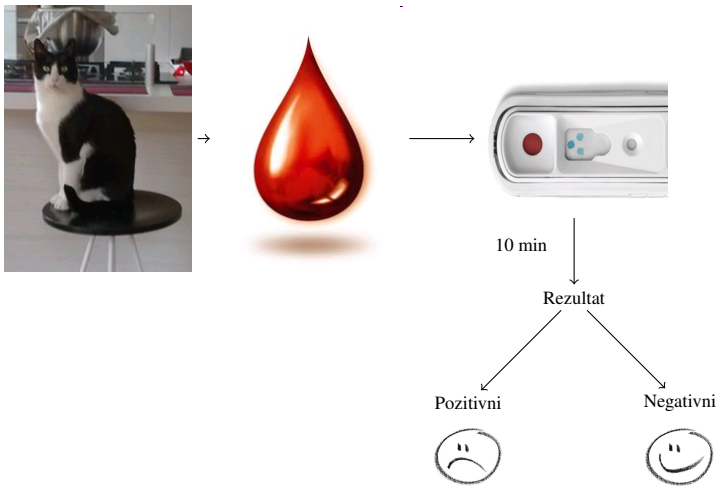
Biostatistika
Veterinarska fakulteta



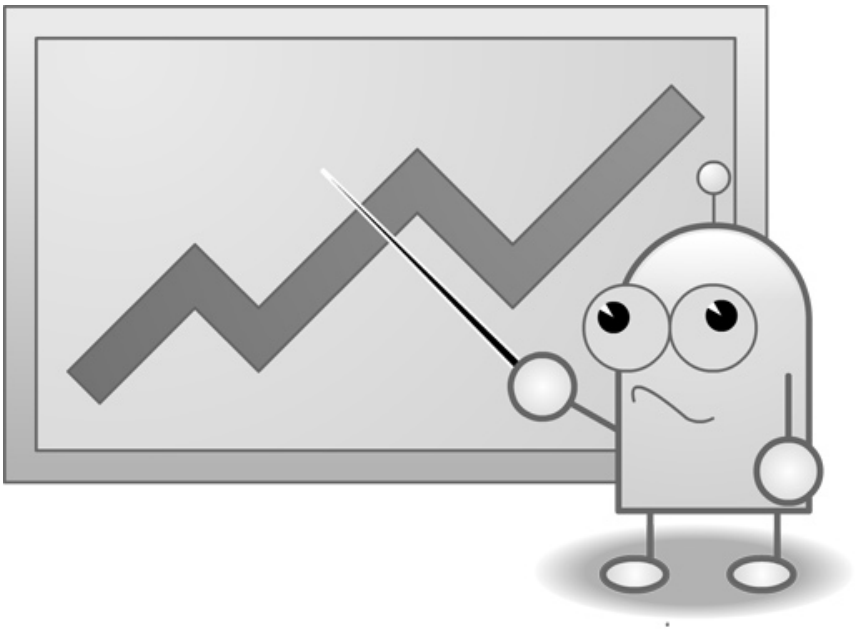
Zapiski s predavanj

Lara Lusa
Inštitut za biostatistiko in medicinsko informatiko
Medicinska fakulteta, Univerza v Ljubljani
lara.lusa@mf.uni-lj.si

Poglavje 1

Opisna statistika

Statistika v prvem letniku



1.1

Opisna statistika je del statistike, ki se ukvarja s povzemanjem podatkov.
Namem poglavja je spoznati:

- grafične in tabelarne prikaze podatkov;
- opisne mere središčnosti in razpršenosti;
- vrste spremenljivk.

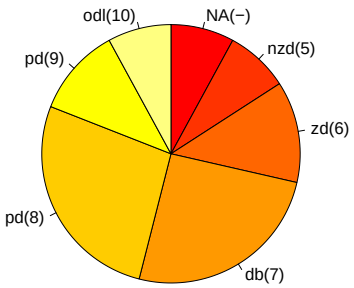
Ogledali si bomo rezultate izpitov pri predmetu biostatistika (akademsko leto 2012/13). Ocene so bile:

```
odl(10),NA(-),zd(6),pd(8),nzd(5),pd(8),db(7),nzd(5),odl(10),pd(8),
db(7),pd(8),odl(10),db(7),pd(8),db(7),pd(9),pd(8),pd(8),pd(8),
odl(10),db(7),zd(6),db(7),pd(8),db(7),pd(8),NA(-),nzd(5),pd(8),
pd(9),pd(8),NA(-),db(7),zd(6),pd(8),zd(6),nzd(5),db(7),db(7),
pd(8),db(7),pd(9),db(7),db(7),db(7),pd(8),NA(-),odl(10),pd(9),
NA(-),pd(9),zd(6),db(7),zd(6),zd(6),db(7),zd(6),pd(9),pd(8),
nzd(5),pd(9),pd(8)
```

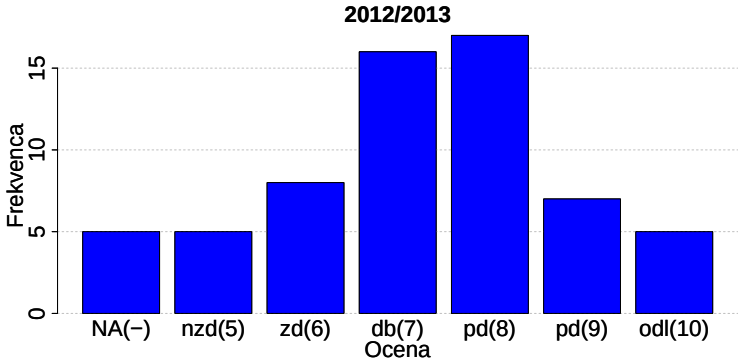
V tem poglavju si bomo ogledali, kako lahko povzamemo in predstavimo te rezultate s pomočjo grafikonov, tabel in opisnih mer.

Kako je šlo lansko leto

Kolač, strukturni krog, pita (pie chart)



Stolpični diagram (bar plot)



Frekvenčna tabela (frequency table)

	Frekvenca	Relativna frekvenca
NA(-)	5	0.08
nzd(5)	5	0.08
zd(6)	8	0.13
db(7)	16	0.25
pd(8)	17	0.27
pd(9)	7	0.11
odl(10)	5	0.08

1.2

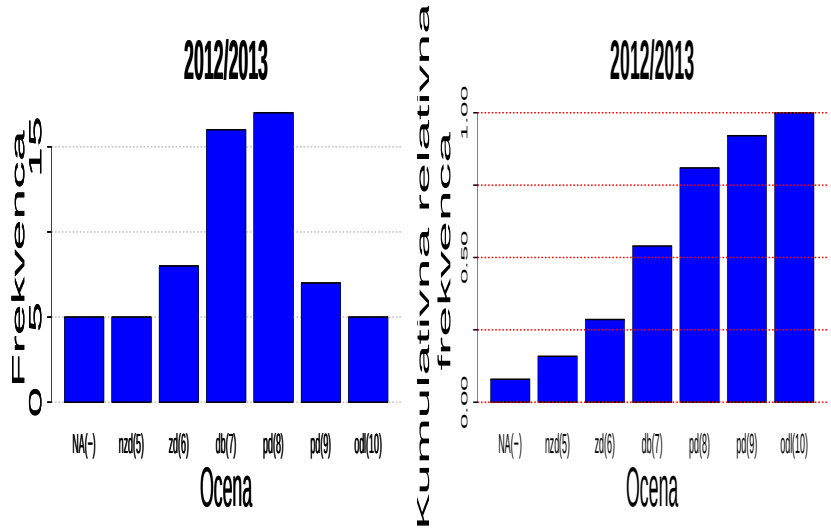
Kako je šlo lansko leto

Kaj lahko odčitamo?

- **Velikost vzorca (n)**= Število študentov = 63.
- **Frekvenca** (pogostost) za odl(10) = 5.
- **Relativna frekvenca** za odl(10) = 5 / 63 = 0.08 = 0.08.
- **Kumulativna relativna frekvenca** za zd(6) = (0.08+0.08+0.13) = 0.29.

1.3

Kako je šlo lansko leto



Kaj lahko odčitamo?

- Najpogostejša ocena je bila pd(8): *modus*.
- Srednja ocena je bila db(7): *mediana* (50. percentil, 2. kvartil).
- Četrtina študentov je dobila oceno zd(6) ali manj: *1. kvartil, 25. percentil*.
- Četrtina študentov je dobila oceno pd(8) ali več: *3. kvartil, 75. percentil*.
- Polovica študentov izmed tistimi, ki so bili najbližji srednje ocene, je dobilo med zd(6) in pd(8): *interkvartilni razmik, IQR*.

1.4

Povzetek

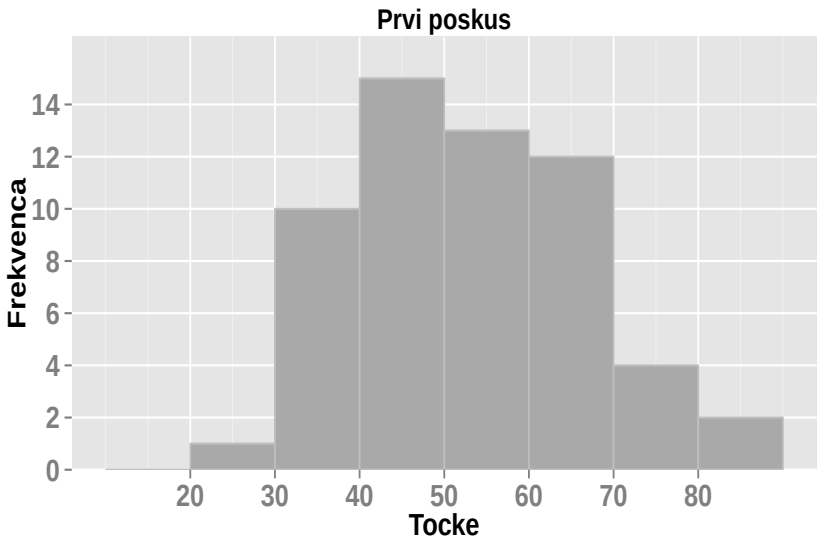
Kaj smo gledali?

- Kaj je *populacija*?
- Kaj je *vzorec*?
- Kaj so *statistične enote*?
- Katero *spremenljivko* smo merili?
- Katere *vrste* je spremenljivka, ki smo jo merili?
- Kateri *grafični prikazi* smo uporabili?
- S katerimi *opisnimi statističnimi merami* smo povzeli rezultate?
- Ali bi bilo smiselno izračunati *aritmetično povprečje* ocen?

1.5

Koliko točk so dosegli na prvem poskusu?

Histogram



Kaj lahko odčitamo?

- Koliko študentov je doseglo 70 točk ali več na prvem poskusu? 6
- Koliko študentov je doseglo manj kot 40 točk? 11
- Koliko študentov je doseglo med 40 in 49 točk? 15
- Kolikšen je delež študentov, ki je opravilo izpit na prvem poskusu? 0.81
- Približno kolikšno je bilo povprečno število točk? 52.3
- Približno kolikšna je bila varianca števila točk? 181.3 točk^2
- ... in standardni odklon? 13.5 točk.

1.6

Opisne mere

Mere središčnosti

- Povprečje:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  (n=velikost vzorca)
- Mediana: 50. percentil.
- Modus: najpogostejša vrednost.

Mere razpršenosti

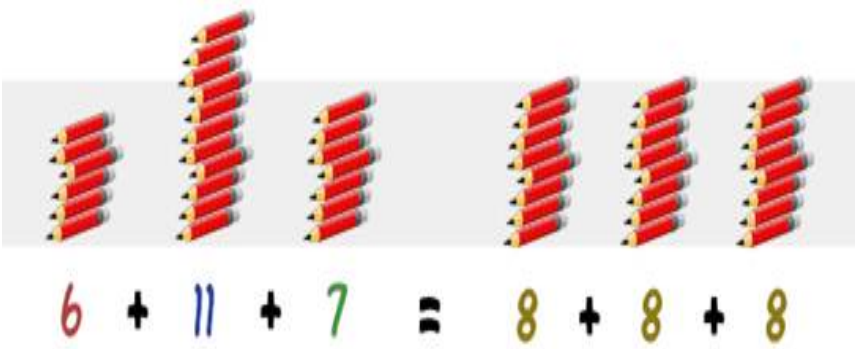
- (Vzorčna) varianca:

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

- (Vzorčni) standardni odklon: $s = \sqrt{(s^2)}$.
- Interkvartilni razmik: 25. do 75. percentila.
- Razpon: najmanjša do največje vrednosti.

1.7

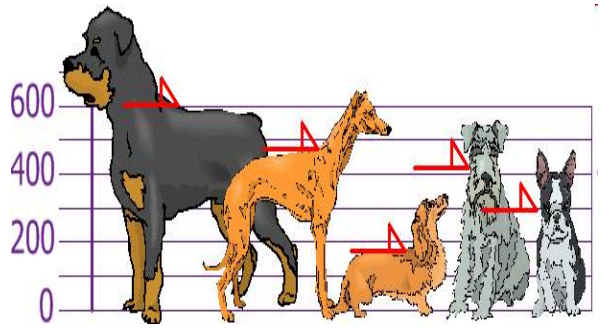
Zakaj rabimo tudi mere razpršenosti?



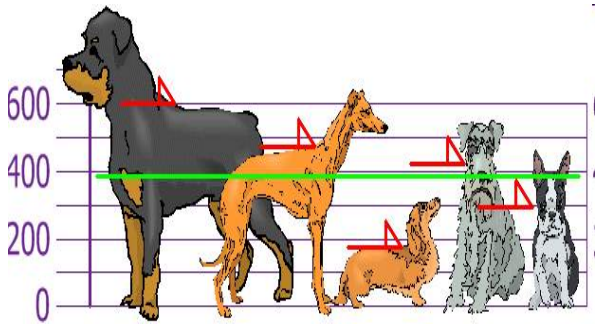
Povprečje	8	8
Standardni odklon	2.65	0
Razpon	5	0

Vir: `http://www.mathsisfun.com/mean.html`

Kako izračunamo (vzorčni) standardni odklon?

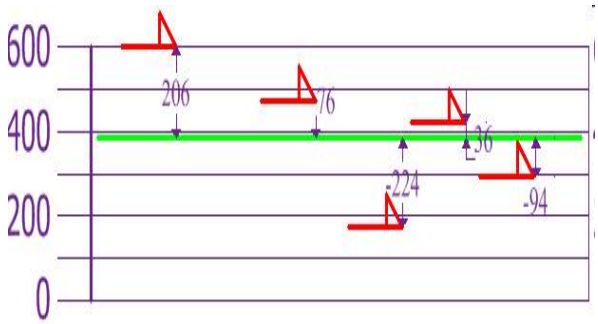


Vrednosti (v mm, n=5):
600, 470, 170, 430, 300



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

$\bar{x} = \frac{600+470+170+430+300}{5}.$
$= \frac{1970}{5} = 394$ mm



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

$s^2 = \frac{206^2+76^2+(-224)^2+36^2+(-94)^2}{4}.$
$= \frac{108,520}{4} = 27,130\ mm^2$
$s = \sqrt{27,130} = 164.7mm$

1.9 Vir: http://www.mathsisfun.com/data/standard-deviation.html

Mere: ali jih lahko izračunamo za vsako spremenljivko?

Spremenljivka	*Število mladičev*	*Stadij raka*	*Spol*
Vrednosti	3, 1, 7, 2, 2	I, IV, III, II, I	ž, m, m, m, ž
Urejene vrednosti	1, 2, 2, 3, 7	I, I, II, III, IV	
Povprečje	3		
1.10 Mediana	2	II	
Modus	2	I	m
Standardni odklon	2.3		
Interkvartilni razmik	(2;3)	(I; III)	
Razpon	(1;7)	(I; IV)	

Mere: ali so občutljive na skrajne vrednosti?

