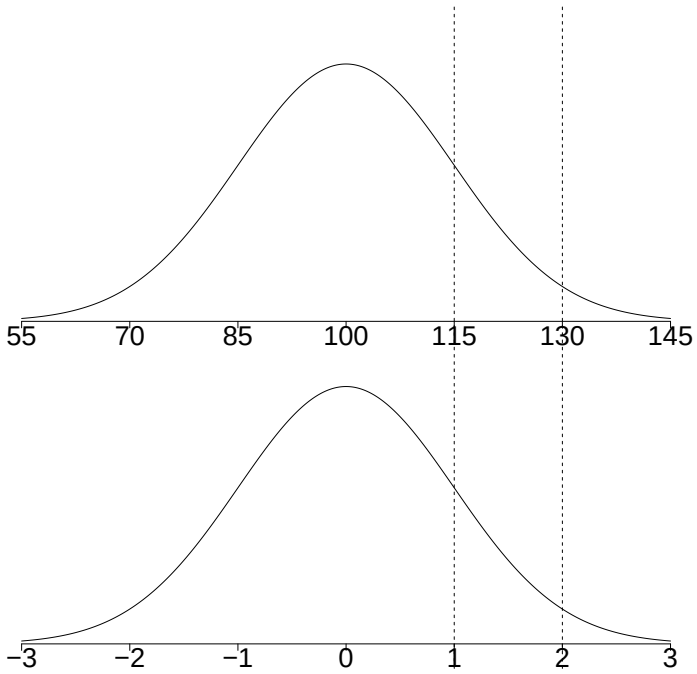


$$
\begin{aligned}
X &= 130 \\
Z &= \frac{X - \mu}{\sigma} \\
&= \frac{130 - 100}{15} \\
&= 2
\end{aligned}
$$

Zanima nas torej ploščina med 1 in 2 pod standardizirano normalno krivuljo

Standardizirana normalna porazdelitev - primer

- Zanima nas delež ljudi z IQ med 115 in 130
- $\text{IQ} \sim \text{N}(100, 15^2)$

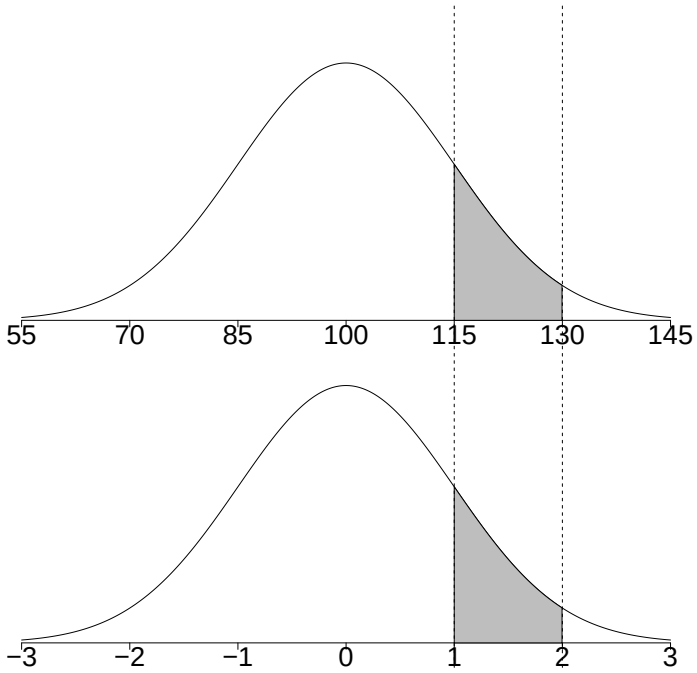


$$
\begin{aligned}
X &= 130 \\
Z &= \frac{X - \mu}{\sigma} \\
&= \frac{130 - 100}{15} \\
&= 2
\end{aligned}
$$

Zanima nas torej ploščina med 1 in 2 pod standardizirano normalno krivuljo

Standardizirana normalna porazdelitev - primer

- Zanima nas delež ljudi z IQ med 115 in 130
- $\text{IQ} \sim \text{N}(100, 15^2)$



$$
\begin{aligned}
X &= 130 \\
Z &= \frac{X - \mu}{\sigma} \\
&= \frac{130 - 100}{15} \\
&= 2
\end{aligned}
$$

Zanima nas torej ploščina med 1 in 2 pod standardizirano normalno krivuljo

Standardizirana normalna krivulja - računanje verjetnosti

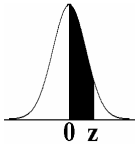
- Integral lahko izračunamo le numerično
- Lahko si pomagamo s tabelami
- Tabele se razlikujejo glede na to, kateri del ploščine je tabeliran
- Dovolj je poznati le ploščine za pozitivne z-je.

Standardizirana normalna krivulja - računanje verjetnosti

- Integral lahko izračunamo le numerično
- Lahko si pomagamo s tabelami
- Tabele se razlikujejo glede na to, kateri del ploščine je tabeliran
- Dovolj je poznati le ploščine za pozitivne z-je.
- Naslednja slika prikazuje primer take tabele

Površina pod standardizirano normalno krivuljo

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441



	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Skupina bolnikov ima levkocite ($10^9/l$) normalno porazdeljene s povprečjem 10 in standardnim odklonom 2.

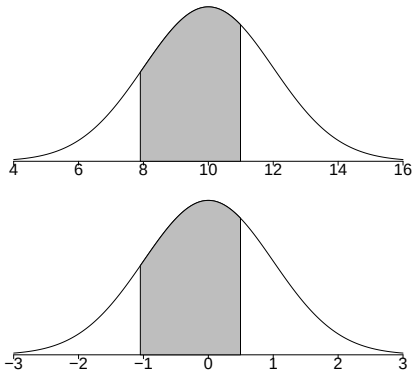
Primer uporabe tabele za izračune - $X \sim \mathcal{N}(10, 2^2)$

Koliko je $P(7{,}9 < X \leq 11)$

Za izračun uporabimo dejstvo, da je $Z = \frac{X-\mu}{\sigma}$ standardizirano normalno porazdeljena!

$$x_1 = 7{,}9 \quad \Rightarrow \quad z_1 = \frac{7{,}9 - 10}{2} = -1{,}05$$

$$x_2 = 11 \quad \Rightarrow \quad z_2 = \frac{11 - 10}{2} = 0{,}5$$



$$\begin{aligned} P(Z \leq 0{,}5) &= P(0 < Z \leq 0{,}5) + P(Z \leq 0) = 0{,}1915 + 0{,}5 = 0{,}6915 \\ P(Z \leq -1{,}05) &= 1 - P(Z \leq 1{,}05) = 1 - (0{,}3531 + 0{,}5) = 0{,}1469 \end{aligned}$$

$$P(7{,}9 < X \leq 11) = 0{,}6915 - 0{,}1469 = 0{,}5446$$

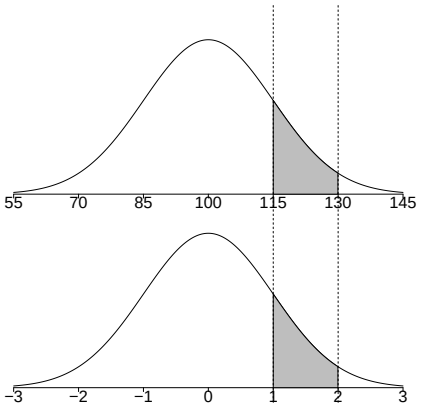
Primer uporabe tabele za izračune - $X \sim \mathcal{N}(100, 15^2)$

Kakšen delež ljudi ima IQ med 115 in 130?

Za izračun uporabimo dejstvo, da je $Z = \frac{X-\mu}{\sigma}$ standardizirano normalno porazdeljena!

$$x_1 = 115 \quad \Rightarrow \quad z_1 = \frac{115 - 100}{2} = 1$$

$$x_2 = 130 \quad \Rightarrow \quad z_2 = \frac{130 - 100}{2} = 2$$



$$P(Z \leq 1) = P(0 < Z \leq 1) + P(Z \leq 0) = 0{,}3413 + 0{,}5 = 0{,}8413$$

$$P(Z \leq 2) = P(0 < Z \leq 2) + P(Z \leq 0) = 0{,}4772 + 0{,}5 = 0{,}9772$$

$$P(1 < Z < 2) = 0{,}9772 - 0{,}8413 = 0{,}1359$$

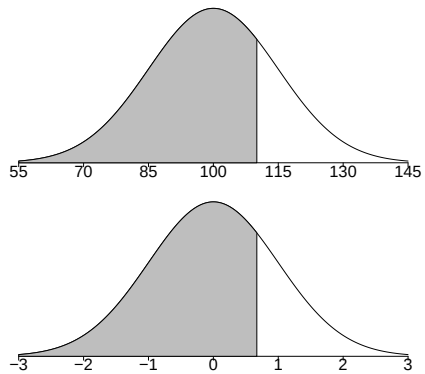
Primer uporabe tabele za izračune - $X \sim \mathcal{N}(100, 15^2)$

Koliko je tretji kvartil za X?

- Določiti moramo torej vrednost a, za katero velja

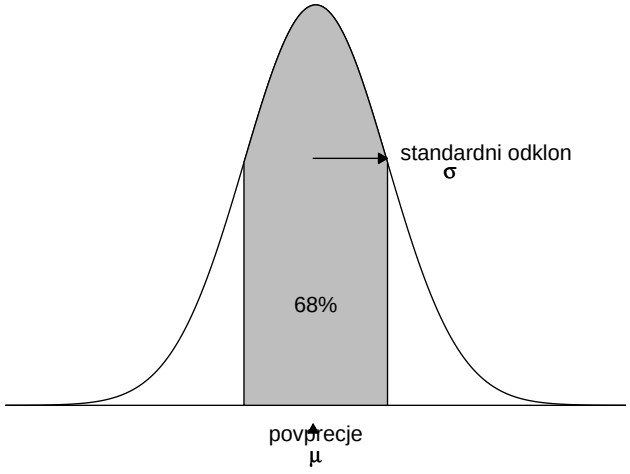
$$P(X \leq a) = 0{,}75$$

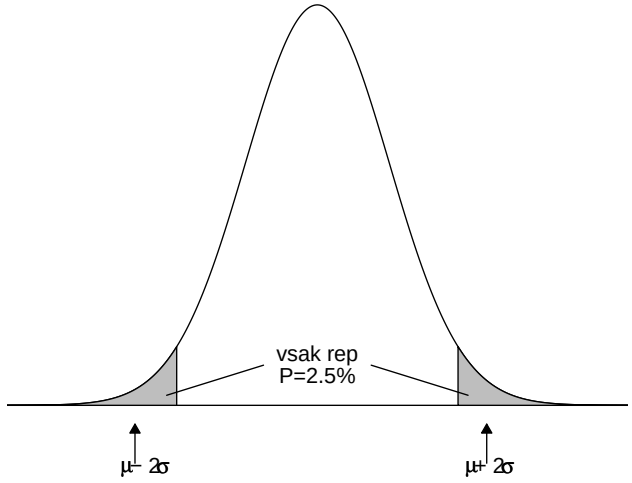
- V tabeli iščemo z, do katerega je ploščina 0,25: $\approx 0{,}67$



- $(a - 100)/15 = 0{,}67 \Rightarrow a = 100 + 15 \cdot 0{,}67 = 110{,}05$

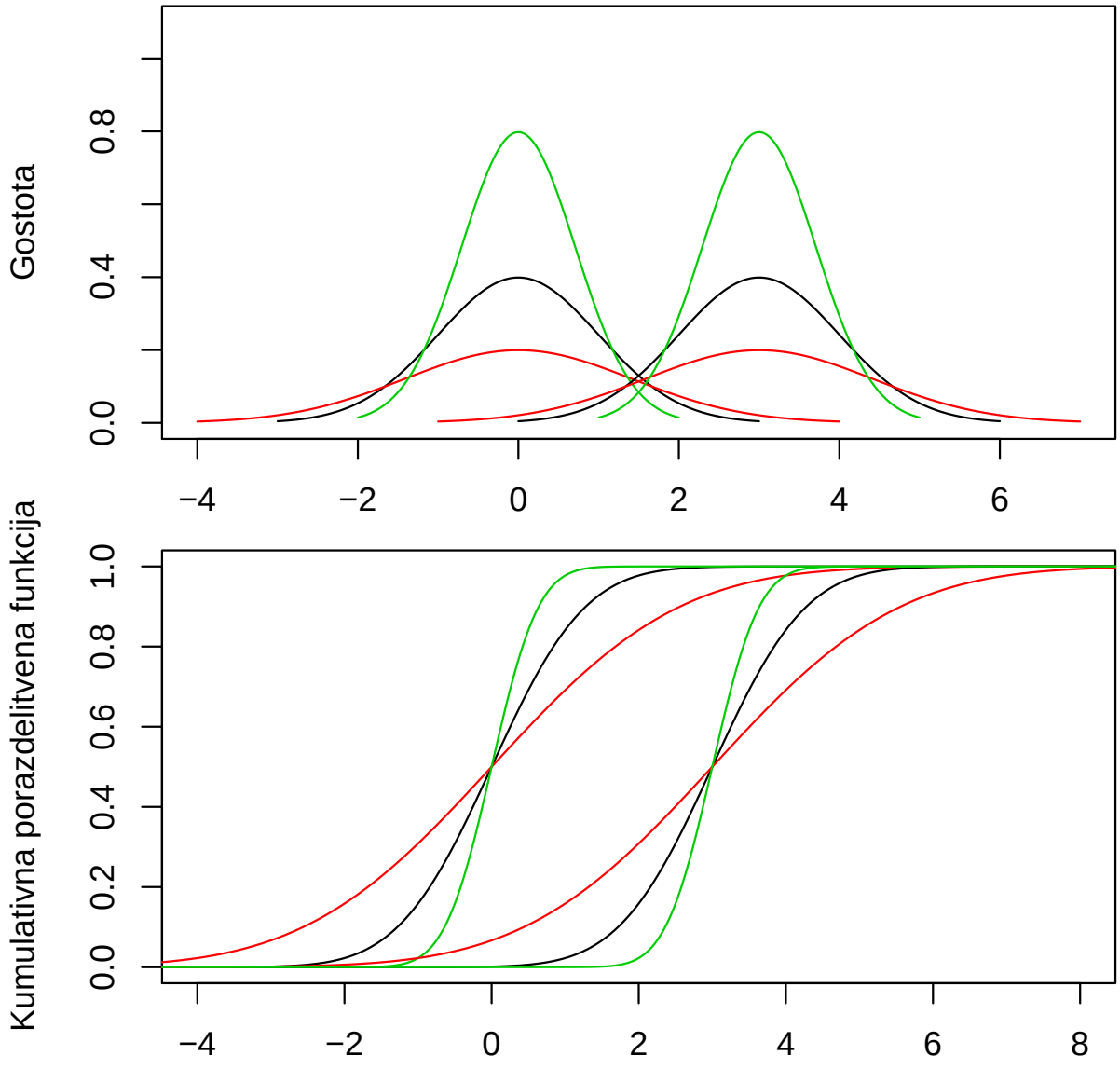
Nekatere pogosto uporabljane vrednosti





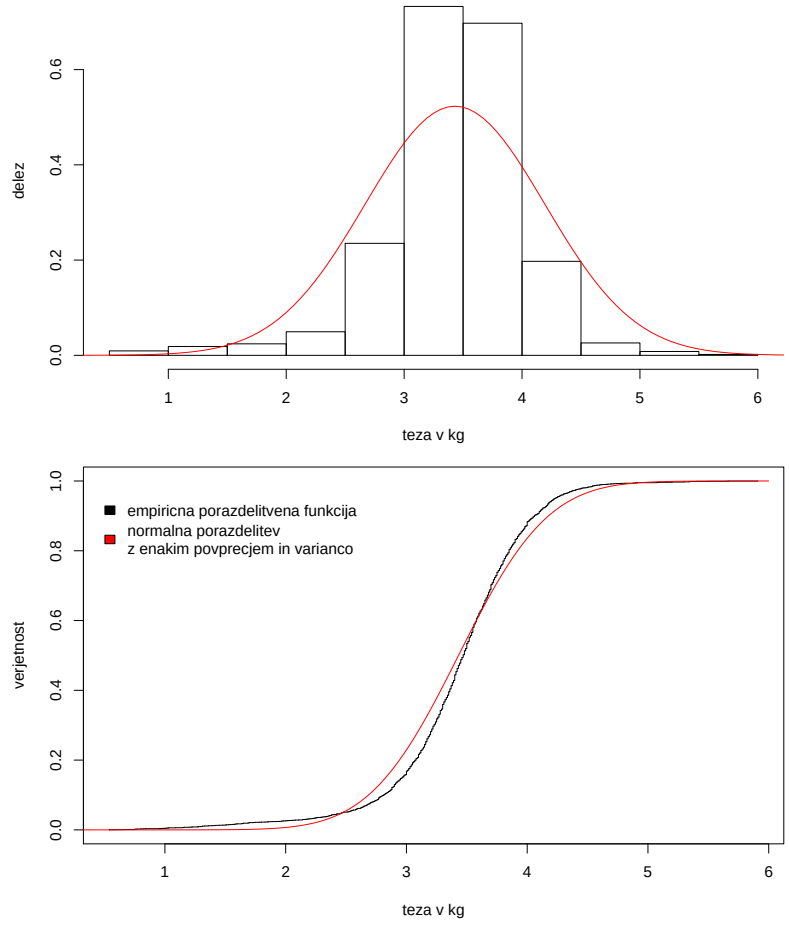
x	$P(X > \mu + x) + P(X < \mu - x)$
0	1
σ	0,3174
$1{,}96 \cdot \sigma$	0,05
$2 \cdot \sigma$	0,0455
$3 \cdot \sigma$	0,0027
$4 \cdot \sigma$	0,00006334

Kumulativna porazdelitvena funkcija



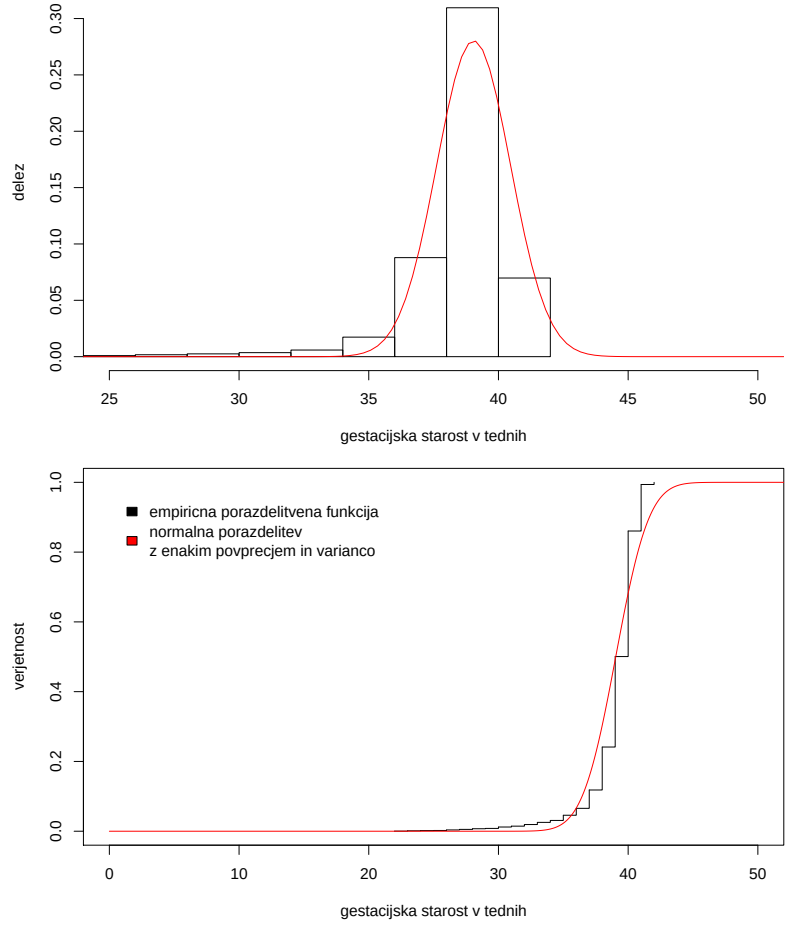
Grafično preverjanje normalnosti

Zanima nas ali je porodna teža normalno porazdeljena



Grafično preverjanje normalnosti

Zanima nas ali je gestacijska starost normalno porazdeljena



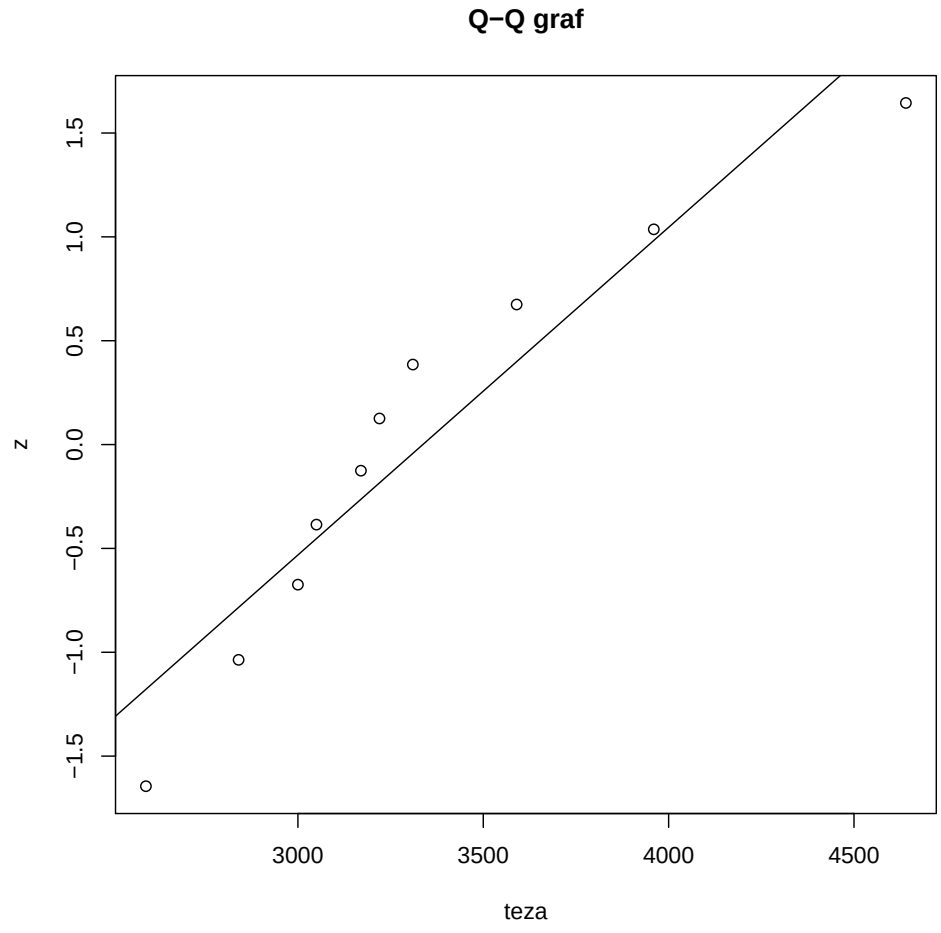
Grafično preverjanje normalnosti

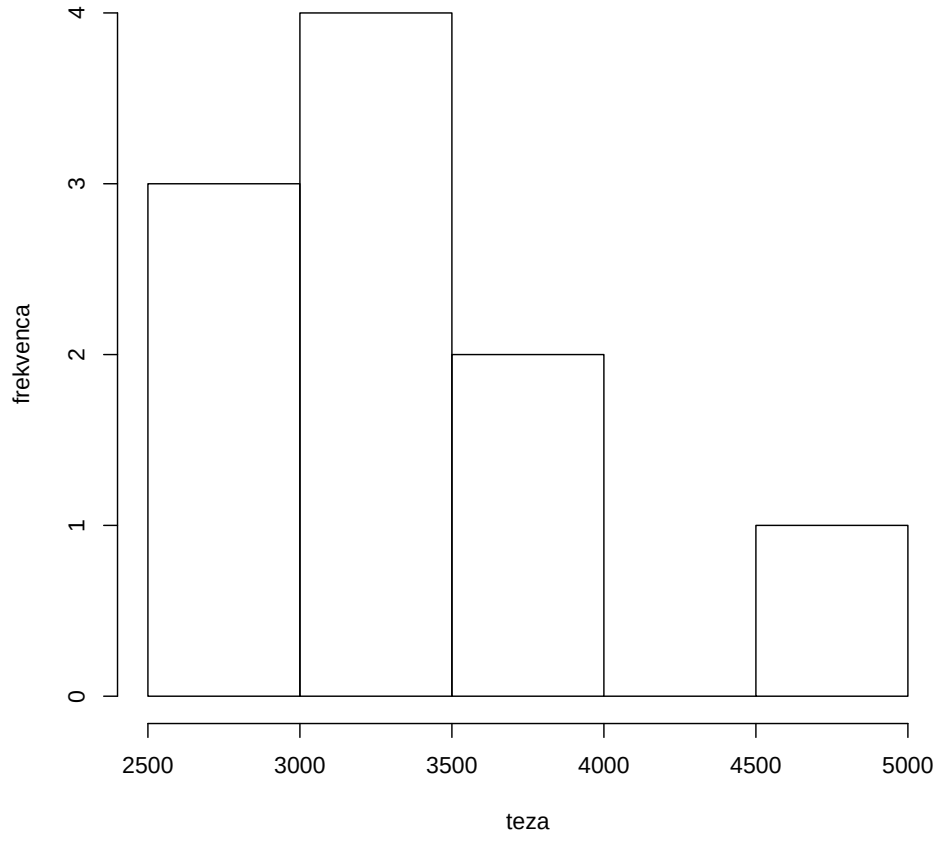
Q-Q graf

- Na abscisno os nanesemo izmerjene vrednosti
- Na ordinatno pa ustrezne z vrednosti, ki bi pripadale x-om glede na njihov rang ob predpostavki, da je X normalno porazdeljena spremenljivka
- Graf mora biti približno linearen, ker je teoretično $z = (x - \mu)/\sigma$.

Grafično preverjanje normalnosti - primer

- 10 naključno izbranih vrednosti porodnih tež dečkov, rojenih v 38. tednu nosečnosti:
  2590, 2840, 3000, 3050, 3170, 3220, 3310, 3590, 3960, 4640.
- Prvi vrednosti pripada tisti z, pod katerim leži 1/10 vseh vrednosti
- Drugi vrednosti pripada tisti, pod katerim je 2/10 vseh vrednosti
- in tako naprej.
- Ker bi za zadnjo vrednost dobili 1 $(= 10/10)$, stvar nekoliko popravimo in namesto i/10 uporabljamo $(i - 0{,}5)/10$


Q−Q graf
z
teza
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
3000
3500
4000
4500


Histogram teze
frekvenca
0
1
2
3
4
2500
3000
3500
4000
4500
5000
teza

Danes seveda z računalniškimi programi poljubno vsoto izračunamo dovolj hitro. Normalna aproksimacija take izračune pohitri, predvsem pa poenostavi našo predstavo - za več kot en standarni odklon je oddaljenih približno 32 % podatkov, ipd.

