Uvod v statistiko - 2. del

Janez Stare

Medicinska fakulteta, Ljubljana

Ljubljana, 2018

Naj bodo $X_1, X_2, \ldots, X_n$ neodvisne in **normalno** porazdeljene z

$$E(X_i) = \mu_i \quad \text{in} \quad Var(X_i) = \sigma^2$$

potem je

$$Y = a_0 + a_1 X_1 + \cdots + a_n X_n$$

tudi **normalno** porazdeljena. Kot pomemben poseben primer imamo:

Če je $X_1, X_2, \ldots, X_n$ slučajni vzorec velikosti n iz $\mathcal{N}(\mu, \sigma^2)$, potem

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

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 dveh skupin

Merimo IQ na vzorcu volivcev, ki podpirajo trenutno slovensko vlado in na vzorcu, ki podpira opozicijo. Izračunamo povprečje v vsaki skupini. Seveda ne pričakujemo, da bi bili povprečji enaki, možni razlogi za razliko pa so:

- Obstaja dejanska razlika v povprečni vrednosti IQ med skupinama volivcev.
- Gre za **slučajno variiranje**
- Rezultati so **pristranski** zaradi vplivov drugih dejavnikov (moteči dejavniki)

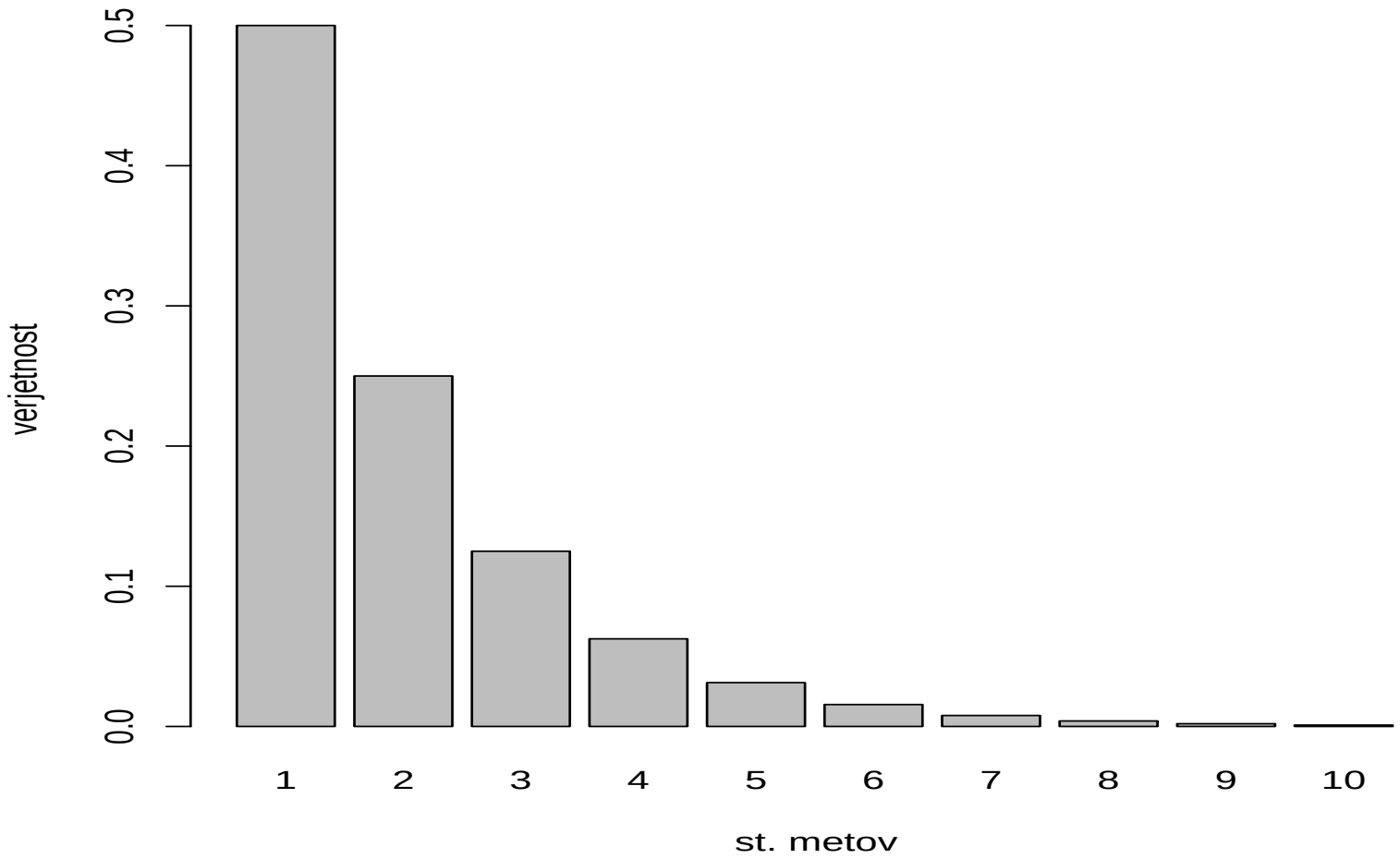
Pristranost lahko odpravimo (ali zmanjšamo) z ustreznim **načrtom študije** pa tudi v statistični analizi lahko **pristranost popravimo**. Pristranost naj torej ne bi bila verjeten razlog za razliko med skupinama.

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

Če je bila študija dobro načrtovana in statistična analiza kaže, da slučajna variabilnost ni verjetna razlaga za opaženo razliko, potem **sklepamo**, da gre najverjetneje za dejansko razliko v IQ med obema skupinama volivcev.

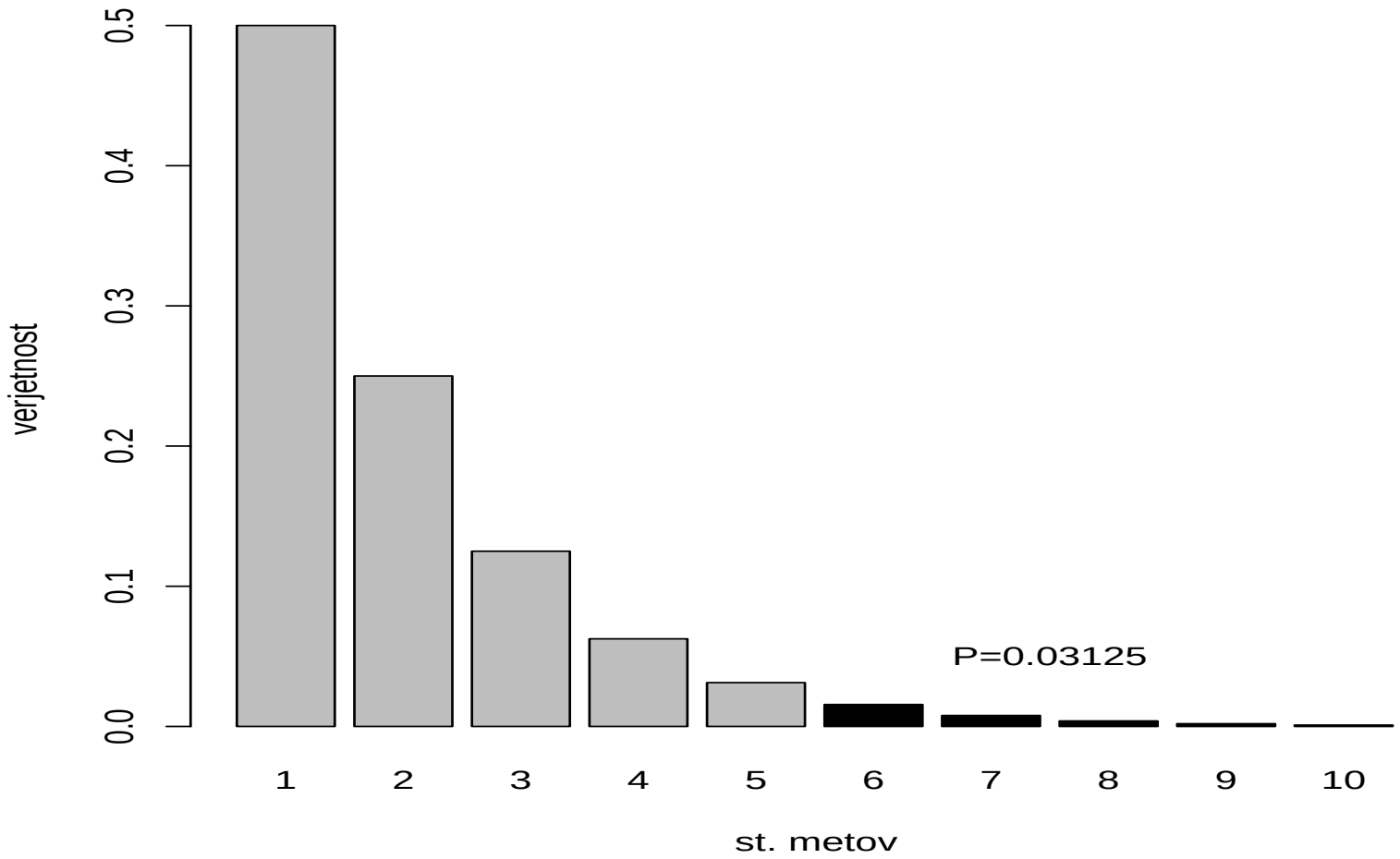
Primer: met kovanca

Na sliki je verjetnost pojava prve cifre v k-tem metu.



Primer: met kovanca

Na sliki je verjetnost pojava prve cifre v k-tem metu.



ANALIZA ENEGA VZORCA IZ NORMALNO PORAZDELJENE POPULACIJE

Primer:

Oglejmo si še enkrat porodne teže desetih dečkov (v kilogramih):

$3{,}310, 3{,}880, 3{,}460, 3{,}490, 3{,}160, 3{,}250, 2{,}630, 4{,}370, 3{,}530, 3{,}280.$

Vprašanje: Ali se povprečna teža otrok 35-letnih mater razlikuje od skupnega povprečja 3,348 kilograma?

Statistični model

Podatki predstavljajo slučajni vzorec (neodvisne, enako porazdeljene meritve) $x_1, x_2, \ldots, x_{10}$ iz normalne porazdelitve s povprečjem μ in varianco σ^2.

Povzemimo **predpostavke**:

1. Meritve so **neodvisne**.
2. Vse meritve izhajajo iz **iste porazdelitve** (torej isto povprečje in ista varianca).
3. Spremenljivka, ki nas zanima, je v populaciji **normalno porazdeljena**.

Preverjanje predpostavk je **pomemben** del statistične analize. Pogosto si pri tem pomagamo z različnimi grafi. Poznavanje postopka merjenja je lahko zelo koristno pri iskanju morebitnih odstopanj od predpostavk.

Preverjanje predpostavk

1. **Neodvisnost** največkrat preverimo tako, da preverimo postopek vzorčenja. Če gre za naključno vzorčenje, se praviloma nimamo česa bati, izjeme pa žal so. V našem primeru bi bila predpostavka o neodvisnosti na primer kršena, če bi bili med podatki tudi podatki o dvojčkih.
2. **Ista porazdelitev**? To se zdi precej očitno, a če so podatki zbirani skozi čas, se stvari lahko spremenijo. Priporočljivo jih je narisati kot funkcijo časa in ugotoviti, če gre za kakšne trende.
3. Ali je predpostavka o **normalnosti** smiselna? Grafično jo preverimo s histogramom ali, še bolje, s q-q grafom. Obstajajo pa tudi formalni testi, a o tem kasneje.

Recimo sedaj, da bi bila predpostavka o normalnosti izpolnjena. Katera normalna porazdelitev bi se najbolje prilegala našim podatkom?

No, če so podatki videti normalni, potem je naravno, da so simetrično porazdeljeni okrog svoje aritmetične sredine in da je najbolj smiselna ocena variance kar vzorčna varianca. Torej

$$\hat{\mu} = \bar{x} = 3{,}436 \text{ kg}$$

$$\hat{\sigma}^2 = s^2 = 0{,}20916 \text{ kg}^2$$

Strešica nad oznako populacijskega parametra v splošnem pomeni, da gre za oceno na vzorcu. Nekatere ocene, kot sta vzorčno povprečje in vzorčna varianca, imajo še posebne oznake.

Vrnimo se zdaj k našemu problemu: ali je **vzorčno variiranje** možna razlaga za odstopanje vzorčnega povprečja 3,436 kilograma od populacijskega povprečja 3,348 kilograma?

Preverjamo torej **predpostavko** (**hipotezo**): $\mu = 3{,}348$

Statistični test:

Predpostavimo, da je povprečna porodna teža otrok 35-letnih mater 3,348 kilograma in si postavimo tole **vprašanje**: Kako velika odstopanja od te vrednosti lahko pričakujemo na vzorcih?

Iz predpostavk sledi, da je $\bar{x}$ ena vrednost iz

$$\mathcal{N}(3{,}348,\, \sigma^2/10)$$

in torej

$$z = \frac{\bar{x} - 3{,}348}{\frac{\sigma}{\sqrt{10}}}$$

ena vrednost iz standardizirane normalne porazdelitve.

Pri dani vrednosti $\bar{x}$ torej lahko povemo, kako verjetne so vrednosti, ki so večje (manjše) od $\bar{x}$.

Popravek: lahko povemo, **če** poznamo σ! V praksi bo to redko (pravzaprav nikoli) res in zdi se, da smo v slepi ulici. No, največ kar lahko naredimo je, da namesto σ vstavimo oceno standardnega odklona na vzorcu. Torej

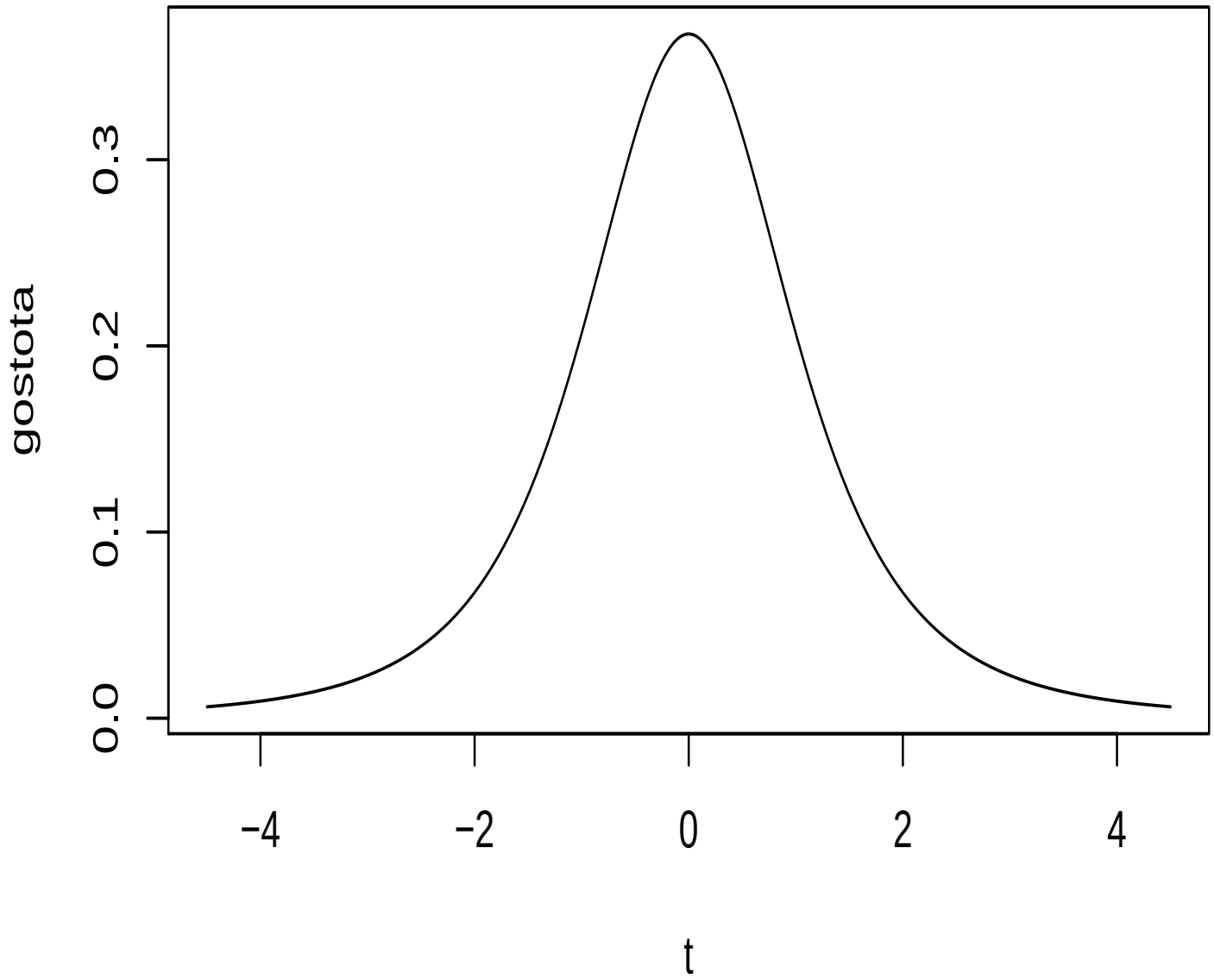
$$t = \frac{\bar{x} - 3{,}348}{\frac{s}{\sqrt{10}}} = \frac{3{,}436 - 3{,}348}{\frac{0{,}4573401}{\sqrt{10}}} = 0.608$$