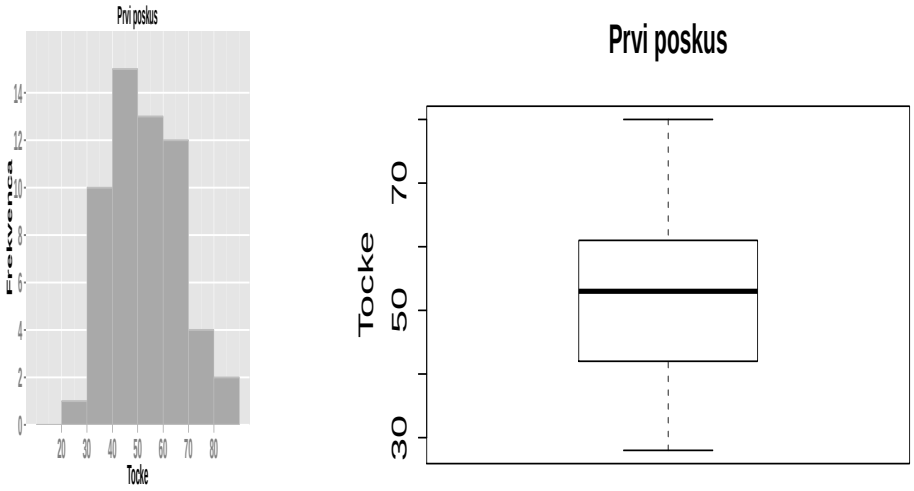
Spremenljivka	Število mladičev	Število mladičev
Vrednosti	3, 1, *7*, 2, 2	3, 1, *20*, 2, 2
Urejene vrednosti	1, 2, 2, 3, 7	1, 2, 2, 3, 20
Povprečje	3	*5.6*
Mediana	2	2
Modus	2	2
Standardni odklon	2.3	*8.1*
Interkvartilni razmik	(2;3)	(2;3)
Razpon	(1;7)	*(1;20)*

Katere mere bi bilo smiselno uporabiti?

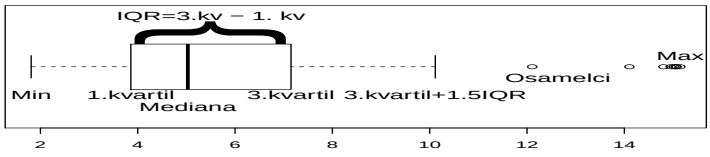
1.11

Različni grafični prikazi za številske spremenljivke

Histogram **Okvir z ročaji (boxplot)**



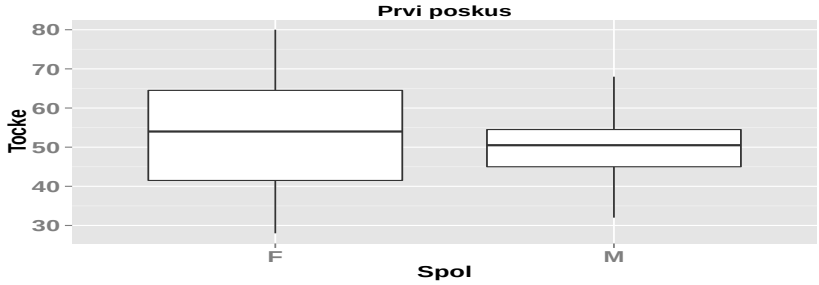
Kaj lahko odčitamo?



1.12

Primerjava dveh skupin

Ali so bila dekleta boljša od fantov?



Ali bodo letos fantje boljši od deklet?

	F	M
Povprečje	53.40	49.00
Mediana	54.00	50.50
Standardni odklon (s)	14.40	9.90
Interkvartilni razmik (IQR)	23.00	9.50
Velikost vzorca (n)	43.00	14.00

Ali so razlike dovolj velike, da ne morejo biti plod naključja?

1.13

Povzetek

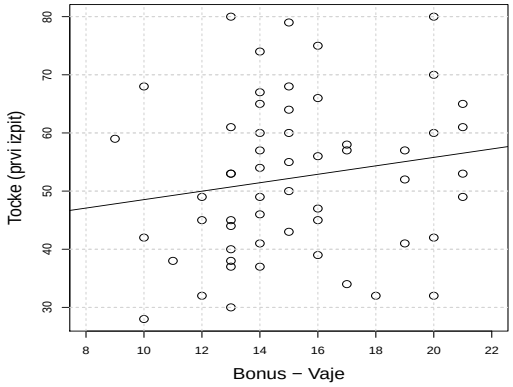
Kaj smo gledali?

- Kaj je *populacija*?
- Kaj je *vzorec*?
- Kaj so *statistične enote*?
- Katero *spremenljivko* smo merili?
- Katere *vrste* je spremenljivka, ki smo jo merili?
- Kateri *grafični prikazi* smo uporabili?
- S katerimi *opisnimi statističnimi merami* smo povzeli rezultate?
- Ali bi bilo smiselno izračunati *aritmetično povprečje* ocen?

1.14

Povezanost med številskimi spremenljivkami

Razsevni diagram (scatter plot)

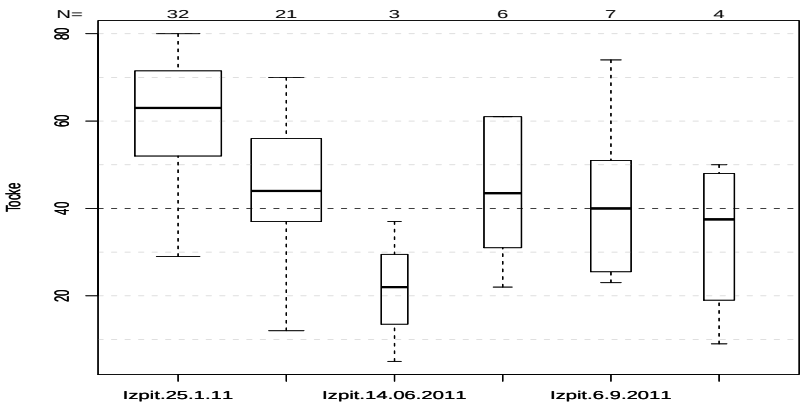


Ali se splača sodelovati?

- Kaj predstavlja vsaka točka?
- Ali so imeli študentje z večjim bonusom bolši uspeh?

1.15

Kdaj se splača priti?



1.16

Opisna statistika

Opisna statistika je skupina statističnih metod, ki se ukvarjajo s povzemanjem pridobljenih podatkov. Te metode iščejo opisne (meta) podatke o populaciji in njenih sestavnih delih, da bi ustvarile pregledni opis.

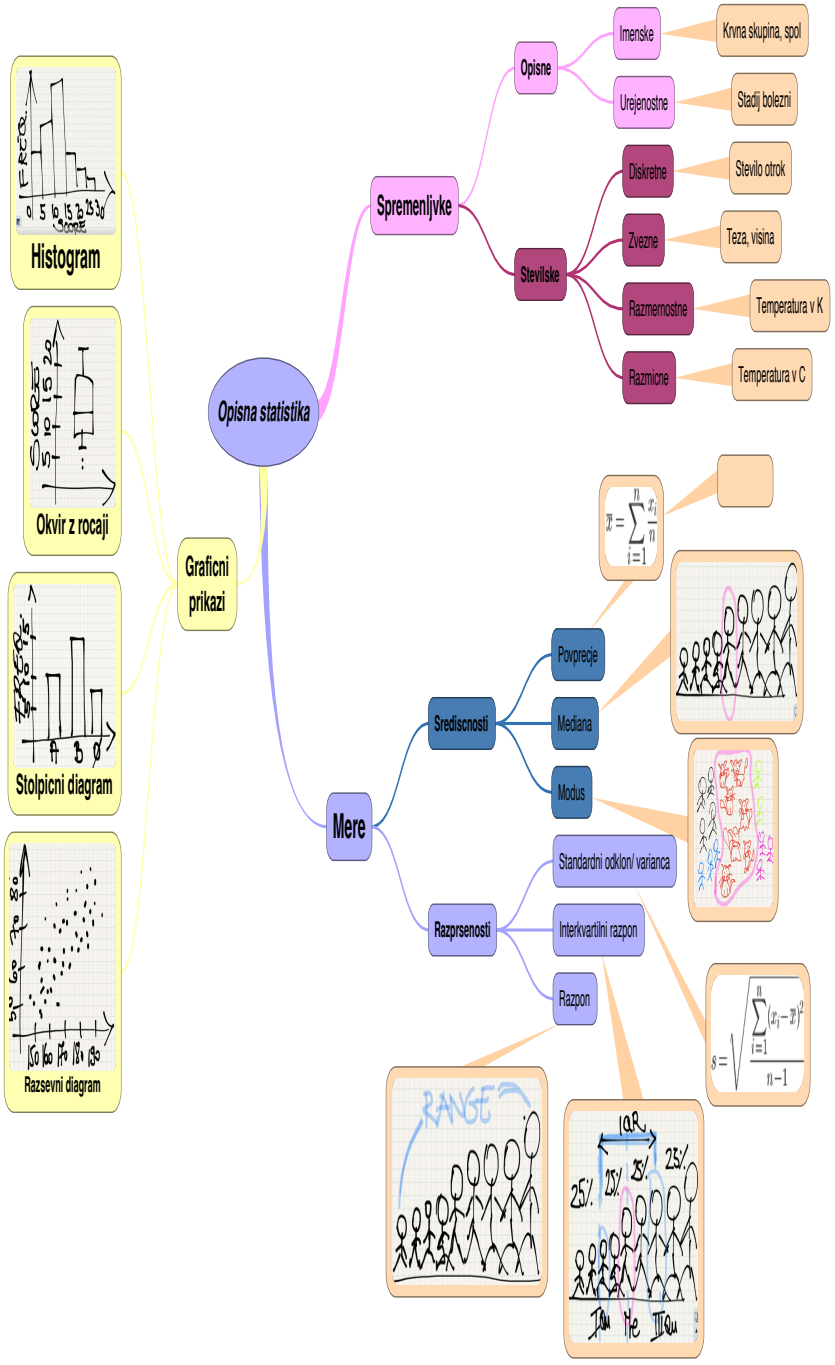
(Wikipedia, Juni 2013)

Kako?

- grafikoni
- tabele
- statistični povzetki
  - mere središčnosti (*central tendency measures*)
  - mere razpršenosti (*variability measures*)

Izbira primerne metode je odvisna od *vrste podatkov*

1.17



1.18

Spremenljivke so lahko **opisne** ali **številske**. Vrednosti opisnih spremenljivk so kategorije (imena); vrednosti številskih spremenljivk so številke.

Opisne spremenljivke so lahko **urejenostne** (lahko rangiramo vrednosti, primer je stopnja izbrazbe: osnovna šola, srednja šola, itd) ali **imenske** (ne moremo urediti vrednosti, primer je spol: moški ali ženski).

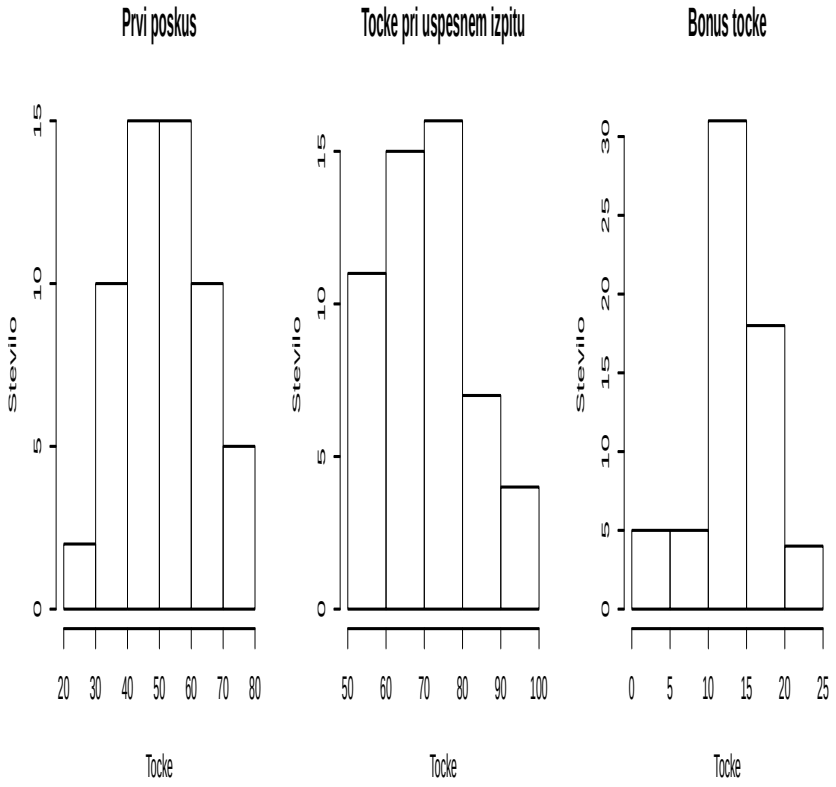
Številske spremenljivke so lahko **razmernostne** (imajo absolutno ničlo in je smiselno izračunati razliko in tudi razmerje med vrednostimi, primer je višina), ali **razmične (intervalne)** (je smiselno izračunati le razliko med vrednostimi, ker nimajo absolutne ničelne vrednosti, primer je temperatura merjena v °C).

Za številske spremenljivke lahko izračunamo vse mere središčnosti in razpršenosti navedene v tem poglavju, čeprav modus ponavadi ni zelo informativna mera, razpon pa je preveč občutljiv na skrajne vrednosti ter je odvisen od velikosti vzorca. Za simetrično porazdeljene spremenljivke ponavadi poročamo aritmetično povprečje in standardni odklon, za asimetrično porazdeljene spremenljivke mediano in interkvartilni razmik, ker sta meri manj občutljivi na skrajne vrednosti. Ti dve meri se lahko izračunata tudi za opisne urejenostne spremenljivke. Za opisne imenske spremenljivke lahko (od navedenih mer) izračunamo le modus.

Grafični prikazi, ki jih lahko uporabimo, za predstavitev podatkov so: **histogram** (primeren za številske spremenljivke), **okvir z ročaji** (primeren za številske spremenljivke, zelo uporaben, ko želimo neposredno primerjati skupine), **stolpični diagram** (primeren za imenske spremenljivke). **Kolač** (strukturni krog) je primeren samo za opisne spremenljivke, ki imajo majhno število možnih vrednostih in je za prestavitev podatkov manj učinkovit od stolpičnega diagrama. Človeško oko namreč lažje loč višine kot površine, zato uporaba strukturnega kroga ni zaželena.

Porazdelitve številskih spremenljivk

Ali imajo vse isto obliko?



Približno simetrična	Pozitivno asimetrična asimetrična v desno	Negativno asimetrična asimetrična v levo
$\bar{x} = 52.32$	$\bar{x} = 71.23$	$\bar{x} = 13.87$
Mediana=53	Mediana=71	Mediana=14
$\bar{x} \approx$ Mediana	$\bar{x} >$ Mediana	$\bar{x} <$ Mediana

1.19

Poglavje 2

Diagnostični testi

Diagnostični testi se uporabljajo v veterinarski praksi za hitro in preprosto ugotavljanje zdravstvenega stanja preiskovanca. Na primer, veterinar lahko s hitrim testom krvi diagnosticira, ali ima mačka mačjo levkozo. Diagnostični testi žal niso popolni in lahko pri diagnozi pride do napak. Za vsak diagnostični test proizvajalec poroča, kolikšna je ocenjena verjetnost (lažno pozitivnih in lažno negativnih) napak.

Namem poglavja je spoznati:

- osnove verjetnosti;
- mere s katerimi ovrednotimo diagnostične teste;
- kako ovrednotiti verjetnost, da ima preiskovanec s pozitivnim testom bolezen.

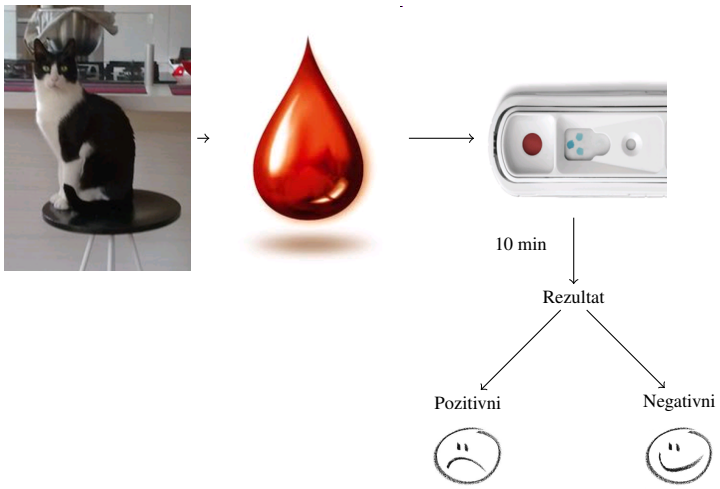
Uporabili bomo primer diagnostičnega testa za diagnozo mačje levkoze, ki se rutinsko uporablja v veterinarski praksi.