Normalna porazdelitev kot aproksimacija za binomsko

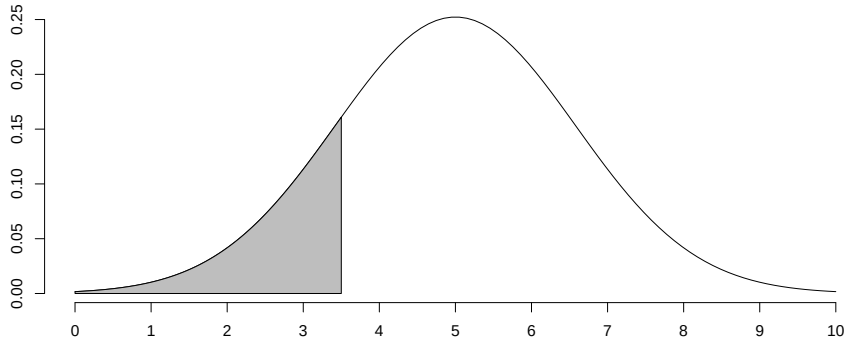
Primer - $X \sim Bin(10,0,5)$

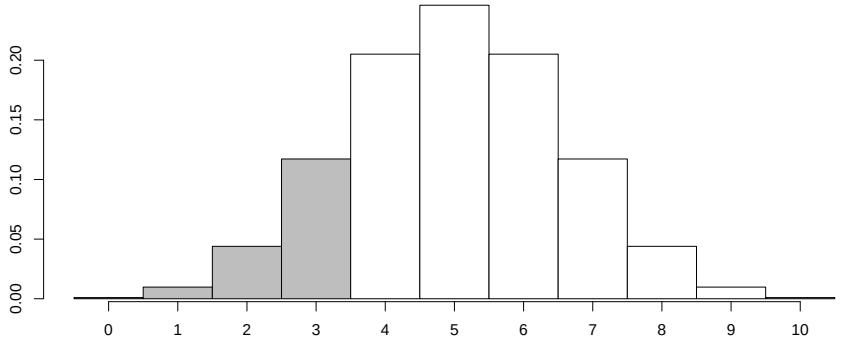
- Koliko je $P(X \leq 3)$?
- Točen odgovor:

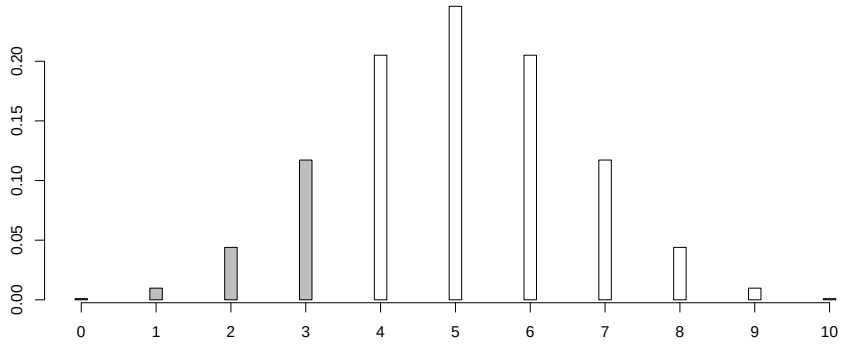
$$P(X \leq 3) = \sum_{k=0}^{3} P(X = k) = 0,1719$$

- Tak postopek dolgotrajen pri velikih n
- Lahko poenostavimo z **normalno aproksimacijo**

Normalna aproksimacija







Normalna aproksimacija

Vsota višin (verjetnosti) prvih štirih stolpcev
$=$ vsota ploščin prvih štirih pravokotnikov v histogramu
$\approx$ ploščina pod normalno krivuljo za x **manjše od 3,5**

Izračun

- $E(X) = n \cdot \pi = 5$ in $Var(X) = n \cdot \pi \cdot (1 - \pi) = 2{,}5$
- $Y \sim \mathcal{N}(5, 2{,}5)$ ($Z \sim \mathcal{N}(0,1)$)
- $$P(X \leq 3) \approx P(Y \leq 3{,}5) = P\left(Z \leq \frac{3{,}5 - 5}{\sqrt{2{,}5}}\right) = 0{,}1714$$
- Normalna aproksimacija dovolj dobro deluje, če se obe vrednosti spremenljivke pojavljata s frekvenco večjo od 5

Normalna aproksimacija

Primer - met kovanca, $X =$ število grbov

- 100 metov
  $E(X) = 50$
  $SD(X) = \sqrt{100 * 0{,}5 * (1 - 0{,}5)} = 5$
- Verjetnost, da bo $X < 40$ ali $X > 60$ je $< 5\%$

- 10000 metov
  $E(X) = 5000$
  $SD(X) = \sqrt{10000 * 0{,}5 * (1 - 0{,}5)} = 50$
- Verjetnost, da bo $X < 4900$ ali $X > 5100$ je $< 5\%$

Centralni limitni izrek

Recimo, da je populacija normalno porazdeljena

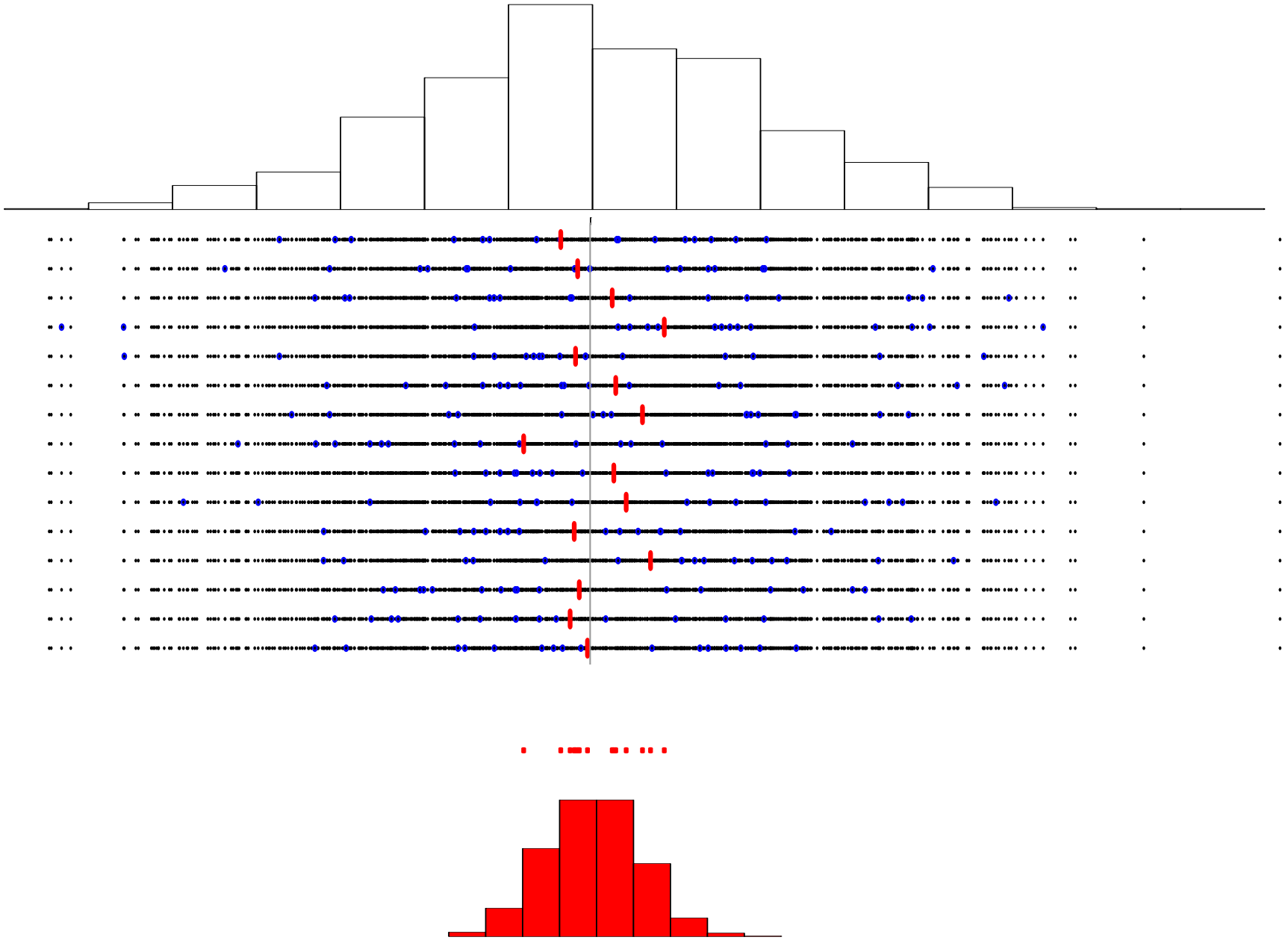
Vemo že:

- Vsota neodvisnih normalnih spremenljivk je normalno porazdeljena
- $E(\overline{X}) = \mu$, $Var(\overline{X}) = \frac{\sigma^2}{n}$

Imejmo vzorec velikosti n iz normalne porazdelitve

- Vrednosti na vzorcu so neodvisne normalne slučajne spremenljivke
- Vsako izmed njih pomnožimo z $\frac{1}{n}$ - še vedno so normalno porazdeljene
- Njihova vsota je prav tako normalno porazdeljena
- Pričakovana vrednost vzorčnega povprečja je μ (populacijsko povprečje)
- Standardni odklon vzorčnega povprečja je $\frac{\sigma}{\sqrt{n}}$

Centralni limitni izrek



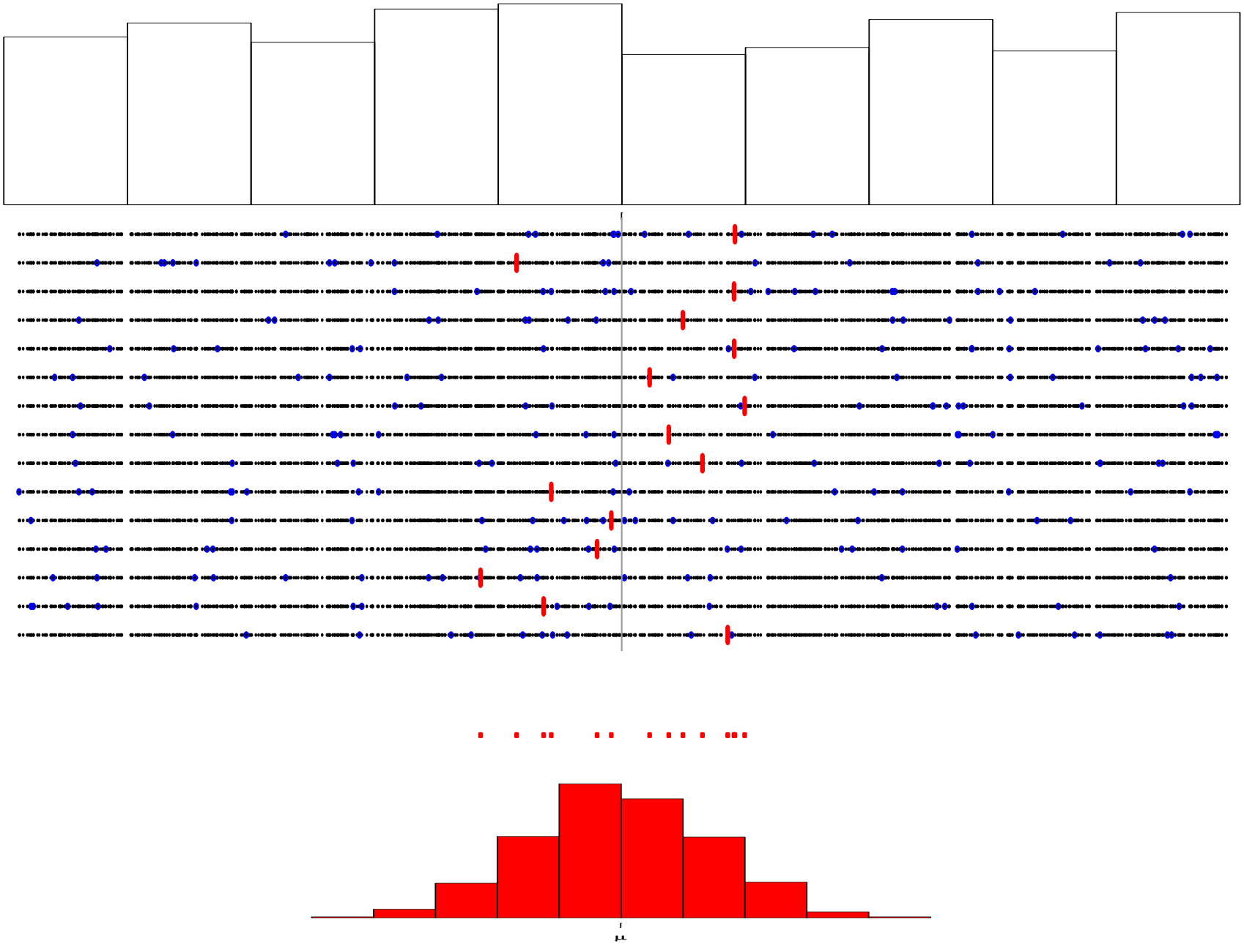
Centralni limitni izrek

Porazdelitev vzorčnih povprečij, porazdelitve populacije ne poznamo

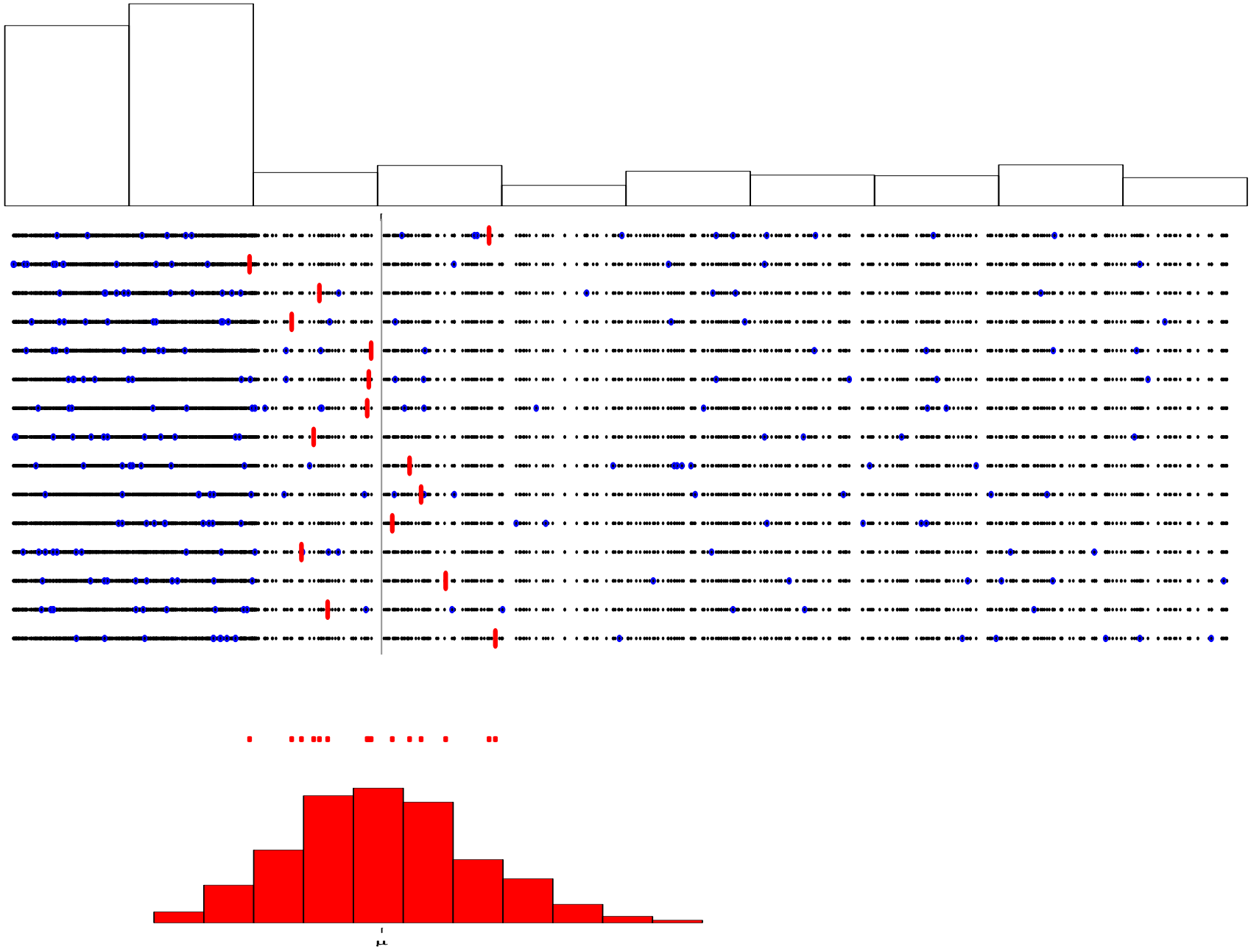
Še vedno velja:

- $E(\overline{X}) = \mu$
  Pričakovana vrednost vzorčnega povprečja je μ (populacijsko povprečje)
- $Var(\overline{X}) = \frac{\sigma^2}{n}$
  Standardni odklon vzorčnega povprečja je $\frac{\sigma}{\sqrt{n}}$

Centralni limitni izrek



Centralni limitni izrek



Centralni limitni izrek

Porazdelitev vzorčnih povprečij

- povprečja vzorcev so približno normalno porazdeljena okoli populacijske vrednosti
- standardna napaka: $\uparrow$ večja variabilnost v populaciji (σ)
  $\downarrow$ večji vzorec (n)

$$SE = \frac{\sigma}{\sqrt{n}}$$

Opombe:

- Čim večji je n, tem boljši je približek.
- stadnardnemu odklonu vzorčnega povprečja pravimo standardna napaka

Vzorčenje

- V praksi bomo imeli praviloma opravka z vzorci, hoteli pa bomo nekaj povedati o populaciji.
- Populacija je množica enot, ki nas zanima, vzorec je del te množice.
- Vzorec mora biti reprezentativen

Vzorčenje

Reprezentativnost vzorca - Primer

- Presejalni test za Downov sindrom v 3. mesecu nosečnosti
- Zanima nas, kakšen delež Downovih sindromov ni bil odkrit
- Imamo podatke iz zdravstvenega doma Domžale
- Tveganje za Downov sindrom je močno povezano s starostjo, ženske nad 36 let imajo možnost brezplačnega pregleda
- Pregled v ZD Domžale je samoplačniski
- Reprezentativnost vzorca (glede na starost) je vprašljiva

Vzorčenje

Reprezentativnost vzorca - Primer

- Predsedniške volitve 2007
- Problem anketiranja po stacionarnem telefonu
- Rezultat odvisen od števila poskusov pri isti osebi
- Vzporedne volitve: zelo velik vzorec, manjkajo podatki po 18. uri

Reprezentativnost vzorca je bistvenega pomena

- Majhen vzorec: večji intervali pri ocenah, manjša verjetnost pri zavrnitvi domneve.
- Pristran vzorec: ne moremo delati zaključkov, ne moremo izračunati verjetnosti
- Če vemo za kakšno pristranost gre, jo lahko v določenih situacijah popravimo v statistični analizi

Vzorčenje

Naključno vzorčenje

- Iz populacije **vzorčimo naključno**
- Vsaka enota populacije enako verjetnost, da bo izbrana
- Imeti moramo seznam vseh enot, ki so lahko izbrane v vzorec
- Tabela naključnih števil, računalnik

Izbiro slučajnega vzorca lahko izvedemo na več načinov. Najprej pa moramo imeti seznam vseh enot, ki so lahko izbrane v vzorec. Temu rečemo **vzorčni okvir**. Če izbiramo iz manjše populacije, lahko vsaki enoti dodelimo številko in številke naključno vlečemo iz na primer **klobuka**.

Druga možnost je uporaba **tabel naključnih števil**. Te tabele so sestavljene iz naključnih števk, ki so ponavadi zaradi večje preglednosti grupirane v skupine po nekaj števk.

Tabela naključnih števil

39634 62349 74088 65564 16379 19713 39153 69459 17986
24537 14595 35050 40469 27478 44526 67331 93365 54526
22356 93208 30734 71571 83722 79712 25775 65178 07763
82928 31131 30196 64628 89126 91254 24090 25752 03091
39411 73146 06089 15630 42831 95113 43511 42082 15140
34733 68076 18292 69486 80468 80583 70361 41047 26792
78466 03395 17635 09697 82447 31405 00209 90404 99457
72570 42194 49043 24330 14939 09865 45906 05409 20830
01911 60767 55248 79253 12317 84120 77772 50103 95836
22530 91785 80210 34361 52228 33869 94332 83868 61672
65358 70469 87149 89509 72176 18103 55169 79954 72002
20582 72249 04037 36192 40221 14918 53437 60571 40995
55006 10694 41692 40581 93050 48734 34652 41577 04631
49184 39295 81776 61885 50796 96822 82002 07973 52925
75467 86013 98072 91942 48917 48129 48624 48248 91465
54898 61220 18721 67387 66575 88378 84299 12193 03785
49314 39761 99132 28775 45276 91816 77800 25734 09801

Vzorčenje s pomočjo računalnika

Uporaba **generatorjev slučajnih števil** v raznih programskih okoljih. Na primer v okolju R nam bo ukaz

```
sample(1:5000,20)
```

generiral 20 slučajnih števil izmed 5000 števil:

4134 1089 4412 794 3006 1845 609 687 1203 4244 1964 2435
3840 1033 3999 3245 1858 4157 112 4496

Sistematično vzorčenje

- Imamo seznam vseh možnih enot, vendar le prvega izberemo naključno
- Ostale izberemo tako, da izberemo vsakega *n*-tega.
- Na primer: če imamo v vzorčnem okviru 500 enot, naš vzorec pa naj bi bil velik 100 enot, je vzorčni delež 1/5, se pravi, da bomo izbrali eno enoto iz vsakih 5 enot
- Možna slaba stran: periodičnost v podatkih

Vzorčenje

Stratificirano vzorčenje

Vzorčni okvir razdelimo na sloje (stratume), na primer po starostnih skupinah, in v vsakem sloju vzorčimo naključno.

Večstopenjsko vzorčenje

Vzorčenje poteka v več fazah. Na primer: najprej vzorčimo iz mest, potem iz seznamov zdravnikov v mestih in na koncu iz seznamov bolnikov posameznih bolnikov.

Randomizacija

Bistvo: Izbira načina zdravljenja odvisna od slučaja in nobenih drugih vplivov.

Cilj: skupine naj bi bile primerljive v vsem, razen v načinu zdravljenja.

Pomembno: osebo najprej vključimo v študijo in šele potem randomiziramo!

Kako randomizirati?

Možnosti: kovanec, kocka, karte, kolo sreče.
Tabela slučajnih števil:

1. v dve skupini (sodo, liho)
2. v tri skupine (1,2,3; 4,5,6 in 7,8,9, 0 zanemarimo)
3. če hočemo enake skupine: ko napolnimo eno skupino, preostale v drugo - ne najbolje
4. lahko slučajno izberemo najprej prvo skupino, ostali so druga

Uvod v statistično sklepanje

- Zbrani podatki (ponavadi) predstavljajo **vzorec** vseh možnih podatkov iz **populacije** (prave ali hipotetične)
- Namen statistične analize je povedati nekaj o populaciji na osnovi podatkov iz vzorca

Primer - primerjava učinkovitosti dveh načinov zdravljenja

- Recimo, da učinek zdravljenja merimo z neko numerično spremenljivko
- V vsaki skupini izračunamo povprečno vrednost
- Seveda ne bomo dobili povsem enakih vrednosti, možni razlogi za razliko pa so:
  1. Zdravljenji sta **različno učinkoviti**
  2. Gre za **slučajno variiranje**
  3. (Rezultati so **pristranski** zaradi vplivov drugih dejavnikov - moteči dejavniki)

Pristranost lahko odpravimo (ali zmanjšamo) z ustreznim **načrtom študije** pa tudi v statistični analizi lahko **pristranost popravimo**. Pristranost naj torej ne bi bila verjeten razlog za razliko med skupinama.

Cilj statistične analize je **oceniti** razliko v učinku zdravljenja in ugotoviti, ali je slučajno variiranje možna razlaga za nastalo razliko.

Če je bila študija dobro načrtovana in statistična analiza kaže, da naključno variiranje ni ustrezna razlaga za nastalo razliko, lahko sklepamo, da gre najverjetneje za dejansko razliko v učinkovitosti zdravljenja.

Uvod v statistično sklepanje

- Cilj statistične analize je **oceniti** razliko v učinku zdravljenja in ugotoviti, ali je slučajno variiranje možna razlaga za nastalo razliko
- Če je bila študija dobro načrtovana in statistična analiza kaže, da naključno variiranje ni ustrezna razlaga za nastalo razliko, lahko sklepamo, da gre najverjetneje za dejansko razliko v učinkovitosti zdravljenja

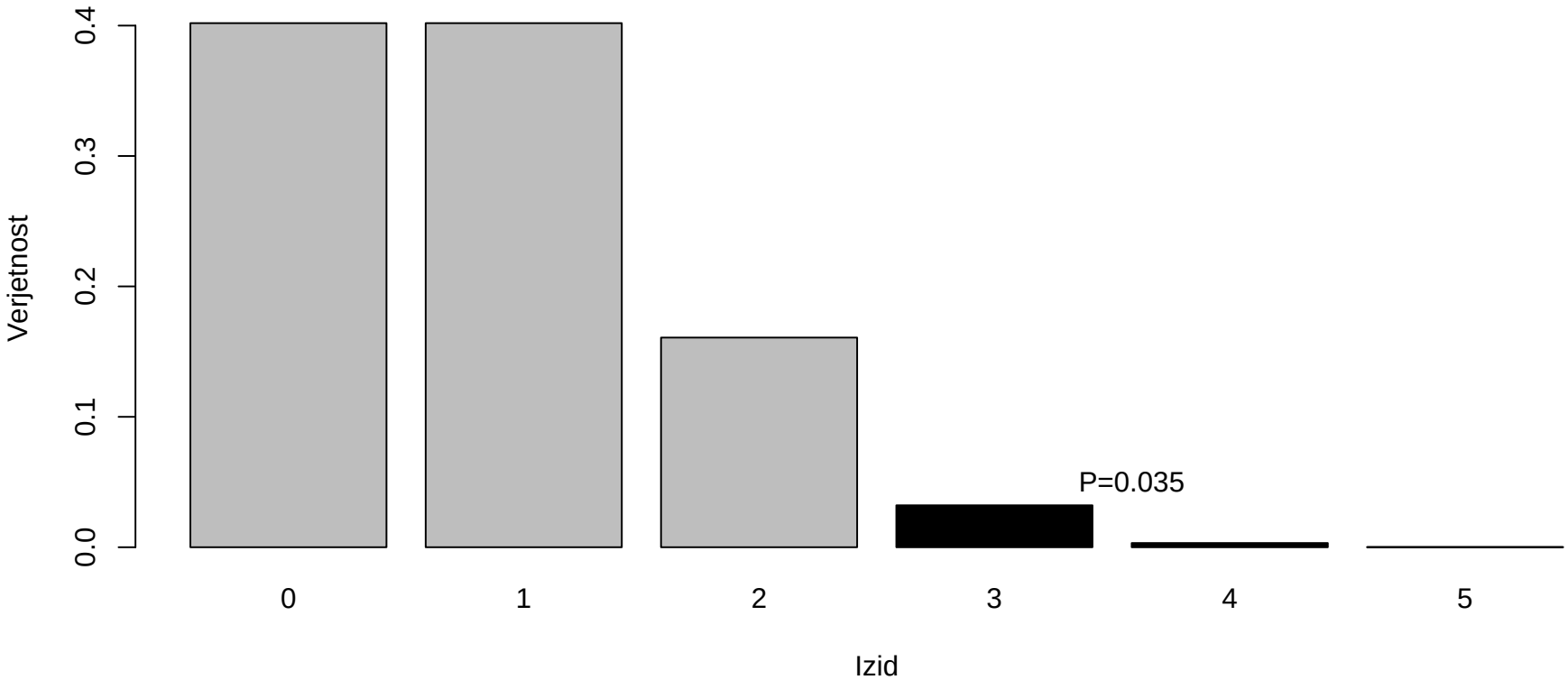
Ideja statističnega sklepanja

- determinističnost $\Leftrightarrow$ variabilnost
- populacija $\Leftrightarrow$ vzorec
- odločanje na podlagi verjetnosti

Statistično sklepanje

- Ničelna domneva, stopnja tveganja
- Izračun verjetnosti = porazdelitev testne statistike (predpostavke)
- Podatki, ‘odločitev’