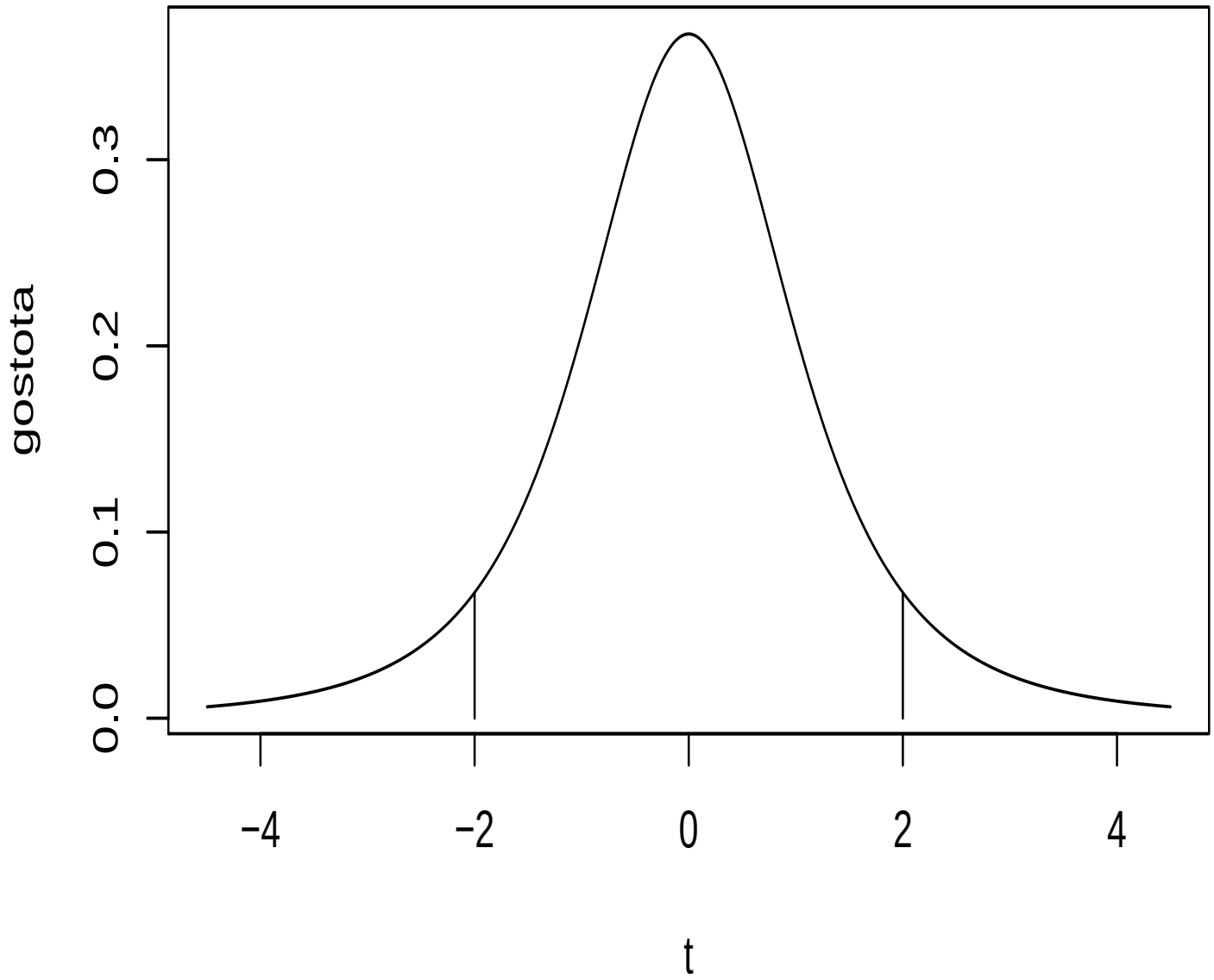
Na levi smo previdno napisali t namesto z, ker nismo prepričani, če je izraz na desni še vedno normalno porazdeljen. Izkaže se, da **res ni**! Intuitivno si to razložimo recimo takole: če imata dva vzorca enako vzorčno povprečje, imata še vedno lahko različni varianci in takrat bosta vrednosti izraza na desni različni. Ob fiksni σ seveda ne bi bili. Nadomestitev σ z s prispeva k dodatni variabilnosti.

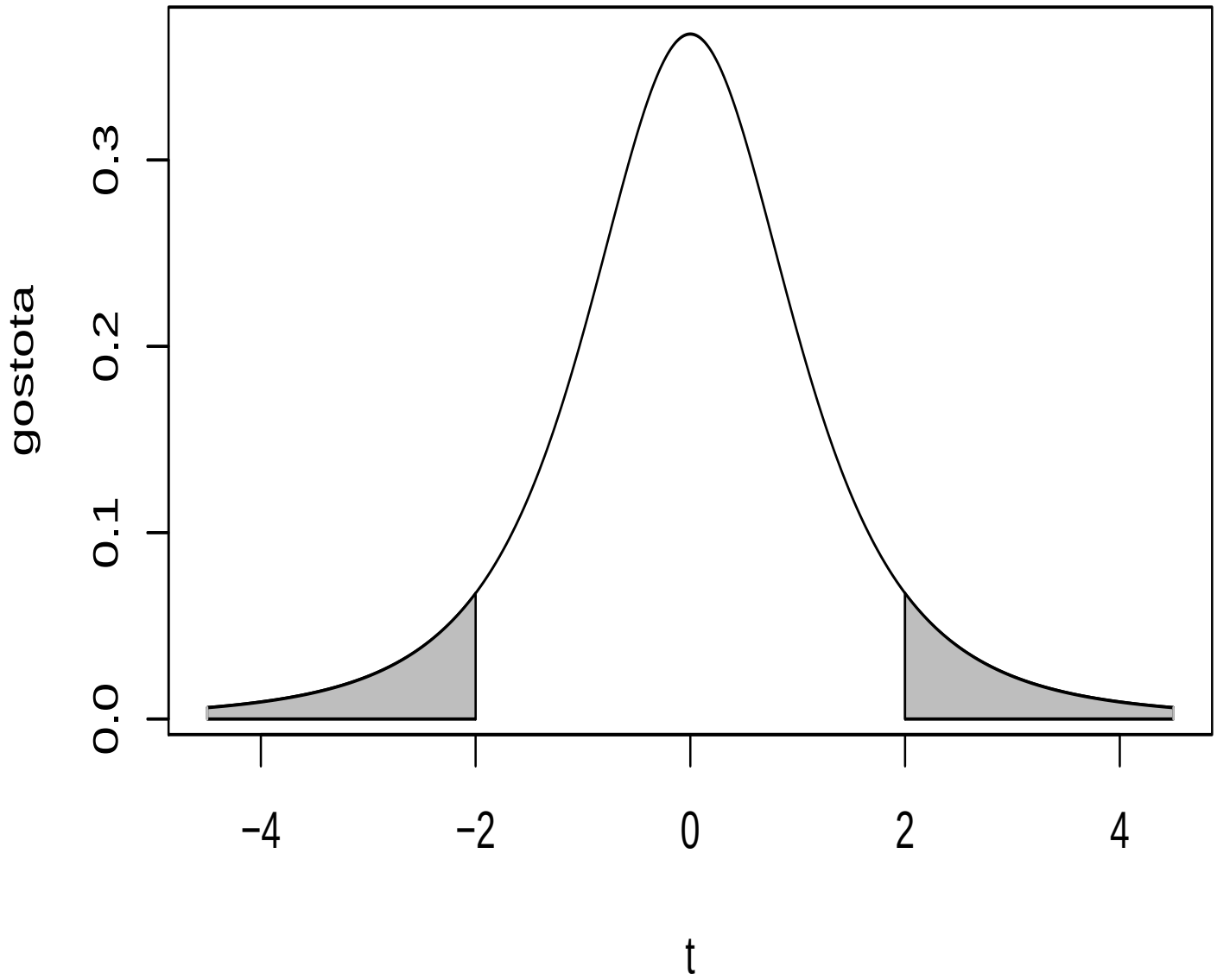
Izkaže se, da je izraz

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

porazdeljen po tako imenovani **Studentovi *t*-porazdelitvi** z $n - 1$ stopinjami prostosti. To pomeni, da je dejanska porazdelitev odvisna od velikosti vzorca! V našem primeru imamo torej $10 - 1 = 9$ stopinj prostosti. K *t* porazdelitvi se bomo še vrnili, tu povejmo še to, da gre za simetrično porazdelitev, ki je podobna normalni, ima pa bolj debele repe.


gostota
0.0
0.1
0.2
0.3
−4
−2
0
2
4
t


gostota
0.0
0.1
0.2
0.3
−4
−2
0
2
4
t


gostota
0.0
0.1
0.2
0.3
−4
−2
0
2
4
t

Verjetnost, da so vrednosti t pri treh stopinjah prostosti za več kot 1,96 oddaljene od 0 je 0,145. To je precej več kot 5% pri normalni porazdelitvi, z naraščanjem stopinj prostosti pa se ta razlika manjša. Ustrezna verjetnost pri dvajsetih stopinjah prostosti je 0,064, pri petdesetih pa 0,056.

No, in zdaj še k našemu primeru. Za t smo dobili 0,608, ustrezna verjetnost (vrednost p) pa je 0,56. Če ničelna hipoteza velja, potem lahko dobimo vrednosti, ki so vsaj za $3{,}436 - 3{,}348 = 0{,}088$ oddaljene od povprečja v 56% primerov.

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

Strnimo:

Izhajali smo iz **statističnega modela**, ki pravi, da so podatki slučajen vzorec neodvisnih meritev iz normalne porazdelitve.

Pri tem smo **ocenili**: $\hat{\mu} = \bar{x}$ in $\hat{\sigma}^2 = s^2$

in **testirali ničelno hipotezo**, da je

$$H_0 : \mu = \mu_0$$

s **testno statistiko**

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

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

Verjetnost, da t-statistika po absolutni vrednosti preseže izračunani t, torej

$$2 \cdot P(t_{n-1} \geq |t|),$$

imenujemo **stopnja tveganja** ali **vrednost** p.

Sklepamo potem takole:

Če je vrednost p manjša od α ničelno hipotezo **zavrnemo**.

Pri tem je α neka izbrana vrednost, ki se nam zdi dovolj majhna. Imenujemo jo **stopnja značilnosti** in je najpogosteje enaka 0,05, včasih tudi 0,01.

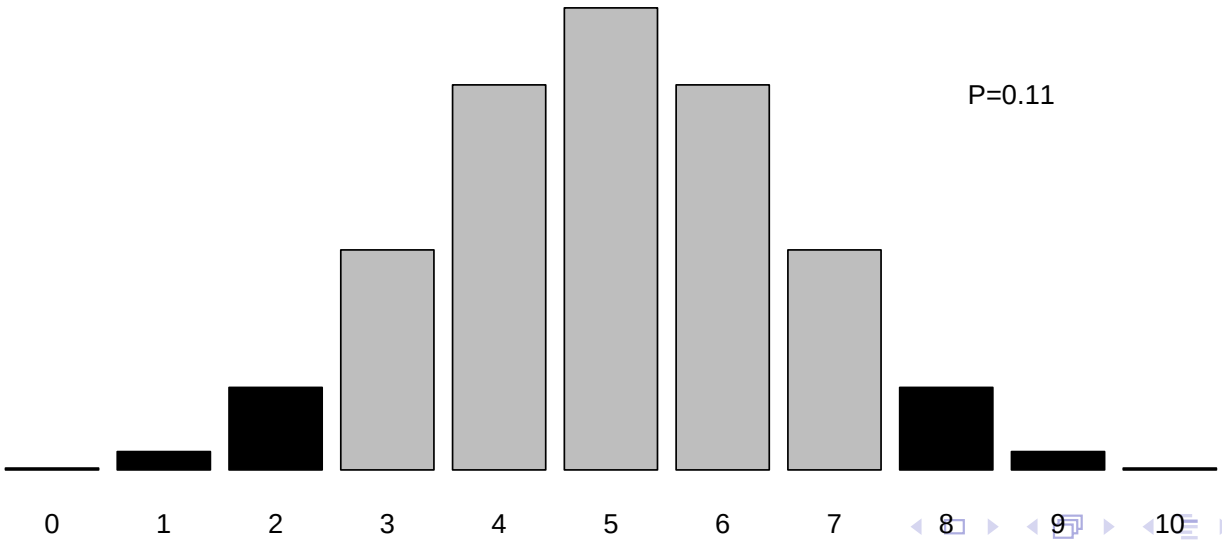
Odmik od pričakovane vrednosti ob veljavni ničelni hipotezi je **statistično značilen**, če je vrednost p manjša od α.

Videli smo, da pri zavračanju ničelne hipoteze lahko naredimo napako, verjetnost te napake pa poznamo. Kaj pa če ničelne hipoteze ne zavrnemo? Ali naj jo potem sprejmemo.

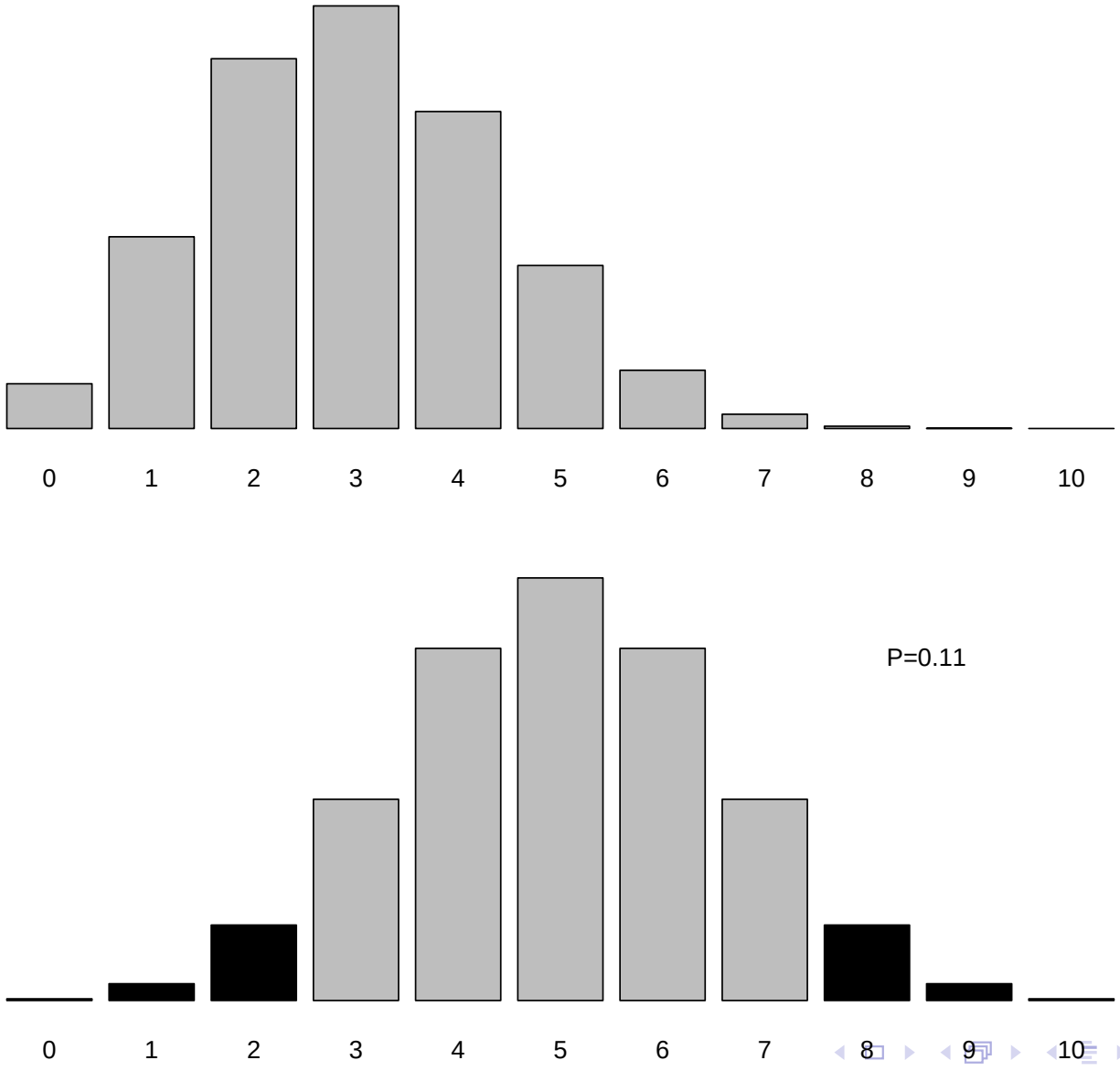
Oglejmo si **primer**, v katerem preverjamo, ali ima vlada res 50% podporo.