FeLV: Mačja levkoza

- Povzroča jo virus FeLV (*Feline Lukemia Virus*)
- Prizadene:
  - imunski sistem
  - organske sisteme (ledvice, jetra, prebavila, itd)
- Lahko povzroči
  - levkemijo (rak belih krvnih telesc)
  - linfosarkome (rak z izvorom v limfnem tkivu)
- Nalezljiva in pogosta bolezen
- Pogosto ostane virus neaktiven za veliko let
- Okužba in prenos
  - neposreden stik z okuženo zival (slina, urin, mleko, blato, kri)
  - Ugrizi, skupno hranjenje, transfuzije krvi, dojenje, itd
- Preventiva: cepljenje
- Diagnoza: hitri test krvi FeLV(/FIV, combo)

1.1

Diagnoza FeLV-a



1.2

Ocenimo verjetnost, da ...

Uganka

Najdete odraslo mačko in se odločite, da jo boste posvojili. Mačka nima kliničnih znakov FeLV-a. Veterinar predlaga, da jo testirate za FeLV. Vam razlaga, da je pregled krvi zanesljiv diagnostični test.

- Verjetnost pozitivenega testa za okuženo mačko je 0.92.
- Verjetnost negativnega testa za zdravo mačko je 0.99.
- Pogostost FeLV-a v asimptomatični mačiji populaciji je 2%.

Test je pozitiven

Kolikšna je verjetnost, da ima mačka FeLV?

- 1
- 0.99
- 0.92
- 0.85
- 0.65
- 0.08
- 0.02
- 0.01
- 0

1.3

Rezultati diagnostičnega testa

Kontingenčna tabela

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Kaj lahko odčitamo?

- Število resnično pozitivnih rezultatov: 35
- Število resnično negativnih rezultatov: 482
- Število lažno pozitivnih rezultatov: 6
- Število lažno negativnih rezultatov: 3
- Pogostost bolezni: P(Bolezen=Da)=38 / 526 = 0.07

1.4

Rezultati diagnostičnega testa

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Napovedna točnost (*predictive accuracy*)

$$\text{P(Test=Bolezen)}=\frac{35+482}{35+6+3+482}=\frac{517}{526}=0.98$$

1.5

Rezultati diagnostičnega testa

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Občutljivost testa (*sensitivity*)

$$\text{P(Test+|Bolezen+)}=\frac{35}{35+3}=\frac{35}{38}=0.92$$

Pogojna verjetnost s formulami: $\text{P(Test+|Bolezen+)}=\frac{\text{P(Test+} \cap \text{Bolezen+)}}{\text{P(Bolezen+)}}=\frac{35/526}{38/526}$

1.6

Rezultati diagnostičnega testa

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Specifičnost testa (*specificity*)

$$\text{P(Test-|Bolezen-)}=\frac{482}{6+482}=\frac{482}{488}=0.99$$

1.7

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Rezultati diagnostičnega testa

Pozitivna napovedna vrednost (*positive predictive value, PPV*)
P(Bolezen+|Test+)=$\frac{35}{35+6} = \frac{35}{41} = 0.85$

P(Bolezen+|Test+)≠P(Test+|Bolezen+)

1.8

Rezultati diagnostičnega testa

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

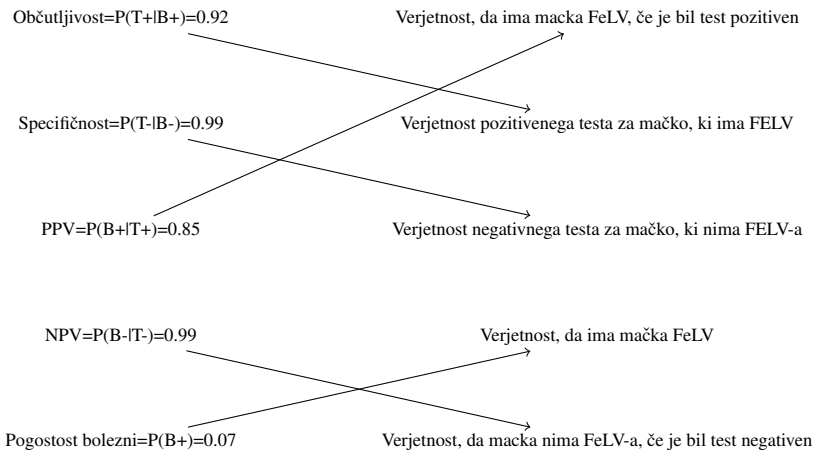
Negativna napovedna vrednost (*negative predictive value, NPV*)
P(Bolezen-|Test-)=$\frac{482}{482+3} = \frac{482}{485} = 0.99$

P(Bolezen-|Test-)≠P(Test-|Bolezen-)

1.9

Povzetek rezultatov

Najdite par



1.10

Ocenimo verjetnost, da ...

Uganka

Najdete odraslo mačko in se odločite, da jo boste posvojili. Mačka nima kliničnih znakov FeLV-a. Veterinar predlaga, da jo testirate za FeLV. Vam razlaga, da je pregled krvi zanesljiv diagnostični test.

- Verjetnost pozitivenega testa za okuženo mačko je 0.92.
- Verjetnost negativnega testa za zdravo mačko je 0.99.
- Pogostost FeLV-a v asimptomatični mačiji populaciji je 2%.
- Pogostost bolezni na vzorcu je bila 7%.

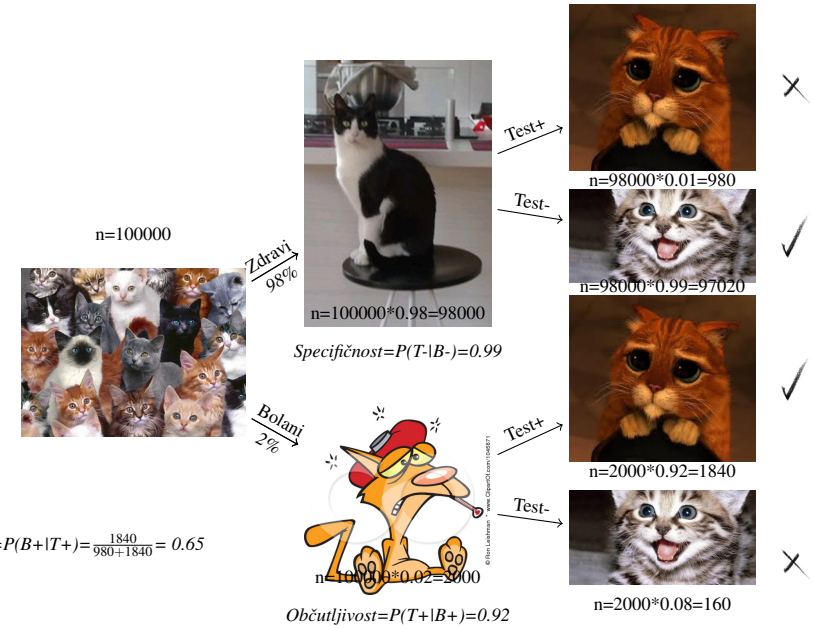
Test je pozitiven

Kolikšna je verjetnost, da ima mačka FeLV?

- 1
- 0.99
- 0.92
- *0.85 ???*
- 0.65
- 0.08
- 0.02
- 0.01
- 0

1.11

Napovedne vrednosti, če je pogostost $p = 0.02$



1.12

Napovedne vrednosti s formulami

Pozitivna napovedna vrednost

$$PPV = \mathrm{P(B+|T+)} = \frac{\mathrm{p}\cdot\mathrm{sens}}{\mathrm{p}\cdot\mathrm{sens}+(1-\mathrm{p})\cdot(1-\mathrm{spec})}$$

Negativna napovedna vrednost

$$NPV = \mathrm{P(B\text{-}|T\text{-})} = \frac{(1\text{-}\mathrm{p})\cdot\mathrm{spec}}{(1\text{-}\mathrm{p})\cdot\mathrm{spec}+\mathrm{p}\cdot(1\text{-}\mathrm{sens})}$$

Zakaj?

$$P(E|H) = \frac{P(H|E)P(E)}{P(H)} \text{(bayesov izrek) in}$$
$$P(H) = P(H|E+)P(E+) + P(H|E-)P(E-),$$

če $P(E+ \cap E-) = 0$ in $P(E+ \cup E-) = 1$

1.13

Osnove verjetnosti

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Produkt dogodkov

P(Test+ $\cap$ Bolezen+)=$\frac{35}{526} = 0.07$

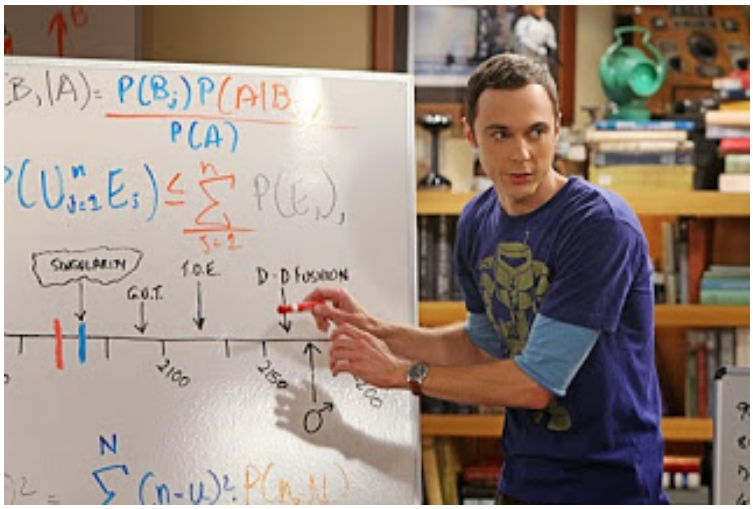
P(Test+ $\cap$ Bolezen+)=P(Bolezen+)P(Test+|Bolezen+)=$\frac{38}{526} \cdot \frac{35}{38} = 0.07$
$\neq P(Bolezen+)P(Test+) = \frac{38}{526} \cdot \frac{41}{526} = 0.006$ *Dogodka nista neodvisna*

P(Test+ $\cap$ Bolezen+)=P(Test+)P(Bolezen+|Test+)=$\frac{41}{526}\frac{35}{41}$ =0.07

1.14

Bayesov izrek

$$P(Bolezen+|Test+) = \frac{P(Test+|Bolezen+)P(Bolezen+)}{P(Test+)}$$



1.15

Osnove verjetnosti

Bayesov izrek

$$P(Bolezen+|Test+) = \frac{P(Test+|Bolezen+)P(Bolezen+)}{P(Test+)}, \text{ker}$$

P(Test+ $\cap$ Bolezen+)=P(Test+|Bolezen+)P(Bolezen+) in

P(Test+ ∩ Bolezen+)=P(Bolezen+|Test+)P(Test+)

1.16

Osnove verjetnosti

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Unija dogodkov

P(Test+ ∪ Bolezen+)=P(Test+) + P(Bolezen+) - P(Bolezen+ ∩ Test+)=
$=\frac{35+6}{526}+\frac{35+3}{526}-\frac{35}{526}=0.08$

P(Bolezen+ ∩ Test+) $\neq 0$ *Dogodka nista nezdružljiva*

1.17

Osnove verjetnosti

		Bolezen		
		+	-	Skupaj
Test	+	35	6	41
	-	3	482	485
	Skupaj	38	488	526

Popolna verjetnost

P(Test+)=P(Test+|Bolezen+)P(Bolezen+)+P(Test+|Bolezen-)P(Bolezen-)
=P(Test+ ∩ Bolezen+) + P(Test+ ∩ Bolezen-)

Velja, ker sta Bolezen+ in Bolezen- *nasprotna* dogodka.
P(Bolezen+ ∩ Bolezen -)=0 (nemogoč dogodek)
P(Bolezen+ ∪ Bolezen -)=1 (gotov dogodek)

1.18

Poglavje 3

Binomska porazdelitev

To poglavje se ukvarja z verjetnostjo; uvedli bomo binomsko porazdelitev. Binomska porazdelitev je teoretična verjetnostna porazdelitev. S pomočjo verjetnostne porazdelitve lahko določimo verjetnost vseh možnih izidov nekega dogodka. Vemo, na primer, da je pri metu poštenega kovanca verjetnost, da pade na eno od strani enaka za obe strani (0.5). Poznamo verjetnostno porazdelitev, ker smo našteli vse možne (nezdružljive) izide in njihovo verjetnost.

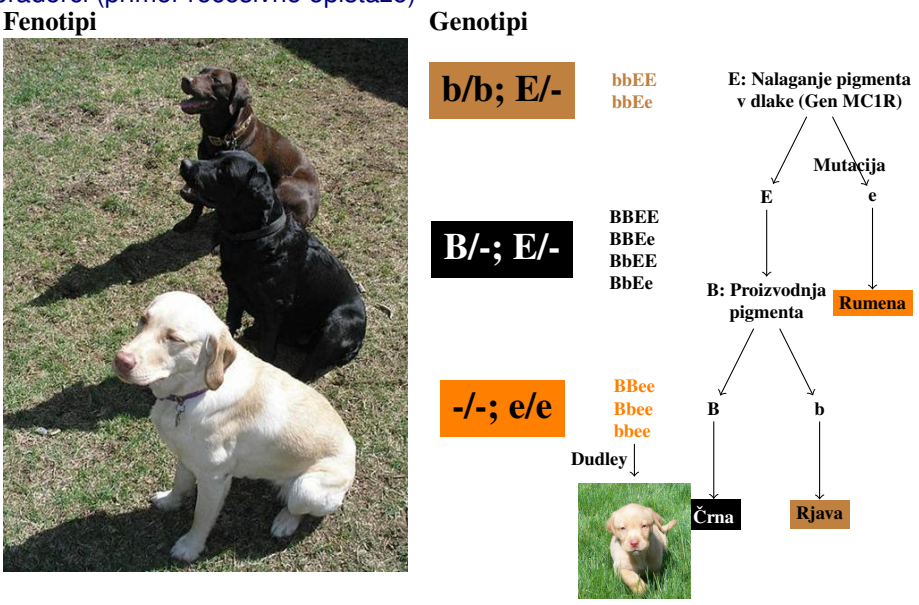
Binomska porazdelitev je uporabna pri poskusih, kjer opazujemo več dogodkov. Vsak dogodek je neodvisen od ostalih in ima samo dva možna izida (na primer, uspeh ali neuspeh); vsi dogodki imajo isto verjetnost uspeha (verjetnost uspeha se iz meta v met ne spreminja). S pomočjo binomske porazdelitve lahko določimo, kolikšna je verjetnost, da se pojavi neko število uspehov (k), pri določnem številu opaženih dogodkov (n). Binomsko porazdelitev določita dva parametra: število poskusov (n) in verjetnost uspeha pri posameznem poskusu (π).

Na primer, če mečemo pošten kovanec desetkrat (n=10), bomo lahko izračunali verjetnost, da dobimo natanko pet grbov ($k = 5, \pi = 0.50$) (in bo mogoče presenetljivo spoznati, da verjetnost ni enaka 0.5, ampak je približno 0.25!). Binomska porazdelitev bo uporabna tudi za izračun verjetnosti, da se v leglu s petimi labradorci rodi natanko 5 črnih mladičev. Pomagali si bomo z genetiko in s statistiko!

Namen poglavja je

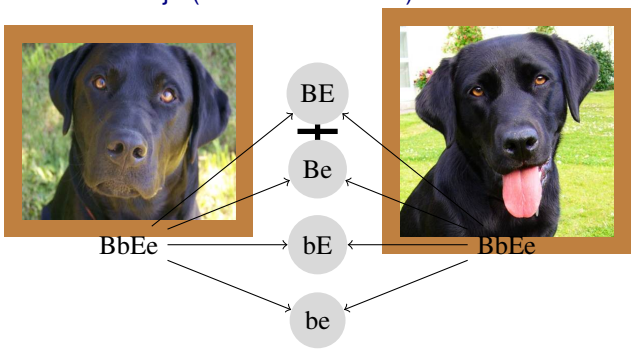
- spoznati binomsko porazdelitev;
- razumeti kdaj jo je smiselno uporabiti in kako.