5 metov kocke

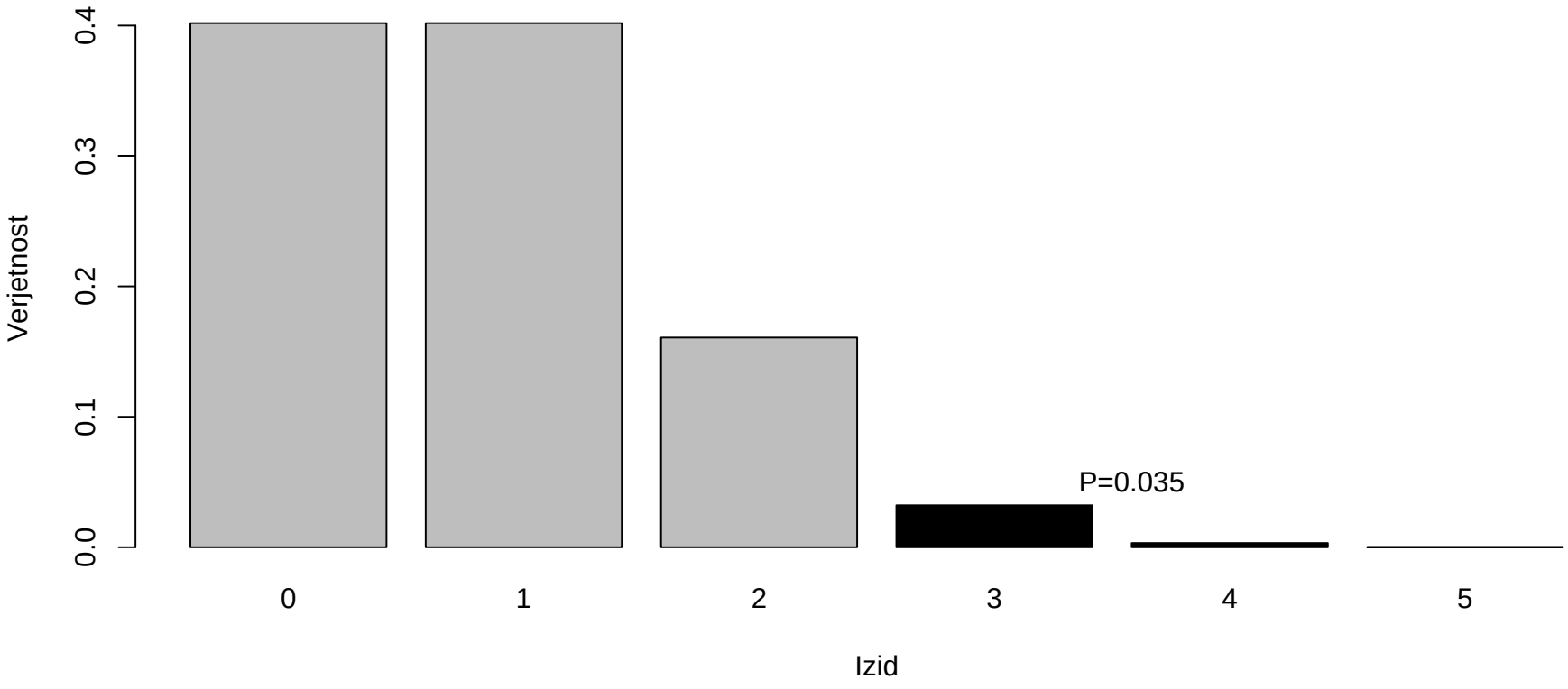


5 metov kocke

Statistično sklepanje

- Ničelna domneva, stopnja tveganja
  H_0: Kocka je poštena (verjetnost šestice je $\frac{1}{6}$)
  (Alternativna domneva: šestica pade prepogosto)
  stopnja tveganja $\alpha = 0{,}05$
- Izračun verjetnosti (predpostavke), postavitev mej
  Slučajna spremenljivka $X =$‘število šestic’ je binomsko porazdeljena
  Vrednosti nad 2 imajo skupno verjetnost manjšo od 0,05: $P(X > 2) < \alpha$
- Podatki, ‘odločitev’
  Padlo je 5 šestic. $P(X \geq 5) = 0{,}0001$
  Verjetnost, da ničelna domneva drži, je zelo majhna $\rightarrow$ ničelno domnevo zavrnemo

5 metov kocke



Primer - Učinkovitost zdravila

Zdravilo za hipertenzijo

- Imamo zdravilo za hipertenzijo A, ki po 3 tednih jemanja zniža krvni tlak pod 160 mmHg pri 70% bolnikov
- Zanima nas učinkovitost zdravila B
- Zdravilo B bomo preizkusili na 30 bolnikih
- $X =$‘število ozdravljenih bolnikov’ je slučajna spremenljivka
- Pričakujemo, da bo zdravilo delovalo pri 21 bolnikih ($E(X) = 21$)
- Ker je X slučajna spremenljivka, so naključna odstopanja pričakovana
- Kaj lahko rečemo, če ozdravi 20 pacientov? Kaj lahko rečemo, če ozdravi 15 pacientov?

Primer - Učinkovitost zdravila

Ničelna domneva, stopnja tveganja

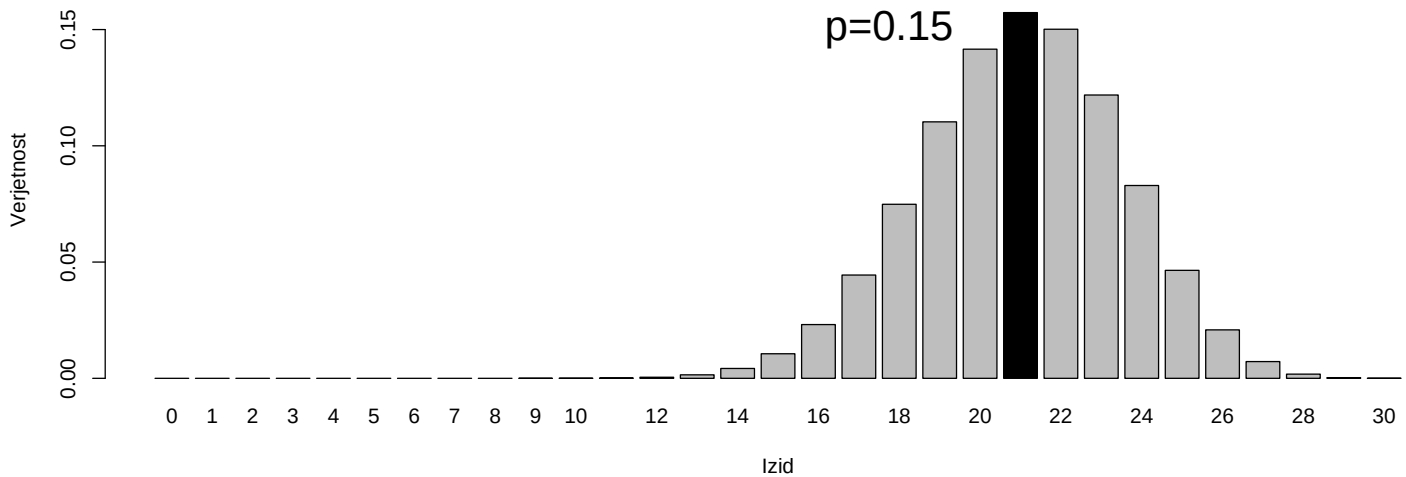
- Vzamemo vzorec 30 bolnikov s krvnim tlakom nad 160 mmHg, dajemo jim zdravilo B
- Ničelna domneva: Zdravilo B je enako učinkovito kot zdravilo A
- Alternativna domneva: Zdravilo B je manj učinkovito od zdravila A
- Stopnja tveganja je $\alpha = 0{,}05$ - rezultat želimo podati s 95 % gotovostjo

Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7; \alpha = 0{,}05$

Izračun verjetnosti (predpostavke), postavitev mej

- Slučajna spremenljivka $X =$‘število ozdravljenih bolnikov’ je binomsko porazdeljena
- (predpostavke: bolniki so med seboj neodvisni, vsi imajo enako verjetnost ozdravitve)

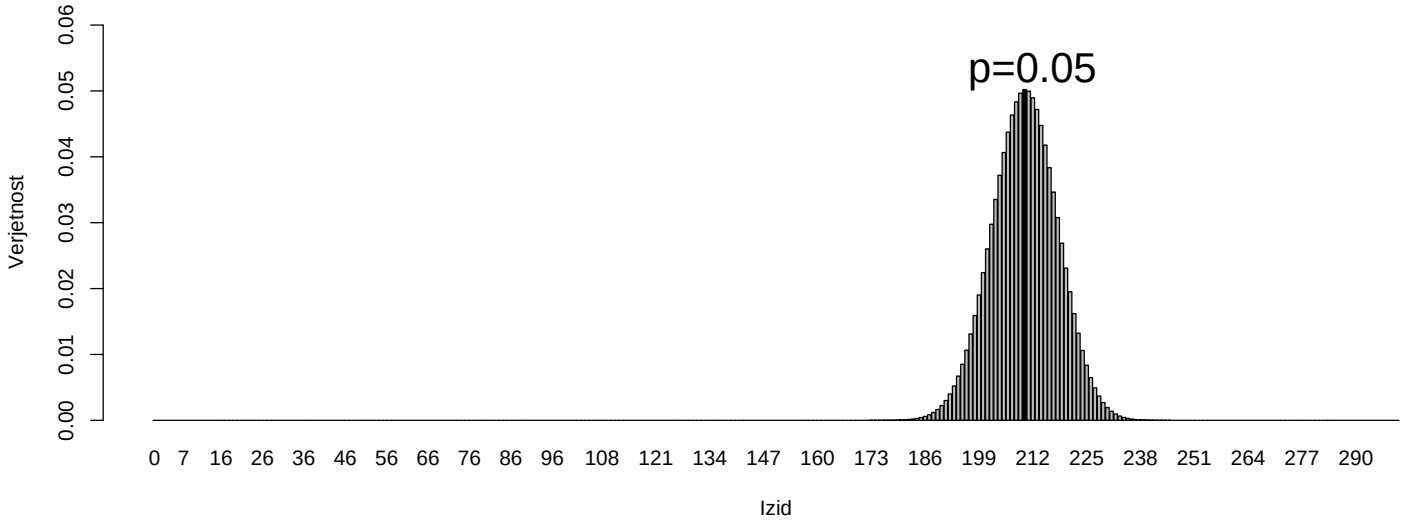


Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7; \alpha = 0{,}05$

Izračun verjetnosti (predpostavke), postavitev mej

- Slučajna spremenljivka $X =$‘število ozdravljenih bolnikov’ je binomsko porazdeljena
- (predpostavke: bolniki so med seboj neodvisni, vsi imajo enako verjetnost ozdravitve)

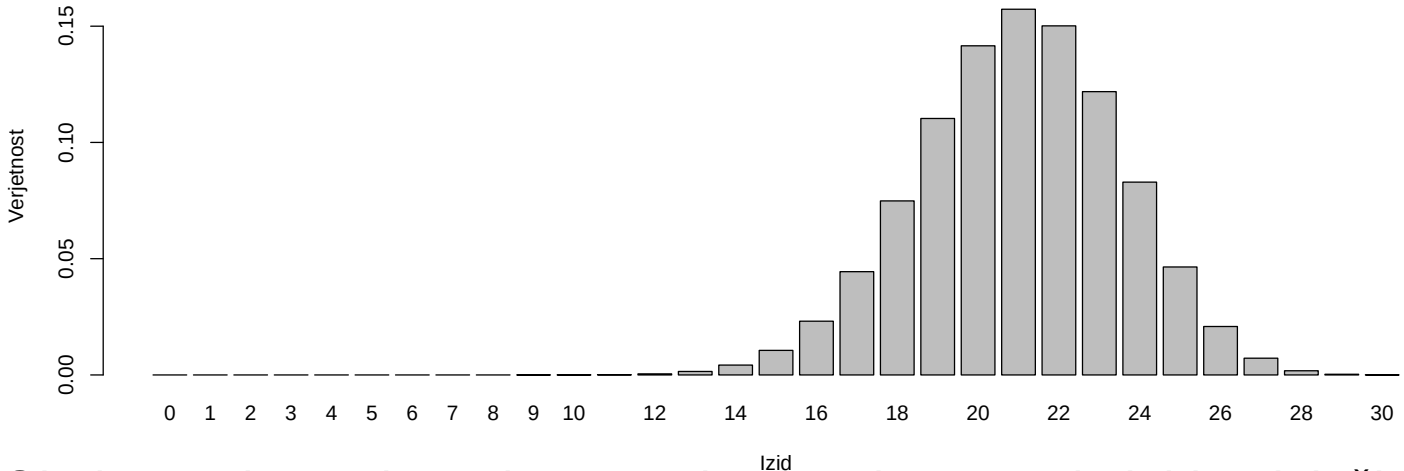


Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7; \alpha = 0{,}05$

Izračun verjetnosti (predpostavke), postavitev mej

- Slučajna spremenljivka $X =$‘število ozdravljenih bolnikov’ je binomsko porazdeljena
- (predpostavke: bolniki so med seboj neodvisni, vsi imajo enako verjetnost ozdravitve)



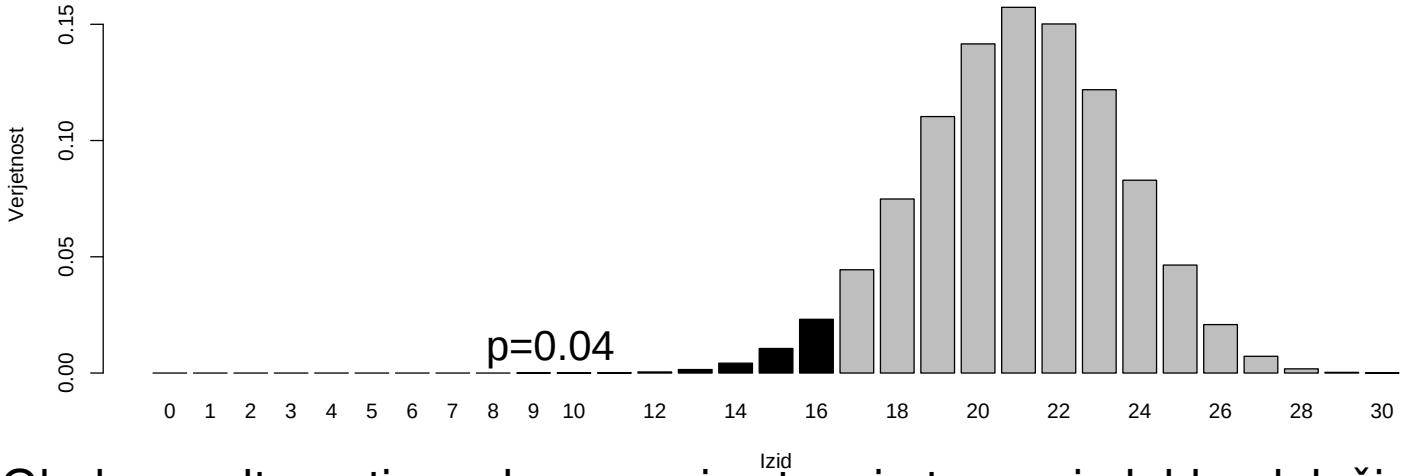
- Glede na alternativno domnevo in stopnjo tveganja lahko določimo interval zavrnitve

Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7; \alpha = 0{,}05$

Izračun verjetnosti (predpostavke), postavitev mej

- Slučajna spremenljivka $X =$‘število ozdravljenih bolnikov’ je binomsko porazdeljena
- (predpostavke: bolniki so med seboj neodvisni, vsi imajo enako verjetnost ozdravitve)



- Glede na alternativno domnevo in stopnjo tveganja lahko določimo interval zavrnitve
- Če bo ozdravljenih 16 ali manj bolnikov, bomo ničelno domnevo zavrnili

Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7$; $\alpha = 0{,}05$; $X \sim Bin(0{,}7, 30)$

Podatki, odločitev

- Krvni tlak se spusti pod željeno mejo pri 15 bolnikih
- Vrednost $X = 15$ je v območju zavrnitve ničelne domneve
- Verjetnost, da je ničelna domneva pravilna je manjša od $\alpha = 0{,}05$
- Ničelno domnevo zavrnemo, sprejmemo alternativno domnevo - zdravilo B je manj učinkovito od zdravila A
- Natančneje: $P(X \leq 15) = 0{,}02$ - tveganje ob zavrnitvi ničelne hipoteze je 0,02

Primer - Učinkovitost zdravila

$H_0 : \pi = 0{,}7$, $\alpha = 0{,}05$, $X \sim Bin(0{,}7, 30)$

Podatki, odločitev

- Krvni tlak se spusti pod željeno mejo pri 20 bolnikih
- Vrednost $X = 20$ je v območju, kjer ničelne domneve ne zavrnemo
- Verjetnost, da je do odstopanja (od pričakovane vrednosti 21) prišlo zaradi naključne variabilnosti je velika ($> \alpha$)
- Ničelne domneve ne zavrnemo, ne moremo trditi, da je zdravilo B manj učinkovito od zdravila A
- Natančneje: $P(X \leq 20) = 0{,}41$

Primer - Učinkovitost zdravila

Recimo, da imamo v vzorcu le 3 bolnike

- $X \sim Bin(0{,}7, 3)$
- $P(X = 0) = 0{,}03$, $P(X = 1) = 0{,}19$, $P(X = 2) = 0{,}44$
- Ničelno hipotezo bi lahko zavrnili le, če zdravilo ne bi učinkovalo pri nobenem bolniku
- Za primerjavo: pri 30 bolnikih, je bil že 0,53 (16/30) premajhen delež

Primer - Učinkovitost zdravila

Opombe

- Najzahtevnejši del bo vedno ugotavljanje porazdelitve slučajne spremenljivke
- Pri tem bomo morali včasih narediti več predpostavk
- Prav nam bo prišel centralni limitni izrek, ki pravi, da je povprečje približno normalno porazdeljeno (ne glede na porazdelitev populacije)
- Tokrat nas je zanimalo le, ali je zdravilo B manj učinkovito - temu pravimo enostranska alternativna domneva

Analiza enega vzorca iz normalno porazdeljene populacije

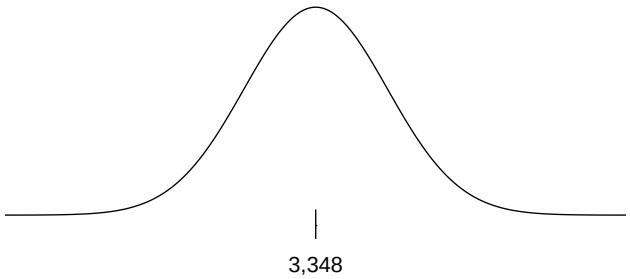
Primer - porodne teže 35-letnih mater

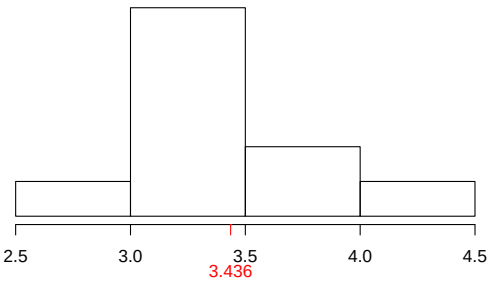
Vzorec 10 dojenčkov (v g):

3310, 3880, 3460, 3490, 3160, 3250, 2630, 4370, 3530, 3280

Vprašanje: Ali se povprečna teža dojenčkov 35-letnic razlikuje od skupnega povprečja?

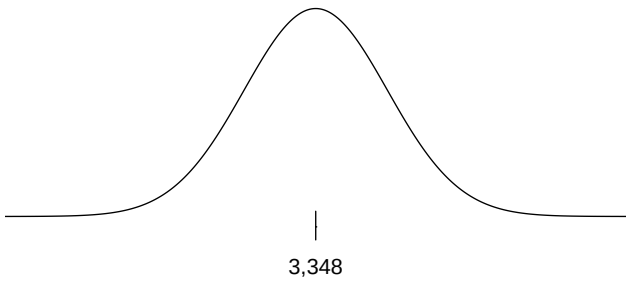
(Vemo, da je povprečna porodna teža 3348 kg s standardnim odklonom 500 in da so teže normalno porazdeljene)

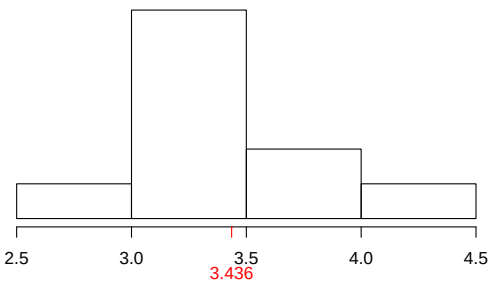




Primer - porodne teže 35-letnih mater

- Celotna populacija: $\mu_0 = 3348$ ($\sigma_0 = 500$)
- Vzorčno povprečje: $\overline{x} = 3436$ ($s = 457$)
- Ali je **vzorčno variiranje** možna razlaga za odstopanje vzorčnega povprečja?





Primer - porodne teže 35-letnic

Statistično sklepanje

- Ničelna domneva, stopnja tveganja
  H_0 : Vzorec izhaja iz populacije s povprečjem 3348
  Populacija 35-letnic ima povprečje $\neq$ 3348
  Stopnja tveganja $\alpha = 0{,}05$
- Porazdelitev testne statistike (predpostavke)
  Najprej predpostavke
- Podatki, ‘odločitev’

Primer - porodne teže 35-letnic

$H_0 : \mu = 3348$

Porazdelitev testne statistike, predpostavke

- Populacija normalno porazdeljena
- $\sigma = 500$
- Podatki predstavljajo slučajni vzorec (neodvisne, enako porazdeljene meritve) $x_1, x_2, \ldots, x_{10}$

Primer - porodne teže 35-letnic

$H_0 : \mu = 3348$

Porazdelitev testne statistike, preverjanje predpostavk

1. **Neodvisnost**
   - Preverimo postopek vzorčenja (ali je naključno?)
   - Primer: predpostavka kršena, če so med podatki tudi podatki o dvojčkih.
2. **Ista porazdelitev**?
   Lahko nastopijo težave pri zbiranju skozi čas (grafično)
3. **Normalnost**? Preverimo grafično (histogram, s q-q grafom)
4. **Varianca**? Obstajajo testi, zaenkrat samo približno ugotavljamo smiselnost, temu se kasneje ognemo

Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

- Predpostavimo: povprečna porodna teža otrok 35-letnih mater: $\mu = 3348g$, $\sigma = 500g$
- Kako velika odstopanja od te vrednosti lahko pričakujemo na vzorcih?

Spomnimo se:

- Vzorčno povprečje je normalno porazdeljeno
- Pričakovana vrednost

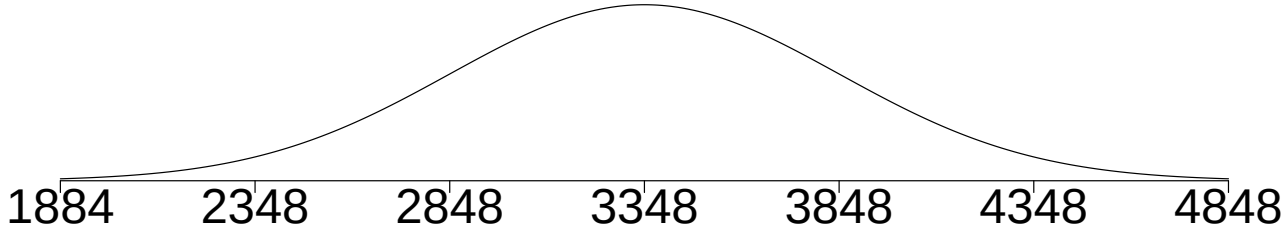
$$E(\overline{X}) = \mu$$

- Standardni odklon

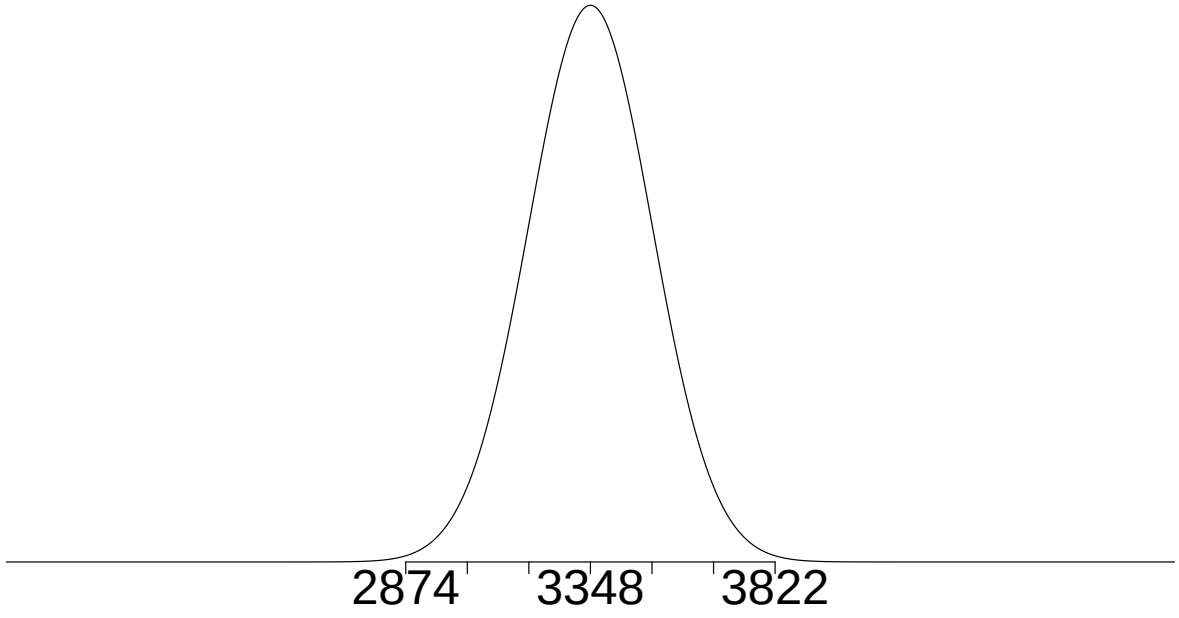
$$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Porazdelitev testne statistike