Ničelna hipoteza je torej: $H_0 : \pi = 0{,}5$

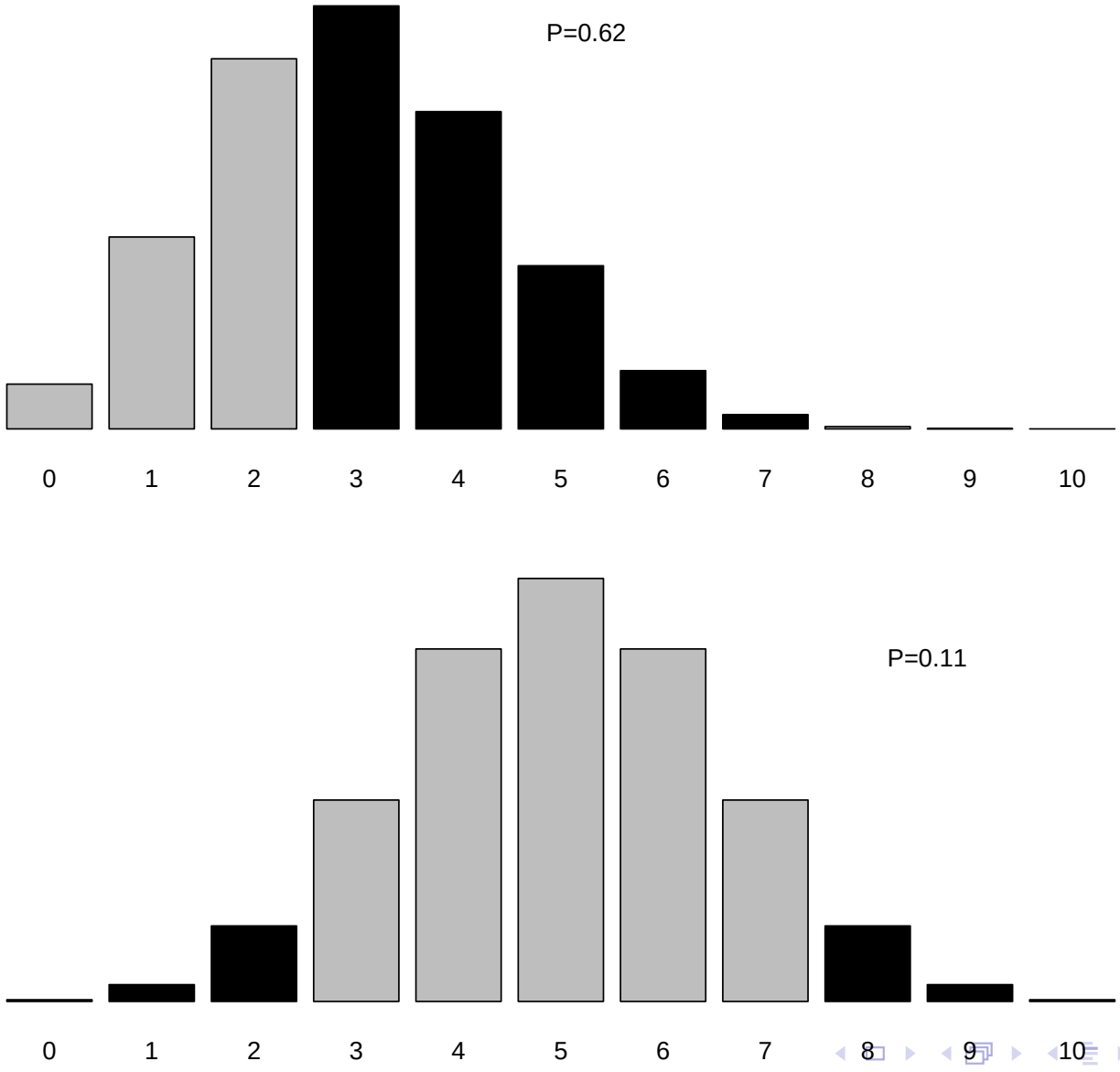
Primer: vlada ima 50% podporo



Primer: vlada ima 50% podporo



Primer: vlada ima 50% podporo



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

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

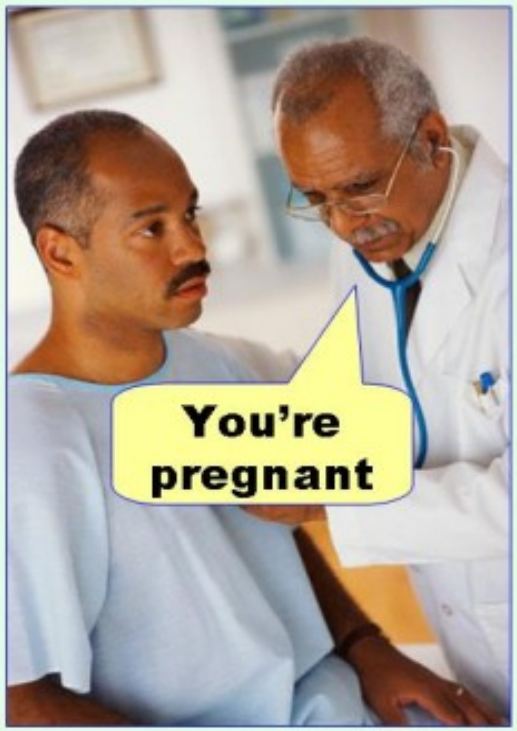
Verjetnost napake 1. vrste poznamo (stopnja tveganja), verjetnosti napake 2. vrste pa ne.

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.

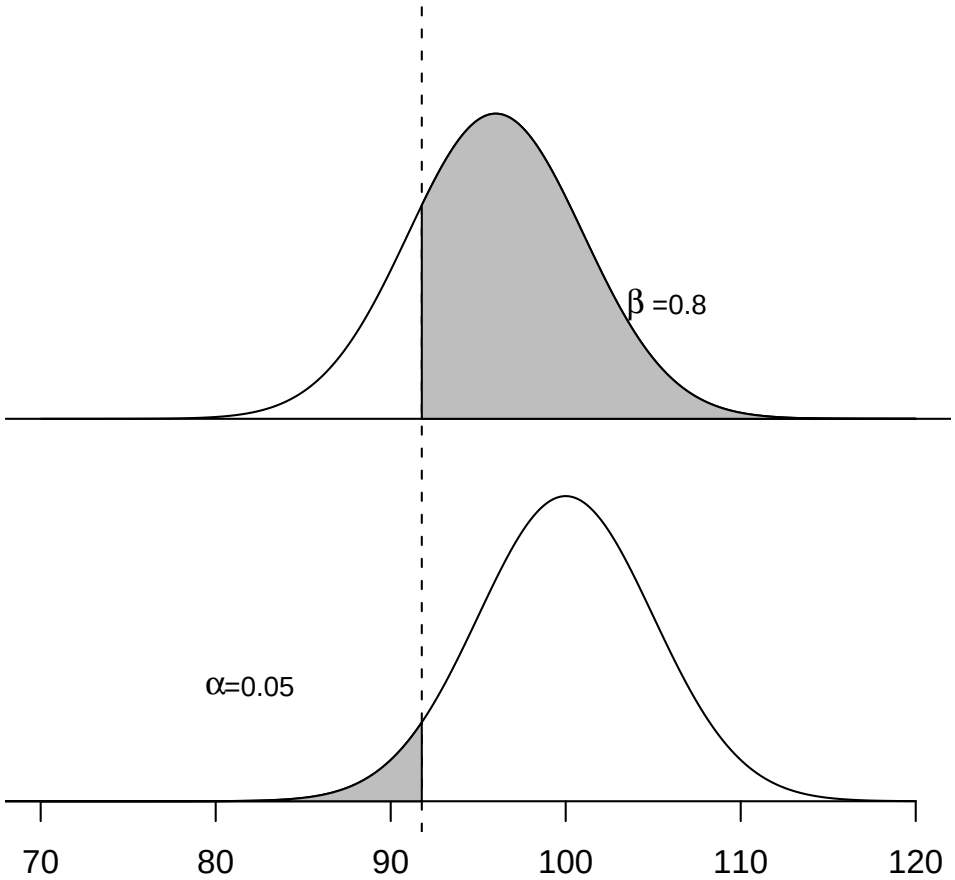
Opozorilo! Ničelne hipoteze (skoraj) **nikoli ne sprejemamo**! Vse, kar rečemo, je, da podatki ne nasprotujejo ničelni hipotezi.

Type I error
(false positive)

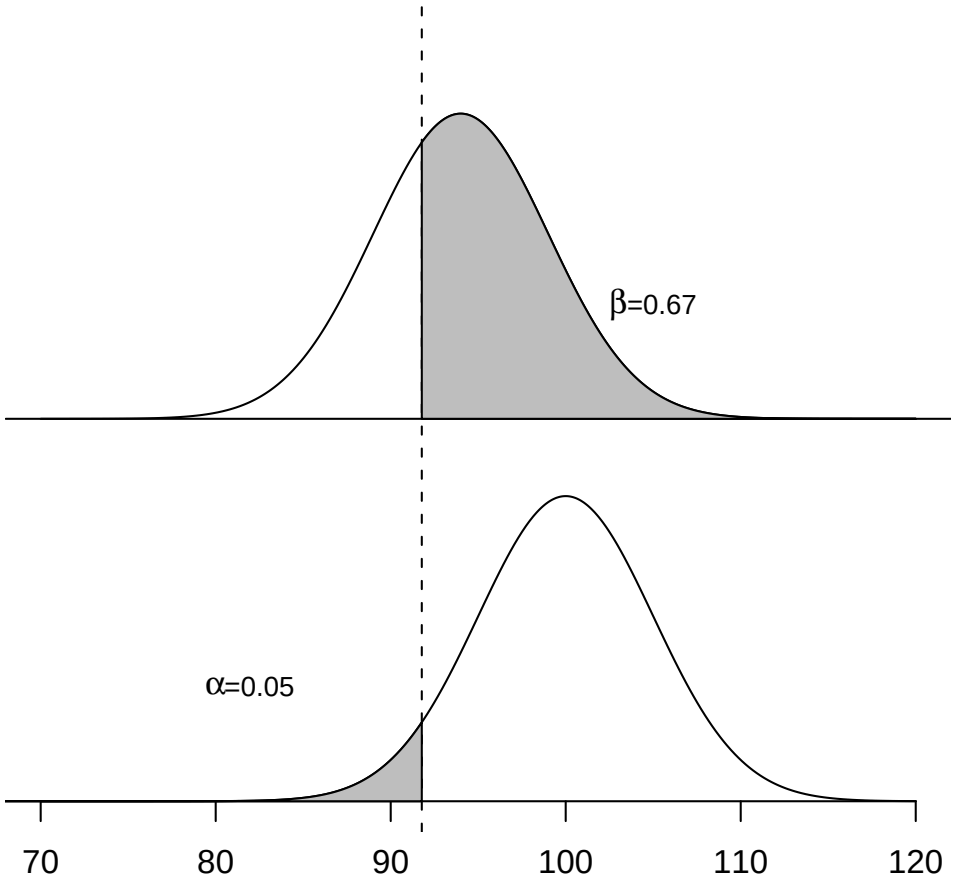


Type II error
(false negative)