Labradorci (primer recesivne epistaze)



3.1

Dihibridno križanje (Punnetov kvadrat)



	BE	Be	bE	be
BE	BBEE	BBEe	BbEE	BbEe
Be	BBeE	BBee	BbeE	Bbee
bE	bBEE	bBEe	bbEE	bbEe
be	bBeE	bBee	bbeE	bbee

Barva dlak mladičev
P(Črna)=$9/16 = 0.5625$
P(Rjava)=$3/16 = 0.1875$
P(Rumena)=$4/16 = 0.25$

3.2

Starši so: BbEe+BbEe, rodi se pet mladičev

Kolikšna je verjetnost, da se v leglu s 5 mladiči rodi natanko 5 črnih (B) mladičev?
P(1.=B ∩ 2.=B ∩ 3.=B ∩ 4.=B ∩ 5.=B)=[1]
=P(1.=B)·P(2.=B)· P(3.=B)· P(4.=B)·P(5.=B)=[2]
=$(9/16)^5 = 0.0563$

S formulami

n: število poskusov (5: leglo)
Uspeh=črni mladič
K: število uspehov; je *slučajna spremenljivka*
k: število uspehov, ki nas zanima (5: črni mladiči)
π=P(B): verjetnost uspeha pri posameznem poskusu (9/16)
P(K=5|n=5, π=9/16)=$(9/16)^5 = \pi^k$

3.3

Starši so: BbEe+BbEe, rodi se pet mladičev

Kolikšna je verjetnost, da se v leglu s 5 mladiči ni niti enega črnega mladiča?
P(1.≠B ∩ 2.≠B ∩ 3.≠B ∩ 4.≠B ∩ 5.≠B)=[3]
=P(1.≠B)·P(2.≠B)· P(3.≠B)· P(4.≠B)·P(5.≠B)=[4]
=$(1-9/16)^5 = 0.016$

S formulami

n=5, k=0, π=9/16
P(K=0|n=5, π=9/16)=$(1-9/16)^5 = (1-\pi)^n$

3.4

[1] dogodki so neodvisni
[2] vsi dogodki imajo isto verjetnost, $P(1.=B) = \ldots = P(5.=B) = \pi = 9/16$
[3] dogodki so neodvisni
[4] vsi dogodki imajo isto verjetnost, $P(1.\neq B) = 1 - P(1.=B) = 1 - \pi = 1 - 9/16$

Starši so: BbEe+BbEe, rodi se pet mladičev

Kolikšna je verjetnost, da se rodi natanko en mladič črne (B) barve? ($P(B) = \pi$)

	X	X	X	X	$\cup$	$\pi(1-\pi)^4$	+
X		X	X	X	$\cup$	$(1-\pi)\pi(1-\pi)^3$	+
X	X		X	X	$\cup$	$(1-\pi)^2\pi(1-\pi)^2$	+
X	X	X		X	$\cup$	$(1-\pi)^3\pi(1-\pi)$	+
X	X	X	X			$(1-\pi)^4\pi$	=
						$5\pi(1-\pi)^4$ =	
						$5 \cdot 9/16 \cdot (1-9/16)^4$ =0.103	

S formulami

n=5, k=1, π=9/16

P(K=1|n=5, π=9/16)=$5 \cdot 9/16 \cdot (1-9/16)^4 = n \cdot \pi^k (1-\pi)^{n-k}$

3.5

Starši so: BbEe+BbEe, rodi se pet mladičev

Kolikšna je verjetnost, da se rodita natanko dva mladiča črne (B) barve? ($P(B) = \pi$)

		X	X	X	$\cup$	$\pi^2(1-\pi)^3$	+
	X		X	X	$\cup$	$\pi^2(1-\pi)^3$	+
	X	X		X	$\cup$	$\pi^2(1-\pi)^3$	+
...	...	...	...	...	$\cup$	$\pi^2(1-\pi)^3$	+
...	...	...	...	...	$\cup$	$\pi^2(1-\pi)^3$	+
X	X	X				$\pi^2(1-\pi)^3$	=
						$10\pi^2(1-\pi)^3$ =	
						$10 \cdot (9/16)^2 \cdot (1-9/16)^3$ =0.265	

S formulami

n=5, k=2, π=9/16

P(K=2|n=5, π=9/16)=$10 \cdot (9/16)^2 \cdot (1-9/16)^3$ =

$= \binom{n}{k} \cdot \pi^k (1-\pi)^{n-k}$

3.6

Binomski simbol

Binomski simbol

Na koliko načinov lahko izberemo k elementov izmed n (brez ponavljanja)?

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

Število kombinacij izbora k elementov iz množice n elementov brez možnosti ponavljanja elementov.

Imamo pet mladičev. Na koliko različnih načinov lahko izberemo 2 mladiča?

$n = 5, k = 2$

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \cdot 4}{2} = 10$$

3.7

Binomski simbol - dodatni primeri

Binomski simbol
Na koliko načinov lahko izberemo k elementov izmed n (brez ponavljanja)?

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

Tri izmed 5?
$n = 5, k = 3$

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} =$$

$$= \frac{5 \cdot 4}{2} = 10$$

Eden izmed 5?
$n = 5, k = 1$

$$\binom{5}{1} = \frac{5!}{1!(5-1)!} =$$

$$= \frac{5}{1} = 5$$

Nobeden izmed 5?
$n = 5, k = 0$

$$\binom{5}{0} = \frac{5!}{0!(5-0)!} =$$

$$= \frac{5!}{5!} = 1,$$

ker $0! = 1$

3.8

Starši so: BbEe+BbEe, rodi se pet mladičev

Kolikšna je verjetnost, da se rodi natanko k črnih mladičev?
$\pi = 9/16, n = 5$

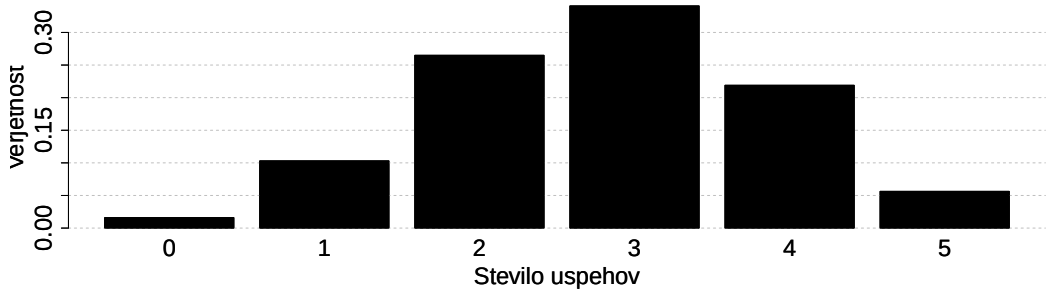
Verjetnostna porazdelitev	Vrednost	Formula	Spošna formula
P(K=0)	0.016	$(1-\pi)^5$	$\binom{5}{0}\pi^0(1-\pi)^{5-0}$
P(K=1)	0.103	$n \cdot \pi \cdot (1-\pi)^4$	$\binom{5}{1}\pi^1(1-\pi)^{5-1}$
P(K=2)	0.265	$\binom{5}{2}\pi^2 \cdot (1-\pi)^3$	$\binom{5}{2}\pi^2(1-\pi)^{5-2}$
P(K=3)	0.341	$\binom{5}{3}\pi^3 \cdot (1-\pi)^2$	$\binom{5}{3}\pi^3(1-\pi)^{5-3}$
P(K=4)	0.219	$n \cdot \pi^4 \cdot (1-\pi)$	$\binom{5}{4}\pi^4(1-\pi)^{5-4}$
P(K=5)	0.056	π^5	$\binom{5}{5}\pi^5(1-\pi)^{5-0}$
P(K=k|n, π)			$\binom{n}{k}\pi^k(1-\pi)^{n-k}$

3.9

Binomska (verjetnostna) porazdelitev (P(K=k|n, π))

P(K=k|n, π)

$$P(K=k|n,\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

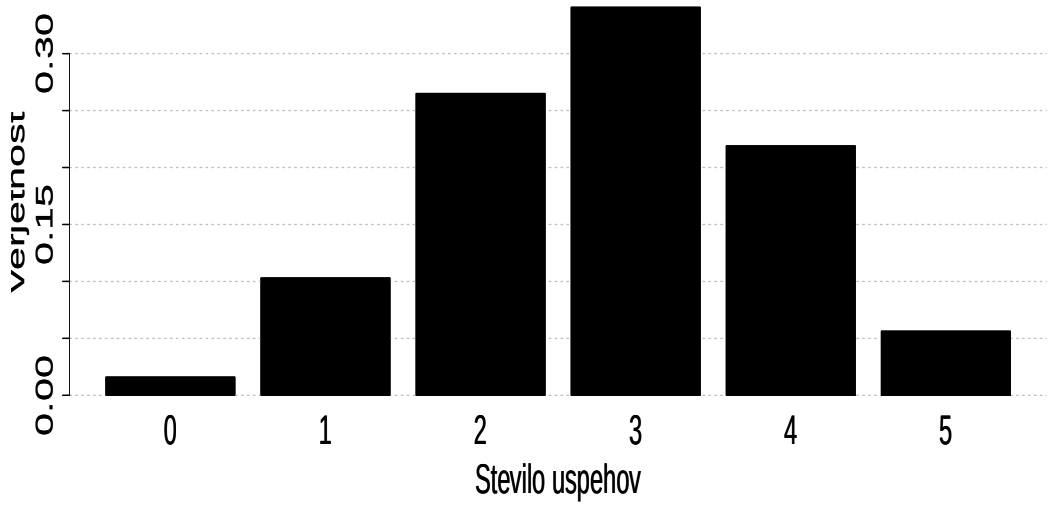


Parametra, ki določata porazdelitev: $n = 5, \pi = 9/16$
$K \sim Bin(n, \pi)$

3.10

Binomska (verjetnostna) porazdelitev (P(K=k|n, π))

P(K=k|n, π)



$K \sim Bin(n = 5, \pi = 9/16)$
Kolišen je modus? 3
Kolišno je povprečje? 2.8
$0 \cdot P(K=0) + 1 \cdot P(K=1) + 2 \cdot P(K=2) + 3 \cdot P(K=3) + 4 \cdot P(K=4) + 5 \cdot P(K=5)$
Kolišen je standardni odklon? 1.1
Ali je porazdelitev simetrična? Rahlo negativno asimetrična (v levo), ker $\pi > 0.5$

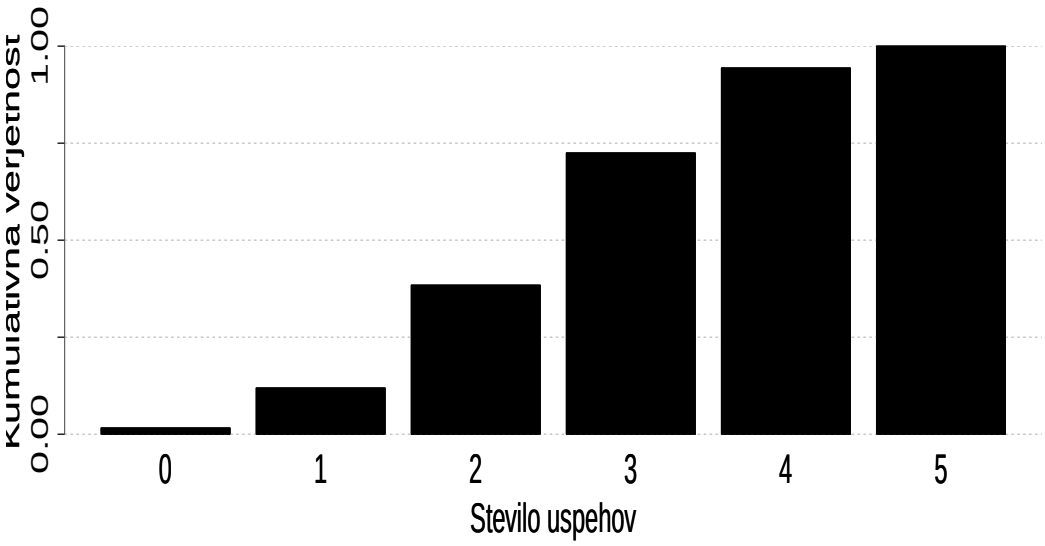
Formule

$E(K) = n \cdot \pi$
$Var(K) = n\pi(1-\pi)$

3.11

Porazdelitvena binomska funkcija ($P(K \leq k|n,\pi)$)

$P(K \leq k|n,\pi)$



$K \sim Bin(n = 5, \pi = 9/16)$

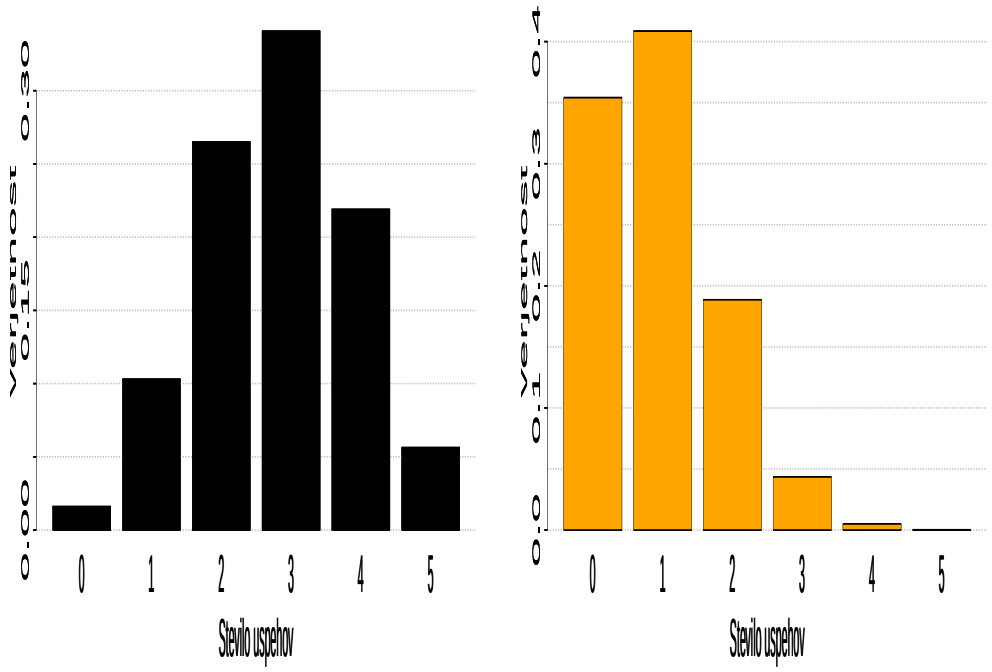
Kolišna je mediana? 3

Kolišen je prvi kvartil? 2

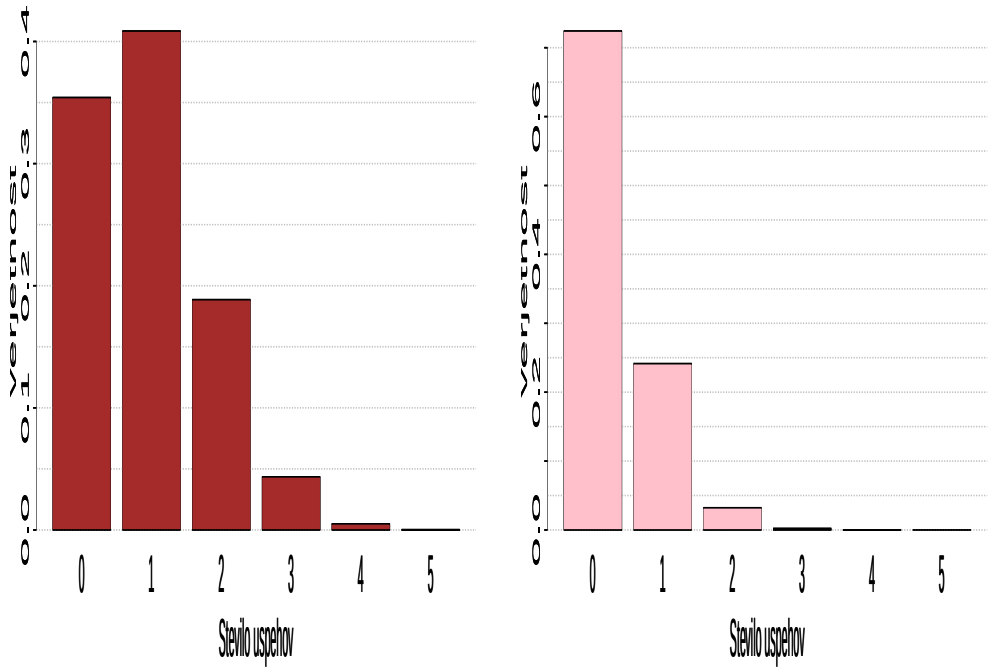
Kolikšen je interkvartilni razmik? od 2 to 4

3.12

Porazdelitve za vse barve



3.13



Binomska porazdelitev - povzetek

Predpostavke za uporabo

- Zanima nas število dogodkov (K) pri n enotah.

- Vsak dogodek je dihotomen (opisna spremenljivka z dvema vrednostima)(da/ne, uspeh/neuspeh, ...)
- Enote so si med sabo neodvisne.
- Vsaka enota ima isto verjetnost, da se bo dogodek zgodil (π).