$X \sim N(3348, 500^2)$

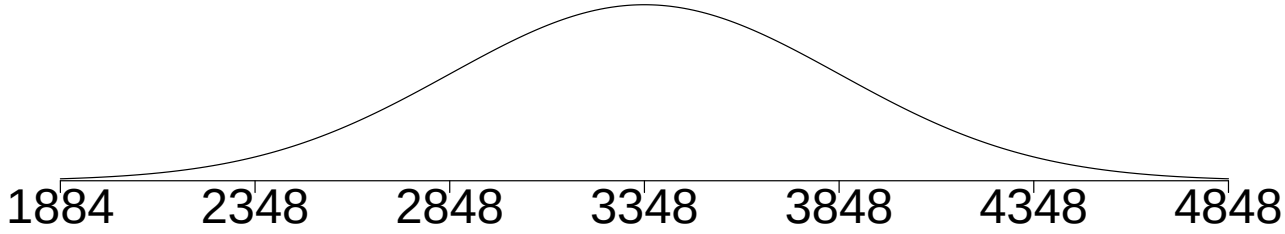


$\overline{X} \sim N(3348, \frac{500^2}{10})$

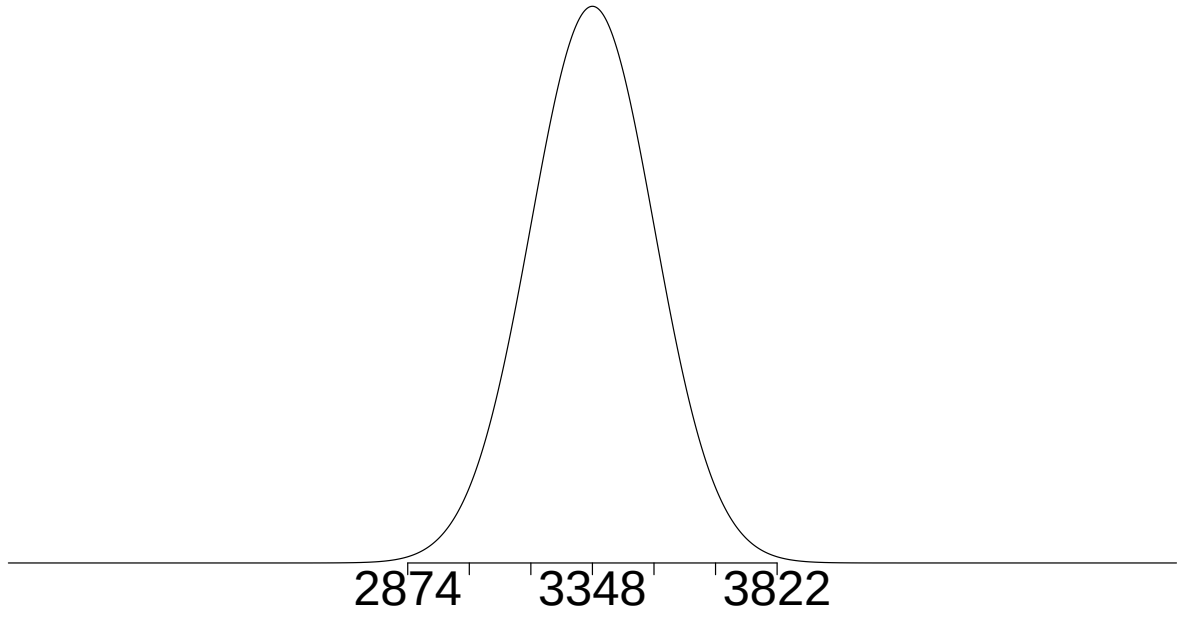


Porazdelitev testne statistike

$X \sim N(3348, 500^2)$



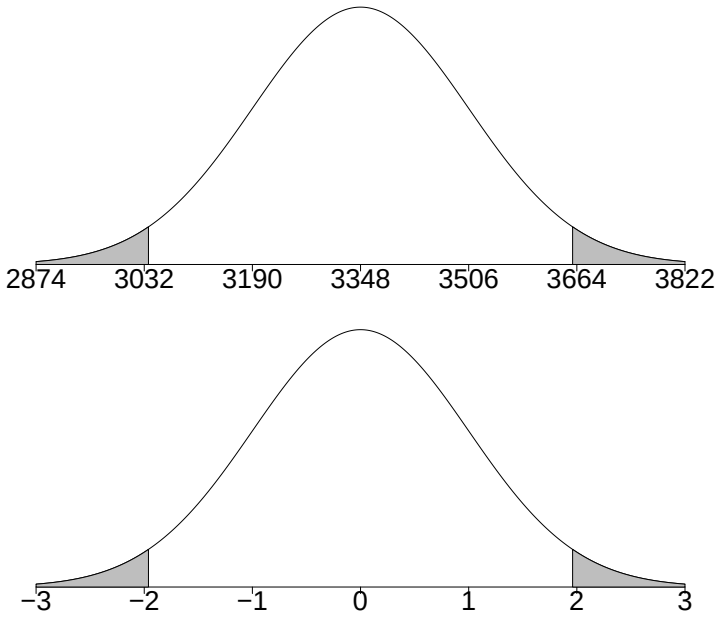
$\overline{X} \sim N(3348, 158^2)$



Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

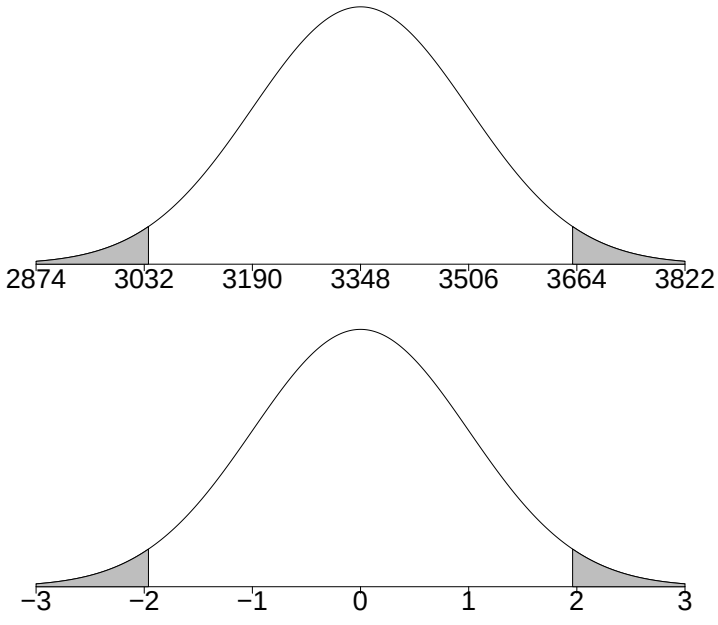
- Iz predpostavk sledi, da je $\overline{X} \sim N(3348, 158^2)$
- Oziroma: $Z = \frac{\bar{X}-3348}{158} \sim N(0,1)$
- S 95 % verjetnostjo torej pričakujemo, da bo vzorčno povprečje na intervalu [3038,3658]



Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

- Iz predpostavk sledi, da je $\overline{X} \sim N(3348, 158^2)$
- Oziroma: $Z = \frac{\bar{X}-3348}{158} \sim N(0,1)$
- S 95 % verjetnostjo torej pričakujemo, da bo vzorčno povprečje na intervalu [3038,3658]
  $[3348 - 1{,}96 \cdot 158, 3348 + 1{,}96 \cdot 158]$
- S 95 % verjetnostjo pričakujemo, da bo Z na intervalu $[-1{,}96, 1{,}96]$



Primer - porodne teže 35-letnic

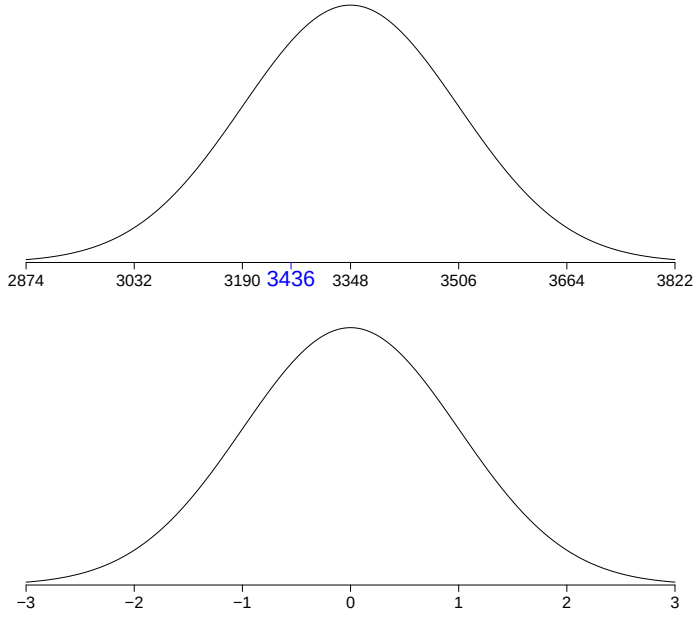
Statistično sklepanje

- Ničelna domneva, stopnja tveganja
  H_0 : Vzorec izhaja iz populacije s povprečjem 3348
  Populacija 35-letnic ima povprečje tež $\neq 3348$
  Stopnja tveganja $\alpha = 0{,}05$
- Porazdelitev testne statistike (predpostavke)
  Predpostavke: normalnost, σ, neodvisnost
  $Z = \frac{\overline{X}-3348}{158} \sim N(0,1)$
- Podatki, "odločitev" $\overline{x} = 3436$, $z = \frac{3436-3348}{158}$

Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

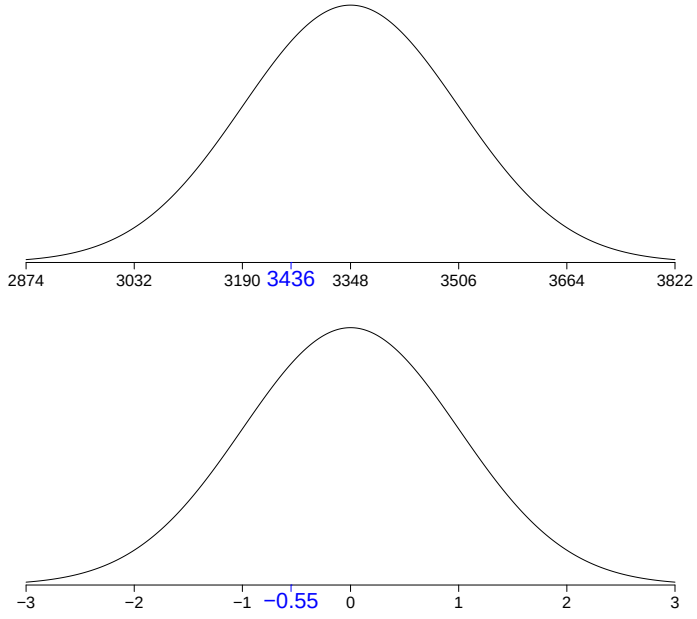
- $\overline{x} = 3436$



Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

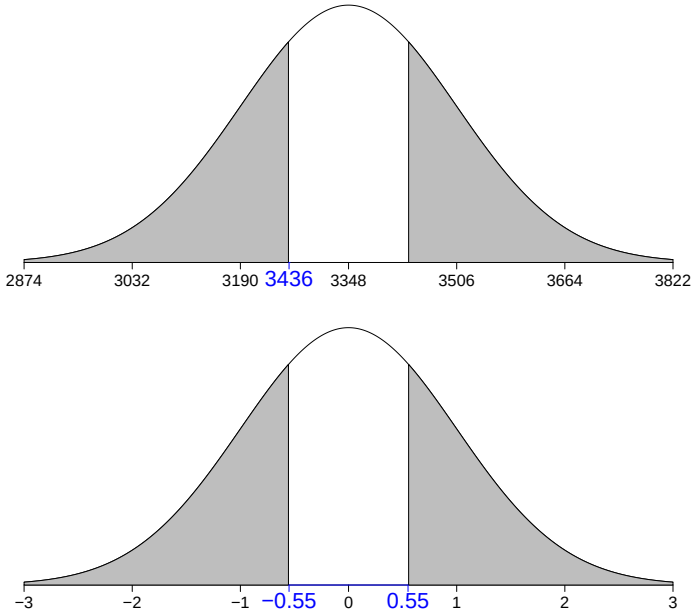
- $\overline{x} = 3436$
- $z = \frac{3436-3348}{158} = -0{,}55$



Porazdelitev testne statistike

Porazdelitev vzorčnega povprečja

- $\overline{x} = 3436$
- $z = \frac{3436-3348}{158} = -0{,}55$
- $P(Z \leq -0{,}55) + P(Z \geq 0{,}55) = 0{,}58$
- Ničelne domneve ne zavrnemo
- Verjetnost, da je do odstopanja prišlo zaradi naključne variabilnosti, je velika



Malo bolj splošno

z-test

- Zanima nas ali vzorec izhaja iz populacije s povprečjem μ_0
- Predpostavimo, da poznamo σ in da je populacija normalno porazdeljena
- Ničelna domneva: $H_0 : \mu = \mu_0$
- Porazdelitev vzorčnih povprečij je normalna $\overline{X} \sim N(\mu_0, \sigma^2)$
- Testna statistika Z je standardiziran odmik od μ_0
- $z = \frac{\overline{x} - \mu_0}{\sigma_0 / \sqrt{n}}$
- Vrednosti $|z|$, ki so večje od 1,96 imajo verjetnost $< 0{,}05$

Opombe

Statistični test

- Pravimo, da je test dvostranski , če nas zanimajo odmiki v katerokoli smer
- Verjetnosti, da ničelna domneva drži, pravimo *p* vrednost
- Ničelno domnevo zavrnemo, če $p < \alpha$
- Če je $p \geq \alpha$ ostajamo pri ničelni domnevi
- Ničelne domneve ne moremo dokazati

Statistična značilnost

testna statistika, p vrednost

- Pri vsakem statističnem testu nas bo zanimala neka slučajna spremenljivka
  primer: standardizirani odmik od povprečja 3348
- Bistveni korak bo poznavanje njene porazdelitve
  primer: $Z \sim N(0,1)$
- V podatkih bomo opazili neko odstopanje od ničelne domneve
  primer: $\overline{x} \neq \mu_0$, $3436 \neq 3348$
- Verjetnost tega odstopanja bomo imenovali *p*-vrednost
  primer: $p = 0{,}58$
- Če je $p < \alpha$ ničelno domnevo zavrnemo, pravimo, da je rezultat statistično značilen

Statistična značilnost

Pomen statistične značilnosti

- Na vzorcu bomo vedno opazili odmik od ničelne domneve
- Zanima nas ali obstajajo razlike tudi v populaciji
- Če je malo verjetno, da bi odmik nastal po naključju pravimo, da je statistično značilen
- Če je odmik statistično značilen, pomeni, da obstaja velika verjetnost, da je odmik tudi v populaciji
- Pozor: ničesar nismo rekli o velikosti odmika!

Statistično sklepanje - z test

Velikost vzorca

- Recimo, da na vzorcu opazimo odmik 88 g. Je to velik odmik?
- Pri odločanju bistveno vlogo igra standardna napaka
- Naš vzorec je bil majhen ($n = 10$), zato je bila standardna napaka velika - 158
  Standardizirani odmik je bil zato $\frac{88}{158} = 0{,}55$
  Pri majhnem vzorcu je tako nihanje pričakovano zaradi naključne variabilnosti
- Kaj bi rekli, če bi enak odmik opazili na vzorcu velikosti 1000?
  $SE = \frac{500}{\sqrt{1000}} = 15{,}8$
  Opaženi odmik je več kot 5 standardnih napak proč od μ_0
  Verjetnost takšnega odmika po naključju je zelo majhna

Naša želja je oceniti populacijsko povprečje iz katerega izhaja vzorec oz. ga primerjati z nekim drugim povprečjem. Ker populacijskega povprečja ne poznamo, je precej nerealno pričakovati, da bomo poznali populacijsko varianco. V prejšnjem primeru (z test) smo predpostavili, da je enaka varianci tiste populacije, s katero našo primerjamo. Tej predpostavki se bomo skušali izogniti.

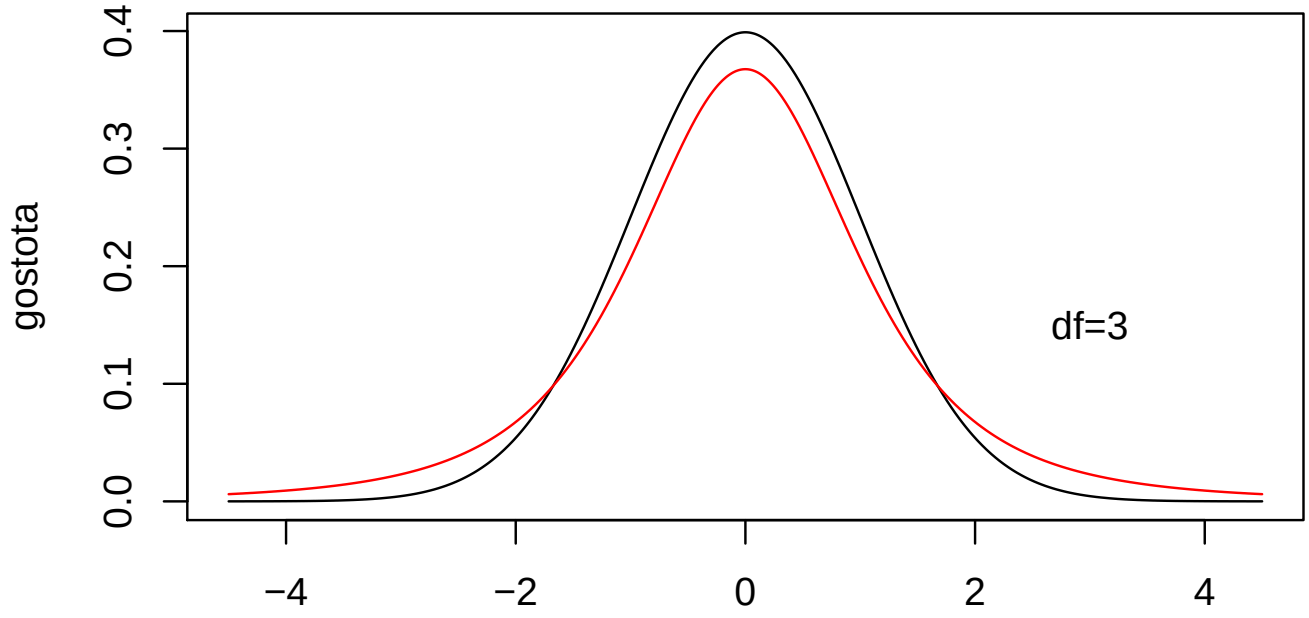
Analiza enega vzorca iz normalno porazdeljene populacije

V praksi bomo σ le redko poznali

- Vrednost σ potrebujemo za izračun standardne napake
- Največ kar lahko naredimo je, da jo ocenimo iz vzorca $\hat{\sigma} = s$
- To uporabimo za izračun $\hat{SE} = \frac{\hat{\sigma}}{\sqrt{n}}$
- Nekaj variabilnosti torej prinese tudi ta ocena
- Vrednost $\frac{\overline{X}-\mu}{\hat{SE}}$ ni porazdeljena normalno
- Intuitivno pričakujemo nekoliko več ‘čudnih’ vrednosti
- Če je vzorec velik, je ocena $\hat{\sigma}$ dobra, zato kakih velikih razlik ne pričakujemo

Porazdelitev *t*

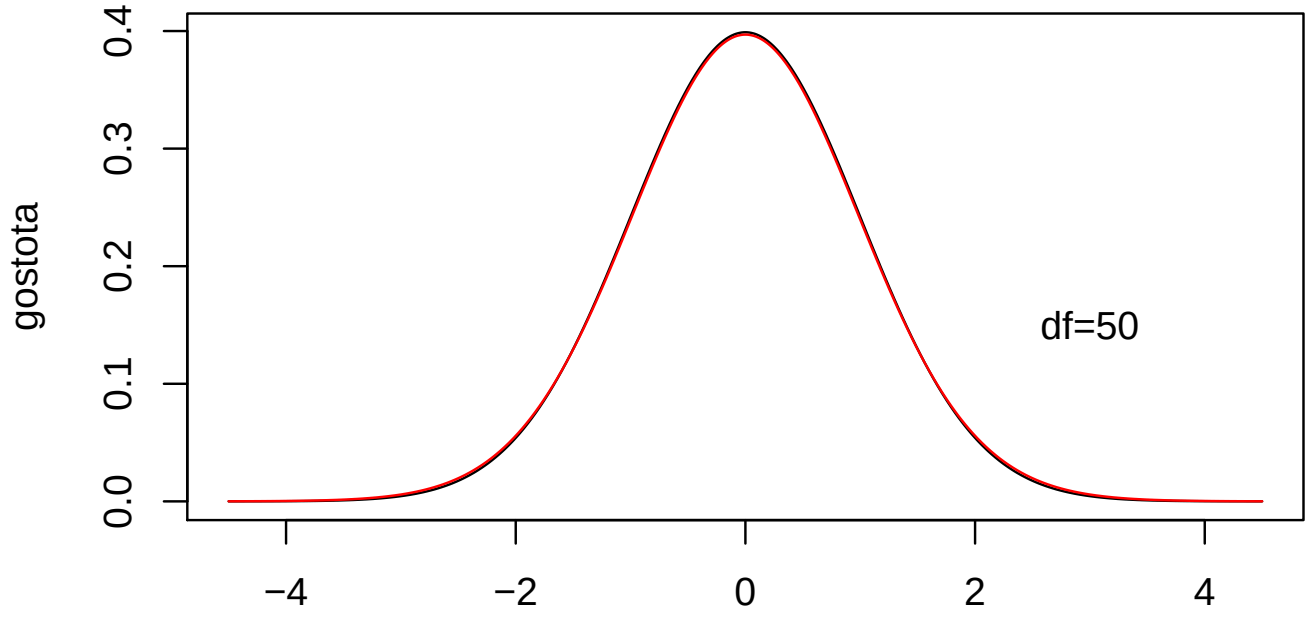
Simetrična porazdelitev, ki je podobna normalni, ima pa bolj debele repe



- Oblika je odvisna od velikosti vzorca
- Uporabljali bomo izraz ‘stopinje prostosti’ ($df = n - 1$)
- Večji kot je vzorec, bolj je *t* porazdelitev podobna normalni

Porazdelitev *t*

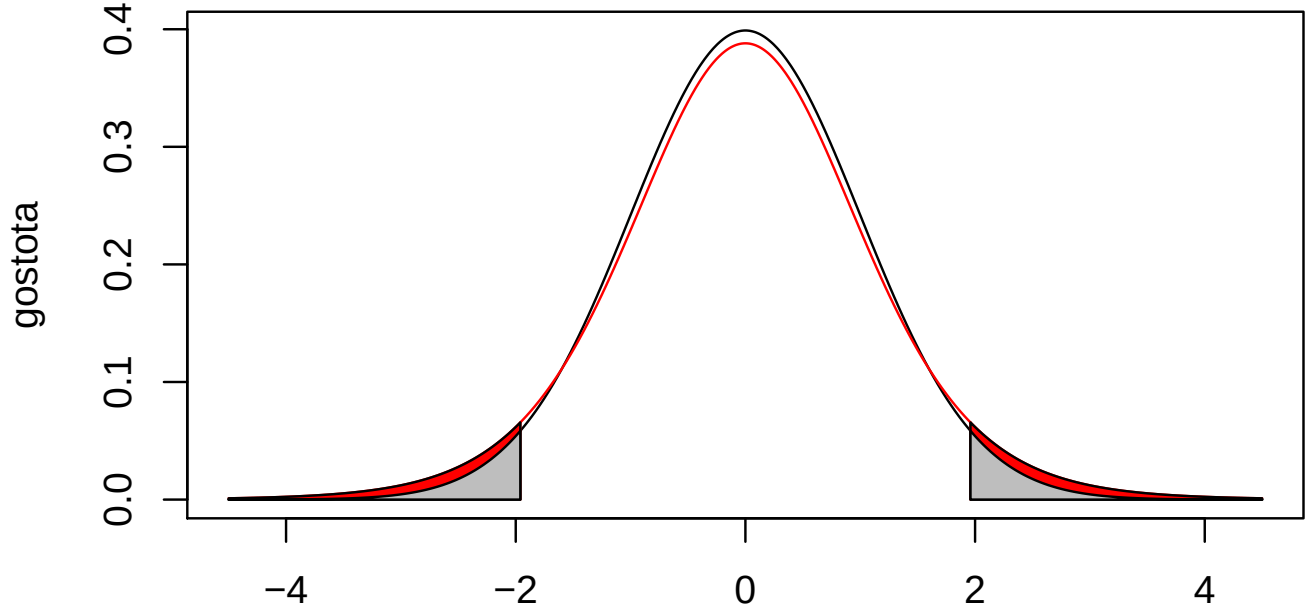
Simetrična porazdelitev, ki je podobna normalni, ima pa bolj debele repe



- Oblika je odvisna od velikosti vzorca
- Uporabljali bomo izraz ‘stopinje prostosti’ ($df = n - 1$)
- Večji kot je vzorec, bolj je *t* porazdelitev podobna normalni

Porazdelitev *t*

Simetrična porazdelitev, ki je podobna normalni, ima pa bolj debele repe



- Verjetnost, da so vrednosti *t* pri devetih stopinjah prostosti za več kot 1,96 oddaljene od 0 je 0,131
- To je precej več kot 0,05 pri normalni porazdelitvi
- Ustrezna verjetnost pri dvajsetih stopinjah prostosti je 0,064 , pri petdesetih pa 0,056

Analiza enega vzorca iz normalno porazdeljene populacije

Izogniti se želimo predpostavkam o σ

- Vrednost variance ocenimo iz vzorca $\hat{\sigma} = s$
- To uporabimo za izračun $\hat{SE} = \frac{\hat{\sigma}}{\sqrt{n}}$
- Vrednost $\frac{\overline{X}-\mu}{\hat{SE}}$ je porazdeljena po Studentovi t porazdelitvi z $n-1$ stopinjami prostosti

Primer 35-letnic, t-test

- Ničelna domneva

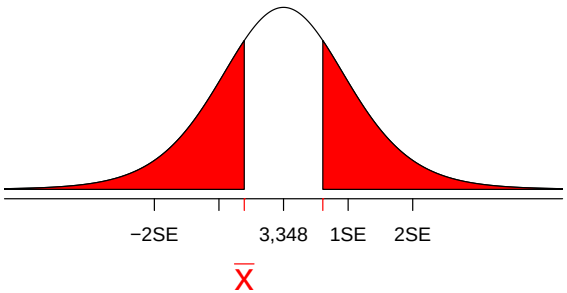
  Vzorec izhaja iz populacije s povprečjem 3348
- Porazdelitev testne statistike, predpostavke

  Predpostavke: normalnost, neodvisnost

  $T = \frac{\overline{X} - 3348}{\frac{457}{\sqrt{10}}} \sim t_9$
- Podatki, 'odločitev '

  Izračun testne statistike $t = 0{,}608$, $p = 0{,}56$

  Podatki ne nasprotujejo predpostavki, da je povprečna teža otrok 35-letnih mater enaka 3348 gramov.



Če ničelna hipoteza velja, potem lahko dobimo vrednosti, ki so vsaj za $3436g - 3348g = 88g$ oddaljene od povprečja v 56 % primerov.

Povzetek

- Izhajali smo iz **statističnega modela**, ki pravi, da so podatki slučajen vzorec neodvisnih meritev iz normalne porazdelitve
- **Ocenili** smo: $\hat{\mu} = \bar{x}$ in $\hat{\sigma} = s$
- **Testiranje ničelne domneve**, da je

$$H_0 : \mu = \mu_0$$

- **Testna statistika**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- t je eno opazovanje iz t-porazdelitve z $n - 1$ stopinjami prostosti

z-test in t-test

- Zanima nas ali vzorec izhaja iz populacije s povprečjem μ_0
- Predpostavimo: populacija normalno porazdeljena, σ poznamo ($= \sigma_0$)

  Populacija normalno porazdeljena, σ ocenimo z vzorca
- Ničelna domneva: $H_0 : \mu = \mu_0$
- Testna statistika $z = \frac{\overline{x} - \mu_0}{\sigma_0 / \sqrt{n}}$

  $t = \frac{\overline{x} - \mu_0}{\hat{\sigma} / \sqrt{n}}$
- Normalna porazdelitev

  t porazdelitev z $n - 1$ stopinjami prostosti
- Vrednosti $|z| > 1{,}96$ imajo verjetnost $< 0{,}05$

  $n = 10$, vrednosti $|t| > 2{,}26$ imajo verjetnost $< 0{,}05$

  $n = 50$, vrednosti $|t| > 2{,}01$ imajo verjetnost $< 0{,}05$

  $n = 500$, vrednosti $|t| > 1{,}96$ imajo verjetnost $< 0{,}05$

Predpostavka normalnosti

- Izhajali smo iz **statističnega modela**, ki pravi, da so podatki slučajen vzorec neodvisnih meritev iz normalne porazdelitve
- Kako preverimo to predpostavko?
- Slika prikazuje 16 vzorcev velikosti 10 iz $\mathcal{N}(3348,500)$

Predpostavka normalnosti

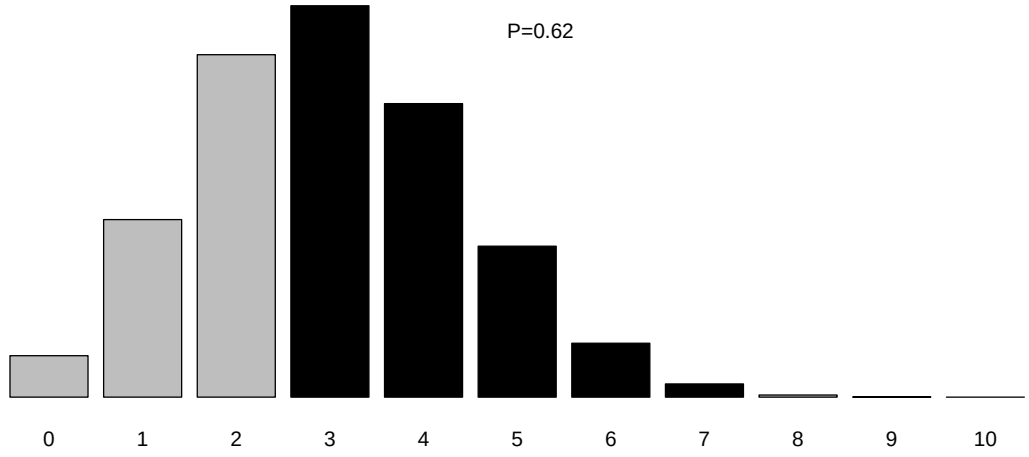
- Ali predpostavko o normalnosti res potrebujemo?
- Ne, če je vzorec dovolj velik
- Centralni limitni izrek: Porazdelitev povprečij vzorcev iz poljubno porazdeljene populacije konvergira proti normalni porazdelitvi
- Torej: pri dovolj velikih vzorcih predpostavka o normalnosti za test *t* ni več tako pomembna
- Vseeno je osnovno vprašanje: ali je primerjava povprečij smiselna?