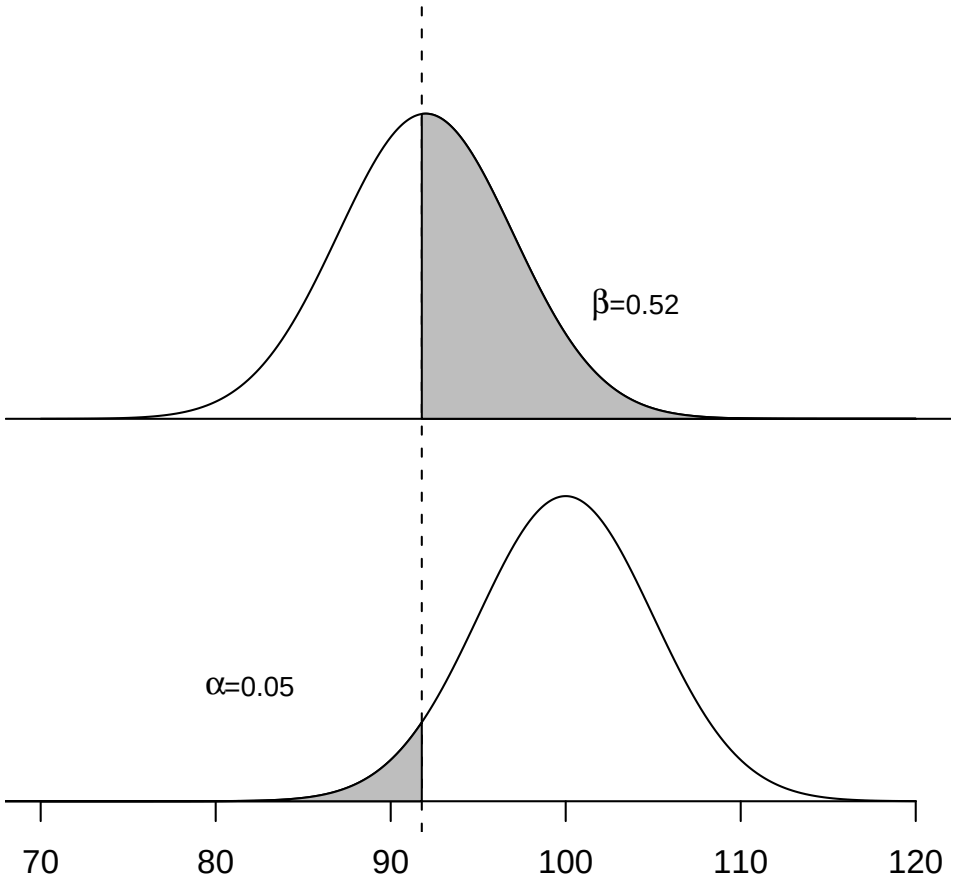
Napaka 2. vrste - vpliv povprečja



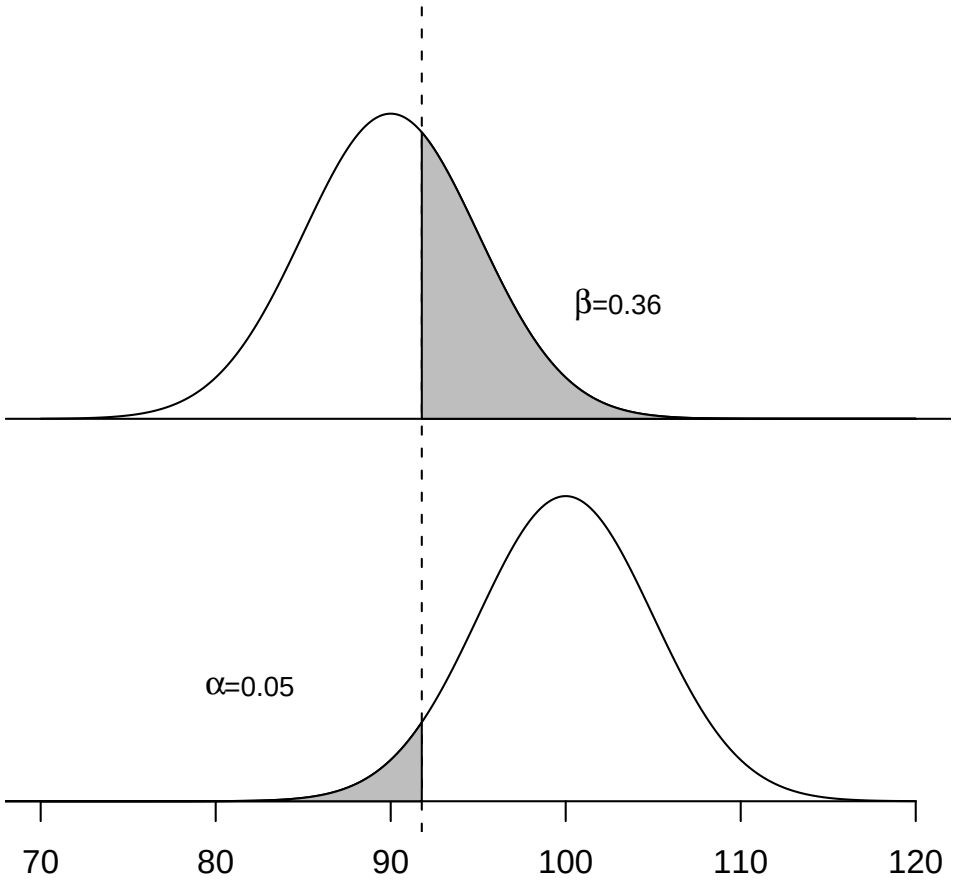
Napaka 2. vrste - vpliv povprečja



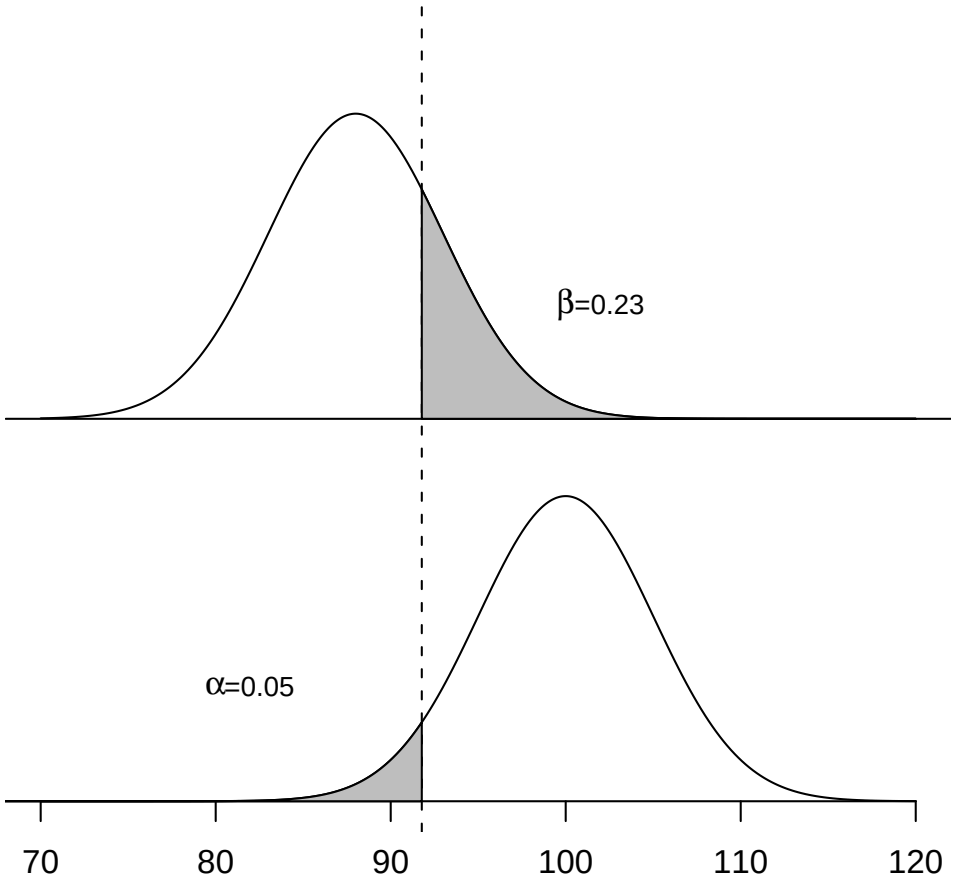
Napaka 2. vrste - vpliv povprečja



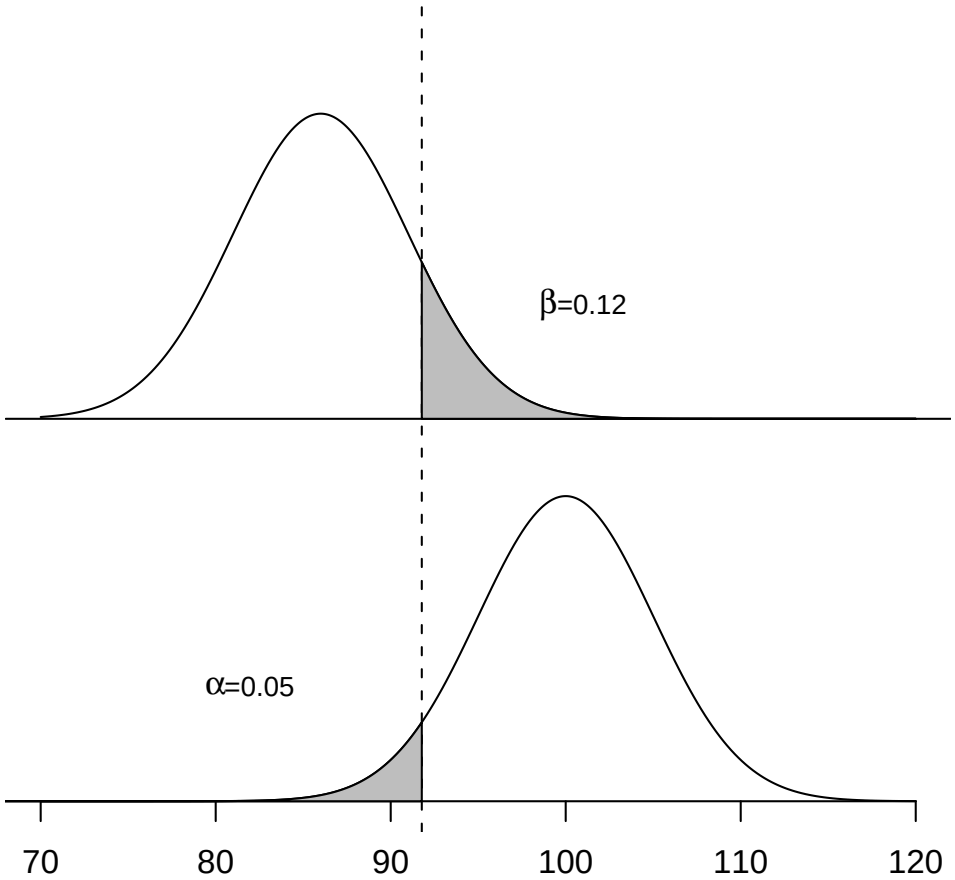
Napaka 2. vrste - vpliv povprečja



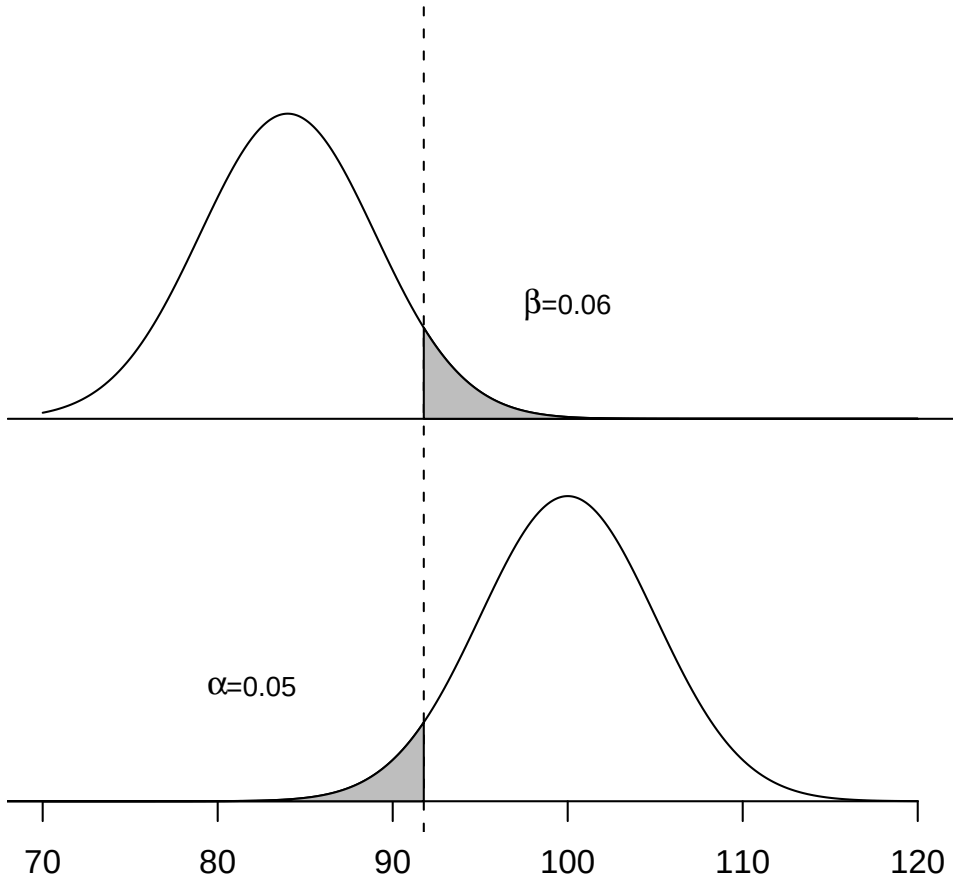
Napaka 2. vrste - vpliv povprečja



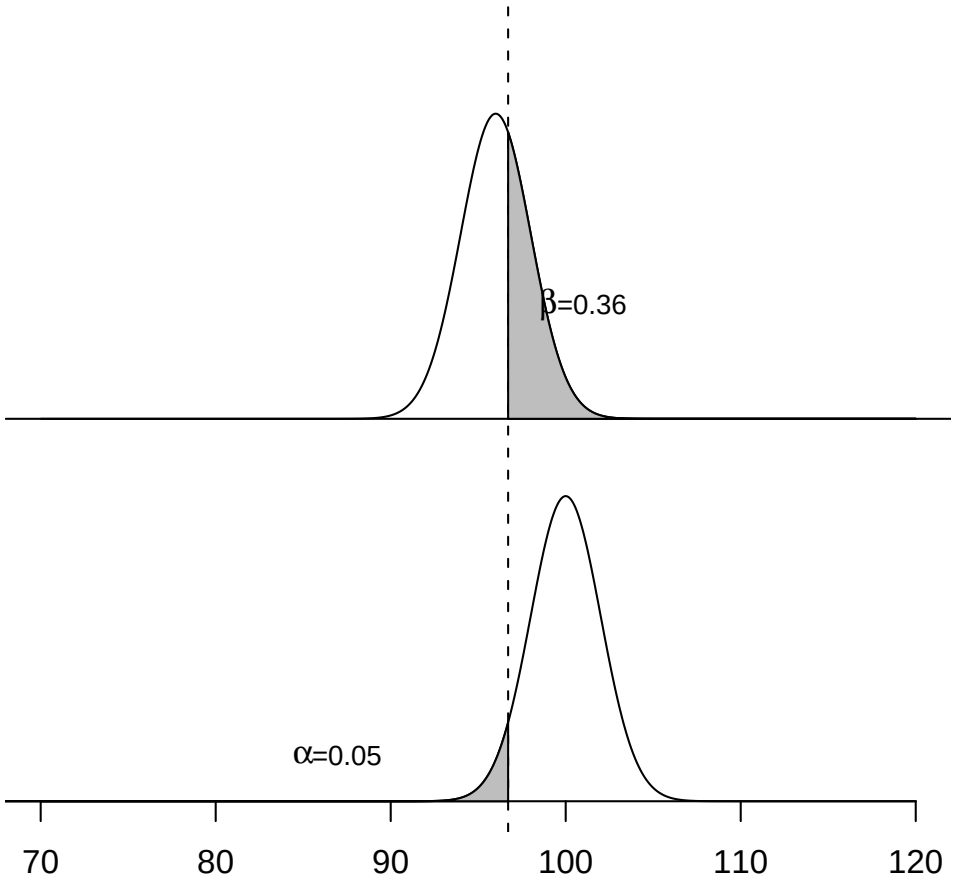
Napaka 2. vrste - vpliv povprečja



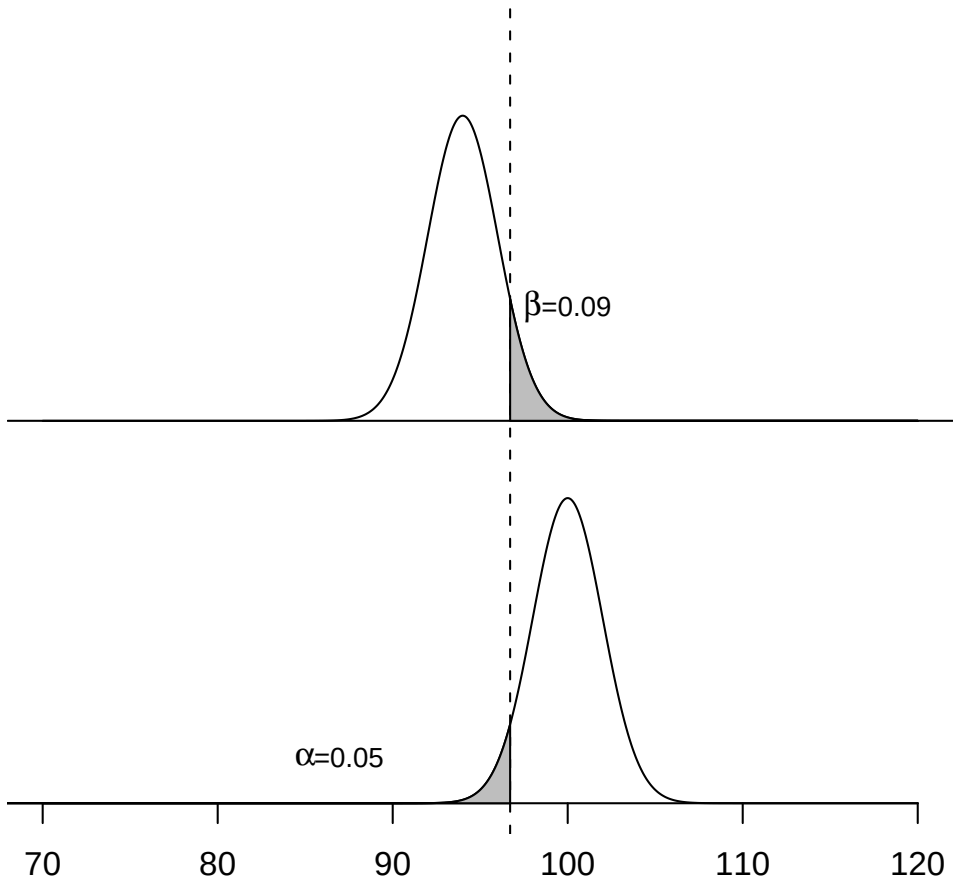
Napaka 2. vrste - vpliv povprečja



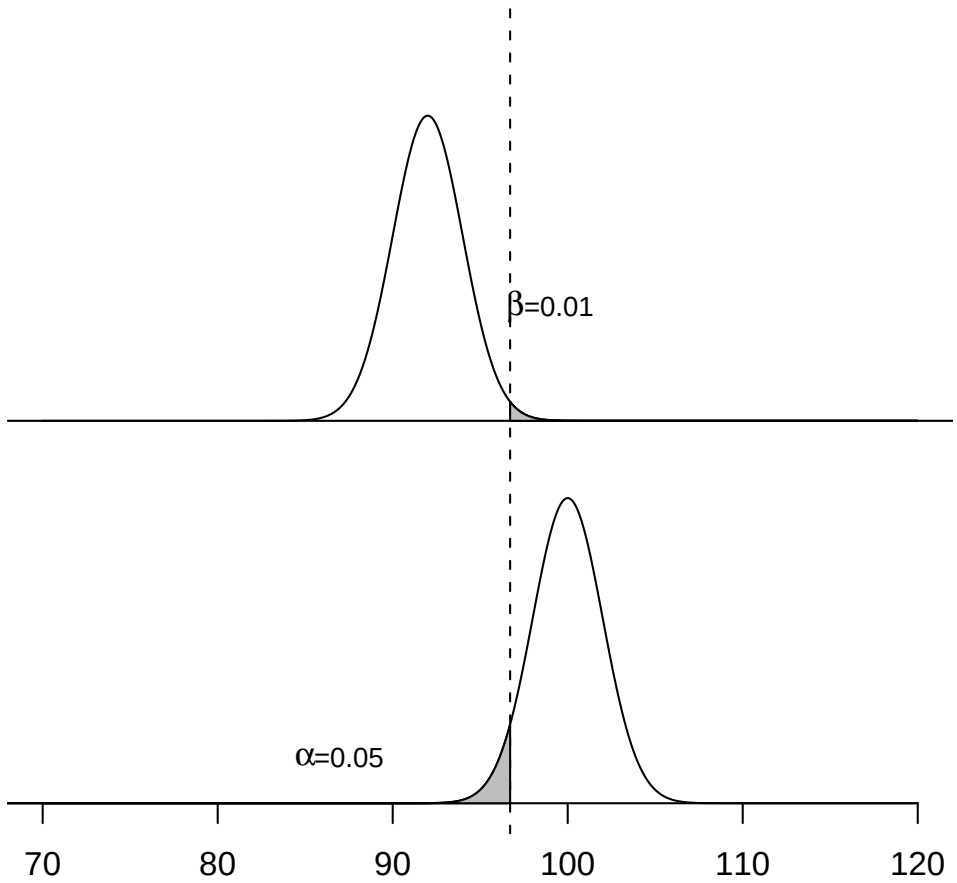
Napaka 2. vrste - vpliv povprečja ob manjši varianci



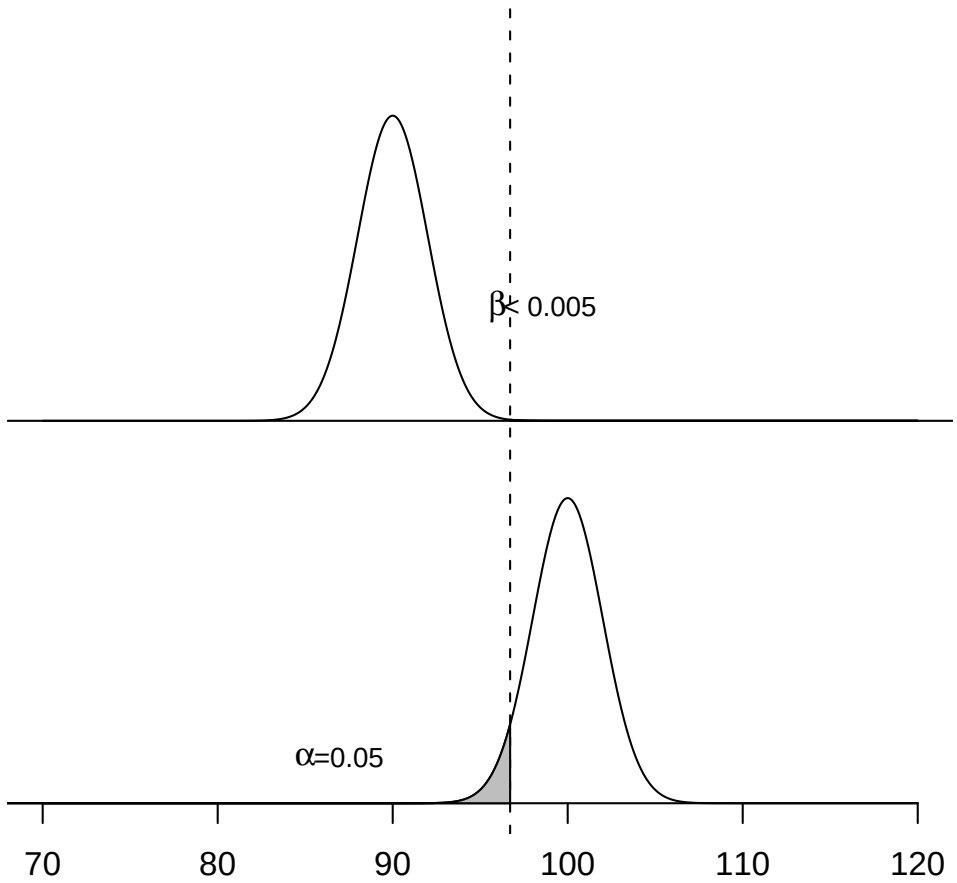
Napaka 2. vrste - vpliv povprečja ob manjši varianci



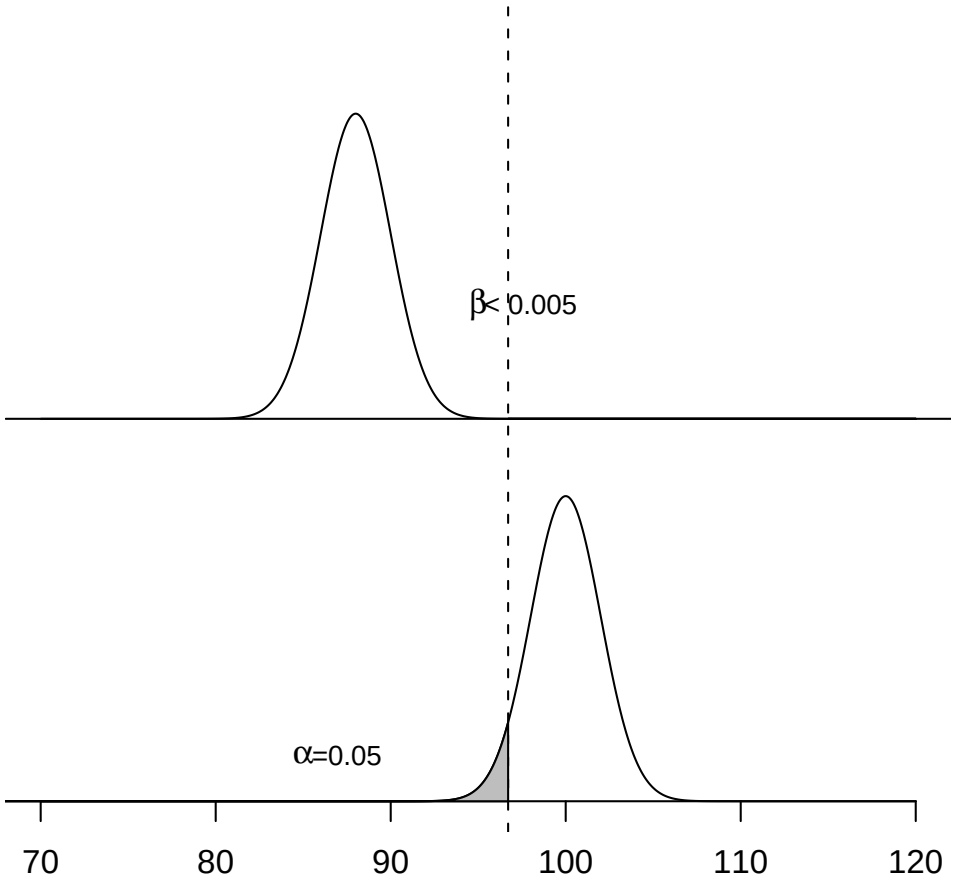
Napaka 2. vrste - vpliv povprečja ob manjši varianci