Dodatni primeri

K: število šestic v 10 metih: $\pi = 1/6$, $n = 10$

K: število punčk za žensko s tremi otroki: $\pi = 0.48$, $n = 3$

K: število študentov, ki bo opravilo izpit brez učenja: (10 vrprašanj, vsaka ima 4 možne odgovore) $\pi = 1/4$, $n = 10$

K: število zmagovalcev na loteriji $\pi = 0.0000001, n = 50,000$

3.14

Poglavje 4

Eksaktni binomski test

V tem poglavju se bomo prvič ukvarjali s sklepno statistiko. Še vedno se bomo osredotočili na opisne spremenljivke (barva labradorcev) in na verjetnost dogodkov (da je mladič labradorca črne barve). Primerjali bomo teoretične rezultate (izračunani s pomočjo binomske porazdelitve) z empiričnimi (z opazovanimi podatki iz vzorca). Ogledali si bomo, ali so odstopanja med teoretičnimi in empiričnimi podatki majhna ali velika in od česa so lahko odvisna. Statistika nam bo pomagala, da bomo lahko ovrednotili velikost teh razlik.

Poskusili bomo tudi oceniti verjetnost dogodkov na podlagi opaženih podatkov. To je uporabno, ko ne poznamo teoretične vrednosti. Ogledali si bomo na primer vzorec, ki vsebuje veliko mladičev labradorcev in bomo ocenili verjetnost, da se rodi mladič črne barve. To bi lahko bilo uporabno, če ne bi poznali genotipa staršev.

Namen poglavja je razumeti:

- razliko med teoretičnimi in empiričnimi vrednosti;
- eksaktni binomski test;
- postopek statističnega skelpanja;
- vrednost p;
- ničelno in alternativno domnevo;
- interval zaupanja.

Uvedli bomo tudi gaussovo porazdelitev, kateri je binomska porazdelitev podobna, ko je število poskusov oziroma velikost vzorca *veliko* (ponavadi je približek dober, če $n\pi > 5$ in $n(1-\pi) > 5$). Srečali bomo ponovno gaussovo porazdelitev, ko se bomo začeli ukvarjati s sklepno statistiko za številske spremenljivke.

Teoretična in empirična porazdelitev

Članek: Templeton in kolegi, 1977; Journal of Heredity

Table I. Mating homozygous dogs

Parents				Offspring			
Genotype	Phenotype	Genotype	Phenotype	No. litters	No. obs.	Coat color	No. exp.
1. *B/B, E/E*	black	x *B/B, E/E*	black	128	859	all black	859
2. *b/b, E/E*	choc.	x *b/b, E/E*	choc.	55	300	all chocolate	300
3. *B/B, e/e*	yellow	x *B/B, e/e*	yellow	4	22	all yellow	22

- Kaj so opazovane (empirične) frekvence? (*No. obs*: število opazovanih)

- Kaj so pričakovane (teoretične) frekvence? (*No. exp*: število pričakovanih)
- Kako so izračunali pričakovane frekvence?
  $E(K) = n \cdot \pi$; K=število mladičev dane barve; n=skupno število mladičev; $\pi = 1$ verjetnost uspeha (mladiča dane barve)
- Ali se teoretične in empirične frekvence *ujemajo*? Zakaj?

4.1

Teoretična in empirična porazdelitev

Članek: Templeton in kolegi, 1977; Journal of Heredity

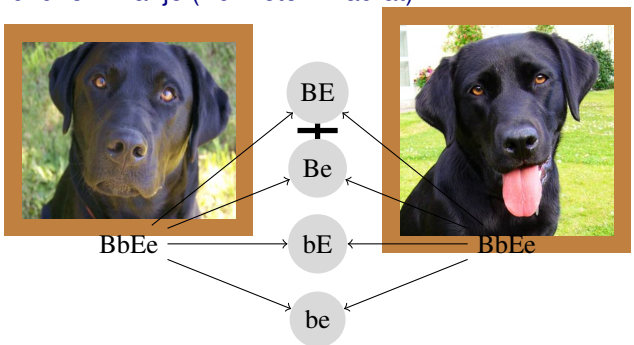
Table III. Various matings involving *B*- and *E*-loci alleles

Parents				Offspring						
Genotype	Phenotype	Genotype	Phenotype	No. litters	No. black obs.	No. black exp.	No. choc. obs.	No. choc. exp.	No. yellow obs.	No. yellow exp.
1. *B/B,E/E*	black	× *B/B,e/e*	yellow	9	68	68	0	0	0	0
2. *b/b,E/E*	choc.	× *B/B,e/e*	yellow	7	61	61	0	0	0	0
3. *B/B,E/e*	black	× *B/B,E/E*	black	3	14	14	0	0	0	0
4. *B/b,E/E*	black	× *B/B,e/e*	yellow	4	30	30	0	0	0	0
5. *B/B,E/e*	black	× *B/B,e/e*	yellow	7	21	18	0	0	15	18
6. *B/b,E/e*	black	× *B/B,e/e*	yellow	3	13	14	0	0	15	14
7. *B/b,E/e*	black	× *B/b,E/e*	black	7	26	28.7	14	9.6	11	12.8
8. *b/b,E/E*	choc.	× *b/b,e/e*	yellow	2	0	0	12	12	0	0

- Ali se teoretične in empirične frekvence *ujemajo*? Zakaj?
- Primer, ki ga dobro poznamo: številka **7: B/b, E/e + B/b, E/e**.
- Ali so odstopanja dovolj velika, da lahko rečemo, da je teoretična porazdelitev napačna?

4.2

Dihibridno križanje (Punnetov kvadrat)



	BE	Be	bE	be
BE	BBEE	BBEe	BbEE	BbEe
Be	BBeE	BBee	BbeE	Bbee
bE	bBEE	bBEe	bbEE	bbEe
be	bBeE	bBee	bbeE	bbee

Barva dlak mladičev

P(Črna)=$9/16 = 0.5625$

P(Rjava)=$3/16 = 0.1875$

P(Rumena)=$4/16 = 0.25$

4.3

Ali so odstopanja velika ali majhna?

B/b, E/e + B/b, E/e

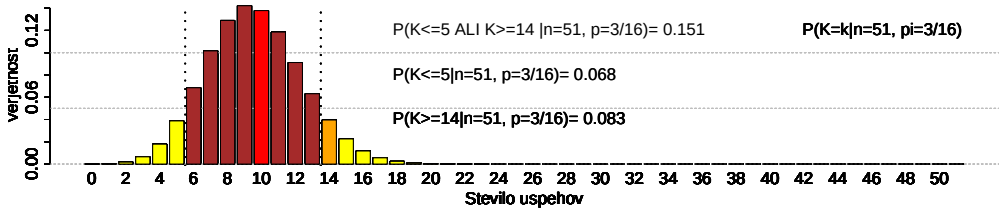
Barva	Pričakovana frekvenca ($n\pi$)	Opazovana frekvenca	Pričakovani delež (π)	Opazovani delež
Črna	28.7	26	0.5625	0.51
Rjava	9.6	14	0.1875	0.27
Rumena	12.8	11	0.25	0.22
Vsota	n=51	n=51	1	1

Kolikšna je verjetnost, da bi lahko opazili taka odstopanja od teoretičnih vrednostih, ali še bolj skrajna odstopanja, če drži, da so starši BbEe+BbEe? Osredotočili se bomo na rjavo barvo *P(Rjava)=3/16*.

4.4

Eksaktni binomski test in vrednost p

Rjavi mladiči - $n = 51$; Domneva: $\pi = 3/16$, k=14



- P($K \leq 5 \cup K \geq 14$ |n=51, π=3/16)= 0.151 je verjetnost, da dobimo na *vzorcu* rezultat, ki smo ga vidli (14 rjavih psov izmed 51) , ali še bolj *skrajni* rezultat (15, 16, ..., 51 ; ampak tudi 5, 4, 3, 2, 1, 0), če drži, da je $\pi = 3/16$.
- Tej verjetnosti pravimo *vrednost p*.

4.5

Ničelna domneva in vrednost p

Interpretacija vrednosti p, p=0.15

- *Ničelna domneva*: $\pi = 3/16$; hipoteza na podlagi katere smo izračunali vrednost p.
- Ničelna domneva: v populaciji, ki jo preučujemo, je verjetnost, da bo mladič rjave barve 3/16.
- Kaj je populacija? Kaj je vzorec?
- Kaj lahko zaključimo na podlagi izračunane vrednosti p (p=0.15)?
  - vrednost p je *visoka*;
  - Ne bi zavrnili domneve, da je $\pi = 3/16$.
  - *Ne bi zavrnili ničelne domneve ($\neq$ sprejeli!!!)* in *rezultat ni statistično značilen*
  - Na podlagi naših podatkov, ne moremo trditi, da je delež rjavih mladičev v populaciji različen od $\pi = 3/16$.
- *Absence of evidence is not evidence of absence*

4.6

Analiza s pomočjo računalnika (program R)

```
binom.test(x=14, n=51, p=3/16)


Exact binomial test

data:  14 and 51
number of successes = 14, number of trials = 51, p-value = 0.1482
alternative hypothesis: true probability of success is not equal to 0.1875
```

```
95 percent confidence interval:
 0.1589 0.4174
sample estimates:
probability of success
                0.2745
```

4.7

Interval zaupanja

95% interval zaupanja za π

- Na vzorcu je bil opažena verjetnost $p = 0.27 = 14/51$.
- 95% interval zaupanja za π (populacijska verjetnost) je: 0.16 do 0.42.
- Interpretacija intervala zaupanja: imamo 95% zaupanje, da je populacijska verjetnost π v intervalu od 0.16 do 0.42

Intervali zaupanja

- π je praviloma neznana in jo želimo oceniti.
- Postopek, s katerim izračunamo 95% interval zaupanja, zagotavlja, da je verjetnost, da bo populacijska verjetnost vključena v intervalu, 0.95.
- Ali je bila ocena natančna ali ne? Kako bi lahko povečali natančnost?

4.8

Isti vzorec, večje zaupanje

```
binom.test(x=14, n=51, p=3/16, conf.level=0.99)


Exact binomial test

data:  14 and 51
number of successes = 14, number of trials = 51, p-value = 0.1482
alternative hypothesis: true probability of success is not equal to 0.1875
99 percent confidence interval:
 0.1314 0.4610
sample estimates:
probability of success
                0.2745
```

4.9

Isti vzorec, enostranski test

```
binom.test(x=14, n=51, p=3/16, conf.level=0.95, alternative="greater")


Exact binomial test

data:  14 and 51
number of successes = 14, number of trials = 51, p-value = 0.0832
alternative hypothesis: true probability of success is greater than 0.1875
95 percent confidence interval:
 0.1742 1.0000
sample estimates:
probability of success
                0.2745
```

4.10

Isti vzorec, enostranski test

```
binom.test(x=14, n=51, p=3/16, conf.level=0.95, alternative="less")


Exact binomial test

data:  14 and 51
number of successes = 14, number of trials = 51, p-value = 0.9563
alternative hypothesis: true probability of success is less than 0.1875
95 percent confidence interval:
 0.0000 0.3954
sample estimates:
probability of success
                0.2745
```

4.11

Večji vzorec, isti p

```
binom.test(x=14*2, n=51*2, p=3/16)


Exact binomial test

data:  14 * 2 and 51 * 2
number of successes = 28, number of trials = 102, p-value =
0.03041
alternative hypothesis: true probability of success is not equal to 0.1875
95 percent confidence interval:
 0.1908 0.3718
sample estimates:
probability of success
                0.2745
```

4.12

Manjši vzorec, (približno) isti p

```
binom.test(x=round(14/2), n=round(51/2), p=3/16)


Exact binomial test

data:  round(14/2) and round(51/2)
number of successes = 7, number of trials = 26, p-value = 0.3117
alternative hypothesis: true probability of success is not equal to 0.1875
95 percent confidence interval:
 0.1157 0.4779
sample estimates:
probability of success
                0.2692
```

4.13

Koliko čudna so odstopanja od teoretične porazdelitve?

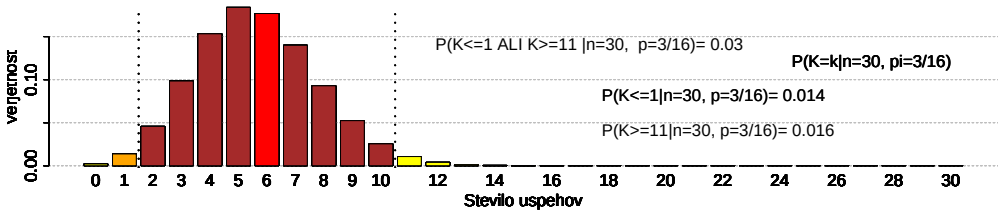
Primer

- Kupite par črnih labradorcev; prodajalec trdi, da imata oba genotip B/b E/e
- Sumite, da je pravi genotip drugačen, oziroma, da je verjetnost, da boste imeli rjavega mladiča različna od 3/16.
- Zbirate podatke! V prvih treh letih se rodi 30 mladičev; samo eden izmed njimi je rjave barve.
- Na vzorcu je delež rjavih mladičev 0.033, v populaciji naj bi bil 0.1875. Ali lahko na podlagi vzorca trdite, da je genotip različen od B/b E/e?

4.14

Eksaktni binomski test in vrednost p

Nadaljevanje: Rjavi mladiči - $n = 30$; Domneva: $\pi = 3/16$, k=1



- P($K \leq 1 \cup K \geq 11$ |n=30, π=3/16)= 0.032 je verjetnost, da dobimo na *vzorcu* rezultat, ki smo ga vidli (1 rjav pes izmed 30) , ali še bolj *skrajni* rezultat (0 ; ampak tudi 11, 12, 13, ... 30), če drži, da je $\pi = 3/16$.
- Tej verjetnosti pravimo *vrednost p*.

4.15

Interpretacija rezultatov

Interpretacija vrednosti p, $p = 0.03$

- *H0: Ničelna domneva*: $\pi = 3/16$; hipoteza na podlagi katere smo izračunali vrednost p. Želimo zavriniti H0.
- *Ha: Alternativna domneva*: $\pi \neq 3/16$; hipoteza, ki jo želimo dokazati.
- Kaj je populacija? Kaj je vzorec?
- Kaj lahko zaključimo na podlagi izračunane vrednosti p (p=0.03)?
  - vrednost p je *majhna*;
  - Zavrnili bi domnevo, da je $\pi = 3/16$.
  - *Zavrnili bi ničelno domnevo* in rekli, da je *rezultat statistično značilen*
  - Na podlagi naših podatkov, lahko trdimo, da je delež rjavih mladičev v naši populaciji različen od $\pi = 3/16$, oziroma, da starši nimajo genotipa B/b E/e.