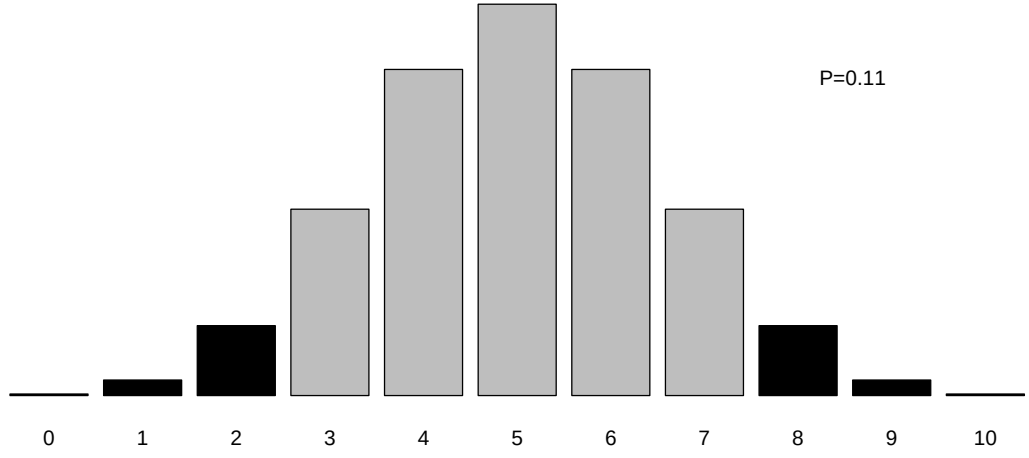
Napake pri statističnem sklepanju

- Pri zavračanju ničelne hipoteze lahko naredimo napako, verjetnost te napake je α
- Kaj če ničelne hipoteze ne zavrnemo? Ali naj jo potem sprejmemo?
- Primer: desetim ljudem damo zdravilo, ki naj bi zdravilo v 50% primerov
- $H_0 : \pi = 0{,}5$

Napake pri statističnem sklepanju

Primer: zdravilo uspešno v 50% primerov





Napaka 1. in 2. vrste

Ker je ničelna hipoteza ali pravilna ali pa nepravilna, lahko naredimo dve napaki:

1. da zavrnemo ničelno domnevo, ki je pravilna (**napaka 1. vrste**),
2. da sprejmemo napačno ničelno domnevo (**napaka 2. vrste**).

Verjetnost napake 1. vrste poznamo (stopnja tveganja), verjetnosti napake 2. vrste pa ne.

Napaka 1. in 2. vrste

Sodišče

- Resnica?
- Obdolženi ni kriv
- Dokazi
- Kriv je le, če to dokažemo onkraj razumnega dvoma (ostane zelo majhna verjetnost, da ni kriv)

- Napaka 1. vrste: obsojen, čeprav ni kriv
- Napaka 2. vrste: zločinec na prostosti

Statistika

- Populacija?
- H_0 : Ni razlik v populaciji
- Na vzorcu je vedno razlika
- Zavrnemo le, če je $p < \alpha$

- Napaka 1. vrste: zavrnemo ničelno domnevo, v populaciji ni razlik
- Napaka 2. vrste: ne zavrnemo ničelne domneve, v populaciji so razlike

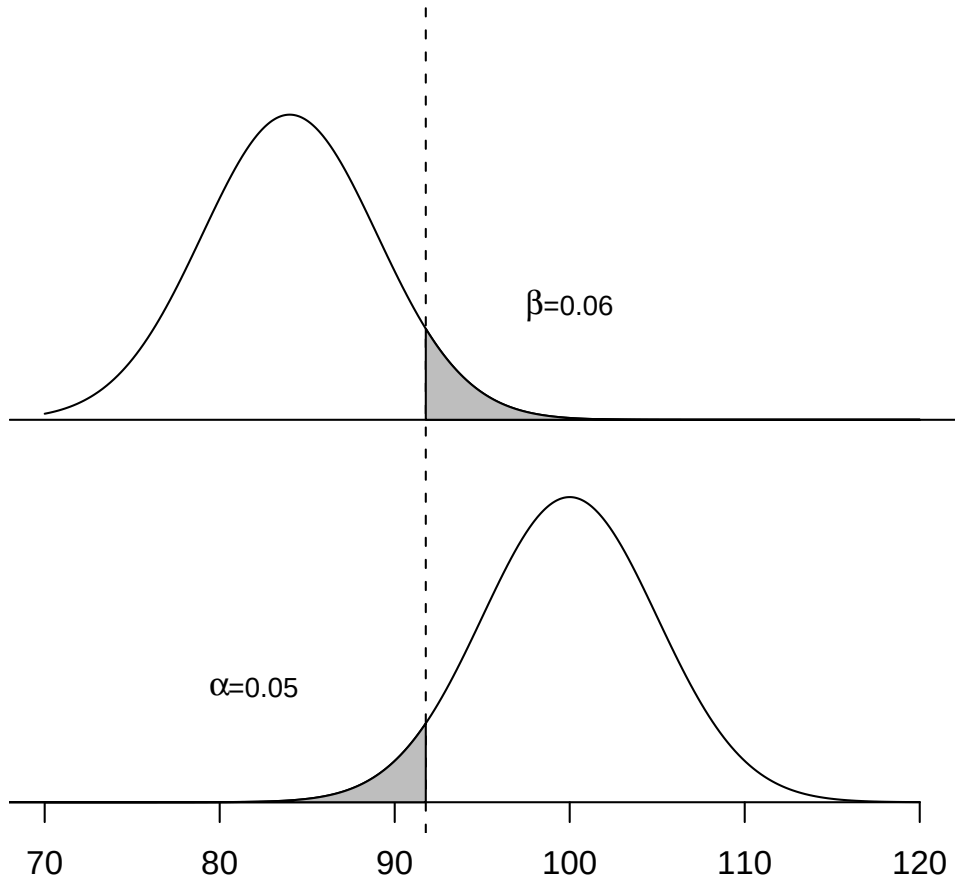
Napaka 1. in 2. vrste

Koliko je verjetnost napake 2. vrste je odvisno od:

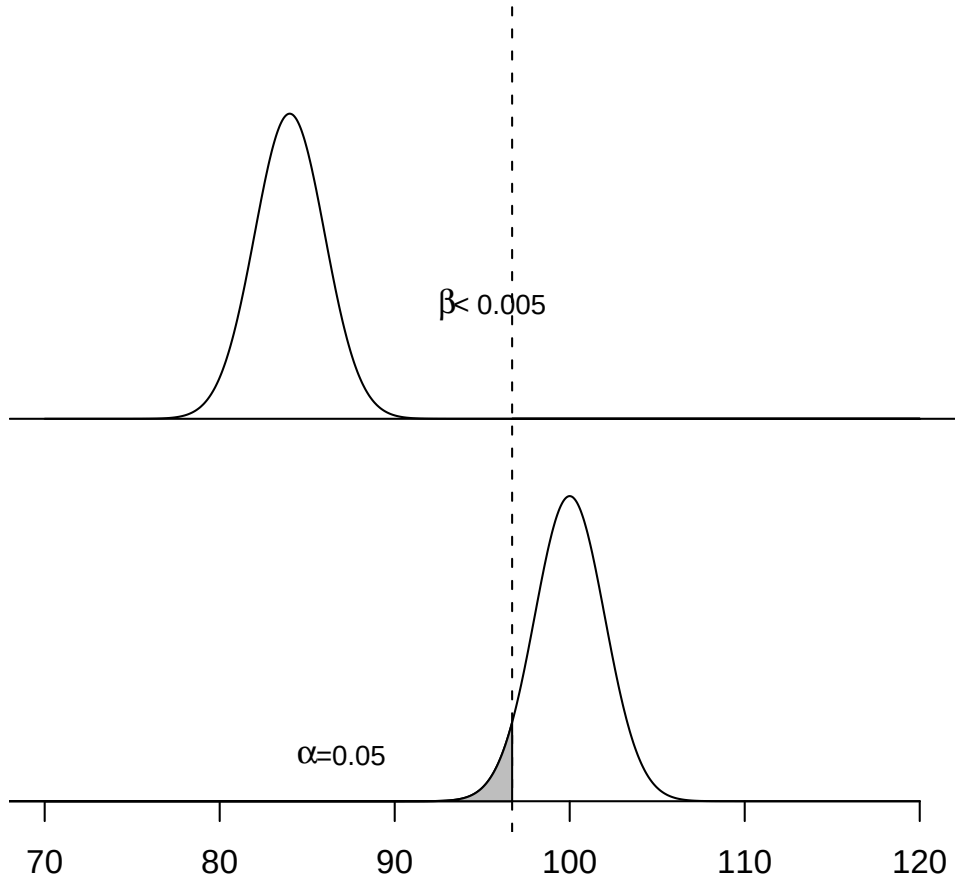
1. od stopnje značilnosti (torej od dopuščene napake 1. vrste),
2. od dejanskega stanja v populaciji (povprečje, varianca, delež, . . .)
3. od velikosti vzorca, ker ta vpliva na standardno napako.

Opozorilo! Ničelne hipoteze (skoraj) **nikoli ne sprejemamo**!
Vse, kar rečemo, je, da podatki ne nasprotujejo ničelni domnevi.

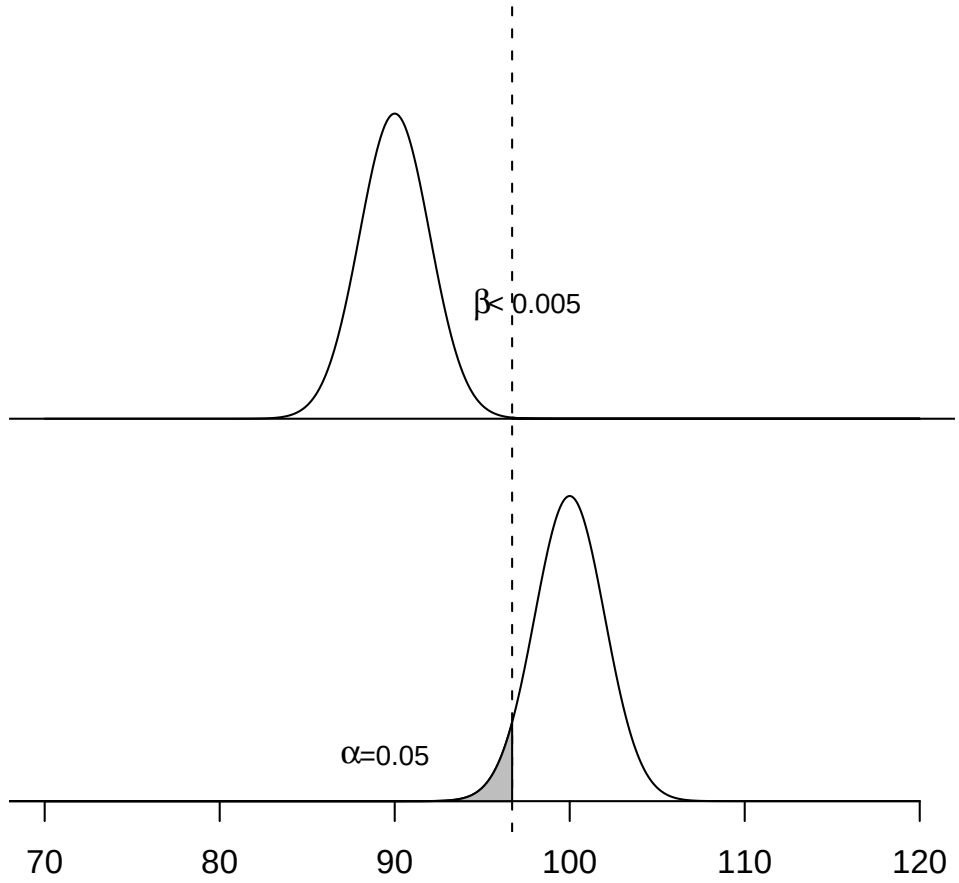
Napaka 2. vrste - vpliv povprečja



Napaka 2. vrste - vpliv povprečja ob manjši varianci



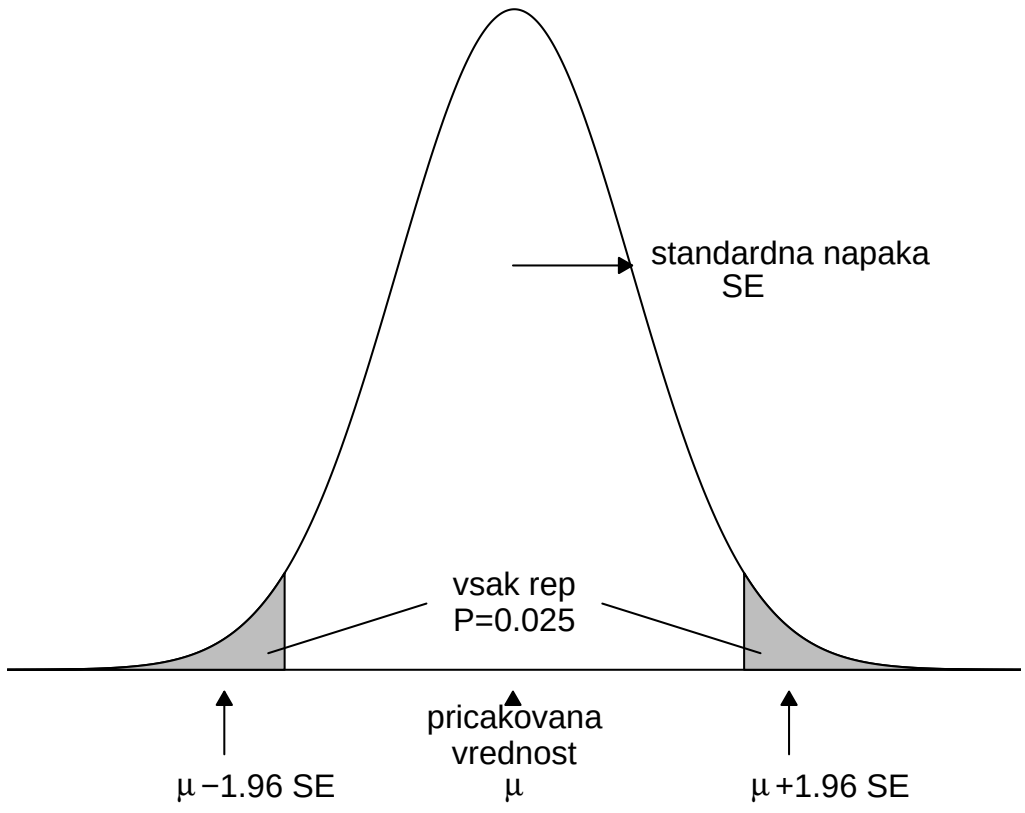
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



Interval zaupanja

Porazdelitev vzorčnih povprečij

$$SE = \frac{\sigma}{\sqrt{n}}$$



Interval zaupanja

Verjetnost

Populacija ima povprečje μ

Vzorčna povprečja $\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Statistika

Okrog $\bar{x}$ tvorimo interval, za katerega upamo, da zajema μ

Na vzorcu ocenimo povprečje $\bar{x}$

Kako širok mora biti interval? Kakšno verjetnost imamo, da nam uspe?

Interval zaupanja

- Obstaja 95% verjetnost, da bosta μ in $\bar{X}$ oddaljena za manj kot 1,96 SE:

$$P(|\mu - \bar{X}| \leq 1{,}96SE) = 0{,}95$$

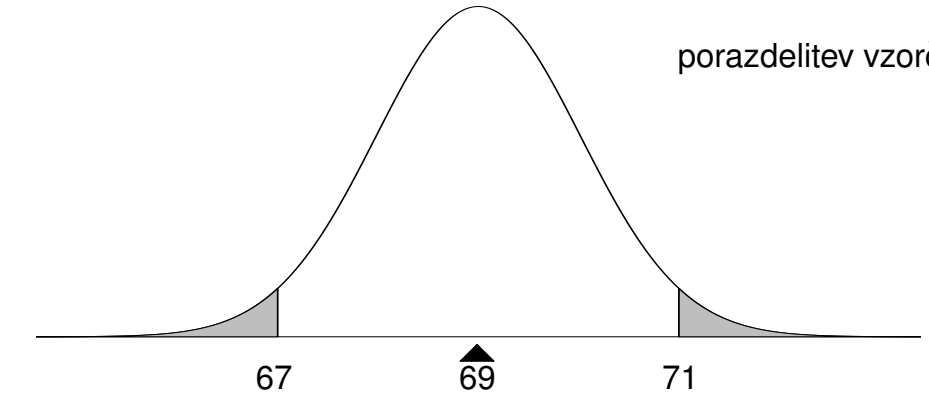
oziroma

$$P(\bar{X} - 1{,}96SE \leq \mu \leq \bar{X} + 1{,}96SE) = 0{,}95,$$

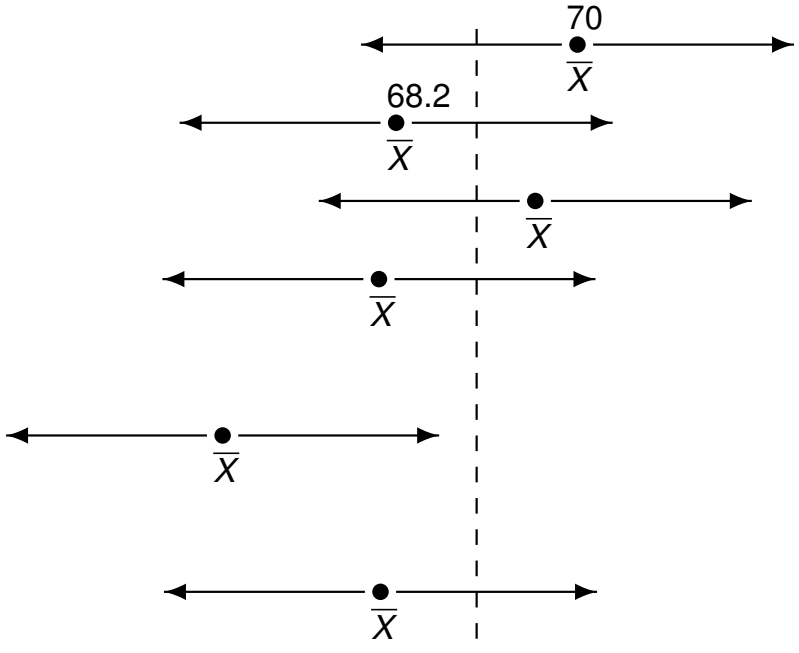
- Torej: razdalja od $\bar{x}$ do vrednosti, ki nas zanima, je v 95 % manj kot $1{,}96SE$
- Če jemljemo večkrat vzorce velikosti n enot iz populacije, bo interval $(\bar{X} - 1{,}96SE, \bar{X} + 1{,}96SE)$ v 95% primerov pokril μ.
- Temu intervalu rečemo 95% **interval zaupanja za vzorčno povprečje**.

Interval zaupanja

Primer: Teža žensk, $\mu = 69$, $\sigma = 12{,}5$, $n = 150 \Rightarrow SE = 1{,}02$



kar vemo mi, statistik pa ne



prvi interval

drugi interval

tretji interval

in tako naprej (zaenkrat vsi vsebujejo μ)

eden redkih zgreškov

petdeseti interval

statistikove ocene intervalov

Interval zaupanja

V praksi je seveda takole:

1. Ocenimo le **en** interval zaupanja.
2. Ciljni μ ni viden.

Interpretacija intervala zaupanja

- Če jemljemo večkrat vzorce velikosti n enot iz populacije, bo interval $(\bar{X} - 1{,}96SE, \bar{X} + 1{,}96SE)$ v 95% primerov pokril μ.
- Katerokoli vrednost v intervalu zaupanja izberemo kot možen μ, ji naš interval ne bo nasprotoval.

Interval zaupanja - σ ne poznamo

- Do sedaj smo intervale zaupanja računali, kot da σ poznamo
- Praviloma varianco ocenimo z vzorca $\Rightarrow$ porazdelitev t
- t_α: tista vrednost porazdelitve t z $n-1$ stopinjami prostosti, zunaj katere je 5% vseh vrednosti te porazdelitve. Velja:

$$P(|\mu - \bar{X}| \leq t_\alpha \frac{s}{\sqrt{n}}) = 0{,}95$$

- Interval zaupanja:

$$(\bar{X} - t_\alpha \frac{s}{\sqrt{n}}, \bar{X} + t_\alpha \frac{s}{\sqrt{n}})$$

- Ti intervali bodo seveda različno dolgi pri različnih vzorcih, a bodo še vedno v 95% pokrili populacijsko vrednost.

Interval zaupanja - primer

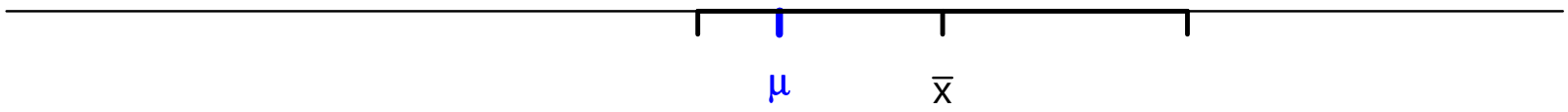
Interval zaupanja za porodno težo otrok 35-letnih mater

- $\bar{x} = 3436$, $s = 460$
- $\hat{SE} = \frac{460}{\sqrt{10}} = 145$
- $t_\alpha = 2{,}26$ (9 stopinj prostosti)
- 95% interval zaupanja za porodno težo

$$[3436 - 2{,}26 \cdot 145,\ 3436 + 2{,}26 \cdot 145] = [3109,\ 3763]$$

Interval zaupanja in *t* test

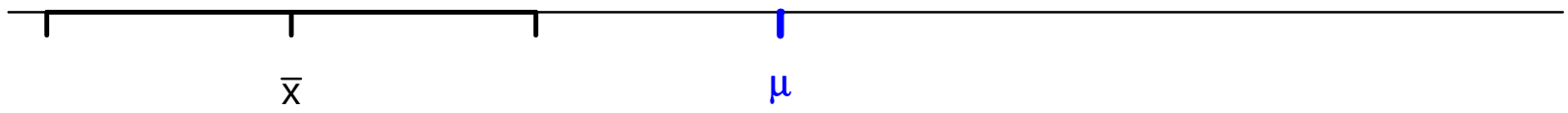
H_0 : vzorec izhaja iz porazdelitve s povprečjem μ



% interval zaupanja $\Leftrightarrow$ stopnja tveganja $\alpha = 0{,}05$

Interval zaupanja in *t* test

H_0 : vzorec izhaja iz porazdelitve s povprečjem μ



95 % interval zaupanja $\Leftrightarrow$ stopnja tveganja $\alpha = 0{,}05$

Opisne spremenljivke - ocenjevanje deležev

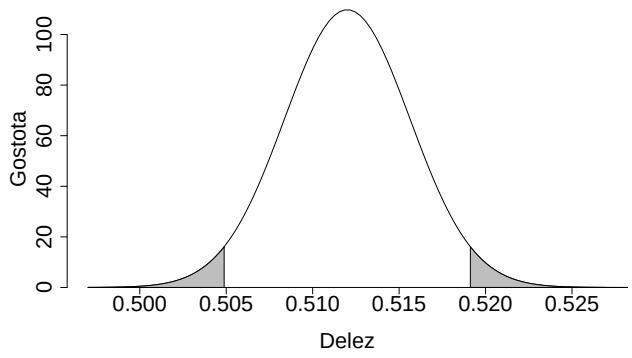
- Radi bi ocenili delež v populaciji oz. testirali neko vrednost.
- Primer:Povprečno razmerje spola novorojenčkov je 105:100, torej je verjetnost, da se rodi deček 0,512. Zanima nas ali je v Sloveniji razmerje kaj drugačno.
- Vemo, da se število dečkov porazdeljuje po binomski
- $X \sim Bin(n,\pi)$, $E(X) = n\pi$, $Var(X) = n\pi(1-\pi)$
- Delež izračunamo kot $\frac{X}{n}$
- $E(\frac{X}{n}) = \pi$, $Var(\frac{X}{n}) = \frac{\pi(1-\pi)}{n}$
- Za testiranje in intervale zaupanja bomo uporabili normalno aproksimacijo

Opisne spremenljivke - ocenjevanje deležev

Primer: porazdelitev spola novorojenčkov v Sloveniji

- Ničelna domneva
  V slovenski populaciji je delež dečkov 0,512
- Testna statistika in njena porazdelitev

Imamo vzorec velikosti 18932
Pod ničelno domnevo pričakujemo, da bo delež dečkov porazdeljen normalno z
$\mu = \pi = 0{,}512$ in
$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}} = 0{,}0036$



- Podatki, 'odločitev'
  Leto 2006: 18932 rojstev, delež dečkov 0,516
  Ničelne domneve ne zavrnemo, $p = 0{,}27$

Delež - test in interval zaupanja

Primer: porazdelitev spola novorojenčkov v Sloveniji

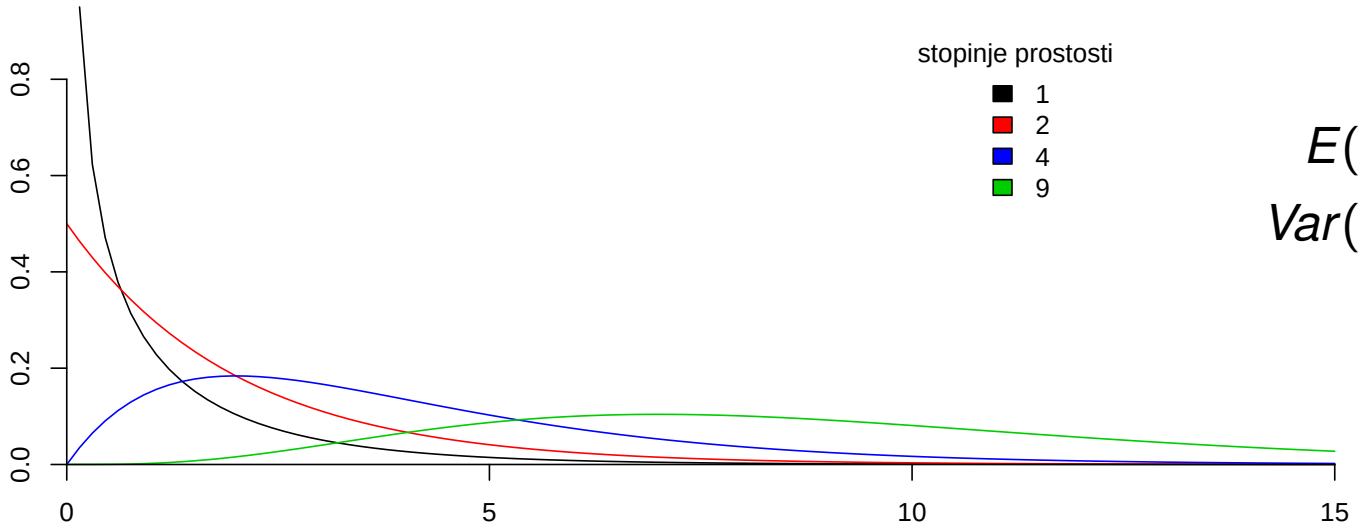
- Ocenjeni delež na vzorcu 0,516, standardni odklon $\sqrt{\frac{0{,}516(1-0{,}516)}{18932}} = 0{,}0036$
- 95 % interval zaupanja: $[0{,}512 - 1{,}96 \cdot 0{,}0036, 0{,}512 + 1{,}96 \cdot 0{,}0036]$
- 95 % interval zaupanja: $[0{,}508, 0{,}523]$
- V 95 % intervalu zaupanja je tudi vrednost 0,512
- Pri različnih rasah in kulturah se delež giblje med 103 in 107 na 100. Naš interval zajema vrednosti od 104 do 107 na 100. Za bolj natančno določitev bi potrebovali večji n
- Opomba: za računanje z deleži moramo vedno poznati tudi velikost vzorca