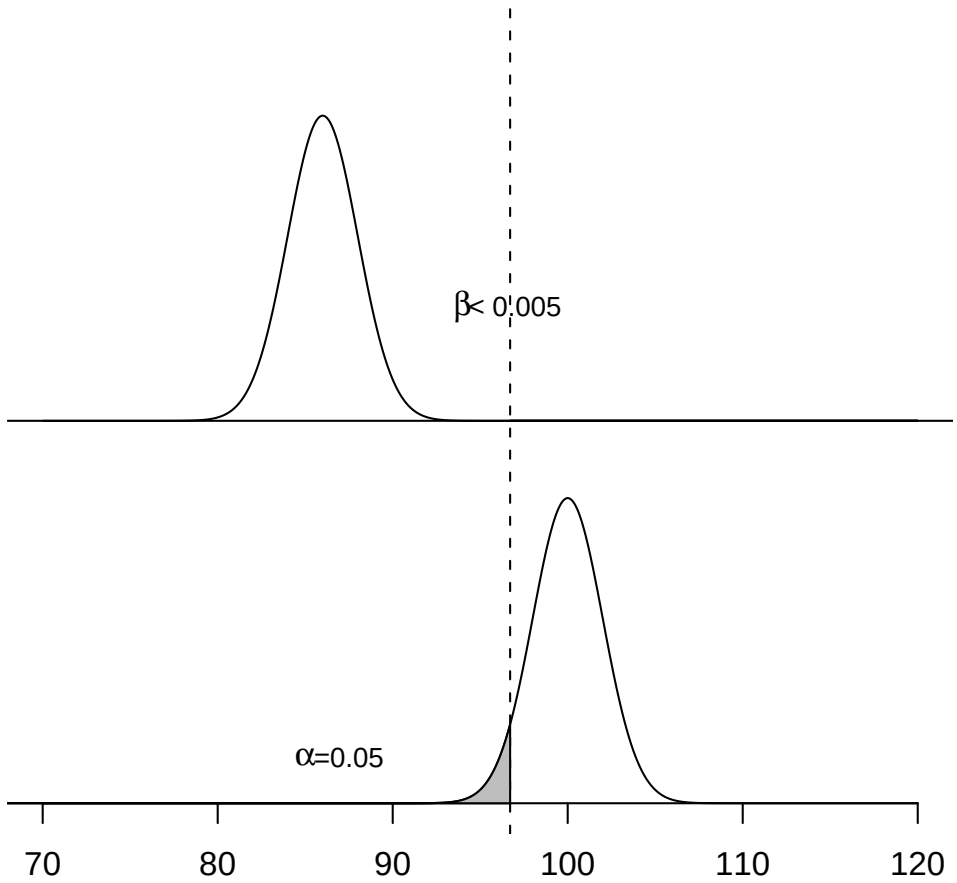
Napaka 2. vrste - vpliv povprečja ob manjši varianci



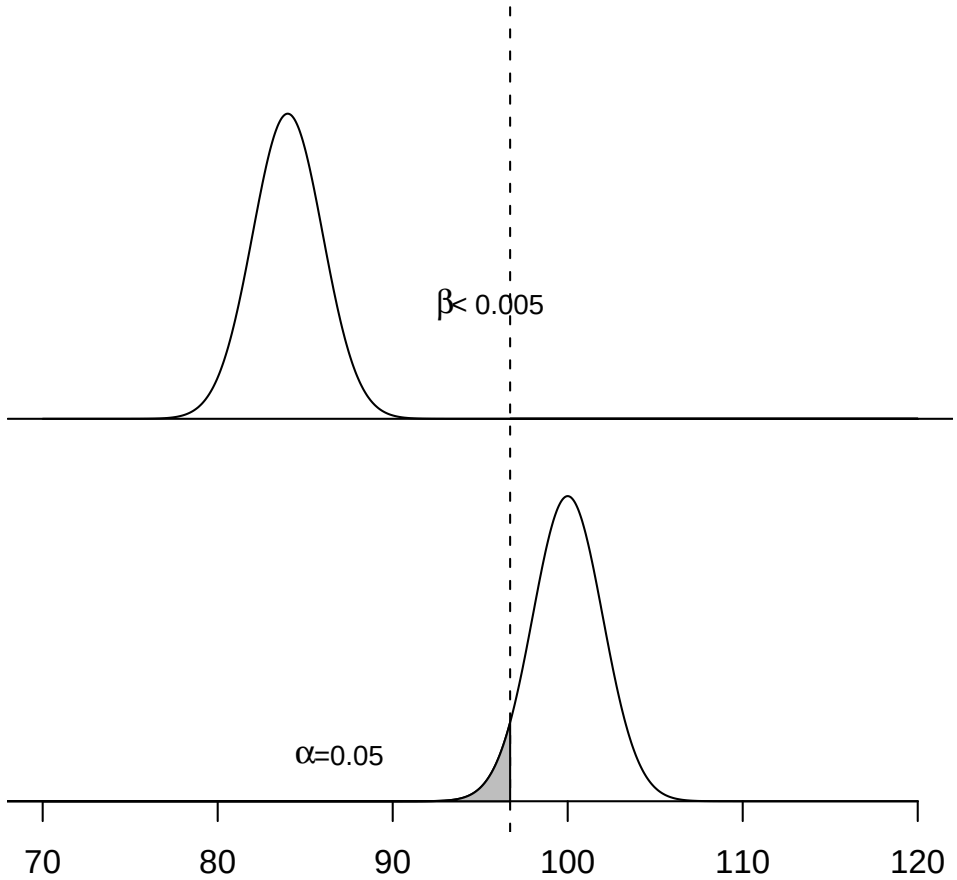
Napaka 2. vrste - vpliv povprečja ob manjši varianci



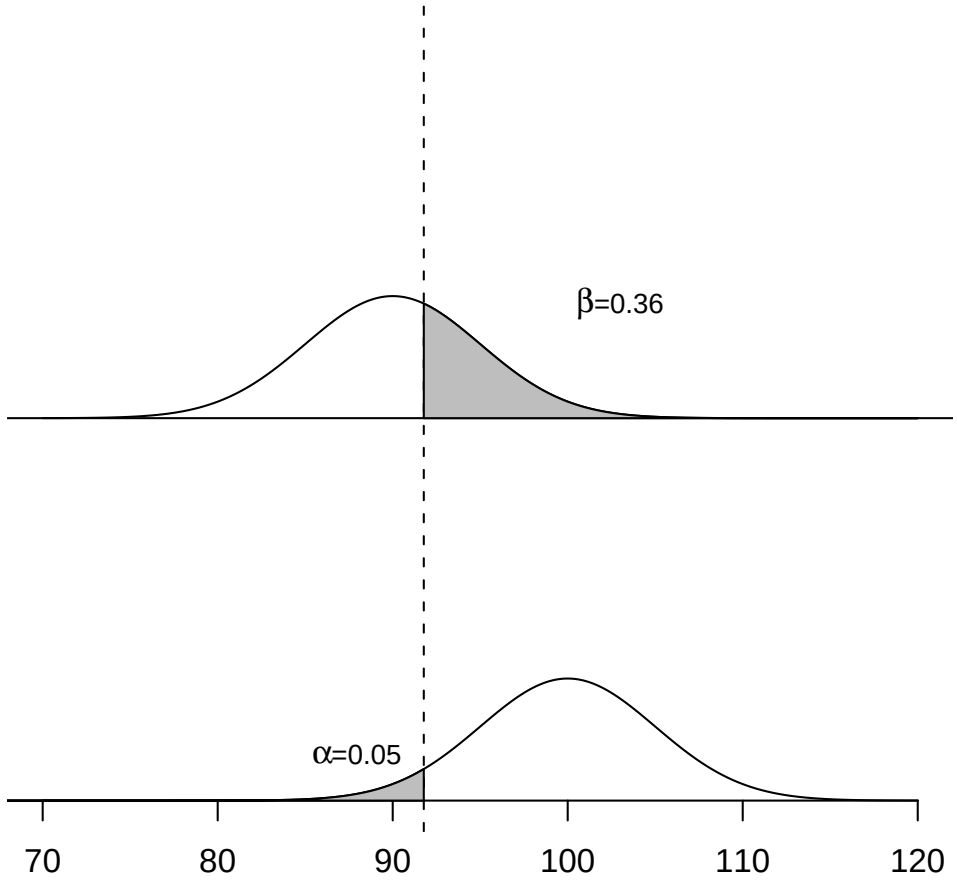
Napaka 2. vrste - vpliv povprečja ob manjši varianci



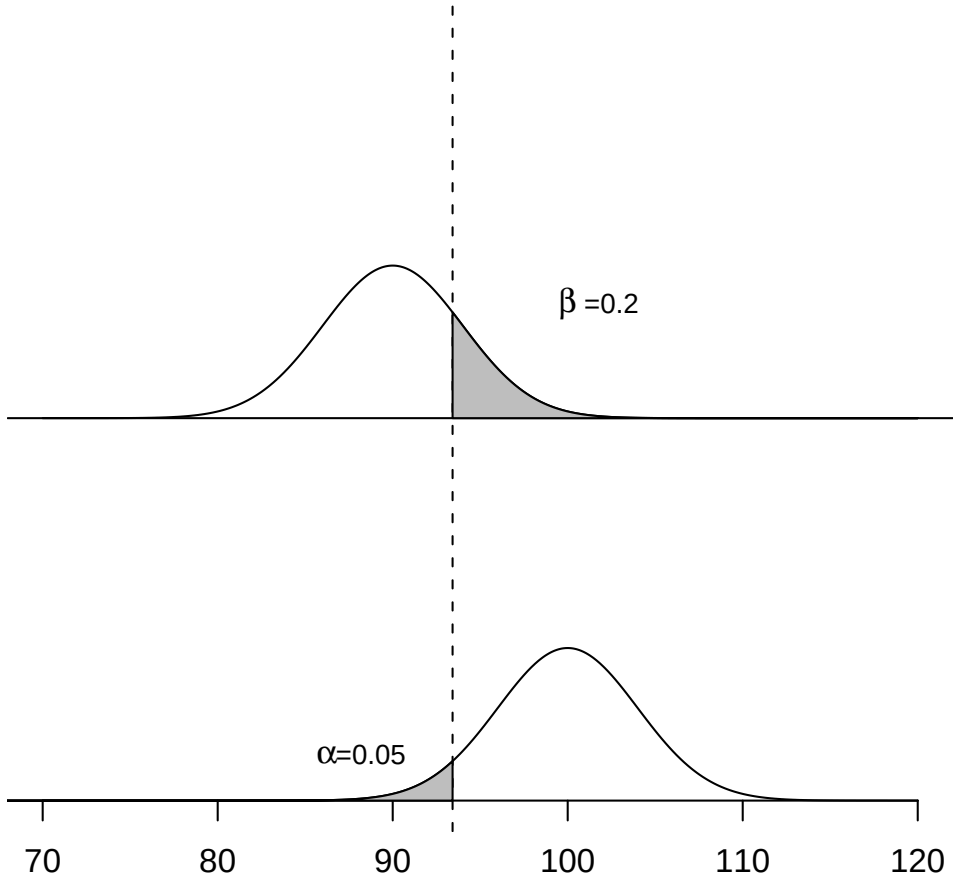
Napaka 2. vrste - vpliv povprečja ob manjši varianci



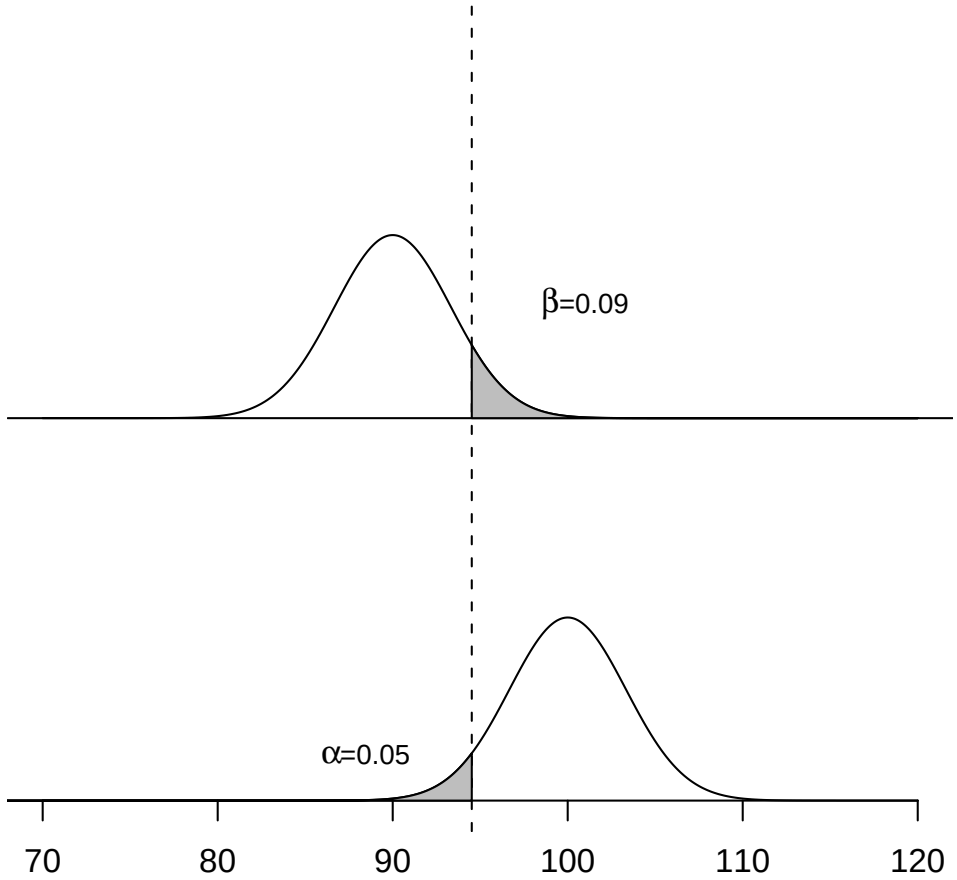
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



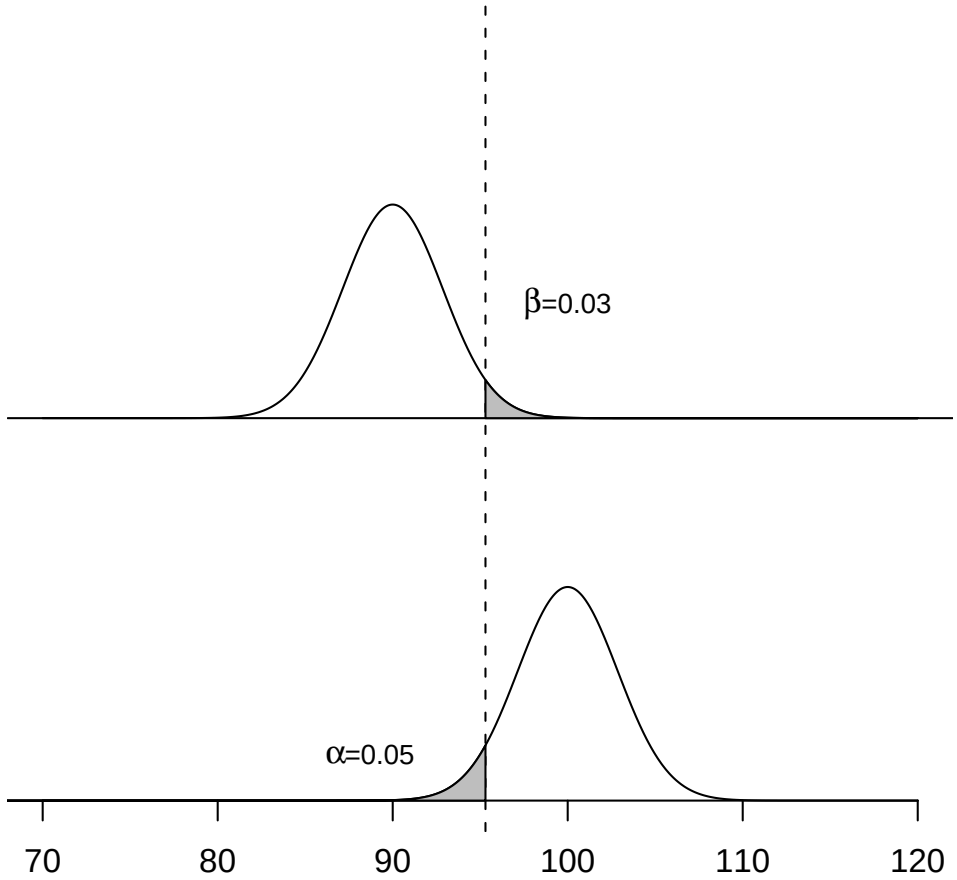
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