4.16

Analiza s pomočjo računalnika (program R)

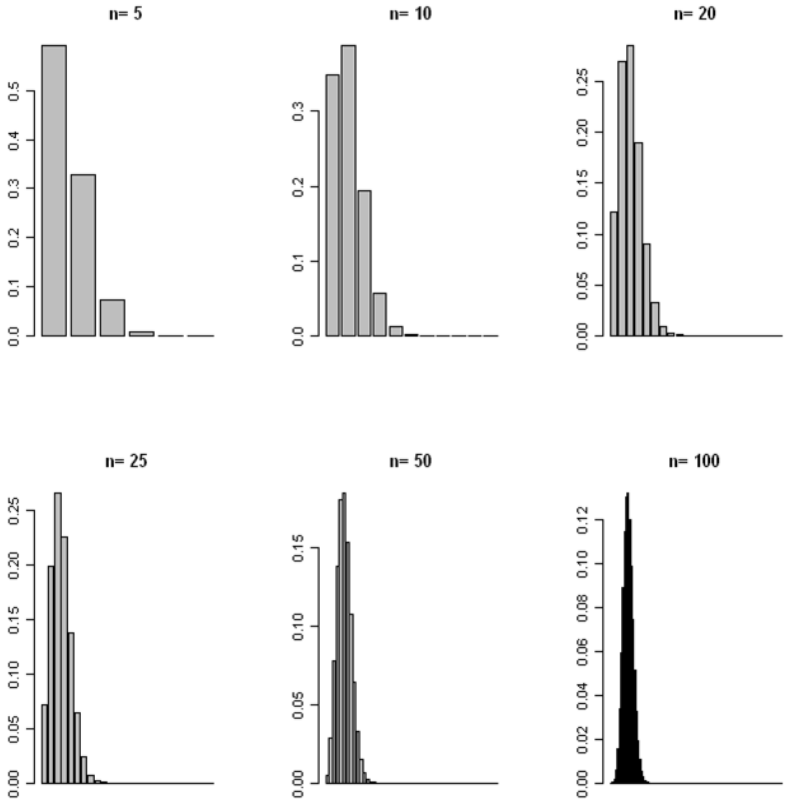
```
binom.test(x=1, n=30, p=3/16)


Exact binomial test

data:  1 and 30
number of successes = 1, number of trials = 30, p-value = 0.03185
alternative hypothesis: true probability of success is not equal to 0.1875
95 percent confidence interval:
 0.0008436 0.1721695
sample estimates:
probability of success
               0.03333
```

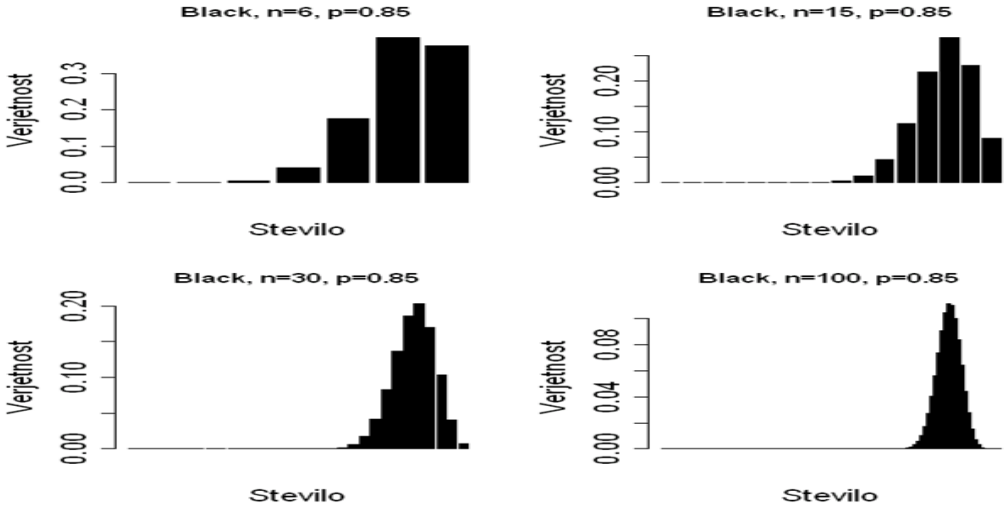
4.17

Binomska porazdelitev: povečamo velikost vzorca



4.18

Binomska porazdelitev: povečamo velikost vzorca



Gaussova porazdelitev

- K: število uspehov
- $K \sim Bin(n, \pi)$; $E(K) = n\pi$, $Var(K) = n\pi(1-\pi)$
- $X \sim N(\mu, \sigma)$: gaussova (normalna porazdelitev) s povrečjem $E(X) = \mu$ in standardnim odkonom $SD(X) = \sigma$
- Če je vzorec *velik*, velja, da $K \sim N(n\pi, \sqrt{n\pi(1-\pi)})$
- *Velik vzorec*: ponavadi je približek dober, ko če $n\pi > 5$ in $n(1-\pi) > 5$

4.19

Gaussova porazdelitev

4.20

Porazdelitev deleža

Porazdelitev

- K: število uspehov
- $K \sim Bin(n, \pi)$; $E(K) = n\pi$, $Var(K) = n\pi(1-\pi)$
- Če je vzorec *velik*, velja, da $K \sim N(n\pi, \sqrt{n\pi(1-\pi)})$
- p=K/n: delež uspehov
- $E(aX) = aE(X)$ in $Var(aX) = a^2 Var(X)$
- $E(K/n) = \pi$, $Var(K/n) = pi(1-\pi)/n$
- Če je vzorec *velik*, velja, da $p = K/n \sim N(\pi, \sqrt{\pi(1-\pi)/n})$

4.21

Poglavje 5

Primerjava deležev

V tem poglavju si bomo ogledali, kako lahko preučujemo povezanost med dvema opisnima spremenljivkama. Ogledali si bomo, kako (in kdaj) so statistiki prvič dokazali, da obstaja povezanost med kajenjem in rakom na pljučih. Pogledali bomo, kako so zasnovali raziskavo, kaj so opazili in s katero statistično metodo so dokazali, da povezanost obstaja (ne samo na vzorcu, ampak tudi v populaciji). Odkrili bomo tudi, zakaj je bila raziskava sporna in ali so lahko dokazali, da kajenje povzroči raka na pljučih.

Namen poglavja je spoznati:

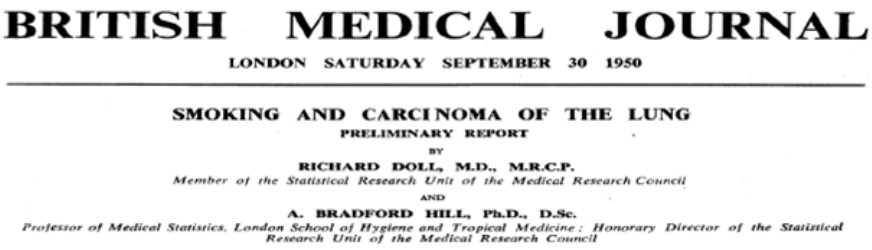
- študijo primerov in kontrol;
- razliko med opazovalnimi in eksperimentalnimi študijami;
- zakaj z opazovalnimi študijami težko dokažemo vzročnost;
- kontingenčno tabelo;
- test hi-kvadrata;
- mere povezanosti (relativno tveganje in razmerje obetov).

Ali je kajenje škodljivo?

"Statistika kaže, da je *večja verjetnost*, da se rak na pljučih in določene druge bolezni pojavijo pri kadilcih kot pri nekadilcih. Ta trditev temelji na *epidemioloških študijah*, ki so s pomočjo vprašalnikov oblikovala zapažanja pri določenih populacijah ali skupinah ljudi. Te študije so pokazale, da je *kajenje povezano z več boleznimi*, javne zdravstvene oblasti pa so na njihovi podlagi sklenile, da je kajenje *vzrok* pljučnega raka in drugih bolezni pri kadilcih."

Iz spletne strani: Tobačna Ljubljana - Kajenje in zdravje. 5.1

Zakaj danes vemo, da je kajenje škodljivo?



BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.

Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.

Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

Prvi dokaz

- Leta *1950*: kaj je bilo takrat že znano o škodljivosti kajenja?
- Zakaj so se osredotočili na *pljučnega raka*?
- Kaj je bil poklic avtorjev?

5.2

Rezultati študije primerov in kontrol *(case control study)*

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

Kaj so naredili?

- Kako pravimo taki tabeli?
- Kaj so merili? Katere vrste so spremenljivke?
- Kako so izvedli raziskavo? Koga so vključili?
- Kaj je populacija?
- Kaj so želeli dokazati?
- Kaj pomeni, da je raziskava *epidemiološka*?
- Kaj mislite, da so zaključili?

5.3

Rezultati študije primerov in kontrol *(case control study)*

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

Kolikšna je verjetnost ...

- ... kajenja za tiste, ki imajo raka na pljučih? $P(Kajenje|Rak)$=688/709=0.97.
- ... kajenja za tiste, ki nimajo raka na pljučih? $P(Kajenje|\overline{Rak})$=650/709=0.92.
- *Na vzorcu* sta bila deleža kadilcev: 0.97 (med tistimi z rakom) in 0.92 (med tistimi brez raka)
- Ali je razlika dovolj velika, da bi jo lahko pričakovali tudi *v populaciji*?
- Kako bi statistično ovrednotili, ali je kajenje povezano z rakom na pljučih? Uporabili bomo *test hi-kvadrata.*

5.4

Test hi-kvadrata

Alternativna domneva (ki jo želimo potrditi!)
Kajenje in rak na pljučih *sta povezani v populaciji*.

Ničelna domneva (ki jo želimo zavrniti!)

- Kajenje in rak na pljučih *sta neodvisni v populaciji*.
- Dve spremenljvki sta neodvisni, ko
  - P(Kajenje|Rak)=P(Kajenje)
  - P(Rak|Kajenje)=P(Rak)
- če sta neodvisni: P(Kajenje ∩ Rak)=P(Kajenje)P(Rak)

5.5

Kako bi se morali porazdeliti osebe, če bi bili spremenljvki *neodvisni*?

Opazovane frekvence (O)

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

Pričakovane frekvence (E)

	Ima raka	Nima raka	Vsota
Kadi	*669*	669	1338
Ne kadi	40	40	80
Vsota	709	709	1418

P(Rak ∩ Kadi)=P(Rak)P(Kadi) = $\frac{709}{1418} \cdot \frac{1338}{1418} = 0.5 \cdot 0.94 = 0.47$
Pričakovana frekvenca za kadilce z rakom: P(Rak ∩ Kadi) $\cdot n$= $0.47 \cdot 1418 = 669$.

5.6

Ali je razlika med pričakovanimi in opazovanimi frekvencami velika?

Testna statistika

Opazovane frekvence (O)

	Ima raka	Nima raka	Vsota
Kadi	688	650	1338
Ne kadi	21	59	80
Vsota	709	709	1418

Pričakovane frekvence (E)

	Ima raka	Nima raka	Vsota
Kadi	669	669	1338
Ne kadi	40	40	80
Vsota	709	709	1418

Kolikšna je verjetnost, da so lahko nastali na vzorcu taki odmiki, ali še večji, med pričakovanimi in opazovanimi frekvencami, če sta spremenljivki v populaciji res neodvisni?
Kolikšna je *vrednost p?*

Povzamemo odmike z eno številko.

Testna statistika: $\chi^2 = \sum \frac{(O-E)^2}{E}$

$\chi^2 = \frac{(688-669)^2}{669} + \frac{(650-669)^2}{669} + \frac{(21-40)^2}{40} + \frac{(59-40)^2}{40} = 19.13$

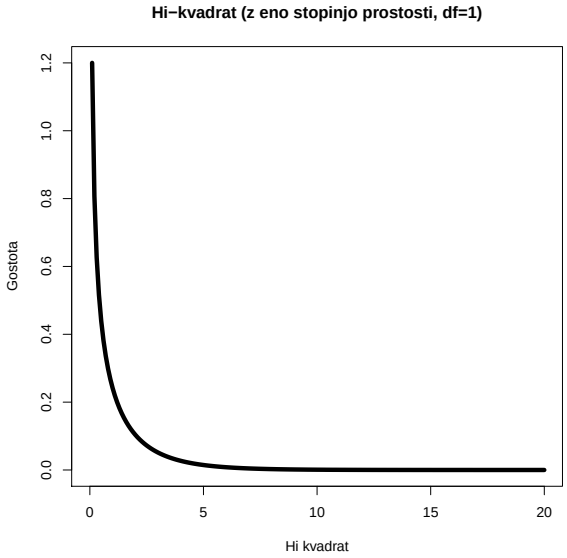
Ali je ta številka velika ali majhna?

Kako določimo vrednost p?

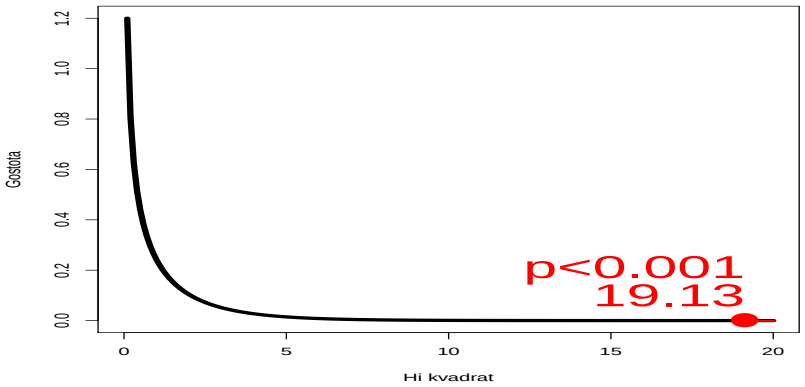
Vrednost p

p=P(χ^2>19.13| H0) =?

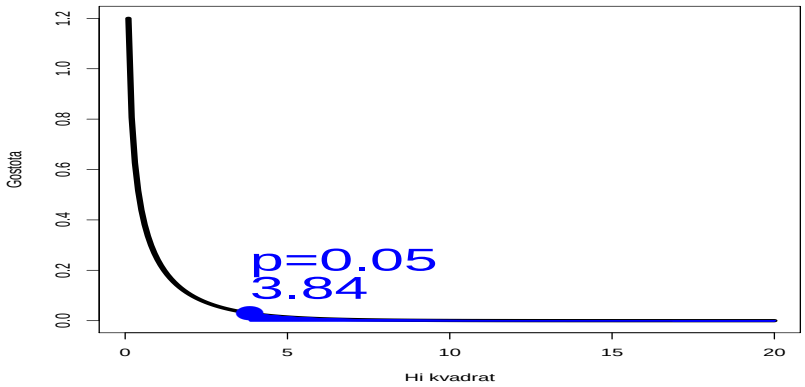
Tokrat je ne znamo *izračunati*. Lahko uporabimo teoretični rezultat, ki ga je dobil leta 1900 Karl Pearson.



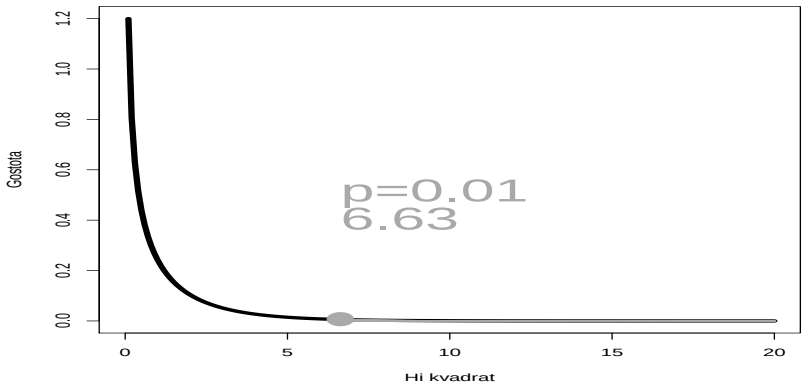
Kako odčitamo vrednost p?



Za katero vrednost testne statistike bi dobili (natanko) p=0.05?



Za katero vrednost testne statistike bi dobili (natanko) p=0.01?



5.8

Kaj lahko sklepamo?

p<0.001

- vrednost p je majhna
- zavrnemo ničelno domnevo, da sta spremenljivki v populaciji neodvisni.
- Sklepamo, da sta kajenje in rak na pljučih povezani v populaciji.
- Ali smo dokazali, da kajenje povzroča rak na pljučih?
- *Video from Joy of Stats*

5.9

Predpostavke za uporabo hi-kvadrata

Predpostavke

- Enote so neodvisne
- Vzorec je *velik*
  - *Rule of thumb* Pričakovane frekvence so večje od 5 za vsaj 80% celic.
- Kaj storiti, če je vzorec *majhen*
  - Uporabimo Yeatesov popravek (*Yates' continuity correction*)

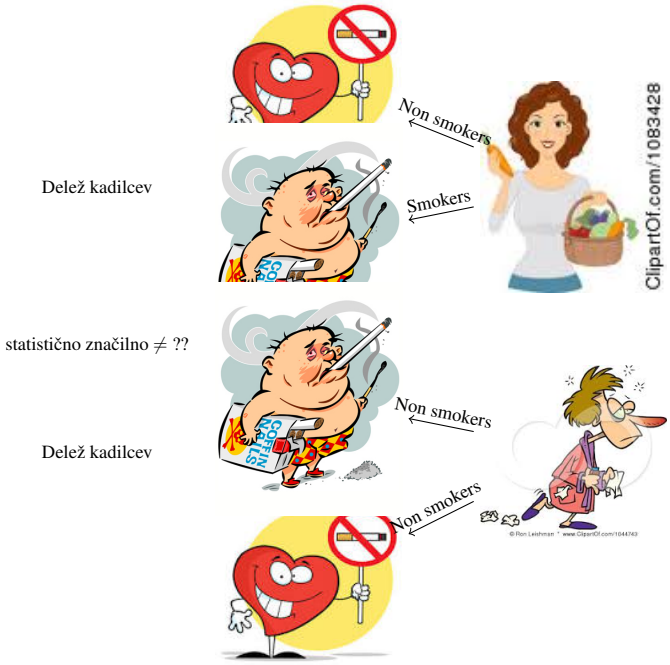
$$\chi^2 = \sum \frac{(|O-E|-0.5)^2}{E}$$

.

  - Uporabimo Fisherjev eksaktni test.