Ocenjevanje variance

- Kaj lahko rečemo o oceni variance na vzorcu?
- Primer: Zanima nas populacija dojenčkov 35-letnic. So njihove teže bolj razpršene?
- Namesto normalne (oz. t) bomo tu potrebovali χ^2 porazdelitev

Porazdelitev χ^2

- $Z \sim \mathcal{N}(0,1) \Rightarrow Z^2 \sim \chi^2_{(1)}$
- $Z_i \sim \mathcal{N}(0,1) \Rightarrow Z_1^2 + \ldots + Z_n^2 \sim \chi^2_{(n)}$



$$\begin{aligned} Y &\sim \chi^2_{(f)} \\ E(Y) &= f \\ Var(Y) &= 2f \end{aligned}$$

Ocenjevanje variance

Velja torej

- Če so X_i vrednosti iz $\mathcal{N}(\mu,\sigma^2)$
- Za vsak i velja $\frac{X_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$
- In $\left(\frac{X_1-\mu}{\sigma}\right)^2 + \ldots + \left(\frac{X_n-\mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$

Velja pa tudi

- $\left(\frac{X_1-\bar{X}}{\sigma}\right)^2 + \ldots + \left(\frac{X_n-\bar{X}}{\sigma}\right)^2 \sim \chi^2_{(n-1)}$

Ocenjevanje variance

Uporaba teh lastnosti

1. Ko varianco ocenjujemo na vzorcu, moramo deliti z $(n-1)$!

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2_{(n-1)} \Rightarrow \frac{1}{\sigma^2}E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] = n-1 \Rightarrow E(s^2)=\sigma^2$$

2. Za test variance lahko uporabimo porazdelitev χ^2

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2_{(n-1)} \Rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

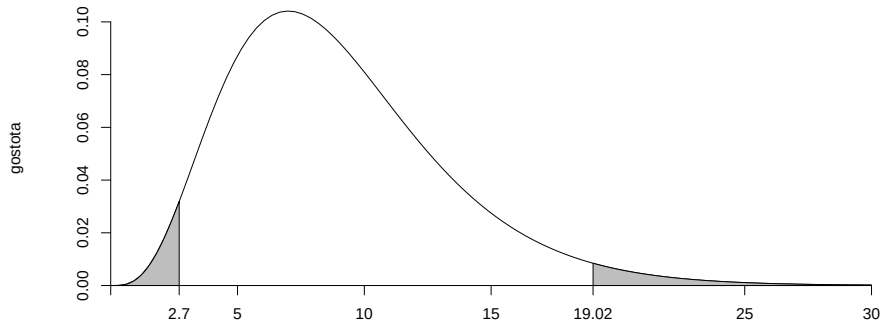
Ocenjevanje variance

Primer - porodne teže 35-letnic

- Ničelna domneva

  Vzorec izhaja iz populacije z varianco 500^2
- Testna statistika in njena porazdelitev

  Pod ničelno domnevo pričakujemo, da bo $\frac{9s^2}{\sigma^2}$ porazdeljen kot $\chi^2_{(9)}$



- Podatki, 'odločitev'

  $s^2 = 457^2$, testna statistika $\chi^2_u = \frac{9 \cdot 457^2}{500^2} = 7{,}51$

  Ničelne domneve ne zavrnemo

Testiranje porazdelitve opisne spremenljivke

Primer - prebivalci tropskega otoka

- Zanima nas ali je porazdelitev krvnih skupin na nekem tropskem otoku enaka kot v ameriški populaciji
- Porazdelitev v populaciji:

A	B	AB	0
0,4	0,1	0,1	0,4

- Porazdelitev na otoku:

A	B	AB	0
195	120	60	125
0,39	0,24	0,12	0,25

Testiranje porazdelitve opisne spremenljivke

Primer - prebivalci tropskega otoka

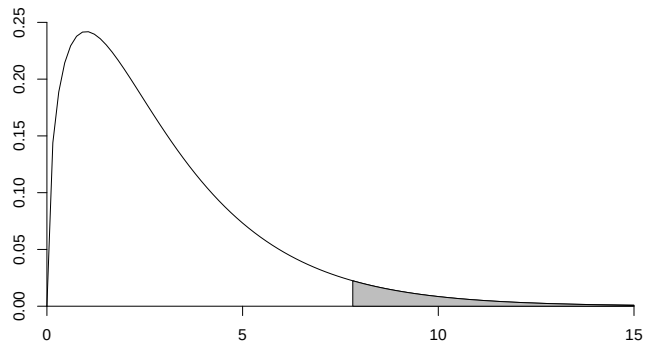
- Ničelna domneva

  Porazdelitev na tropskem otoku je enaka porazdelitvi v ameriški populaciji

- Testna statistika in njena porazdelitev

Pod ničelno domnevo pričakujemo naslednje frekvence (500 otočanov):

A	B	AB	0
200	50	50	200



Testna statistika: $\sum \frac{(\text{opazovane} - \text{pricakovane})^2}{\text{pricakovane}}$

Porazdelitev: χ^2 s 3 stopinjami prostosti

- Podatki, 'odločitev'

  testna statistika $\chi^2_u = 129{,}4$, ničelno domnevo zavrnemo, $p < 0{,}001$

Primerjava dveh skupin

Povezanost med številsko in opisno spremenljivko

številska spremenljivka	$\leftrightarrow$	opisna spremenljivka
starost	$\leftrightarrow$	spol
pritisk	$\leftrightarrow$	zdravilo
teža	$\leftrightarrow$	depresija

- Spremenljivki sta lahko povezani na različne načine
- Testna statistika odvisna od ničelne domneve

Primerjava varianc

- Ničelna domneva

  Vzorca izhajata iz populacij z enako varianco
- Testna statistika in njena porazdelitev

  Pod ničelno domnevo pričakujemo, da bo kvocient varianc $(\frac{\sigma_1^2}{\sigma_2^2})$ enak 1

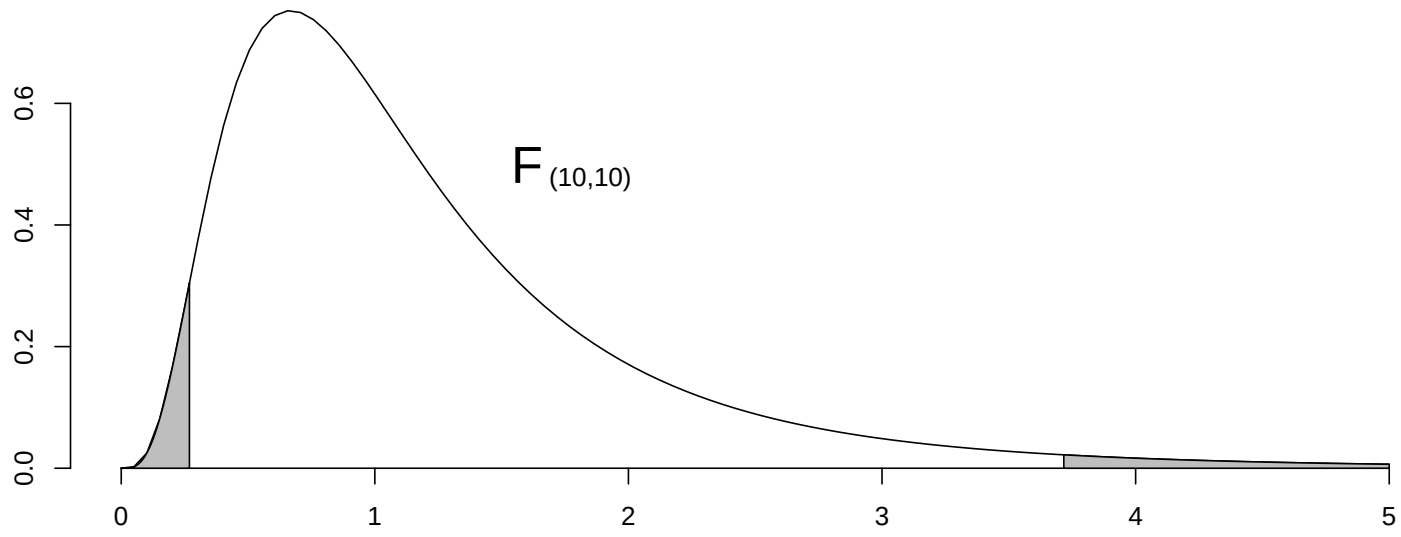
  Poznamo porazdelitev odstopanj od 1: $\mathcal{F}$ porazdelitev
- Podatki, 'odločitev'

  Če so odstopanja $\frac{s_1^2}{s_2^2}$ od 1 'velika' , ničelno domnevo zavrnemo

Porazdelitev $\mathcal{F}$

Kakšna odstopanja pričakujemo?

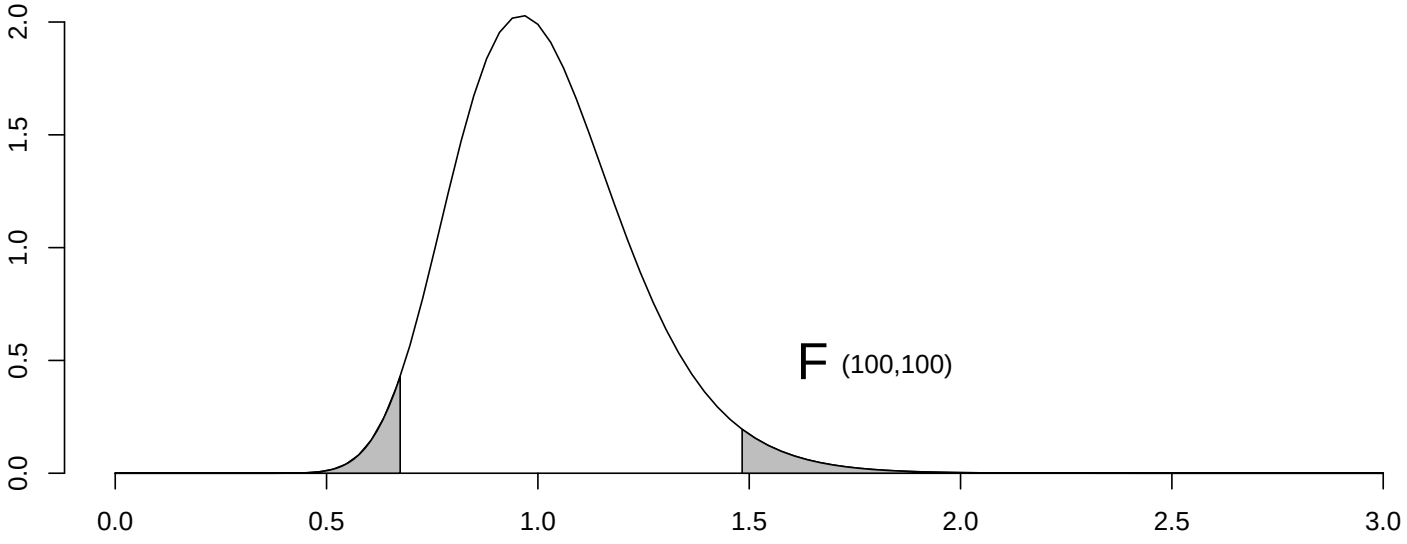
- Dlje kot smo od 1, bolj je ‘sumljivo’
- Kvocient 2 še ne bo sumljiv pri majhnih vzorcih, bo pa pri velikih
- Pomembna je velikost obeh vzorcev - 2 parametra



Porazdelitev $\mathcal{F}$

Kakšna odstopanja pričakujemo?

- Dlje kot smo od 1, bolj je ‘sumljivo’
- Kvocient 2 še ne bo sumljiv pri majhnih vzorcih, bo pa pri velikih
- Pomembna je velikost obeh vzorcev - 2 parametra



Primerjava povprečij dveh skupin

številska spremenljivka $\leftrightarrow$ opisna spremenljivka

starost $\leftrightarrow$ spol

pritisk $\leftrightarrow$ zdravilo

H_0: spremenljivki nista povezani

povprečji skupin v populaciji sta enaki

$$\begin{array}{ccc} \mu_1 & = & \mu_2 \\ \downarrow & & \downarrow \\ \overline{x}_1 & \neq & \overline{x}_2 \end{array}$$

Na vzorcu bo praktično vedno razlika.

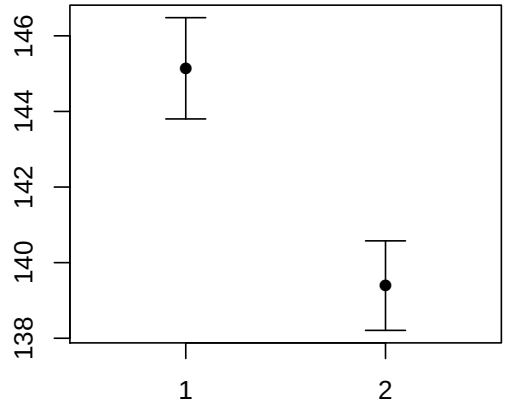
Vprašanje: ali (in s kakšnim tveganjem) lahko trdimo, da v populaciji obstajajo razlike?

Primerjava povprečij dveh skupin

Primer: raziskava PID-PAB

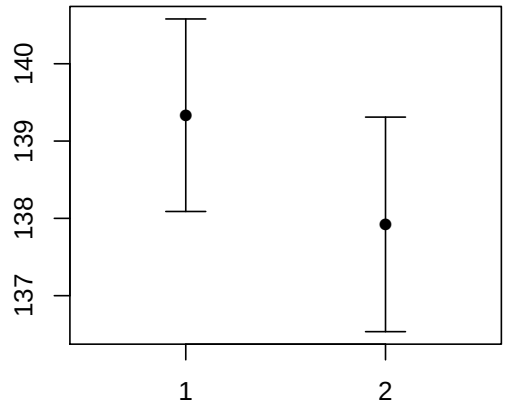
1. Primerjava povprečja sistoličnega tlaka pri bolnikih in kontrolah

	bolniki	kontrole	razlika
n	823	785	
$\bar{x}$	145,1	139,4	5,75
s	19,6	16,9	
$\hat{SE}$	0,68	0,60	



2. Primerjava povprečja sistoličnega tlaka pri kontrolah pri 1. in 2. pregledu

	1.pregled	2.pregled	razlika
n	694	694	
$\bar{x}$	139,3	137,9	1,4
s	16,7	18,6	
$\hat{SE}$	0,63	0,71	



Analiza parnih podatkov

Primer: raziskava PID-PAB

Uvedemo novo slučajno spremenljivko $D_i = X_{i1} - X_{i2}$
Primer 5 bolnikov:

bolnik	1.pregled	2.pregled	razlika
i	X_{i1}	X_{i2}	D_i
1	130	120	10
2	180	160	20
3	140	135	5
4	190	140	50
5	120	155	-35

- Ničelna domneva

  $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_D = 0$
- Testna statistika in njena porazdelitev

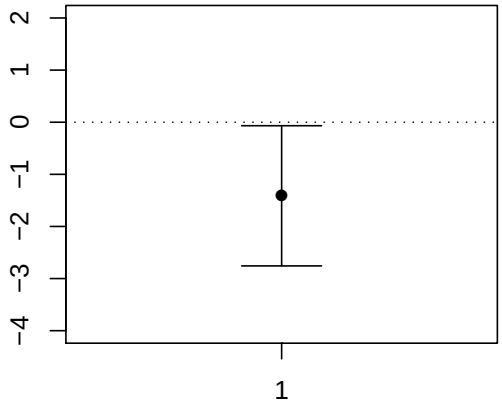
  $\frac{\bar{D}-0}{\hat{SE}_D} \sim t_{693}$
- Podatki, ‘odločitev’

Test t za parna vzorca

Primer: raziskava PID-PAB

Primerjava povprečja sistoličnega tlaka pri kontrolah pri 1. in 2. pregledu

	1.pregled	2.pregled	razlika
n	694	694	
$\bar{x}$	139,3	137,9	1,4
s	16,7	18,6	18,05
$\hat{SE}$	0,63	0,71	0,69



$$t = \frac{\bar{D} - 0}{SE_D} = \frac{1{,}4}{0{,}69} = 2{,}06$$

- Mejna vrednost t_{693} za $\alpha = 0{,}05$: $1{,}96 \Rightarrow p = 0{,}04$
- 95% Interval zaupanja za razliko: $[0{,}07, 2{,}76]$

Test t za neodvisna vzorca

Primer: raziskava PID-PAB

Primerjava povprečja sistoličnega tlaka pri bolnikih in kontrolah

	bolniki	kontrole	razlika
n	823	785	
$\bar{x}$	145,1	139,4	5,75
s	19,6	16,9	
$\hat{SE}$	0,68	0,60	

Kaj vemo o razliki $\bar{X}_1 - \bar{X}_2$:

- Razlika dveh normalnih spremenljivk je normalno porazdeljena
- Variance se seštevajo: $SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Primerjava povprečij dveh neodvisnih vzorcev

Če bi poznali populacijske vrednosti σ

- Ničelna domneva
  $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$
- Testna statistika in njena porazdelitev
  $\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$
- Podatki, 'odločitev'

V praksi seveda populacijski vrednosti σ_1 in σ_2 nista znani.
Težava:
Če ju nadomestimo s s_1 in s_2, porazdelitev ni ne $\mathcal{N}$ ne t_{n-1}

Primerjava povprečij dveh neodvisnih vzorcev

Recimo da v populaciji velja $\sigma_1 = \sigma_2$

Označimo vrednost obeh s σ. Potem velja

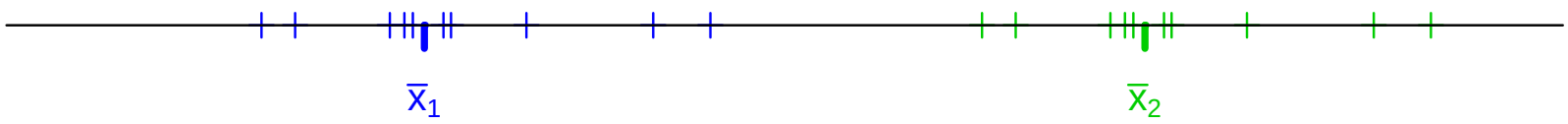
$$\frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

Če v gornjem izrazu σ zamenjam s s, je stvar spet porazdeljena kot t
Vprašanji:

- Kako izračunam s (imam namreč dva vzorca?)
- Število stopinj prostosti?

Test t, $\sigma_1 = \sigma_2 = \sigma$

Ocenjevanje σ na vzorcu



Odmiki od $\bar{x}_1$:

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$$

Odmiki od $\bar{x}_2$:

$$\sum_{i=1}^{n_2}(x_{1i} - \bar{x}_2)^2$$

Seštejemo in delimo s številom stopinj prostosti (eno smo porabili za izračun $\bar{x}_1$, eno za $\bar{x}_2$)

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \end{aligned}$$

Test t za neodvisna vzorca

Nazaj k našemu primeru

Primerjava povprečja sistoličnega tlaka pri bolnikih in kontrolah

	bolniki	kontrole	razlika
n	823	785	
$\bar{x}$	145,1	139,4	5,75
s	19,6	16,9	18,3
$\hat{SE}$	0,68	0,60	0,94

$$s^2 = \frac{822 \cdot 19{,}6^2 + 784 \cdot 16{,}9^2}{823 + 785 - 2} = 335{,}6$$

$$\hat{SE} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{335{,}6} \cdot \sqrt{\frac{1}{832} + \frac{1}{785}} = 0{,}94$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{SE}} = \frac{5{,}75}{0{,}94} = 6{,}09 \Rightarrow p = 1{,}4 \cdot 10^{-9}$$

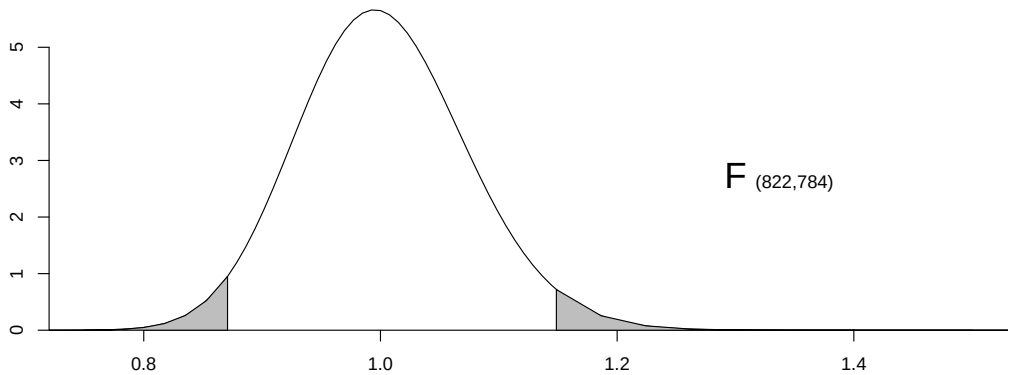
Test *t* za neodvisna vzorca

Izračun je bil narejen ob predpostavki $\sigma_1 = \sigma_2$

Je ta predpostavka smiselna?

$$s_1^2 = 19{,}6^2 \quad s_2^2 = 16{,}9^2$$

$$\frac{s_1^2}{s_2^2} = \frac{19{,}6^2}{16{,}9^2} = 1{,}34$$



Predpostavka o enakosti varianc je le malo verjetna. Kaj lahko storimo?

Test t za neodvisna vzorca, $\sigma_1 \neq \sigma_2$

Za oceno standardne napake uporabimo formulo:

$$\hat{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Porazdelitev $\frac{\bar{x}_1 - \bar{x}_2}{\hat{SE}}$ ni znana, uporabimo približke

V našem primeru:

$$\hat{SE} = \sqrt{\frac{19{,}6^2}{823} + \frac{16{,}9^2}{785}} = 0{,}93 \Rightarrow \; t = 6{,}31$$

Statistični paket SPSS nam da rezultat $p = 3{,}54 \cdot 10^{-10}$

Test t za neodvisna vzorca, povzetek

- Ničelna domneva

  $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

- Testna statistika in njena porazdelitev
  - varianci enaki $\Rightarrow s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$, $\frac{\bar{X}_1-\bar{X}_2}{\hat{SE}} \sim t_{(n_1+n_2-2)}$
  - varianci različni $\Rightarrow \hat{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, za porazdelitev uporabimo približke

- Podatki, 'odločitev'

Primerjava t-testa za odvisne in neodvisne vzorce

Primerjava povprečja sistoličnega tlaka pri kontrolah pri 1. in 2. pregledu

	1.pregled	2.pregled	odvisna	neodvisna
n	694	694		
$\bar{x}$	139,3	137,9	1,4	1,4
$\hat{SE}$	0,63	0,71	0,69	0,95
t			2,06	1,49
p			0,04	0,14

95 % Intervala zaupanja:

Za neodvisna vzorca: $[-0{,}45,\ 3{,}28]$

Upoštevamo odvisnost: $[0{,}07,\ 2{,}76]$

Neparametrični testi

- Uporabimo jih, kadar predpostavka o normalnosti ni smiselna, oz. ni smiselno primerjati povprečij
- Upoštevajo range namesto dejanskih vrednosti
- Ogledali si bomo 2 testa:
  1. Mann-Whitney (za neodvisne vzorce)
  2. Wilcoxon (za odvisne vzorce)

Primeri

- Laboratorijske vrednosti: trigliceridi : Povprečje 1,86, standardni odklon 1,33
- Splošno počutje, telesna dejavnost, ipd

Mann-Whitney oz. Wilcoxonov test vsote rangov

- Združimo vrednosti iz obeh skupin in jih uredimo po velikosti (ranžirna vrsta)
- Seštejemo range v vsaki skupini oz. gledamo njihovo povprečje
- Primerjamo opaženo vrednost s pričakovano

Primer - tekmovanje

Podatki:

Ime	Čas	Skupina	Rang	Ime	Čas	Skupina	Rang
Jože	90*s*	SLO	3	Helmut	87*s*	AUT	2
Ivan	85*s*	SLO	1	Fritz	93*s*	AUT	5
France	92*s*	SLO	4	Hans	95*s*	AUT	6
Lojze	130*s*	SLO	8	Jürgen	97*s*	AUT	7

Mann-Whitney oz. Wilcoxonov test vsote rangov

Primer - tekmovanje

- 8 tekmovalcev, vsota vseh rangov je $\frac{8*9}{2} = 36$
- Minimum v eni skupini je $1 + 2 + 3 + 4 = 10$
- Če sta skupini enaki, je v vsaki vsota 18 ($\frac{n_1 * n_2}{2}$)
- Vseh možnih permutacij (če ni delitve mest) je 8!
- Gledamo manjšo od obeh vrednosti, lahko tabeliramo vse možne izide in verjetnosti
- V našem primeru sta vsoti rangov 16 in 20
- Ničelne domneve ne moremo zavreči
- Pri velikih vzorcih lahko uporabimo normalno aproksimacijo - rezultat približno enak testu *t* na rangih

Mann-Whitney oz. Wilcoxonov test vsote rangov

Primer - trigliceridi pri bolnikih (820) in kontrolah (785)

Podatki:

Vrednost	Skupina	Rang		Vrednost	Skupina	Rang
0,30	2	1	.	9,40	2	1601
0,39	2	2	.	9,41	1	1602
0,40	2	3,5	.	9,42	1	1603
0,40	1	3,5	.	9,51	1	1604
0,49	2	5	.	11,40	1	1605

Primer

- Ničelna domneva
  H_0: Trigliceridi in skupina (bolniki, kontrole) niso povezani
- Testna statistika in njena porazdelitev
  Namesto vrednosti uporabimo range
  Pri velikih vzorcih uporabimo normalno aproksimacijo
- Podatki, 'odločitev'

	Skupina 1	Skupina 2
Vsota rangov	710197,5	583438,5
Povprečni rang	862,9	743,2

$p < 0{,}001$, Ničelno domnevo zavrnemo

Wilcoxonov test predznačenih rangov

- Izračuanamo razlike in njihovim absolutnim vrednostim dodelimo range (ničle izločimo)
- Seštejemo range pri pozitivnih in negativnih razlikah
- Če velja ničelna domneva se vsoti ne bosta bistveno razlikovali

Primer - tekmovanje, 1. in 2. tek

Podatki:

Ime	Čas		Čas	Razlika	Rang
Jože	90s	.	88s	2s	1
Ivan	85s	.	93s	-8s	3
France	92s	.	95s	-3s	2
Lojze	130s	.	97s	33s	4

Vsota pozitivnih rangov $=$ vsota negativnih rangov $=5$

Primerjava dveh neodvisnih deležev

Primer - učinkovitost zdravil

- Primerjamo zdravljenji A in B
- Rezultati:

Zdravljenje	Število pacientov	Uspeh
A	257	216
B	244	180
Skupaj	501	396

- Vprašanje: Ali je tveganje v dveh skupinah različno?
- Oziroma: Ali drži ničelna hipoteza

$$H_0 : \pi_1 = \pi_2 = \pi$$

Primerjava dveh neodvisnih deležev

Primer - učinkovitost zdravil

- O deležih smo že govorili - normalna aproksimacija binomske
- Uporabimo raje pristop, ki omogoča posplošitev na več skupin (opisna z večimi vrednostmi)
- Vrednosti zapišemo v kontingenčno tabelo:

Opazovane frekvence	Izid		
	Neuspeh	Uspeh	Skupaj
Zdravljenje A	41	216	257
B	64	180	244
Skupaj	105	396	501