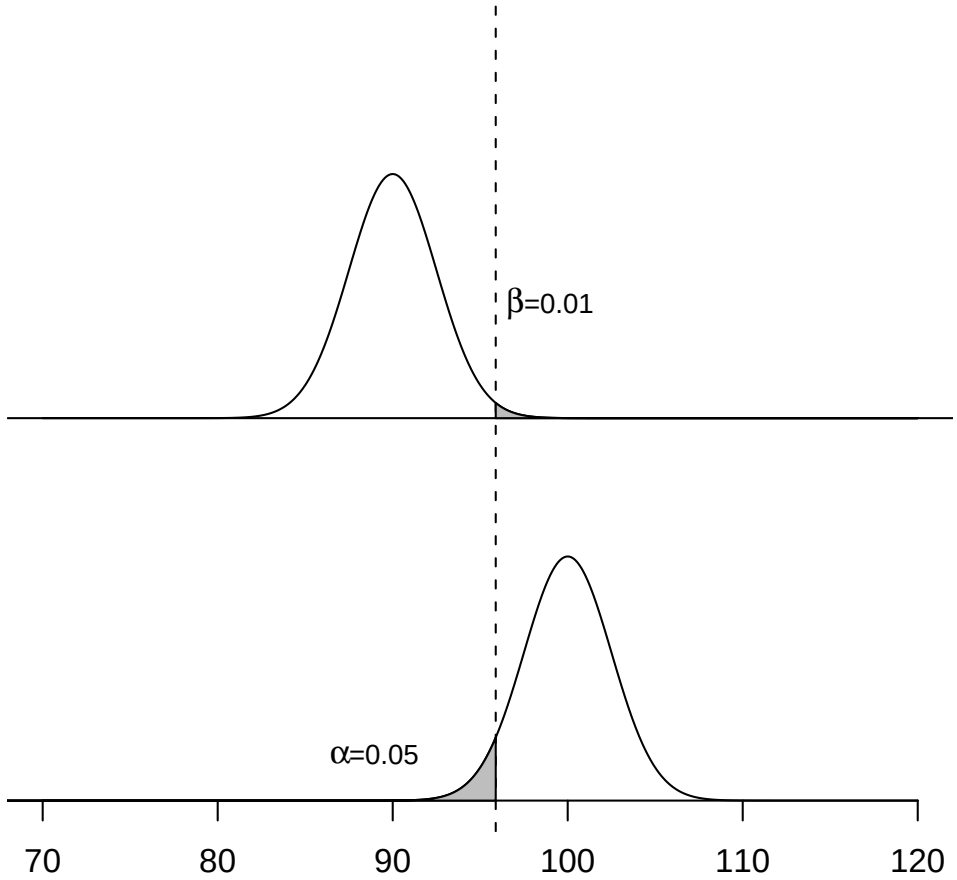
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



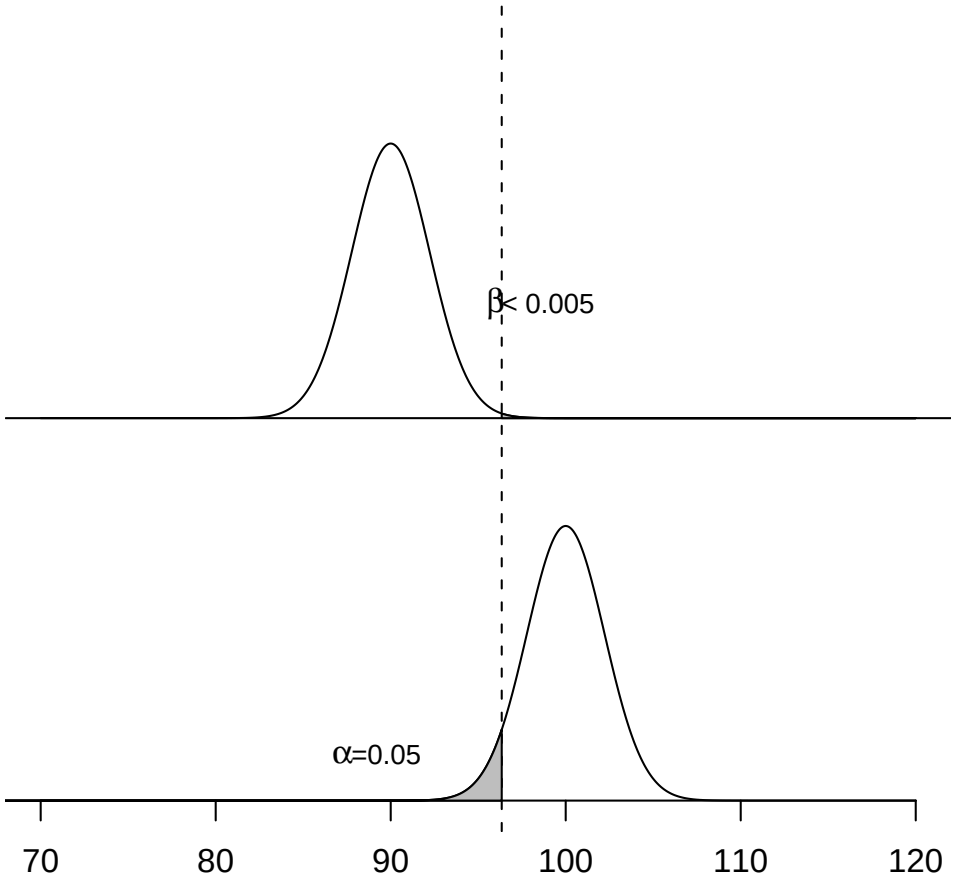
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



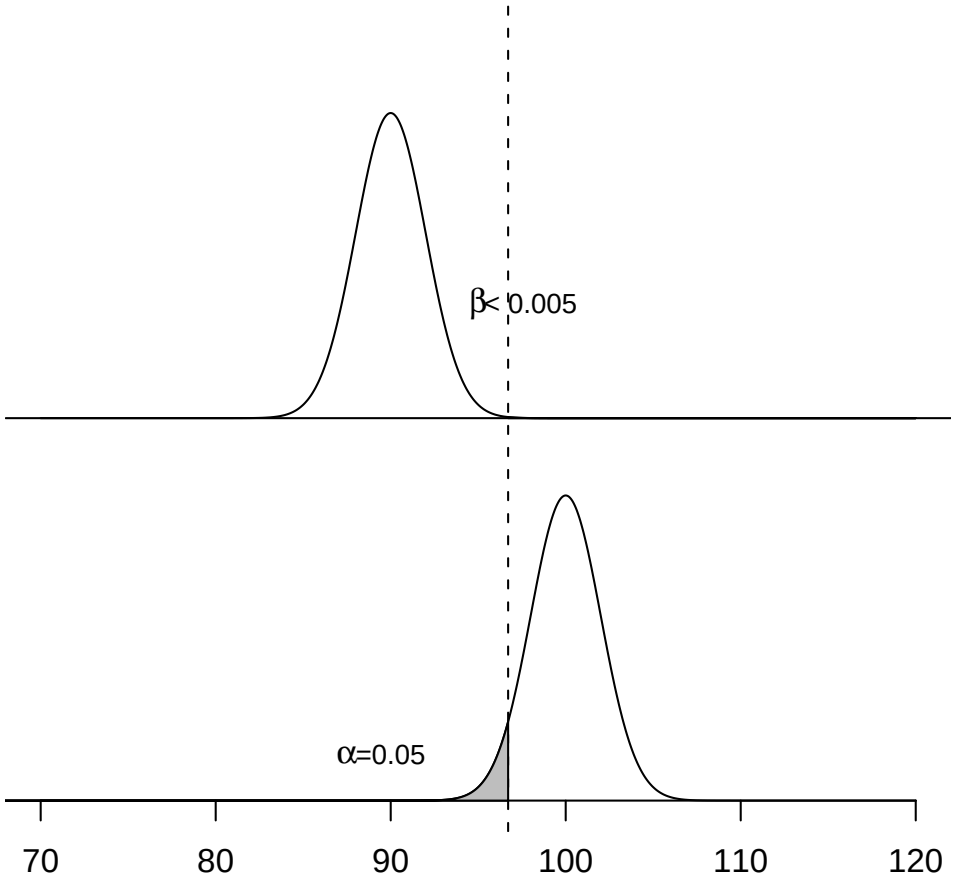
Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



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

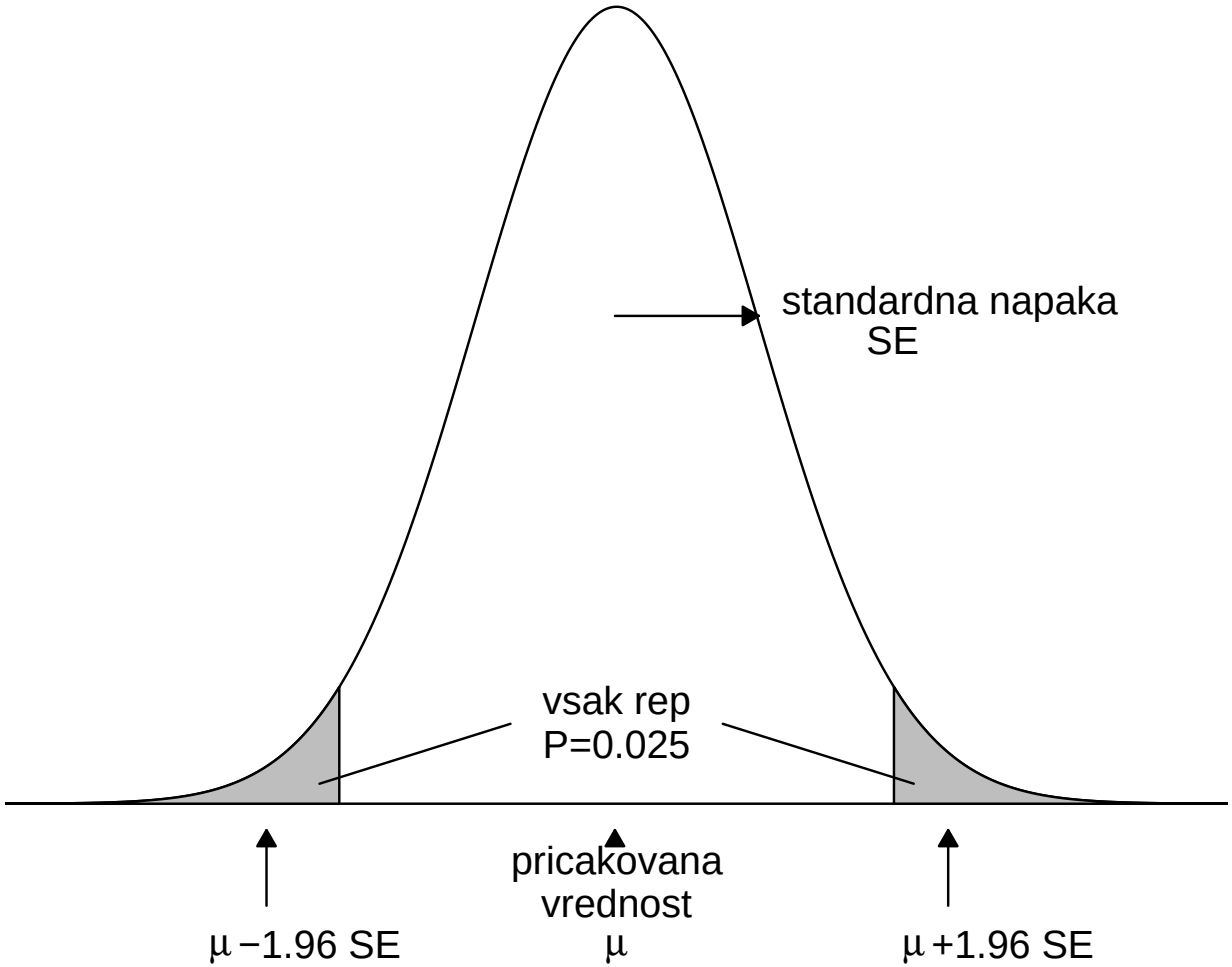


Napaka 2. vrste - vpliv velikosti vzorca (in s tem standardne napake)



INTERVALI ZAUPANJA

Spomnimo se še enkrat, kako so porazdeljena vzorčna povprečja.



Vemo torej, da obstaja 95% verjetnost, da bosta μ in $\bar{X}$ oddaljena za manj kot 1,96 standardne napake. Torej

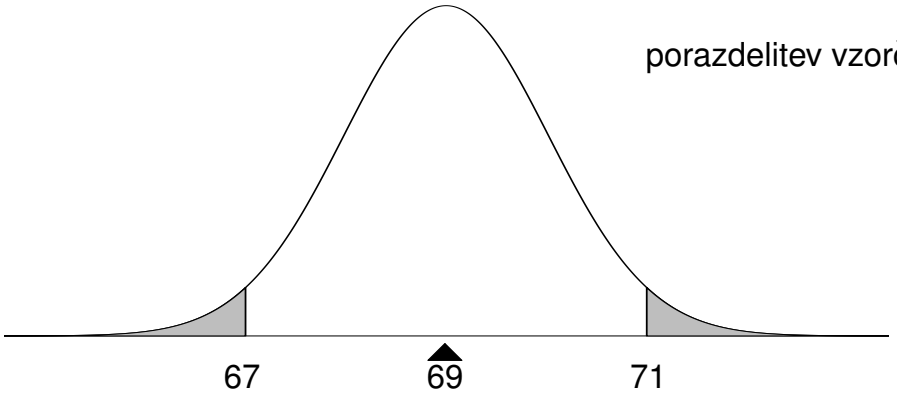
$$P(|\mu - \bar{X}| \leq 1{,}96SE) = 0{,}95$$

oziroma

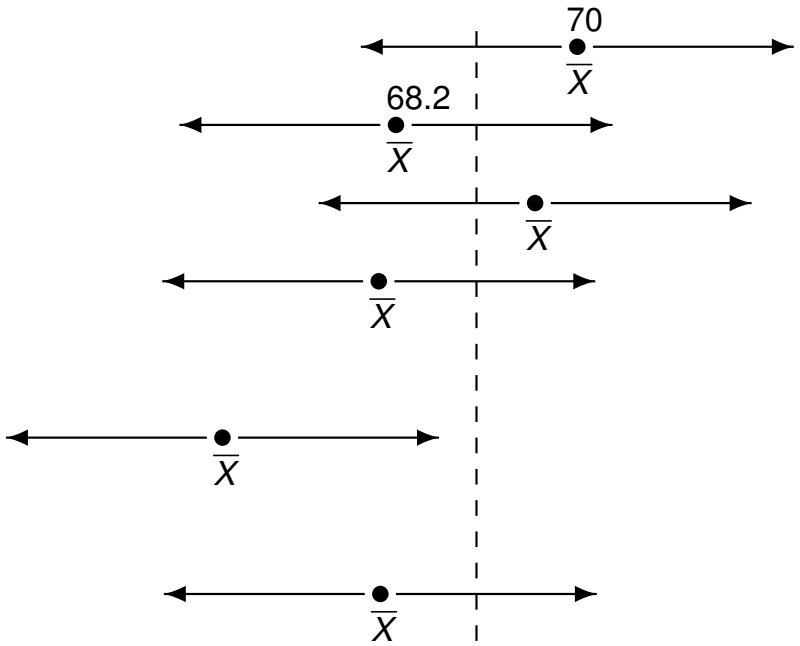
$$P(\bar{X} - 1{,}96SE \leq \mu \leq \bar{X} + 1{,}96SE) = 0{,}95.$$

Pri tem poudarimo, da je spremenljivka v gornjih izrazih $\bar{X}$. Torej, 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**.



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

V praksi je seveda takole:

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

Interpretacija intervala zaupanja

Ena:

Če jemljemo večkrat vzorce velikosti n enot iz populacije, bo interval $(\bar{X} - 1{,}96SE, \bar{X} + 1{,}96SE)$ v 95% primerov pokril μ.

Druga:

Katerokoli vrednost v intervalu zaupanja izberemo kot možen μ, ji naš interval ne bo nasprotoval.

Do sedaj smo intervale zaupanja računali, kot da σ poznamo. To bo seveda redko res in smo torej pred podobnim problemom, kot pri testu enega vzorca. Rešitev je spet porazdelitev t. Zdaj bo pač veljalo

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

kjer je t_α tista vrednost porazdelitve t z $n-1$ stopinjami prostosti, zunaj katere je 5% vseh vrednosti te porazdelitve. Interval zaupanja je potem

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

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

Primer: Interval zaupanja za porodno težo otrok 35-letnih mater.

Ker je bilo naše vzorčno povprečje 3,436kg, varianca pa 0,20916 in torej standardni odklon $s = \sqrt{0{,}20916} = 0{,}45734$, ustrezna t-vrednost pri 9 stopinjah prostosti pa je 2,262, je 95% interval zaupanja za porodno težo

$$(3{,}436 - 2{,}262\frac{0{,}45734}{\sqrt{10}},\ 3{,}436 + 2{,}262\frac{0{,}45734}{\sqrt{10}}) = (3{,}109,\ 3.763)$$

OCENJEVANJE VARIANCE

Videli smo, da pri vzorcu velikosti n, vzetem iz normalne populacije, velja, da je $\bar{X} \sim \mathcal{N}(\mu,\sigma^2/n)$ in spoznali smo, kako to dejstvo uporabimo pri ocenjevanju povprečja in pri statističnem sklepanju o povprečju. Enako želimo tudi za varianco. V ta namen moramo vedeti, kako je porazdeljena vzorčna varianca.