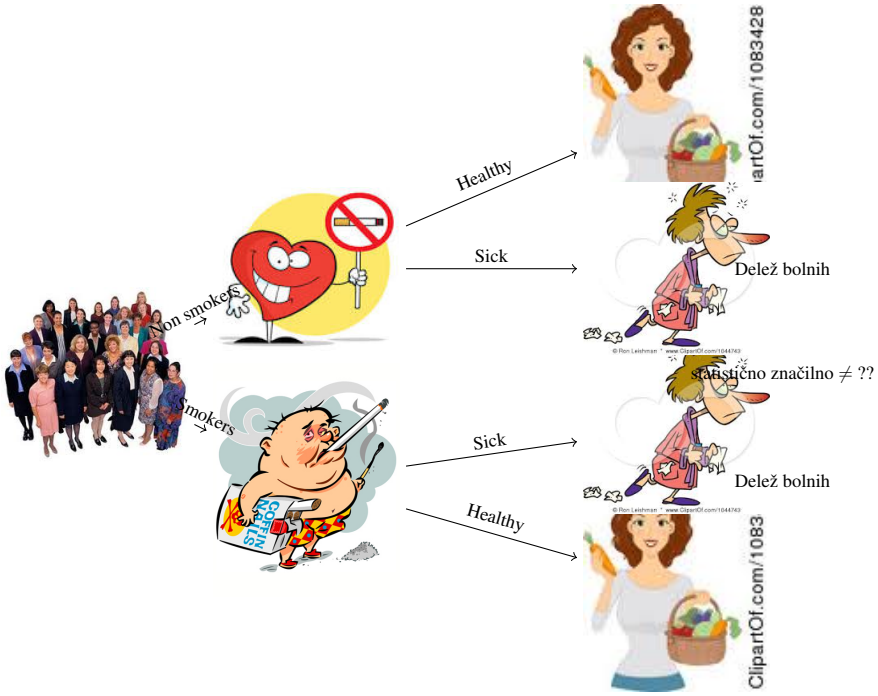
5.10

Retrospektivna študija primerov in kontrol



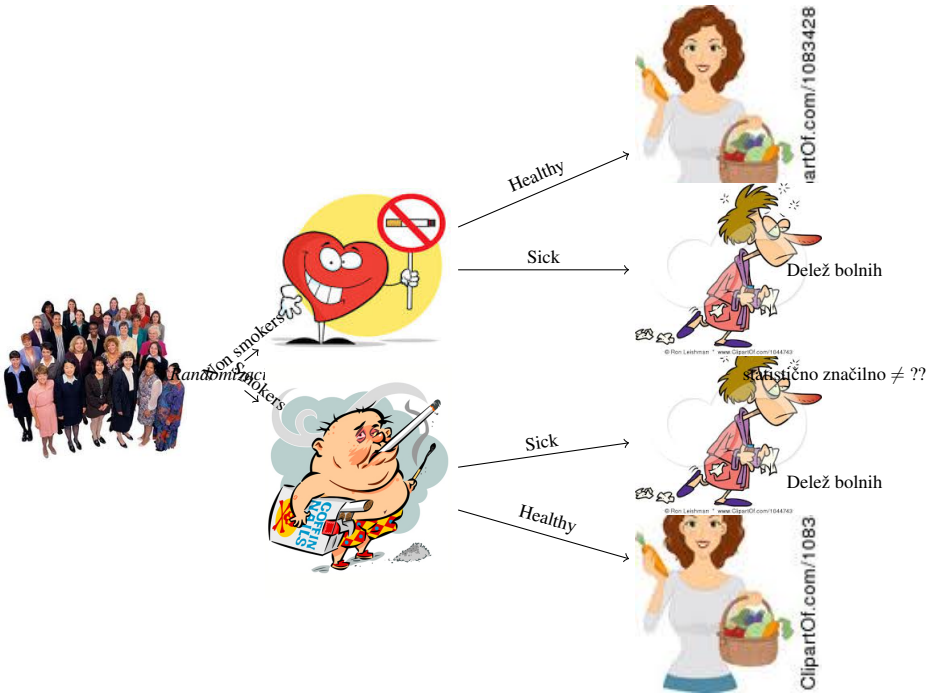
5.11

Prospektivna (cohortna) študija



5.12

Prospektivna randomizirana študija



5.13

Obeti (*odds*)

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

$$Obet(O) = \frac{p}{1-p}$$

Obeti za kajenje za bolnike, ki imajo raka na pljučih

- P(kajenje|RakP)=688/709=0.97
  $O_{RakP} = \frac{P(kajenje|RakP)}{1-P(kajenje|RakP)} = \frac{0.97}{1-0.97} = 32.76 = \frac{a}{c}$.
- Interpretacija: za vsakega nekadilca z rakom na pljučih imamo 32.76 kadilcev z rakom na pljučih.

Obeti za kajenje za bolnike, ki nimajo raka na pljučih

- P(kajenje|Ni RakP)=650/709=0.92
  $O_{NiRakP} = \frac{P(kajenje|NiRakP)}{1-P(kajenje|NiRakP)} = \frac{0.92}{1-0.92} = 11.02 = \frac{b}{d}$.
- Interpretacija: za vsakega nekadilca, ki nima raka na pljučih imamo 11.02 kadilcev, ki nimajo raka na pljučih.

Razmerje obetov (odds ratio) za kajenje za tiste, ki imajo raka na pljučih glede na tiste ki imajo druge bolezni je

$$OR = \frac{O_{RakP}}{O_{NiRakP}} = \frac{32.76}{11.02} = 2.97 = \frac{ad}{bc}$$

Obeti za kajenje so *2.97* večji za bolnike z rakom na pljučih v primerjavi z bolniki, ki imajo druge bolezni.

Dobili bi isto vrednost, če bi izračunali razmerje obetov za raka na pljučih za kadilce glede na nekadilce!

5.14

Relativno tveganje (*relative risk*)

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

Ali bi bilo smiselno izračunati...?*Ne!*

- Verjetnost, da ima bolnik raka na pljučih? P(RakP)=709/1418=0.5.
- Verjetnost, da ima kadilec raka na pljučih? P(RakP|Kadilec)=688/1338=0.51.
- Verjetnost, da ima nekadilec raka na pljučih? P(RakP|Ne kadilec)=21/80=0.26.
- *Relativno tveganje (*relative risk*)* raka na pljučih za kadilce glede na nekadilce?

$$RR = \frac{P(RakP|Kadilec)}{P(RakP|Nekadilec)} = \frac{0.51}{0.26} = 1.96 = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

.

5.15

Obeti in relativno tveganje

	Ima raka	Nima raka	Vsota
Kadi	688 (a)	650 (b)	1338
Ne kadi	21 (c)	59 (d)	80
Vsota	709	709	1418 (n)

*1.96: Relativno tveganje (*relative risk*)* raka na pljučih za kadilce glede na nekadilce?

Imamo 10 krat več kontrol

$$RR = \frac{\frac{a}{a+10*b}}{\frac{c}{c+10*d}} = 2.78$$

$$OR = \frac{a \cdot 10d}{b \cdot 10c} = \frac{ad}{bc} = 2.97$$

Imamo 100 krat več kontrol

$$RR = \frac{\frac{a}{a+100*b}}{\frac{c}{c+100*d}} = 2.95$$

$$OR = \frac{a \cdot 100d}{b \cdot 100c} = \frac{ad}{bc} = 2.97$$

Imamo 1000 krat več kontrol

$$RR = \frac{\frac{a}{a+1000*b}}{\frac{c}{c+1000*d}} = 2.97$$

$$OR = \frac{a \cdot 1000d}{b \cdot 1000c} = \frac{ad}{bc} = 2.97$$

Razmerje obetov in relativno tveganje sta si med sabo zelo podobni, ko je bolezen redka

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{ad}{bc} \cdot \frac{\frac{b}{a+b}}{\frac{d}{c+d}} = OR \cdot \frac{\frac{b}{a+b}}{\frac{d}{c+d}} = OR \cdot \frac{1 - \frac{a}{a+b}}{1 - \frac{c}{c+d}}.$$

5.16

Analiza s pomočjo računalnika (program R)

```
My.table

         Ima raka Nima raka
Kadi          688       650
Ne kadi        21        59
```

```
chisq.test(My.table)


Pearson's Chi-squared test with Yates' continuity correction

data:  My.table
X-squared = 18.14, df = 1, p-value = 2.057e-05

chisq.test(My.table, corr=FALSE)


Pearson's Chi-squared test

data:  My.table
X-squared = 19.13, df = 1, p-value = 1.222e-05
```

5.17

Poglavje 6

Sklepna statistika za številske spremenljivke

V tem poglavju se bomo ukvarjali s sklepno statistiko za eno številsko spremenljivko. Na podlagi podatkov iz vzorca bomo ocenili: povprečno vrednost spremenljivke, interval, ki zajame vrednosti večine enot (95%) in interval zaupanja za povprečno vrednost. Spoznali bomo tudi statistični test (test t), s katerim bomo lahko preverili na podlagi podatkov iz vzorca, ali je povprečna vrednost neke spremenljivke v populaciji enaka dani vrednosti.

Analizirali bomo telesno temperaturo zdravih psov. Ocenili bomo v katerem intervalu lahko pričakujemo temperaturo za večino zdravih psov. Ocenili bomo povprečno temperaturo na vzorcu in podali interval, v katerem bomo z veliko verjetnostjo zajeli populacijsko povprečno temperaturo. Preverili bomo tudi, ali je povprečna temperatura zdravih psov v populaciji 38 stopinj °C.

Ker so temperature podane v stopinjah Fahrenheita in želimo izraziti vse rezultate v Celzijevi lestvici, se bomo ukvarjali tudi s transformacijo podatkov.

Namen poglavja je spoznati

- referenčni razpon za povprečje;
- interval zaupanja za povprečje;
- test t za en vzorec.

Kolikšna je povprečna temperatura psov?

Raziskava (n=102) Celsius=(Fahrenheit-32)/1.8

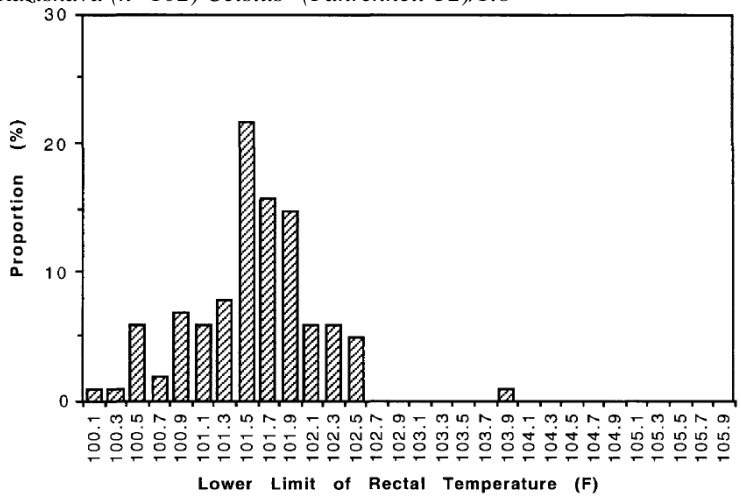


Figure 2.1 Frequency distribution of rectal temperatures in normal dogs.

$\bar{x} = 101.5$ F; $s = 0.58$ F

In Celsius: $\bar{x} =$ (101.5 F -32)/1.8 = 38.61 °C ;

$s =$ (0.58 F)/1.8 = 0.32 °C.

6.1 Vir: R.D.Smith, Veterinary Clinical Epidemiology (II Edition)

Grafični prikaz - kako bi ga lahko izboljšali?

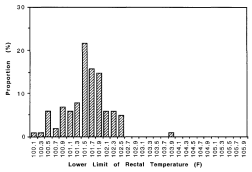
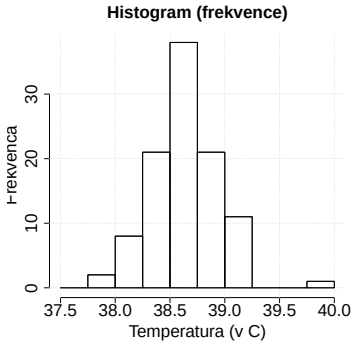


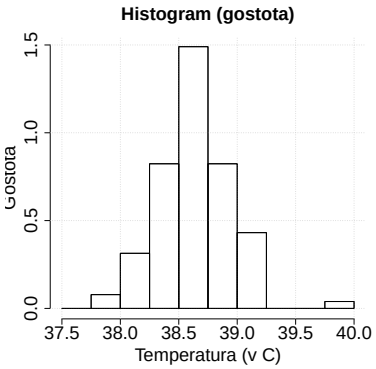
Figure 2.1 Frequency distribution of rectal temperatures in normal dogs.

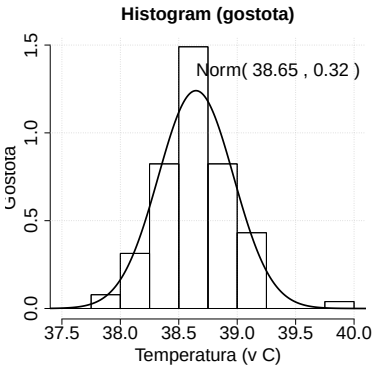
Histogram



6.2

Gostota (*Density*)
Površina histograma=1

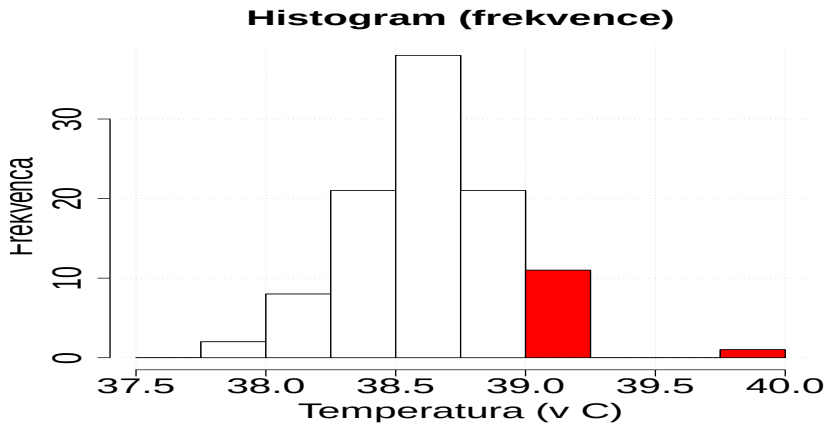




Površina pod krivuljo=1

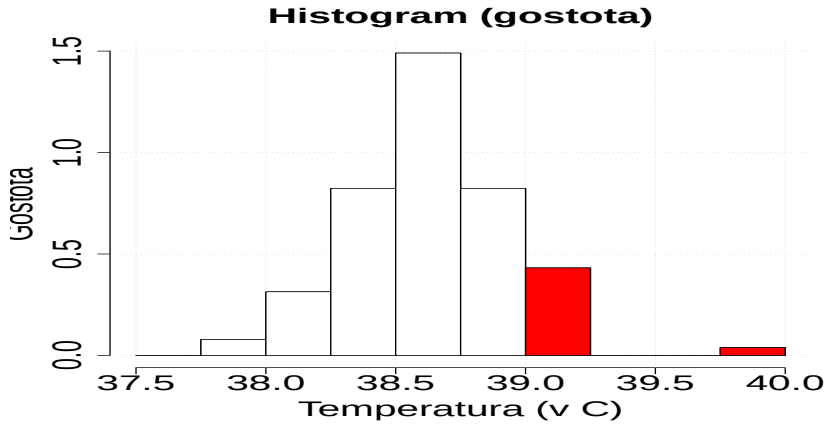
Kolikšna je verjetnost, da je temperatura....