- Vrednosti v tabeli: opazovane frekvence (f_o)
- Vsote po stolpcih in vrsticah: robne frekvence

Primerjava dveh neodvisnih deležev

- Ničelna domneva
  $H_0 : \pi_1 = \pi_2 = \pi$, spremenljivki nista povezani
- Testna statistika in njena porazdelitev
  Pod ničelno domnevo izračunamo pričakovane frekvence (f_p)
  Zanimajo nas odmiki f_o od f_p
- Podatki, 'odločitev'
  Če so odmiki 'veliki', ničelno domnevo zavrnemo

Primerjava dveh neodvisnih deležev

$H_0 : \pi_1 = \pi_2 = \pi$

Izračun pričakovanih frekvenc

Pričakovane frekvence	Izid		
	Neuspeh	Uspeh	Skupaj
Zdravljenje A	53,9	203,1	257
B	51,1	192,9	244
Skupaj	105	396	501

- Slučajna spremenljivka $X =$‘Število uspehov v skupini A’
- $X \sim Bin(257,\pi)$ $(E(X) = n\pi)$
- π ne poznamo, ga ocenimo iz podatkov: $\hat{\pi} = 396/501$
- Pričakovano število uspehov potem: $257 \cdot \frac{396}{501} = 203,1$
- Na enak način izračunamo tudi ostale f_p

Primerjava dveh neodvisnih deležev

$H_0 : \pi_1 = \pi_2 = \pi$

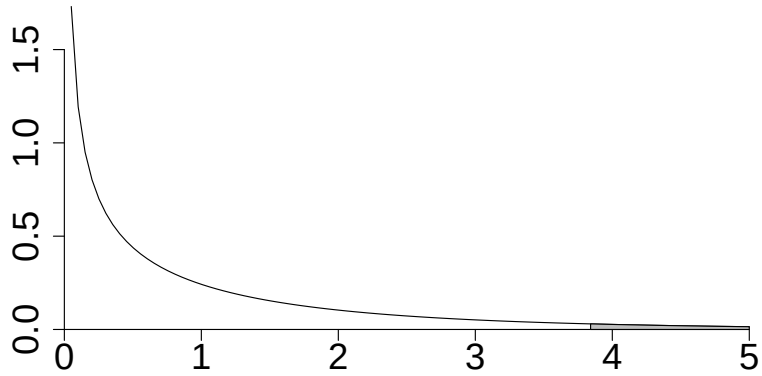
Testna statistika

- Zanimajo nas odmiki $f_p - f_o$
- Vsota odmikov je 0, zato jih kvadriramo
- Odmiki so pomembnejši pri manjših frekvencah - delimo s f_p
- $\sum \frac{(\text{opazovane} - \text{pricakovane})^2}{\text{pricakovane}} = \sum \frac{(f_o - f_p)^2}{f_p}$
- Testna statistika je porazdeljena kot $\chi^2_{(1)}$

Povezanost med dvema opisnima spremenljivkama

- Ničelna domneva
  $H_0 : \pi_1 = \pi_2 = \pi$, spremenljivki nista povezani
- Testna statistika in njena porazdelitev

$\chi^2_o = \frac{(f_o - f_p)^2}{f_p}$
Porazdelitev:
χ^2 z 1 stopinjo prostosti



- Podatki, ‘odločitev’

	Neuspeh	Uspeh
A	$\frac{(53,9-41)^2}{53,9}$	$\frac{(203,1-216)^2}{203,1}$
B	$\frac{(51,1-64)^2}{51,1}$	$\frac{(192,9-180)^2}{192,9}$

$$\chi^2_o = \sum_{i=1}^{4} \frac{(f_{oi} - f_{pi})^2}{f_{pi}} = 7,98$$

$p = 0,0047$
Ničelno domnevo zavrnemo

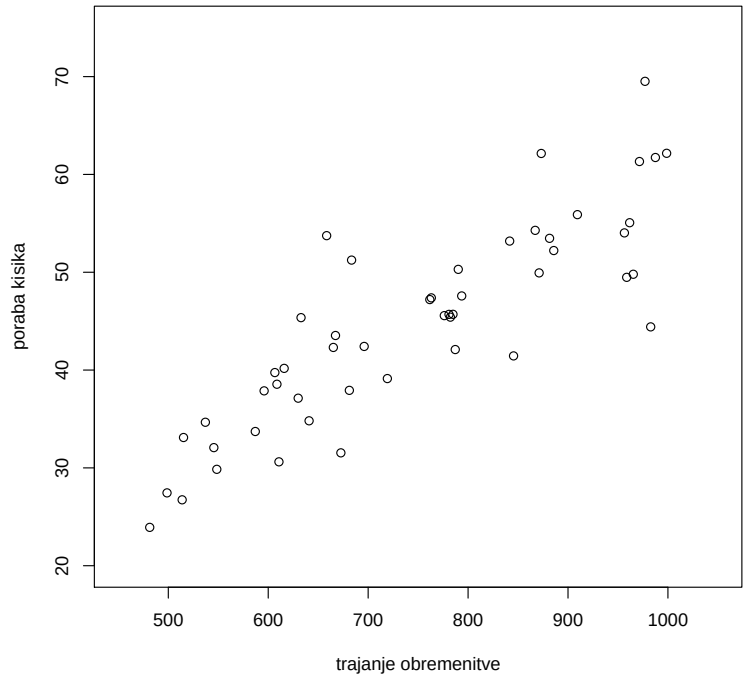
Povezanost med dvema opisnima spremenljivkama

Posplošitev na več skupin

- Če ima katera izmed spremenljivk več možnih vrednosti, bo ideja testa enaka
- V tabeli bo več polj, zato bo tudi več clenov v testni statistiki
- Ker seštevamo več členov, moramo dovoliti večja odstopanja - χ^2 z več stopinjami prostosti
- Število stopinj prostosti:
(število vrstic-1)(število stolpcev -1)
- Pozor: pri večih skupinah hitro naletimo na podskupine, v katerih so pričakovane frekvence premajhne
- Opomba: porazdelitev χ^2 je približek, dovolj dobra je, če so pričakovane frekvence > 5

Korelacija

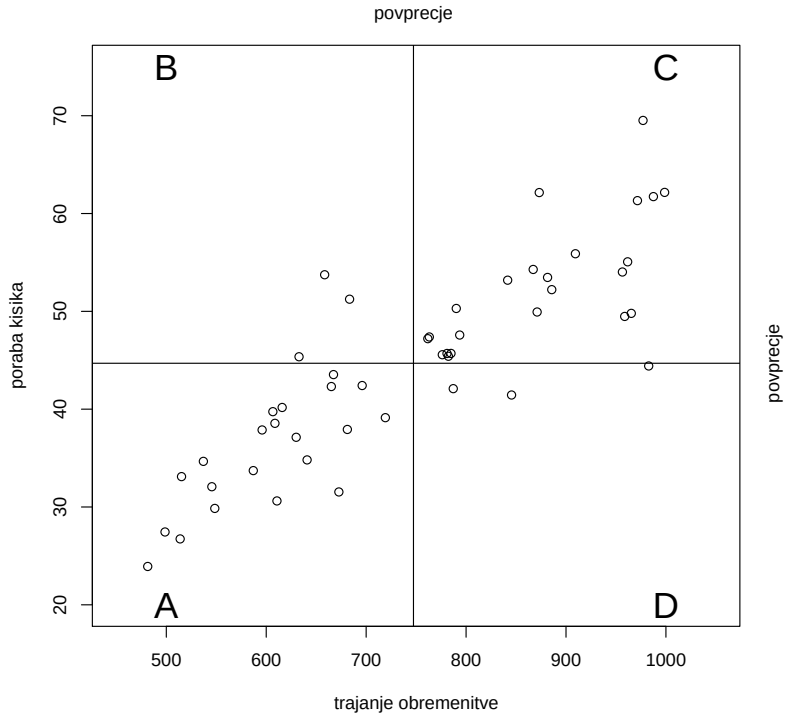
- Zanima nas povezanost med dvema številskima spremenljivkama
- Imamo pare podatkov (x_i, y_i), kjer so x_i vrednosti spremenljivke X, y_i pa vrednosti spremenljivke Y.
- Grafično lahko njuno medsebojno odvisnost prikažemo z razsevnim diagramom



- Stopnjo povezanosti bi radi tudi številsko ovrednotili

Korelacija

$$\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]$$



- pozitivni členi: točke v področjih **A** in **C**
- negativni členi: točke v področjih **B** in **D**

Korelacijski koeficient - mera povezanosti

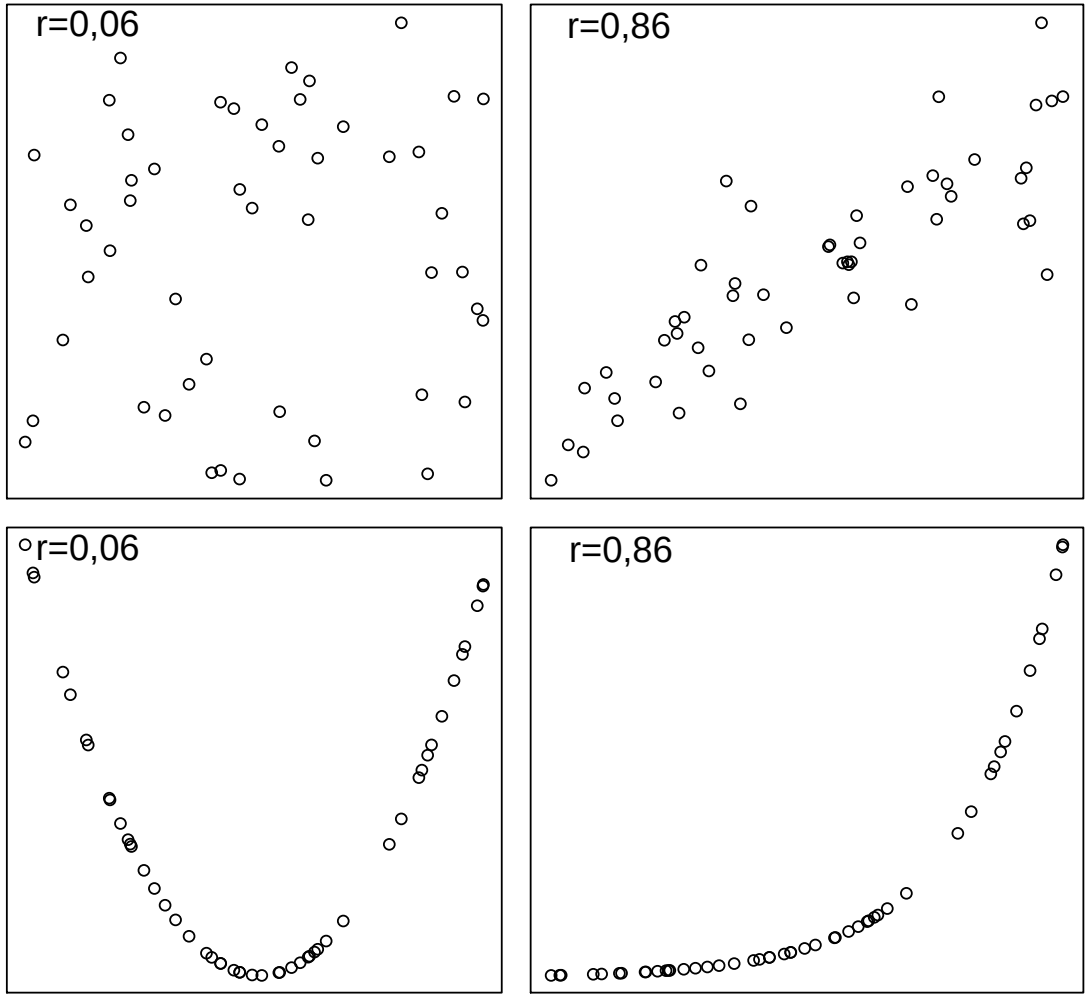
Izraz standardiziramo in definiramo empirični **korelacijski koeficient** (**Pearsonov** korelacijski koeficient)

$$r_{XY} = r = \frac{\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

Lastnosti:

- Korelacijski koeficient nima enot
- $-1 \leq r \leq 1$
- večji $|r|$, večja povezanost
- $r = 1$ ali $r = -1$, kadar vse točke ležijo na premici
- Če je naklon pozitiven, je r pozitiven in obratno

Korelacija



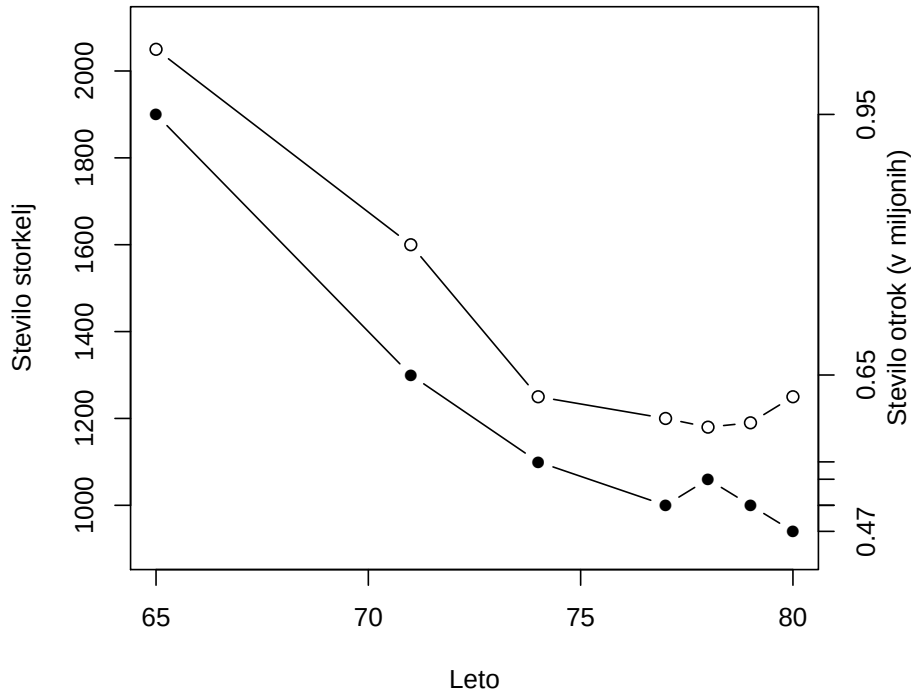
Visok korelacijski koeficient še ne pomeni, da gre za linearen odnos med spremenljivkama! Vedno **narišite podatke!**

Linearna regresija

Vprašanja:

- Ali je povezanost statistično značilna?
- Kako lahko na podlagi vrednosti X napovemo vrednost Y?

Pozor: povezanost $\neq$ odvisnost

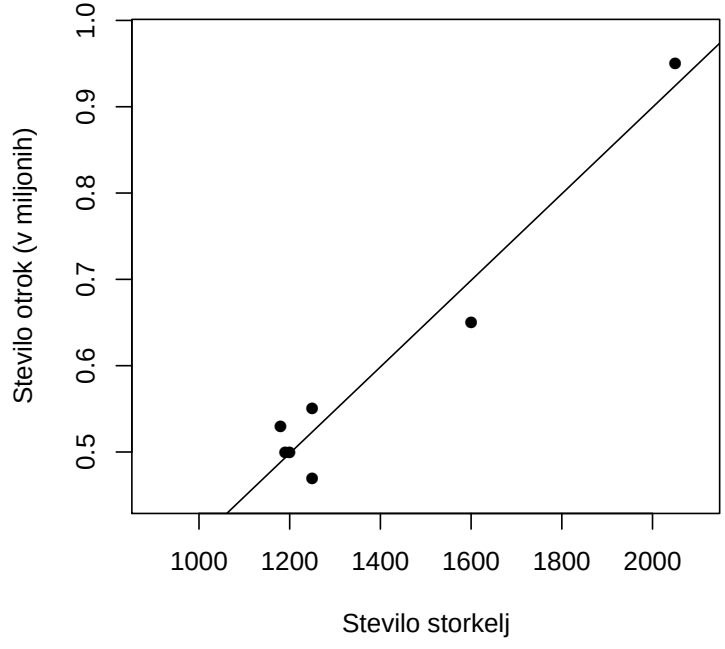


Linearna regresija

Vprašanja:

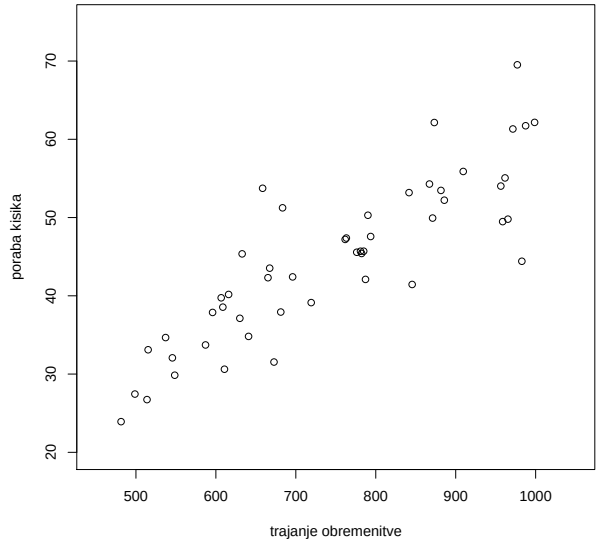
- Ali je povezanost statistično značilna?
- Kako lahko na podlagi vrednosti X napovemo vrednost Y?

Pozor: povezanost $\neq$ odvisnost



Linearna regresija

Ideja

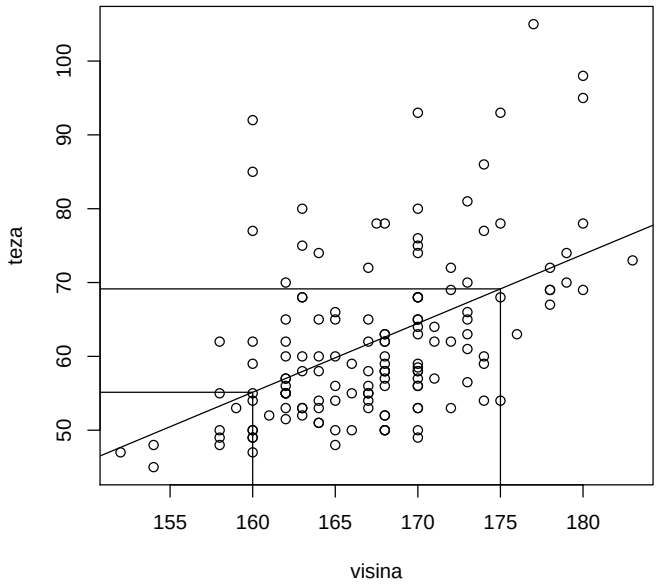


Trend opišemo s premico, ki jo priredimo točkam.

1. Če je naklon enak 0, med spremenljivkama ni povezanosti
   Testiramo ničelno domnevo, da je naklon v populaciji enak 0
2. Zanima nas povprečna poraba kisika na minuto, če obremenitev traja 10 minut
   Zanima nas koliko se poveča poraba kisika na minuto, če obremenitev traja 1 minuto dlje

Linearna regresija - Primer

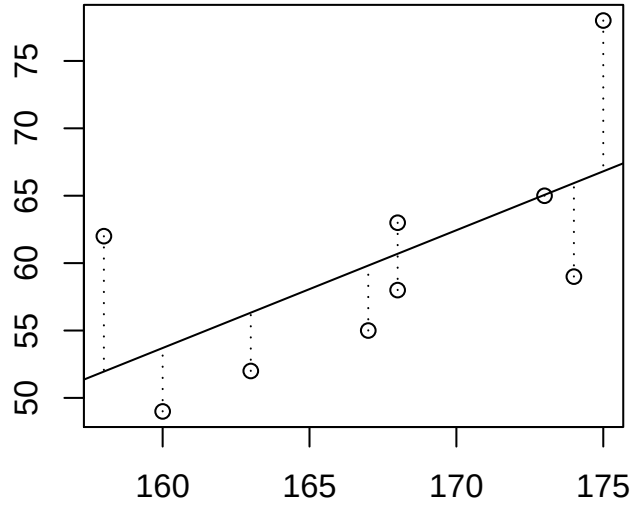
Povezanost med višino in težo



$$
\begin{aligned}
y &= a + bx \\
\text{teza} &= -94 + 0{,}93 * \text{visina} \\
55{,}1 &= -94 + 0{,}93 * 160 \\
69{,}1 &= -94 + 0{,}93 * 175
\end{aligned}
$$

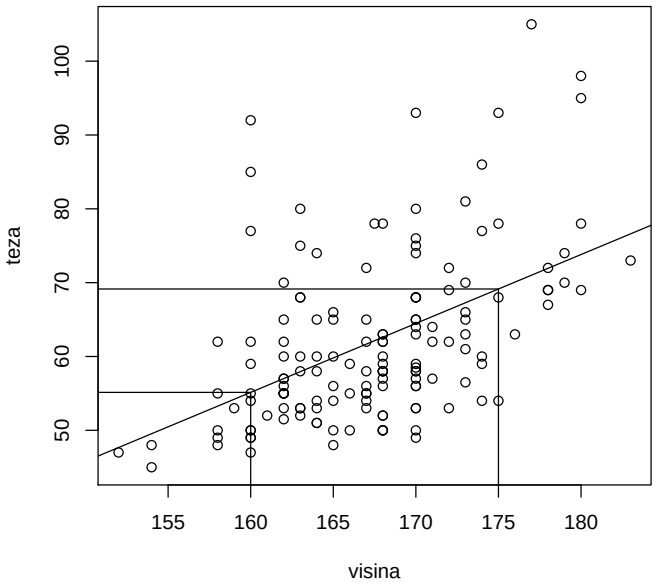
Linearna regresija - metoda najmanjših kvadratov

Kako prirediti premico podatkom?



- Enačba premice: $y = \alpha + \beta x$
- β je naklon premice
- α je presečišče z osjo y (vrednost y, ko je $x = 0$)
- Točke $T(x_i, y_i)$, na premici $T(x_i, \alpha + \beta x_i)$
- Gledamo ostanke: $y_i - (\alpha + \beta x_i)$
- Želimo najti taka α in β, da bo $\sum_i (y_i - (\alpha + \beta x_i))^2$ najmanjša možna

Linearna regresija - interpretacija



$$
\begin{aligned}
y &= a + bx \\
\text{teza} &= -94 + 0{,}93 * \text{visina} \\
55{,}1 &= -94 + 0{,}93 * 160 \\
69{,}1 &= -94 + 0{,}93 * 175
\end{aligned}
$$

Pomen regresijskega koeficienta *b*:

Naklon premice

Razlika odvisne spremenljivke, če je neodvisna različna za 1 enoto.

Razlika v teži, če se višina razlikuje za 1 cm

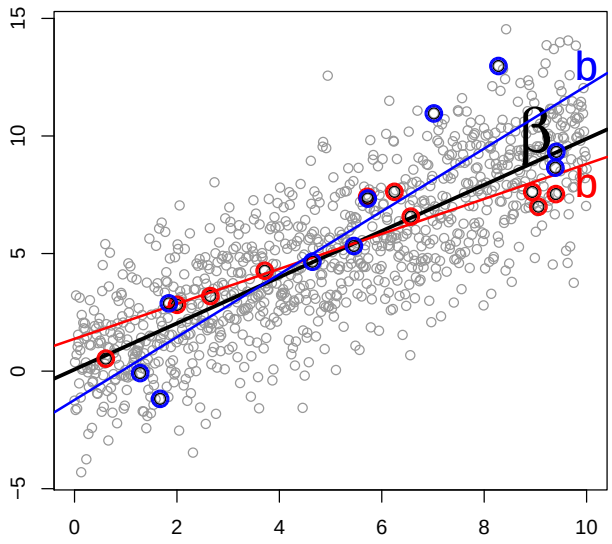
Pomen regresijske konstante *a*:

Presečišče z osjo *y* (vrednost *y*, ko je $x = 0$)

Teža pri višini 0

Pogosto nima smiselne interpretacije, osredotočimo se na *b*!

Linearna regresija - populacija $\leftrightarrow$ vzorec



Populacija

$y = \alpha + \beta x$

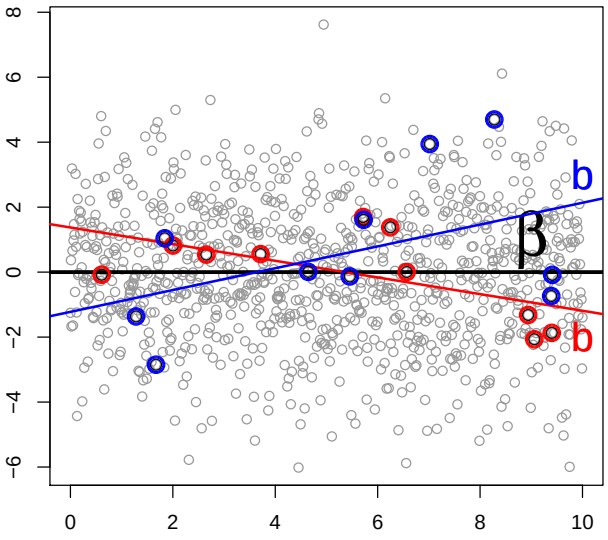
Vzorec

$y = a + bx$

Ocenjevanje populacijske vrednosti β s pomočjo vzorčne vrednosti b

SE_β odvisna od razpršenosti točk okoli premice, velikosti vzorca

Linearna regresija - statistično sklepanje



- $H_0: \quad \beta = 0, \quad \alpha = 0{,}05$
- Testna statistika, predpostavke

$$\frac{\hat{\beta} - 0}{\hat{SE}_\beta}$$

- Podatki, ‘odločitev ’

Linearna regresija - preverjanje predpostavk

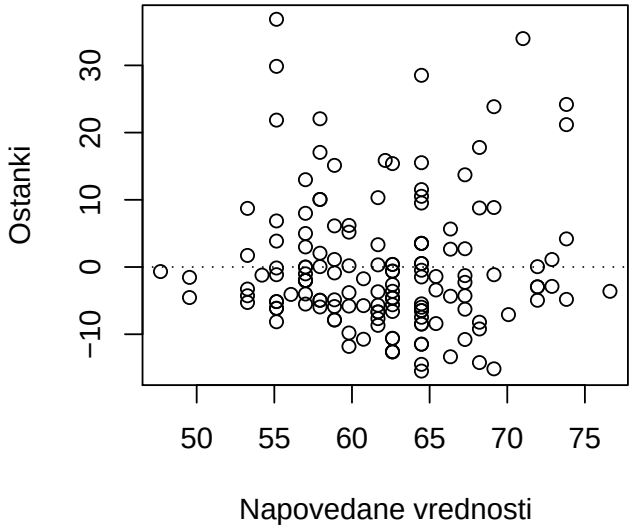
1. Opazovanja, se pravi pari meritev, so **neodvisna**.
2. Regresijska funkcija je **linearna**.
3. Vrednosti variirajo okrog regresijske premice s **konstantno varianco** (homoscedastičnost).
4. Vrednosti Y so okrog regresijske premice **normalno porazdeljene**.

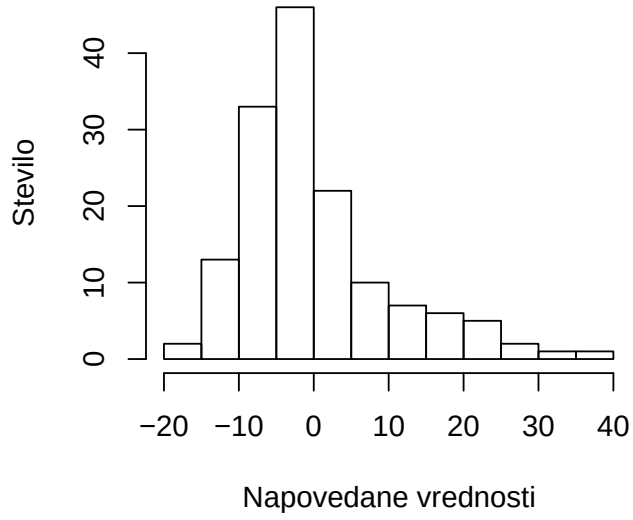
Ustreznost teh predpostavk moramo vedno **preveriti**.