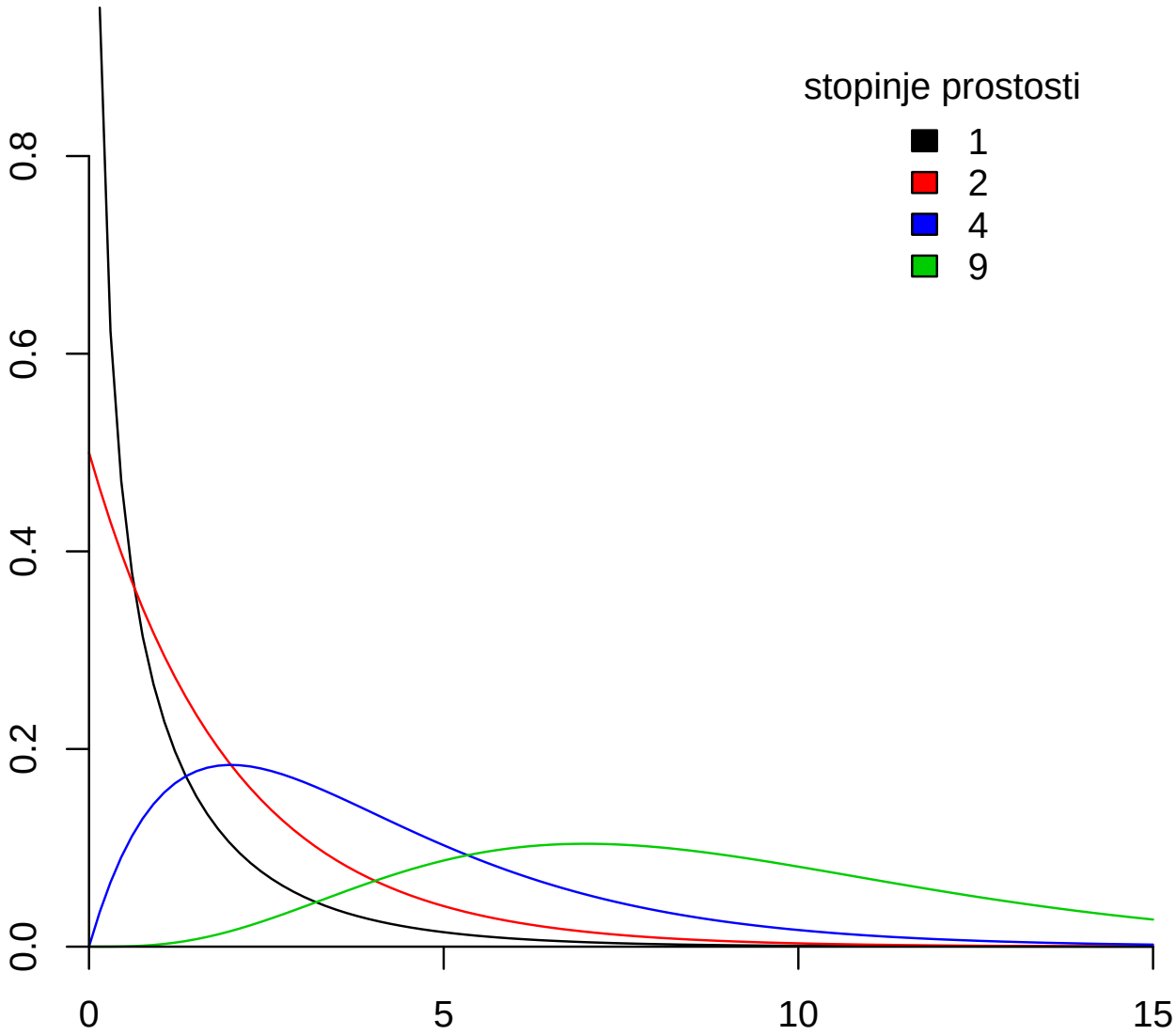
Brez dokazovanja bomo v nadaljevanju navedli nekaj teoretičnih rezultatov, katerih dokazovanje presega okvire tega gradiva. Najprej

Izrek: Če so $Z_1, \ldots, Z_n$ neodvisne standardizirano normalne spremenljivke, je

$$Y = Z_1^2 + \cdots + Z_n^2$$

porazdeljena po porazdelitvi χ^2 z n stopinjami prostosti.

Izraza za gostoto porazdelitve χ^2 ne bomo pisali, nekaj grafov je na naslednji sliki.


stopinje prostosti
1
2
4
9
0.0
0.2
0.4
0.6
0.8
0
5
10
15

Če je Y porazdeljena kot χ^2 z f stopinjami prostosti bomo pisali

$$Y \sim \chi^2(f)$$

in v splošnem velja, da je

$$E(Y) = f \quad \text{in} \quad Var(Y) = 2 \cdot f$$

Naj bo zdaj $X_1, X_2, \ldots, X_n$ slučajni vzorec iz $\mathcal{N}(\mu, \sigma^2)$. Slučajna spremenljivka

$$Y = \left(\frac{X_1 - \mu}{\sigma}\right)^2 + \left(\frac{X_2 - \mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_n - \mu}{\sigma}\right)^2$$

je potem porazdeljena po porazdelitvi χ^2 z n stopinjami prostosti.

Velja pa še tole: če v gornjem izrazu nadomestimo populacijsko povprečje z vzorčnim povprečjem, dobimo slučajno spremenljivko

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2,$$

ki je prav tako porazdeljena po porazdelitvi χ^2, vendar le z $n-1$ stopinjami prostosti. Torej lahko zapišemo

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \sigma^2 \cdot \chi^2(n-1)$$

in potem imamo

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \frac{\sigma^2}{n-1} \cdot \chi^2(n-1),$$

kar pomeni, da je porazdelitev vzorčne variance znana, če poznamo σ^2 (huda zahteva!).

Pomembno pa je, da iz prejšnje enačbe vidimo, da velja

$$E(S^2) = \sigma^2 \quad \text{in} \quad Var(S^2) = \frac{2 \cdot \sigma^4}{n-1}$$

iz česar vidimo, da

1. natančnost ocene narašča z velikostjo vzorca.
2. je pričakovana vrednost vzorčne variance ravno populacijska varianca, kar pa ne bi bilo res, če bi v izrazu za vzorčno varianco delili z n.

Prej smo že odgovorili na vprašanje ali so podatki o desetih porodnih težah otrok 35-letnih mater v skladu s hipotezo, da prihajajo iz populacije z znanim povprečjem. Naredimo to zdaj še za varianco.

Populacijska varianca σ^2 je 0,334514, vzorčna pa 0,20916. Če predpostavimo, da se varianca tež otrok 35-letnih mater ne razlikuje od populacijske (**ničelna hipoteza**), kako to preveriti?

Če je ničelna hipoteza pravilna, potem je vzorčna varianca s^2 eno opazovanje iz porazdelitve

$$\frac{0{,}334514}{10-1} \cdot \chi^2(10-1).$$

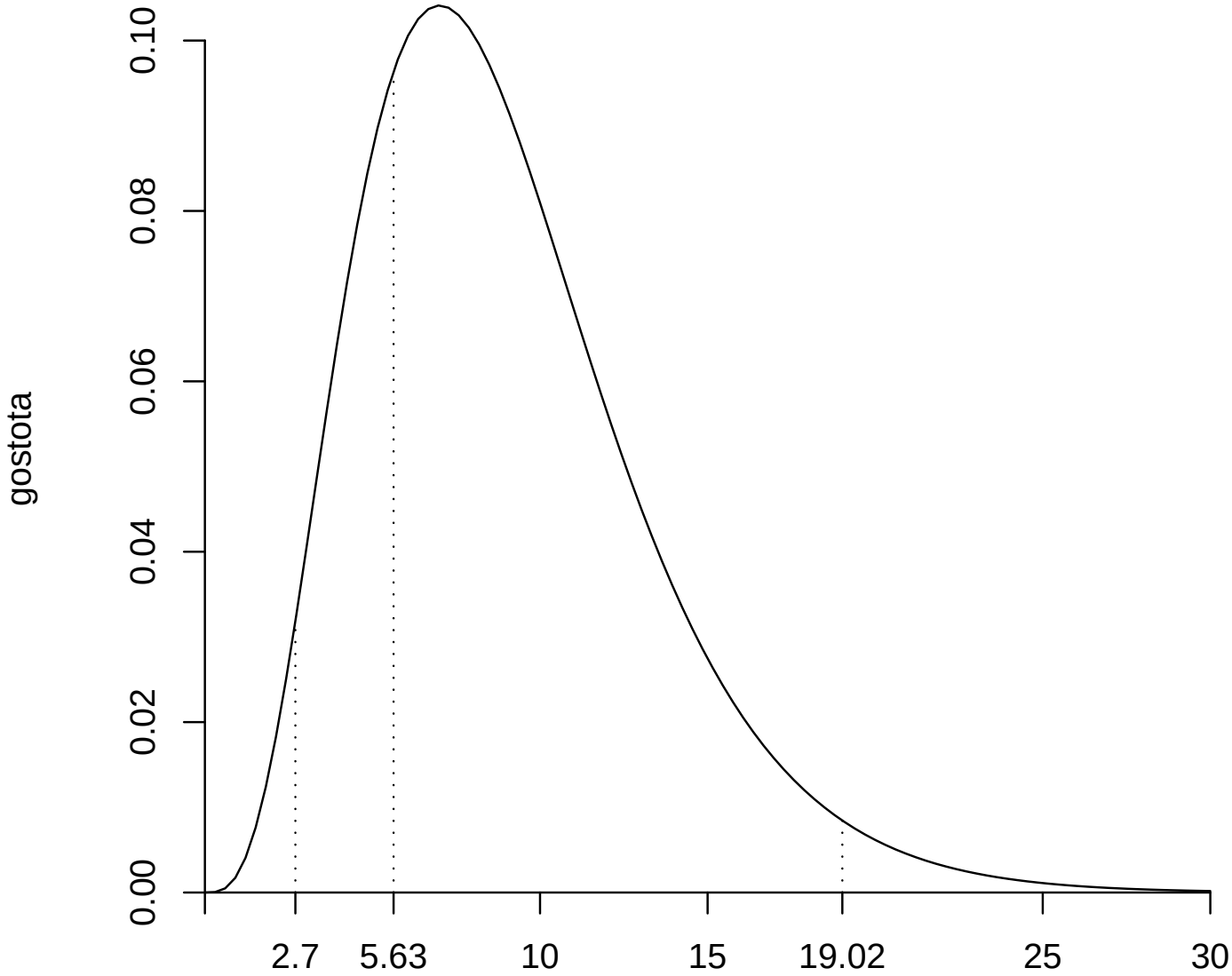
oziroma

$$\chi^2_u = \frac{s^2 \cdot (10-1)}{0{,}334514} = 5{,}63$$

eno opazovanje iz porazdelitve χ^2 z 9 stopinjami prostosti. Z indeksom u smo označili, da gre za ugotovljeno (izmerjeno) vrednost.

Velika vrednost χ^2_u pomeni, da je prava (neznana) varianca **večja** od 0,334514, **majhna** vrednost χ^2_u pa pomeni, da je prava varianca **manjša** od 0,334514.

Na naslednji sliki sta porazdelitvi χ^2 z 9 stopinjami prostosti vrisani vrednosti, zunaj katerih je 5% vseh vrednosti. Poleg tega je vrisana tudi vrednost 5,63, ki smo jo dobili v našem primeru.


gostota
0.00
0.02
0.04
0.06
0.08
0.10
2.7
5.63
10
15
19.02
25
30

Vidimo, da dobljena vrednost ni zelo nenavadna, iz česar sklepamo, da naši podatki ne nasprotujejo ničelni hipotezi. Natančno bi lahko izračunali, da je

$$P(\chi^2(9) \leq 5{,}63) = 0{,}22$$

in

$$P(\chi^2(9) \geq 5{,}63) = 0{,}78$$

Interval zaupanja za varianco

Spomnimo se še enkrat, da velja

$$\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1).$$

Potemtakem bo z verjetnostjo 0,95 veljalo

$$\chi^2_{0,025}(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{0,975}(n-1)$$

iz česar sledi, da je 95% interval zaupanja za varianco

$$\frac{(n-1)s^2}{\chi^2_{0,975}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{0,025}(n-1)}$$

Primer: Interval zaupanja za varianco teže novorojenčkov.

Imamo $10 - 1 = 9$ stopinj prostosti in ustrezni kritični vrednosti porazdelitve χ^2 sta $\chi^2_{0,975} = 19{,}02$ in $\chi^2_{0,025} = 2{,}70$. Torej je, ker je ocena variance na vzorcu enaka 0,20916, interval zaupanja za varianco v populaciji

$$\frac{9 \cdot 0{,}20916}{19{,}02} < \sigma^2 < \frac{9 \cdot 0{,}20916}{2{,}70}$$

$$0.099 < \sigma^2 < 0.697$$

za standardni odklon pa

$$0.315 < \sigma < 0.835$$

Interval za varianco seveda vključuje predpostavljeno vrednost 0,334514, saj ji vzorčna vrednost ni nasprotovala (in torej test ni bil statistično značilen).

VZORČENJE

Naključno vzorčenje

V praksi bomo imeli praviloma opravka z vzorci, hoteli pa bomo nekaj povedati o populaciji. Pri tem je populacija tista množica enot, ki nas zanima, vzorec pa je del te množice. Seveda je za sklepanje o populaciji iz vzorca nujno, da je vzorec za populacijo reprezentativen. To najlažje dosežemo tako, da iz populacije **vzorčimo naključno**. To pomeni, da ima vsaka enota populacije enako verjetnost, da bo izbrana. (Če smo natančni, takšna definicija velja le za končne populacije, v splošnem govorimo o slučajnem vzorcu kot o realizaciji določenega števila neodvisnih, enako porazdeljenih slučajnih spremenljivk.)

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

Tretja, in mnogo bolj preprosta možnost, je 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.

Sistematično vzorčenje

Tu imamo zopet seznam vseh možnih enot, vendar le prvega izberemo naključno, ostale pa tako, da izberemo vsakega *n*-tega.

Najprej določimo vzorčni delež tako, da velikost vzorčnega okvira delimo s predvideno velikostjo vzorca. 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. Prvo enoto naključno izberemo izmed prvih 5 (lahko z uporabo slučajnih števil), potem pa vsako peto.

Možna slaba stran: periodičnost v podatkih.

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

RANDOMIZACIJA

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

RANDOMIZACIJA

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

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

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.

Kako randomizirati? Možnosti: kovanec, kocka, karte, kolo sreče.

Tabela slučajnih števil:

Kako randomizirati? Možnosti: kovanec, kocka, karte, kolo sreče.

Tabela slučajnih števil:

1. v dve skupini (sodo, liho)

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)

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 hocemo enake skupine: ko napolnimo eno skupino, preostale v drugo - ne najbolje

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 hocemo enake skupine: ko napolnimo eno skupino, preostale v drugo - ne najbolje
4. lahko slučajno izberemo najprej prvo skupino, ostali so druga

ANALIZA PARNIH PODATKOV

Primer: Povprečna količina zaužite hrane (v kJ) v desetih predmenstrualnih in desetih pomenstrualnih dneh.

Oseba i	Pred X	Po Y	Razlika D
1	5260	3910	-1350
2	5470	4220	-1250
3	5640	3885	-1755
4	6180	5160	-1020
5	6390	5645	-745
6	6515	4680	-1835
7	6805	5265	-1540
8	7515	5975	-1540
9	7515	6790	-725
10	8230	6900	-1330
11	8770	7335	-1435

Vprašanje: Ali je kakšna razlika v količini zaužite hrane med obema obdobjema?

Možna analiza: Izračunamo razlike med dvema meritvama, se pravi $d_i = y_i - x_i$. Če lahko privzamemo, da je D normalno porazdeljena, uporabimo t-test za en vzorec. Po analogiji z

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

je ustrezni test potem

$$t_{n-1} = \frac{\bar{d}}{s_d/\sqrt{n}},$$

kjer smo za ničelno hipotezo privzeli, da je pričakovana vrednost razlike enaka 0.

Absolutna ali relativna napaka?

Ko gledamo parne podatke (x,y) in se sprašujemo, kako se y-i razlikujejo od x-ov, je naravno, da najprej pomislimo na razlike med njimi. Toda, ali naj gledamo absolutne ali relativne razlike?

$$\textbf{Absolutna razlika} = y - x$$

$$\textbf{Relativna razlika} = \frac{y-x}{x} = \frac{y}{x} - 1$$

Če velja, da je $y \approx x + a$, je razlika precej stabilna in je absolutna razlika ustrezna mera spremembe. Parni t-test, ki smo ga opisali je takrat primeren.

Če pa velja, da je $y \approx b \cdot x$, so stabilni kvocienti in je ustrezneje uporabiti relativno razliko kot mero spremembe. Ker je potem $log(y) \approx log(x) + log(b)$, s parnim testom v takšnem primeru rajši analiziramo razlike med logaritmi.

ANALIZA DVEH NEODVISNIH VZORCEV

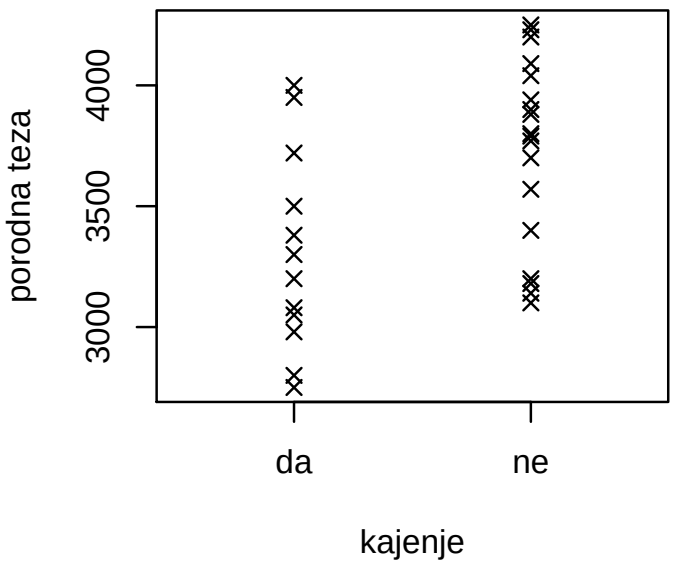
Primer: Kajenje in porodna teža

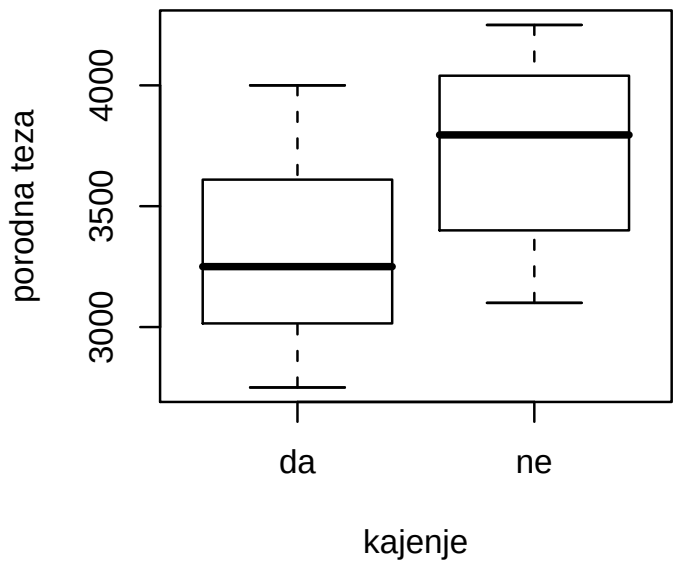
Spodaj so podatki o porodnih težah 30 otrok, posebej za matere kadilke (12) in nekadilke (18).

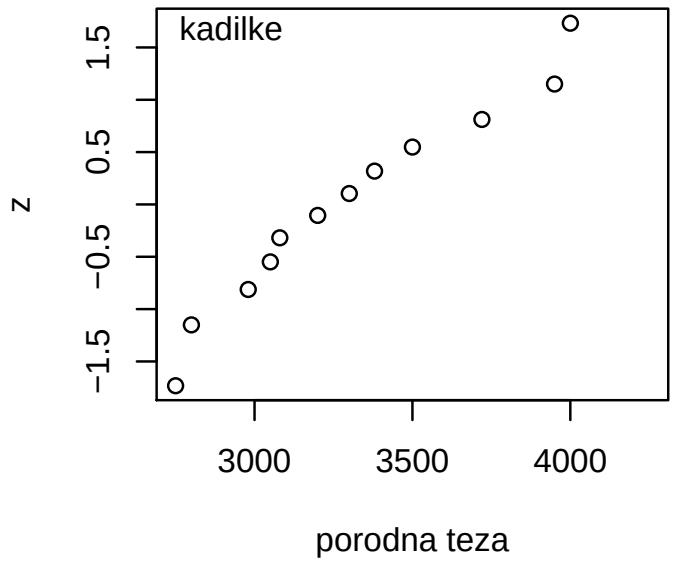
Mati **kadilka**: 3080, 2980, 3500, 2750, 3200, 3050, 4000, 3720, 3300, 3380, 3950, 2800.

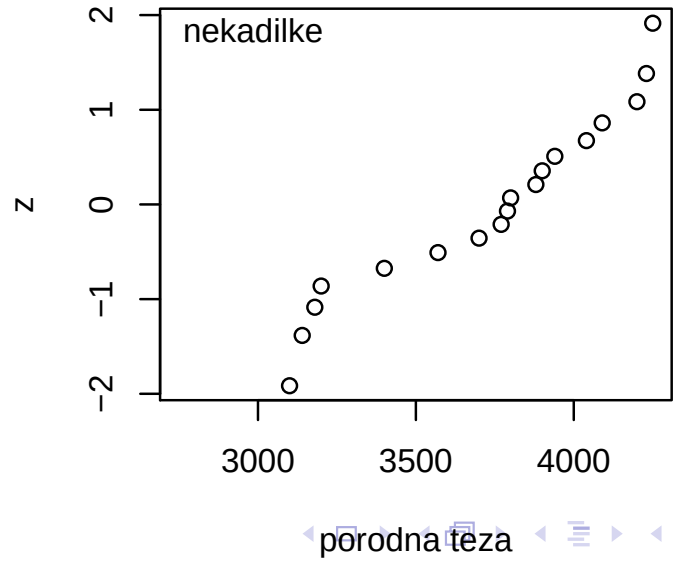
Mati **nekadilka**: 4250, 3900, 3200, 3790, 4040, 3770, 3180, 4230, 3570, 3800, 3700, 3940, 3140, 4200, 3880, 4090, 3400, 3100.

Skupina	n	$\bar{x}$	s	Mediana
kadilke	12	3309	416	3250
nekadilke	18	3732	386	3795


porodna teza
3000
3500
4000
da
ne
kajenje


porodna teza
3000
3500
4000
da
ne
kajenje


kadilke
z
−1.5
−0.5
0.5
1.5
3000
3500
4000
porodna teza


nekadilke
z
−2
−1
0
1
2
3000
3500
4000
porodna teza

Privzeli bomo naslednji **statistični model**:

1. Množici meritev v obeh skupinah sta neodvisni.
2. Meritve v vsaki skupini so slučajen vzorec iz normalne populacije z naslednjimi parametri:

Skupina	Povprečje	Varianca
kadilke	μ_1	σ_1^2
nekadilke	μ_2	σ_2^2

Zapisano drugače:

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Iz tega potem sledi, da sta povprečji porazdeljeni takole

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

razlika povprečij pa takole

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Test za ničelno hipotezo

$$H_0 : \mu_1 - \mu_2 = \delta$$

je potem lahko

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Seveda smo spet pred že znanim problemom - da bi takšen test lahko uporabili, moramo poznati σ_1 in σ_2. In ker ju ponavadi ne, se zdi naravno, da uporabimo ustaljeni recept - nadomestimo populacijski vrednosti z vzorčnima in poglejmo kako je porazdeljen izraz

$$\frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{1}$$

Žal stvar ni enostavna, nekaj več o tem pozneje.

Če pa lahko privzamemo, da velja

$$\sigma_1^2 = \sigma_2^2 = \sigma^2,$$

torej da sta varianci enaki, potem je izraz

$$t = \frac{(\bar{x_1} - \bar{x_2}) - \delta}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

porazdeljen po porazdelitvi t z $n_1 + n_2 - 2$ stopinjami prostosti. Pri tem je s^2 ocena skupne variance, ki jo dobimo takole

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

torej kot uteženo povprečje ocen v obeh vzorcih.

Primer: Kajenje in porodna teža (nadaljevanje)

Privzemimo za zdaj, da sta varianci enaki. Potem dobimo za skupno oceno variance $s^2 = 0{,}1588$, vrednost testa t po formuli (2) pa je -2,85. Ker imamo $n_1 + n_2 - 2 = 28$ stopinj prostosti, dobimo za vrednost$p$ 0,008. Torej hipoteza o enakih porodnih težah pri kadilkah in nekadilkah ne zdrži in jo **zavrnemo**.

Kaj pa če **varianci nista enaki**? Izkaže se, da točna porazdelitev izraza (2) ni znana. Obstajajo približki, ki jih uporabljajo statistični paketi, lahko pa uporabimo tudi enega od naslednjih dveh načinov:

- Podatke transformiramo tako, da sta varianci transformiranih podatkov enaki.
- Uporabimo neparametričen test (kakšen, bomo spoznali kasneje).

Seveda ostane vprašanje: **Kako pa se odločimo, če sta varianci enaki?**

PRIMERJAVA VARIANC

Želimo preveriti ničelno hipotezo: $H_0 : \sigma_1^2 = \sigma_2^2$

Poglejmo si kvocient

$$\frac{s_2^2}{s_1^2}$$

Če ničelna hipoteza drži, pričakujemo, da bo ta kvocient blizu 1, majhne vrednosti bi pomenile, da je $\sigma_2^2 < \sigma_1^2$, velike pa, da je $\sigma_2^2 > \sigma_1^2$. Torej so tako majhne kot velike vrednosti tega kvocienta kritične za ničelno hipotezo. Seveda ne moremo govoriti o majhnih ali velikih vrednostih tega kvocienta dokler ne poznamo njegove porazdelitve. Izkaže se, da je kvocient (ob pravilni ničelni hipotezi!) porazdeljen po porazdelitvi $\mathcal{F}$, ki je določena s parom stopinj prostosti (stopinje prostosti za vsako oceno variance posebej).

V praksi izračunamo

$$F_u = \frac{\text{večja vzorčna varianca}}{\text{manjša vzorčna varianca}}$$

in dobimo ustrezno stopnjo tveganja kot

$$p\text{-vrednost} = 2 \cdot P(F \geq F_u),$$

pri čemer verjetnost na desni izračunamo na osnovi porazdelitve $\mathcal{F}$ s stopinjami prostosti

$$(f_{\text{večja vzorčna varianca}}, f_{\text{manjša vzorčna varianca}})$$

Primer: Kajenje in porodna teža (nadaljevanje)

Imeli smo

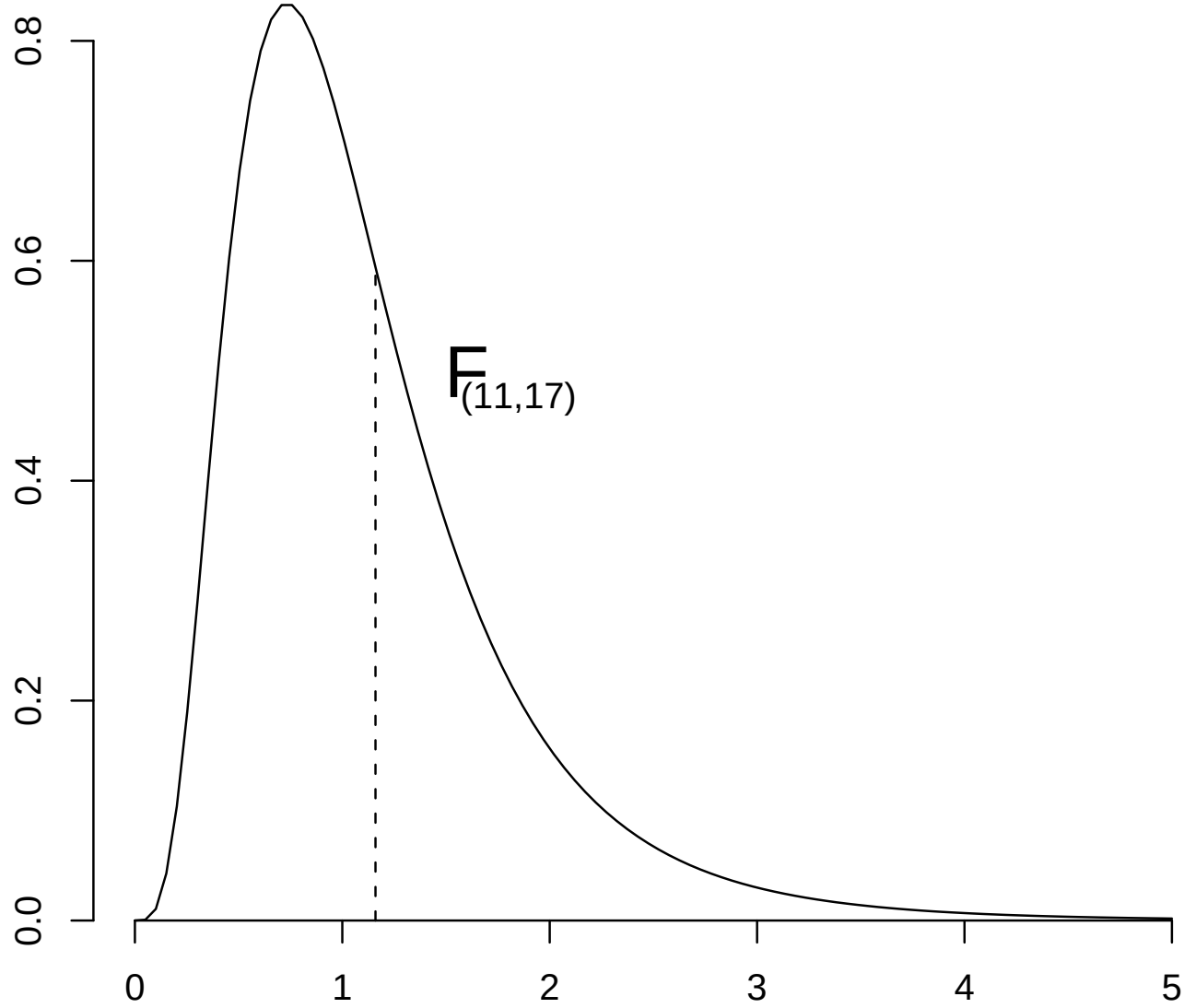
$$s_1^2 = 0{,}1736 \qquad f_1 = 12 - 1 = 11$$
$$s_2^2 = 0{,}1490 \qquad f_1 = 18 - 1 = 17$$

Ker je torej $s_1^2 > s_2^2$, izračunamo

$$F_u = \frac{s_1^2}{s_2^2} = \frac{0{,}1736}{0{,}1490} = 1{,}16$$

Če velja ničelna hipoteza, je ta vrednost eno opazovanje iz porazdelitve $\mathcal{F}$ z (11,17) stopinjami prostosti. Od tam izvemo (program, tabele!), da je

$$p\text{-vrednost} = 2 \cdot P(F \geq 1{,}16) = 2 \cdot 0{,}3794 = 0{,}76$$


0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
F(11,17)

Iz tega sklepamo, da podatki ne nasprotujejo ničelni hipotezi o enakosti varianc. Večina knjig tukaj svetuje, da nadaljujemo, kot da sta varianci enaki. Seveda pa sprejemanje ničelne hipoteze nosi s seboj vse nevarnosti napake 2. vrste. Mi bomo bolj previdni: ker večina računalniških programov tako ali tako izračuna obe verziji *t*-testa, pogledamo pač obe. Če pripovedujeta isto zgodbo, je tako vseeno, če pa ne (en rezultat značilen, drug pa ne) bomo seveda verjeli rezultatu, ki ga dobimo na osnovi predpostavke o **neenakih** variancah. Razlika namreč lahko pride le od tod.

NEPARAMETRIČNE METODE

- Metode za analizo numeričnih spremenljivk temeljijo na predpostavki o normalnosti porazdelitve, ki je bolj ali manj pomembna
- Včasih so kršitve normalnosti hude, zato so potrebne metode, ki se tej predpostavki izognejo
- Običajne metode tudi niso primerne za urejenostne spremenljivke

Wilcoxonov test predznačenih rangov

- Testira ničelno hipotezo, da je porazdelitev slučajne spremenljivke simetrična glede na vrednost 0 (ali kakšno drugo vrednost)
- Tipično uporaben za parne meritve (analogija parnemu t-testu)

Wilcoxonov test predznačenih rangov (nadaljevanje)

Wilcoxonov test predznačenih rangov (nadaljevanje)

- Recimo, da imamo n vrednosti, od tega nekaj negativnih, nekaj ničel in nekaj pozitivnih.

Wilcoxonov test predznačenih rangov (nadaljevanje)

- Recimo, da imamo n vrednosti, od tega nekaj negativnih, nekaj ničel in nekaj pozitivnih.
- Ničle zanemarimo (ostane n' vrednosti), ostale vrednosti pa uredimo po velikosti (od najmanjše do največje) glede na absolutne vrednosti.

Wilcoxonov test predznačenih rangov (nadaljevanje)

- Recimo, da imamo n vrednosti, od tega nekaj negativnih, nekaj ničel in nekaj pozitivnih.
- Ničle zanemarimo (ostane n' vrednosti), ostale vrednosti pa uredimo po velikosti (od najmanjše do največje) glede na absolutne vrednosti.
- Tako urejenim vrednostim dodelimo range od 1 do n'.

Wilcoxonov test predznačenih rangov (nadaljevanje)

- Recimo, da imamo n vrednosti, od tega nekaj negativnih, nekaj ničel in nekaj pozitivnih.
- Ničle zanemarimo (ostane n' vrednosti), ostale vrednosti pa uredimo po velikosti (od najmanjše do največje) glede na absolutne vrednosti.
- Tako urejenim vrednostim dodelimo range od 1 do n'.
- Nato seštejemo range posebej pri pozitivnih (T_+) in negativnih (T_-) vrednostih. Če velja ničelna hipoteza, se ti dve vsoti ne bosta močno razlikovali.

Wilcoxonov test predznačenih rangov (nadaljevanje)

- Recimo, da imamo n vrednosti, od tega nekaj negativnih, nekaj ničel in nekaj pozitivnih.
- Ničle zanemarimo (ostane n' vrednosti), ostale vrednosti pa uredimo po velikosti (od najmanjše do največje) glede na absolutne vrednosti.
- Tako urejenim vrednostim dodelimo range od 1 do n'.
- Nato seštejemo range posebej pri pozitivnih (T_+) in negativnih (T_-) vrednostih. Če velja ničelna hipoteza, se ti dve vsoti ne bosta močno razlikovali.
- Statistični test naredimo tako, da izračunamo verjetnost, da je vsota T_+ ali T_- enaka ali bolj ekstremna od ugotovljene. V tabelah najdemo kritične vrednosti za manjšo od obeh vsot.

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija *t*-testu za neodvisne vzorce).

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija *t*-testu za neodvisne vzorce).

- Vrednosti iz obeh vzorcev skupaj uredimo v ranžirno vrsto.

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija *t*-testu za neodvisne vzorce).

- Vrednosti iz obeh vzorcev skupaj uredimo v ranžirno vrsto.
- Nato seštejemo range pri vrednostih manjšega vzorca (T_1).

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija *t*-testu za neodvisne vzorce).

- Vrednosti iz obeh vzorcev skupaj uredimo v ranžirno vrsto.
- Nato seštejemo range pri vrednostih manjšega vzorca (T_1).
- Če so med skupinama razlike, bo ta vsota precej manjša ali precej večja od pričakovane.

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija *t*-testu za neodvisne vzorce).

- Vrednosti iz obeh vzorcev skupaj uredimo v ranžirno vrsto.
- Nato seštejemo range pri vrednostih manjšega vzorca (T_1).
- Če so med skupinama razlike, bo ta vsota precej manjša ali precej večja od pričakovane.
- Minimalna možna T_1 je $n_1(n_1+1)/2$, maksimalna pa $n_1 n_2 + n_1(n_1+1)/2$, pričakovana T_1 pa $n_1(n_1+n_2+1)/2$.

Wilcoxonov test vsote rangov

Testira ničelno hipotezo, da je porazdelitev dveh slučajnih spremenljivk enaka (analogija t-testu za neodvisne vzorce).

- Vrednosti iz obeh vzorcev skupaj uredimo v ranžirno vrsto.
- Nato seštejemo range pri vrednostih manjšega vzorca (T_1).
- Če so med skupinama razlike, bo ta vsota precej manjša ali precej večja od pričakovane.
- Minimalna možna T_1 je $n_1(n_1+1)/2$, maksimalna pa $n_1 n_2 + n_1(n_1+1)/2$, pričakovana T_1 pa $n_1(n_1+n_2+1)/2$.
- Kritične vrednosti T_1 so tabelirane.