Linearna regresija - grafično preverjanje predpostavk

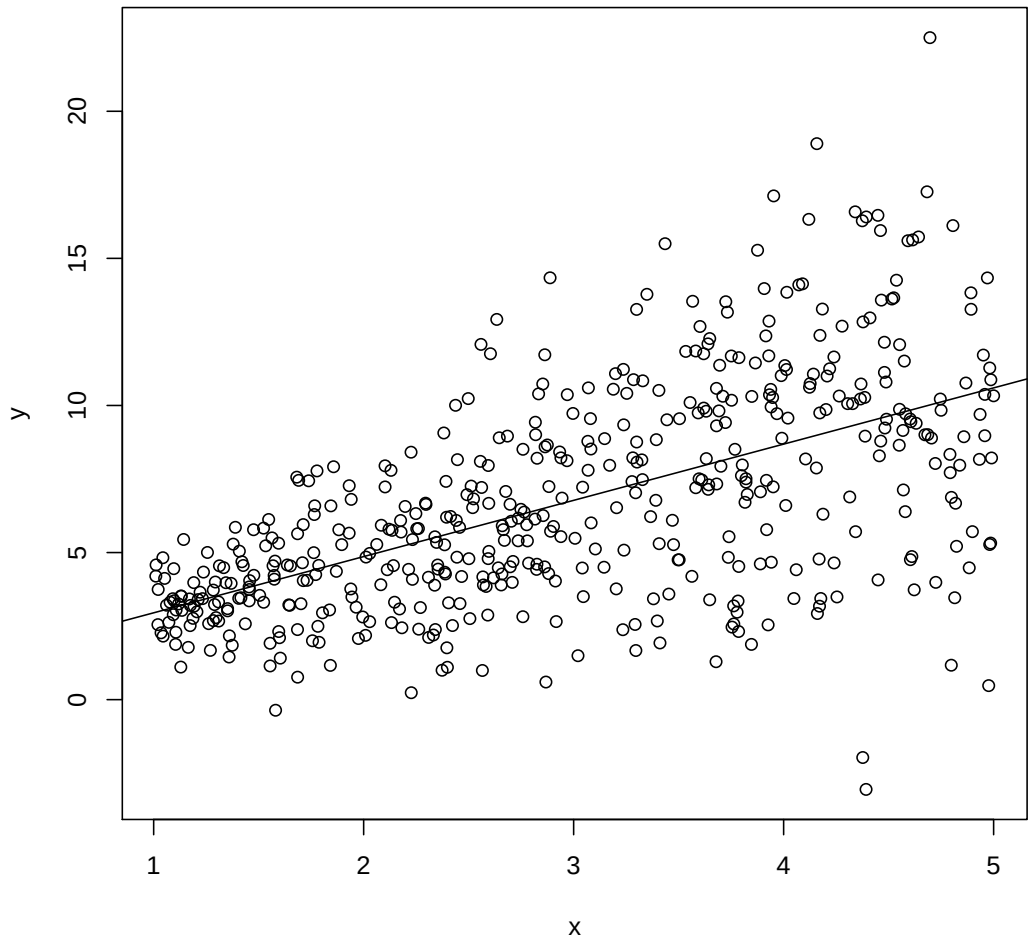
Če je model pravilen, so **ostanki simetrično razporejeni okrog premice in imajo konstantno varianco**.

Teža in višina

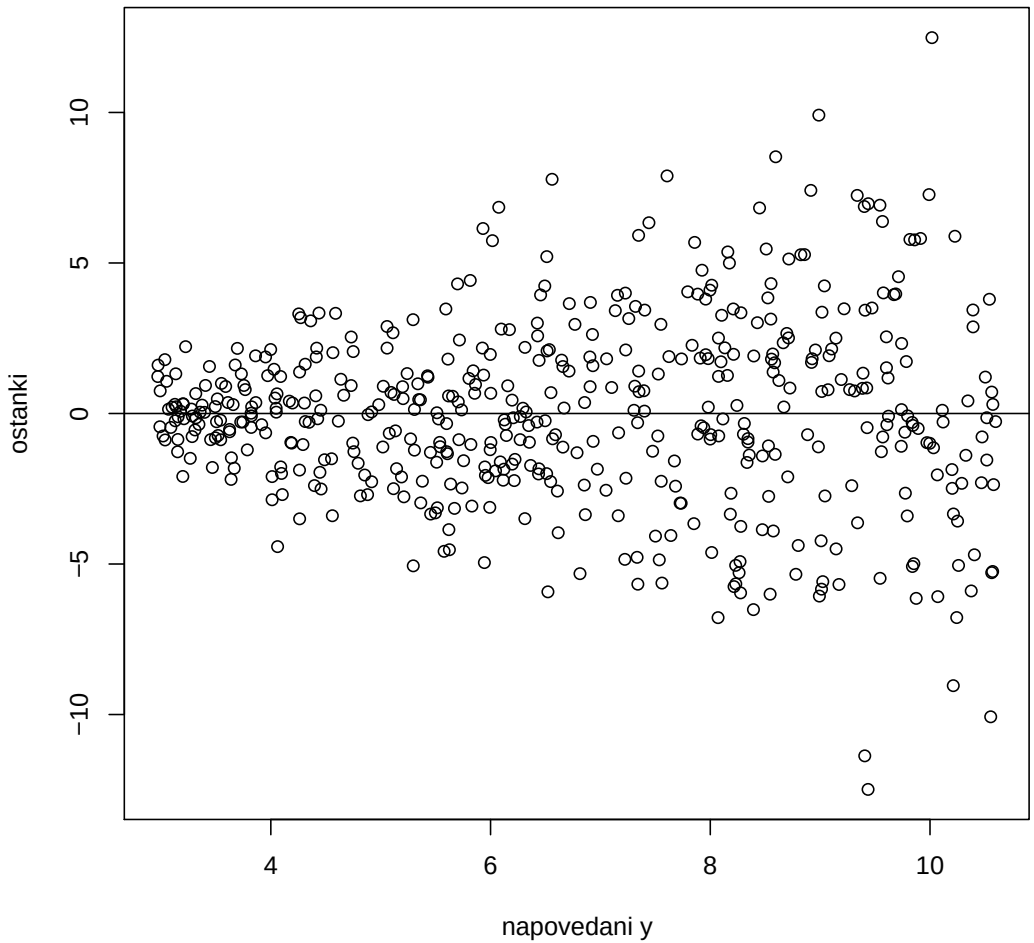




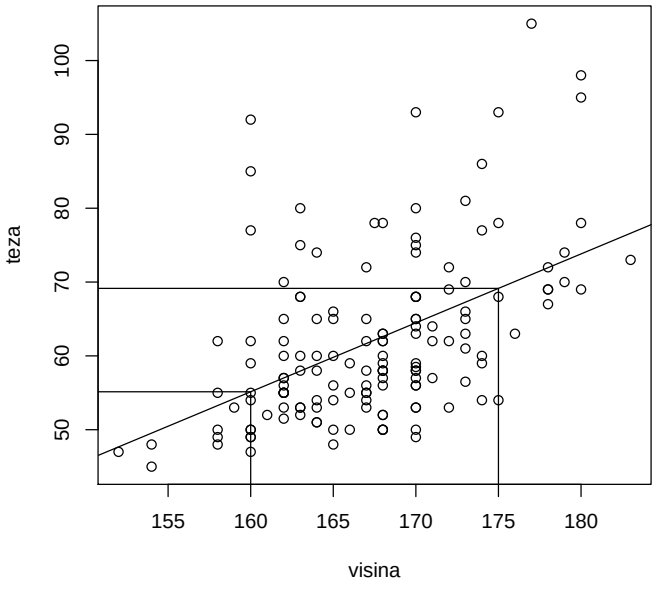
Primer heteroscedastičnih podatkov



Primer - slika ostankov



Regresijski model



$$
\begin{aligned}
y &= a + bx + \varepsilon \\
\text{teza} &= -94 + 0{,}93 * \text{visina} + \varepsilon \\
55{,}1 &= -94 + 0{,}93 * 160 + \varepsilon \\
69{,}1 &= -94 + 0{,}93 * 175 + \varepsilon
\end{aligned}
$$

Predpostavimo torej

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Regresijski model - ocenjevanje σ

- Označimo napovedane vrednosti z $\hat{y}$, torej $\hat{y}_i = a + bx_i$
- Ostanki so potem enaki $y_i - \hat{y}$
- Povprečje ostankov je 0, varianco ocenimo kot

$$\hat{\sigma} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}$$

- Zakaj $n-2$? Zato ker smo a in b ocenili z vzorca

Regresijski model - statistično sklepanje

- Zanima nas, koliko bi se motili, če bi pri istih vrednostih x vsakič dobili drug vzorec y-onov (ki bi sledil istemu modelu)
- Velja: $b = \hat{\beta} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- In zato:

$$SE_\beta^2 = Var(\hat{\beta}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 Var(Y)}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} = \frac{Var(Y)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Velja

$$\frac{\hat{\beta} - \beta}{SE_\beta} \sim \mathcal{N}(0,1)$$

Regresijski model - statistično sklepanje

- V praksi ne poznamo σ in zato ne poznamo SE_β
- Standardno napako ocenimo iz podatkov

$$\hat{SE}_\beta = \frac{\hat{\sigma}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Naša testna statistika

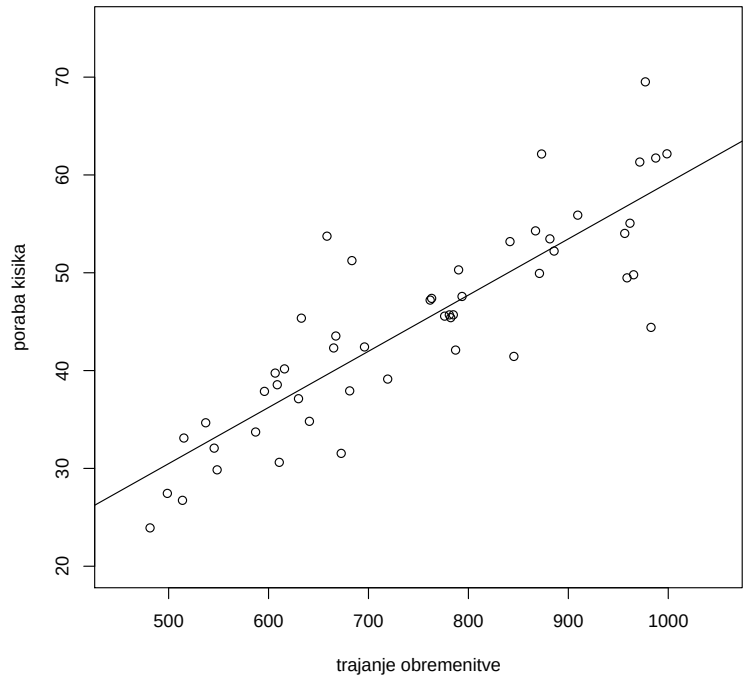
$$t = \frac{\hat{\beta} - 0}{\hat{SE}_\beta}$$

je porazdeljena kot t_{n-2}

Regresijski model - primer

Obremenitveni test

- Na obremenitvenem testu so 50 moškim izmerili **trajanje obremenitve** in **porabo kisika**. Porabo kisika izražamo v mililitrih na kilogram telesne mase na minuto.



- Predpostavke (linearnost, homoscedastičnost, normalno porazdeljeni ostanki) niso huje kršene

Regresijski model - primer

- Ničelna domneva

  V populaciji ni povezanosti med trajanjem obremenitve in porabo kisika, $\beta = 0$

- Testna statistika in njena porazdelitev

  Na podatkih ocenimo $b = 0{,}057$, $\hat{SE}_\beta = 0{,}0049$

$$t = \frac{b - 0}{\hat{SE}_\beta} \sim t_{48}$$

- Podatki, 'odločitev'

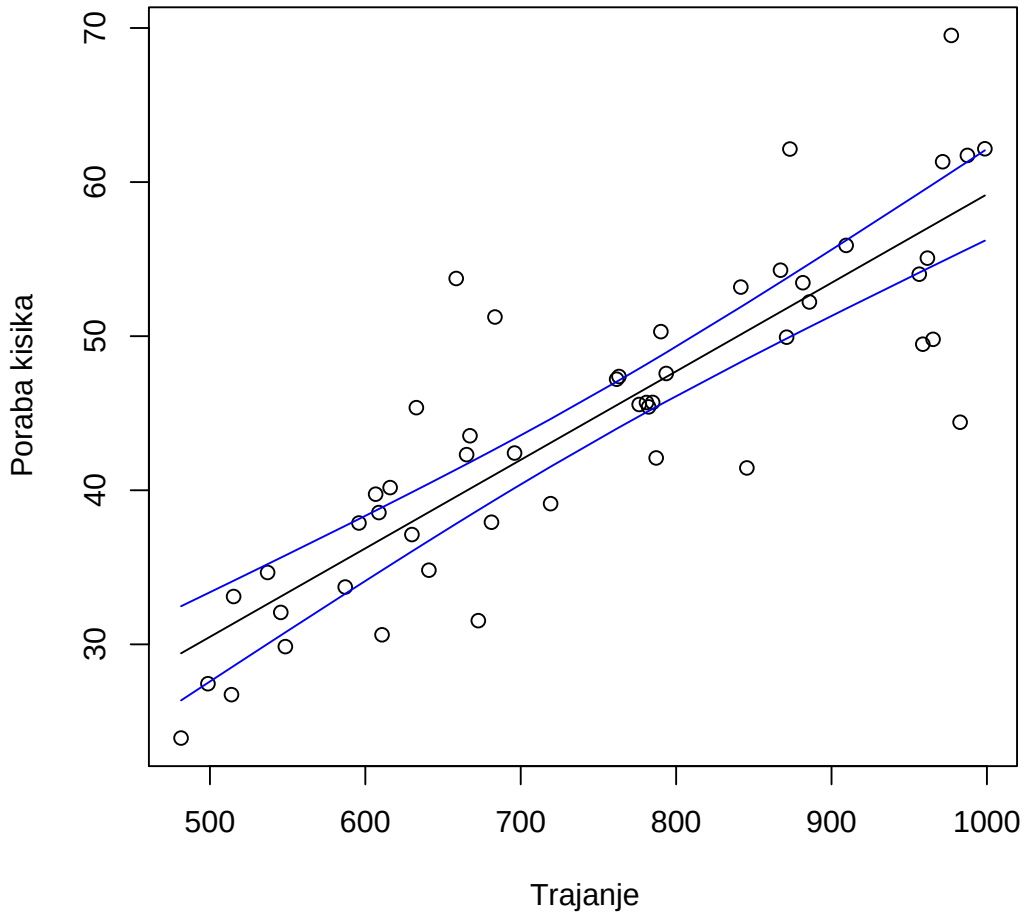
$$t = \frac{0{,}057 - 0}{0{,}0049} = 11{,}5$$

$p < 0{,}001 \Rightarrow$ ničelno domnevo zavrnemo

Regresijski model - primer

- Pokazali smo, da lahko z veliko verjetnostjo (majhnim tveganjem) trdimo, da sta spremenljivki v populaciji povezani.
- Povejmo še nekaj več o tej povezanosti
- Enačba premice na vzorcu je
  $\text{poraba kisika} = 1{,}765 + 0{,}057\text{cas}$
- Na vzorcu je pričakovana vrednost porabe kisika na minuto pri moškem, ki je test opravljal 10 minut enaka
  $1{,}765 + 0{,}057 \cdot 600 = 35{,}96$
- Ocenjujemo, da v populaciji vsaka sekunda trajanja testa v povprečju poveča porabo za 0,057 ml/kg/min. 95% interval zaupanja za to vrednost je $[0{,}047, 0{,}067]$

Regresijski model - interval zaupanja za premico



Linearna regresija - opombe

Absolutna vrednost *b* ni pokazatelj pomembnosti spremenljivke oz. stopnje povezanosti.
Odvisnost od enote, v kateri merimo spremenljivke:

$$55{,}1 = -94 + 0{,}93 * 160$$ ↗ višina v cm

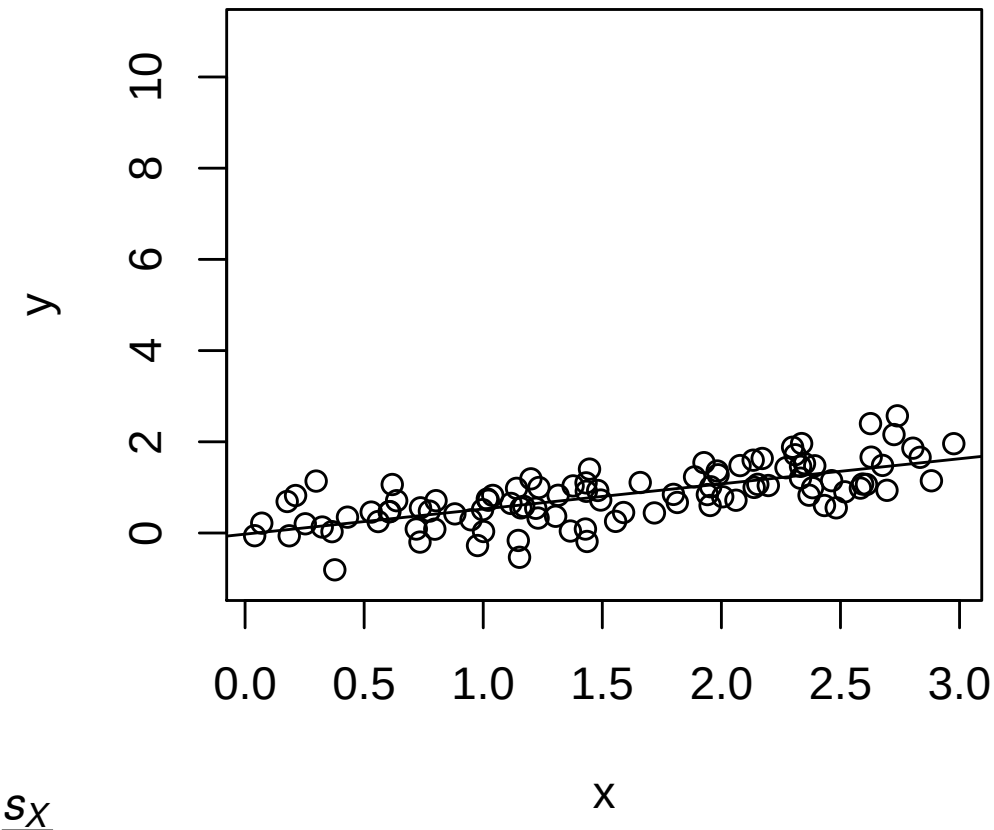
$$55{,}1 = -94 + 9{,}30 * 1{,}6$$ ↘ višina v m

Linearna regresija - opombe

Vsi zaključki veljajo izključno znotraj razpona, v katerem smo imeli podatke!

Regresijski koeficient $\Leftrightarrow$ korelacijski koeficient

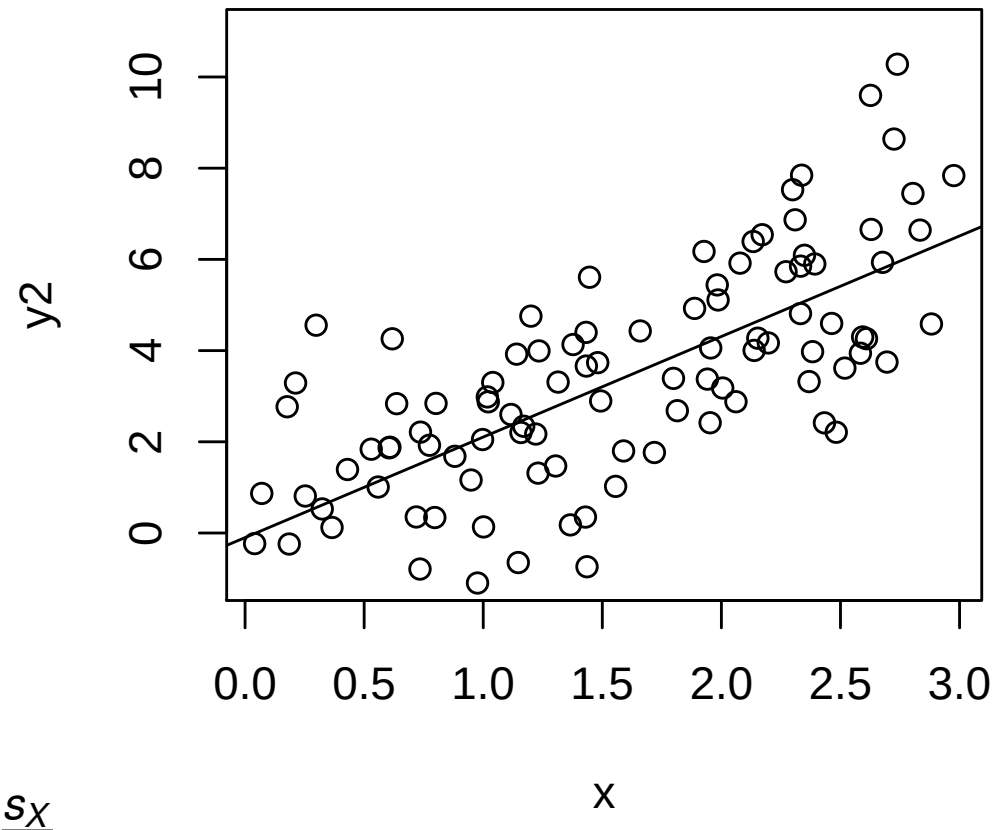
$r = 0{,}7$, $\beta = 0{,}5$



Velja $\beta = r\frac{s_X}{s_Y}$

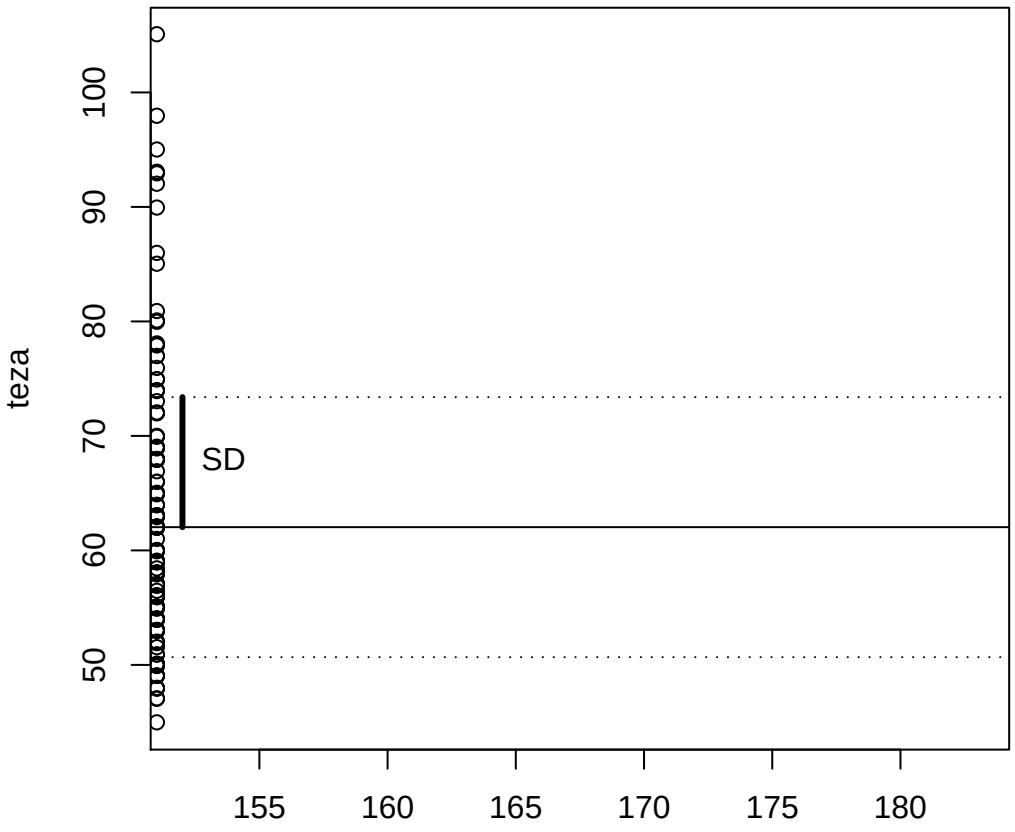
Regresijski koeficient $\Leftrightarrow$ korelacijski koeficient

$r = 0{,}7$, $\beta = 2$



Velja $\beta = r\frac{s_X}{s_Y}$

R^2 - mera pojasnjene variabilnosti

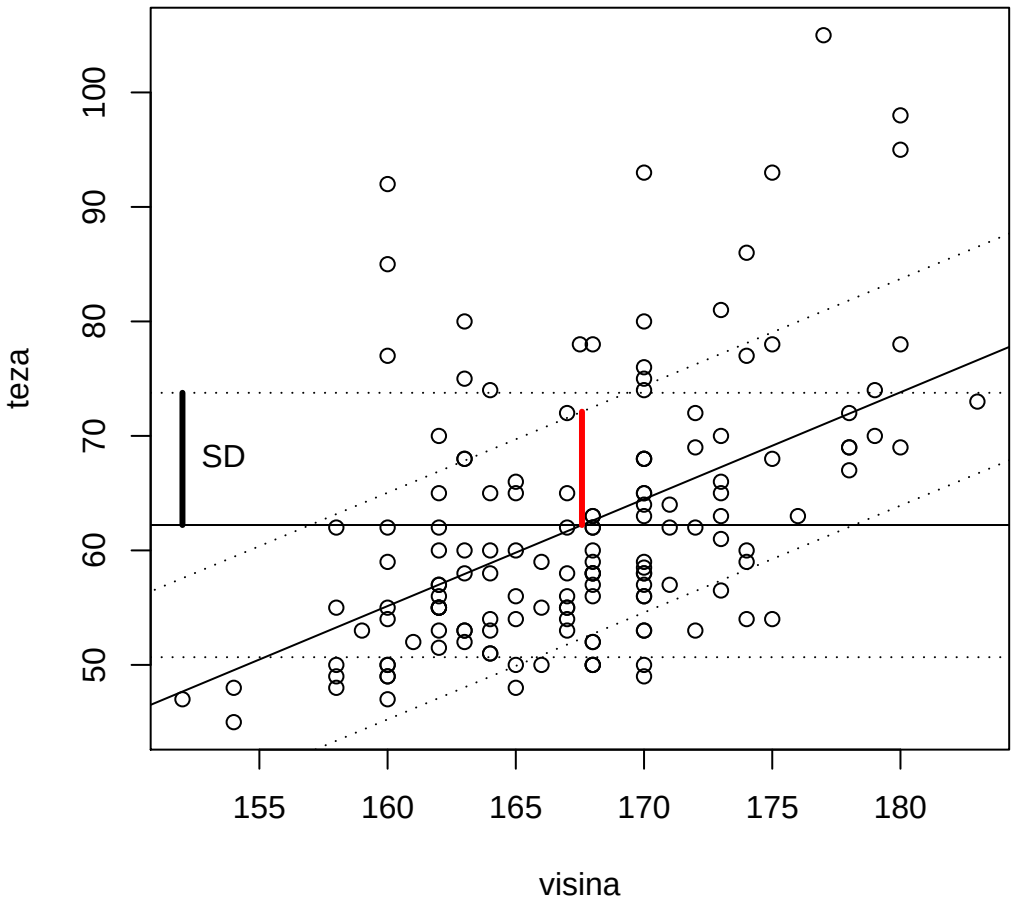


$$R^2 = 1 - \frac{\text{varianca ostankov}}{\text{celotna varianca}}$$

$$= 1 - \frac{9{,}9^2}{11{,}4^2}$$

$$= 0{,}24$$

R^2 - mera pojasnjene variabilnosti



$$R^2 = 1 - \frac{\text{varianca ostankov}}{\text{celotna varianca}}$$

$$= 1 - \frac{9{,}9^2}{11{,}4^2}$$

$$= 0{,}24$$