... $\geq 39\ °C?$



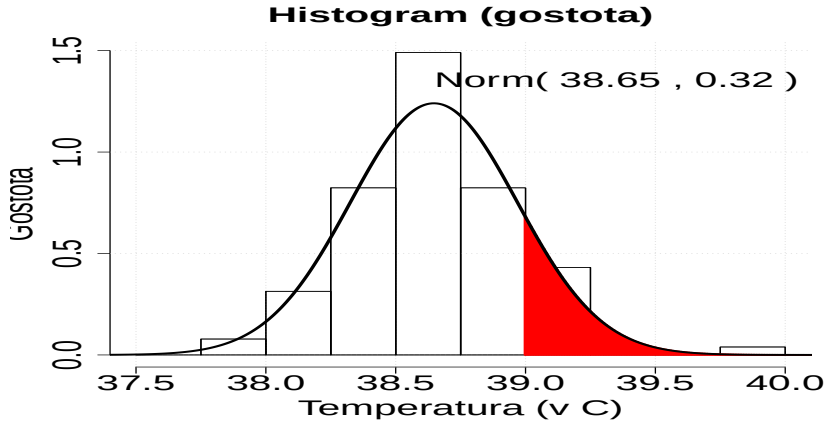
Število psov s temperaturo $\geq$ 39 °C= 12 od 102.
Delež= 0.12

...$\geq 39°C$**?**
Površina histograma=1



Površina= 0.431*0.25 + 0.039*0.25 = 0.12

...$\geq 39C$**?**
Temperatura (v °C) $\sim$ Norm(μ=38.65, σ=0.32)

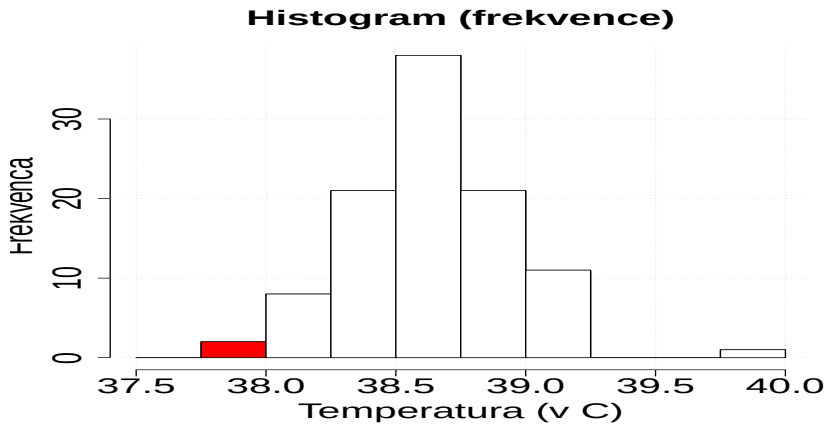


Površina pod krivuljo:
P(Temp$\geq$ 39)=0.14.

6.3

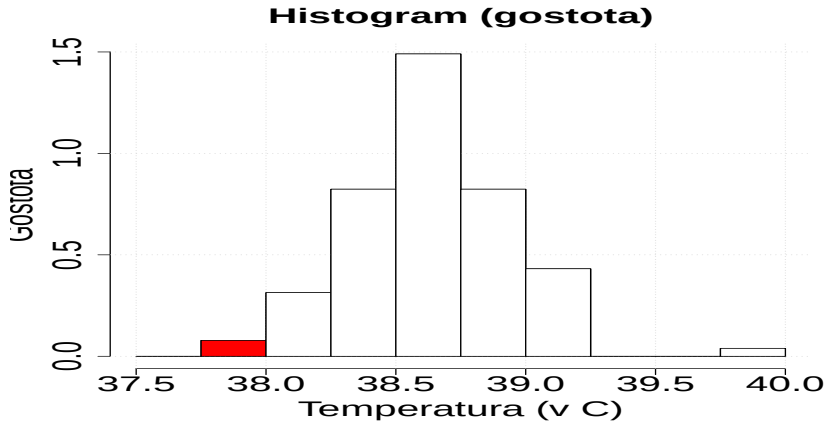
Kolikšna je verjetnost, da je temperatura....

... $< 38\ °C?$



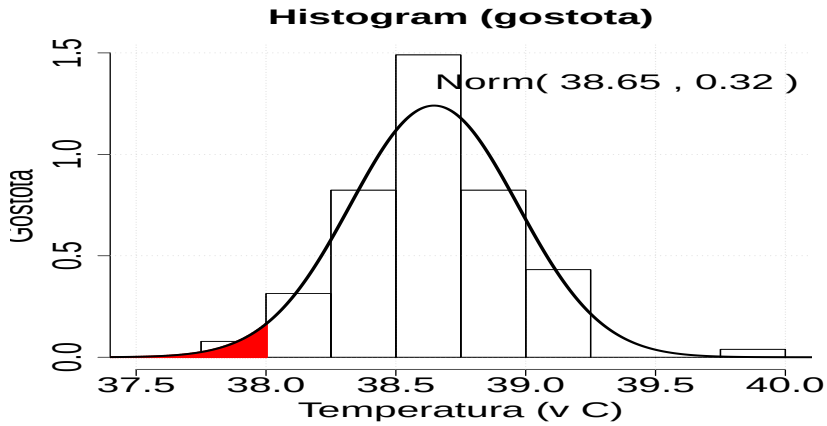
Število psov s temperaturo < 38 °C= 2 od 102.
Delež= 0.02

...$< 38°C$**?**
Površina histograma=1



Površina= 0.078*0.25 = 0.02

...< 38 °**C?**
Temperatura (v °C) $\sim$ Norm(μ=38.65, σ=0.32)

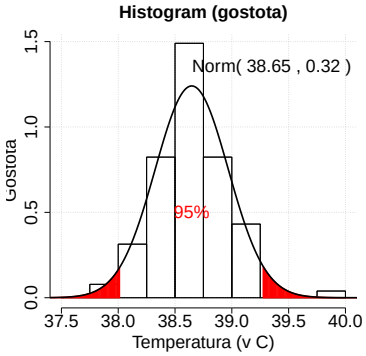


Površina pod krivuljo:
P(Temp< 38)=0.02.

6.4

Kateri je srednji interval, ki vključuje 95% zdravih psov?

Referenčni razpon za normalno temperaturo psov (*reference range*)



Interval od 38 do 39.3.

- Kako smo ga odčitali? Poiskali smo 2.5% in 97.5% percentil gaussove porazdelitve N(μ=38.65, σ =0.32).
- Pomagamo si z računalnikom ali s statitističnimi tabelami.
- Lahko bi ga dobili tudi s formulo: μ*-1.96*σ *do* μ*+1.96*σ *1.96=*$z_{0.975}$
- Ta postopek je veljaven samo, če je spremenljivka *normalno porazdeljena v populaciji*.

Referenčni razpon

Popravek ocene

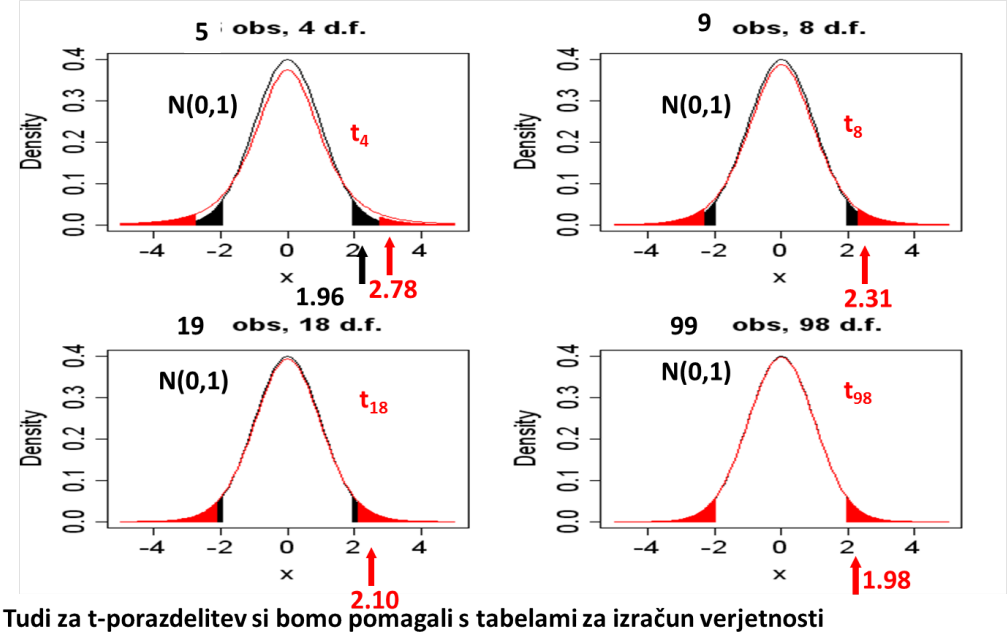
- Uporabili smo formulo: μ*-1.96*σ *do* μ*+1.96*σ
- Problem: predpostavili smo, da sta povprečije in standardni odklon s populacije znani Ni res! Ocenili smo jih z vzorca.
- Formula, ki jo dejansko uporabimo je $\bar{x} \pm z_{0.975} s$, oziroma $\bar{x} + / - z_{0.975} s$.
- Ocenili bi bolj natančno referenčni razpon s formulo

$$\bar{x} \pm t_{0.975,n-1} s$$

- t porazdelitev je zelo podobna standardni normalni porazdelitvi $N(\mu = 0, \sigma = 1)$, ko je vzorec velik.
- $t_{0.975,102-1} = 1.98$.
- $t_{0.975,10-1} = 2.26$.

Standardna normalna porazdelitev in t porazdelitev

Z in t... Ali sta si podobni?



Tudi za t-porazdelitev si bomo pomagali s tabelami za izračun verjetnosti

6.7

Interval zaupanja za povprečje

Povrpečna temperatura zdravih psov

- Na vzorcu so ocenili, da je povprečna temperatura zdravih psov $\bar{x}$ =38.65.
- Koliko natančno smo jo ocenili?
- Določimo *95% interval zaupanja za populacijsko povprečije* μ (95% IZ, ali 95% CI za μ).
- Interval je: 38.58 do 38.71.
- Kako ga interpretiramo?
- Alj je natančen?
- Ali bi pričakovali ožiji ali širši interval, če bi bil vzorec večji?

6.8

Interval zaupanja za povprečje

Kako smo ga izračunali?

- Na vzorcu so ocenili, da je povprečna temperatura zdravih psov $\bar{x}$ =38.65 in da je 95% IZ za populacijsko povprečije: 38.58 do 38.71.
- Forumla:

$$\bar{x} \pm t_{0.975,n-1} \frac{s}{\sqrt{n}}$$

- V čem se formula razlikuje od formule za referenčni razpon?
- $\frac{s}{\sqrt{n}}$ je *standardna napaka* (*standard error, SE*)

6.9

Standardna napaka povprečija

Kaj je in kako jo ocenimo?

- $\hat{SE} = \frac{s}{\sqrt{n}}$.
- Standardna napaka povprečja je standardni odklon vzorčnega povprečja.
- Standardni odkolon za X je $\sigma \rightarrow$ Standardni odkolon za $\bar{X}$ je $\sigma/\sqrt{n}$.
- Standardni odklon je mera variabilnosti neke spremenljivke.
- Standardna napaka je mera natančnosti ocene.

Simulacija - Centralni limitni izrek

Simulacija

- $X \sim (\mu, \sigma)$; $\bar{X} \sim (\mu, \sigma/\sqrt{n})$.
- $X \sim N(\mu, \sigma)$; $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.
- $X \sim (\mu, \sigma)$ z velikim vzorcem; $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$.
- $X \sim N(\mu, \sigma)$; $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- $X \sim N(\mu, \sigma)$; $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$.

6.10

test t za en vzorec

test t; n: velikost vzorca

- $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$
- 95% interval zaupanja za μ : $\bar{x} \pm t_{0.975,n-1} \cdot \hat{SE}; \hat{SE} = s/\sqrt{n}$.
- *Ničelna domneva*: H0: $\mu = \mu_0$ (dana vrednost).
- *Testna statistika*:
$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}},$$
izhaja iz t_{n-1}, če drži ničelna domneva.
- vrednost p: verjetnost, da opazimo testno statistiko, ki smo jo opazili na vzorcu ali vrednost, ki je bolj skrajna.
- *Predpostavke*: neodvisnot enot, gaussova porazdelitev spremenljivke.

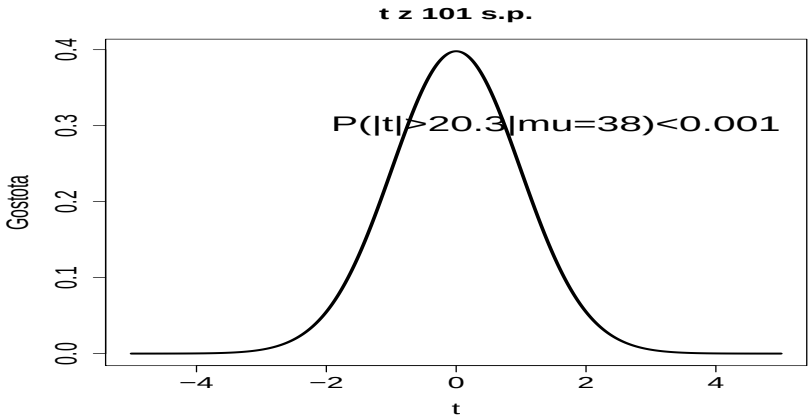
6.11

test t za en vzorec

Temperatura zdravih psov

- n=102; $\bar{x}$=38.65 °C. ; s=0.32 °C.; $\hat{SE}$=0.03 ; $t_{0.975,101} = 1.98$
- 95% interval zaupanja: $38.65 \pm 1.98 \cdot 0.03$: 38.58 do 38.71 °C.
- Ničelna domneva: H0: $\mu = 38$ °C.
- Vrednost testne statistike

$$t = \frac{38.65 - 38}{0.03} = 20.32$$



- P<0.001. Zavrnemo ničelno domnevo. Povprečna temperatura zdravih psov ni 38 °C.

6.12

Analiza s pomočjo računalnika (program R)

```
  t.test(temp.C.normal, mu=38)


One Sample t-test

data:  temp.C.normal
t = 20.32, df = 101, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 38
95 percent confidence interval:
 38.58 38.71
sample estimates:
mean of x
    38.65
```

6.13

Poglavje 7

Sklepna statistika za številske spremenljivke - primerjava skupin

Namen poglavja je spoznati statistične metode za primerjavo ene številske spremenljivke med dvema skupinama. Ukvarjali se bomo s primeri, kjer so skupine neodvisne in kjer obstaja odvisnost. Na primer, ko merimo isto spremenljivko dvakrat pri istih enotah (temperaturo pred in po zdravljenju z antibiotiki).

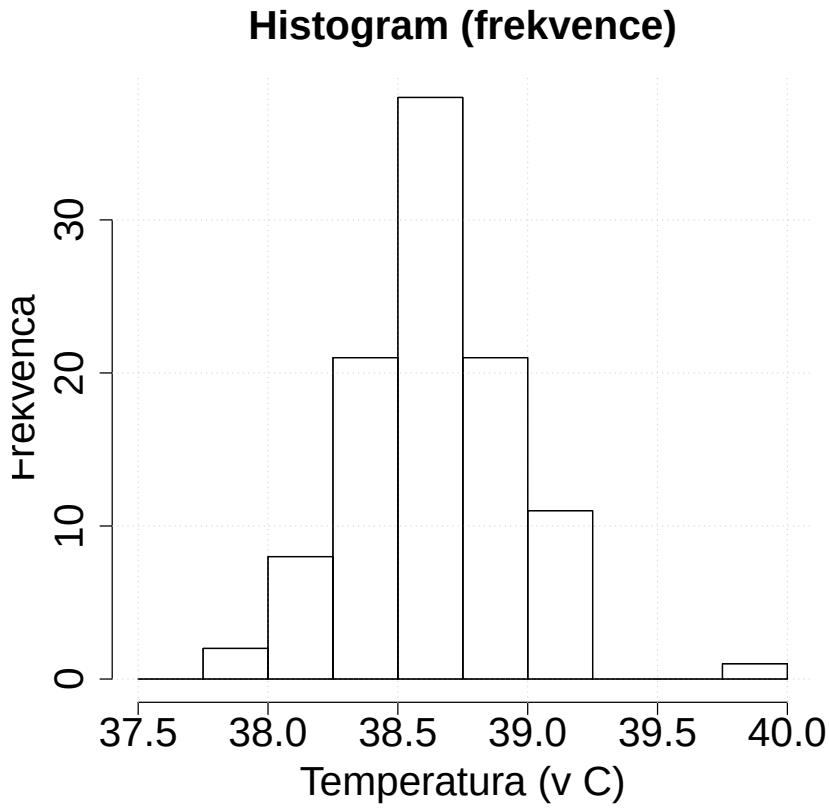
V primeru bomo primerjali povprečno temperaturo zdravih in bolnih psov. Uporabili bomo podatke iz vzorca in bomo preverili, ali je v populaciji temperatura različna med tema dvema skupinama in ocenili velikost razlike. Primerjali bomo tudi povprečno količino mačjega alergena (FelD1) v različnih delih telesa mačk.

V tem poglavju bomo spoznali:

- test t za dva neodvisna vzorca;
- test t za parne podatke;
- interval zaupanja za razliko povprečij.

Kolikšna je povprečna temperatura zdravih psov?

Raziskava



- Povprečna temperatura na vzorcu: $\bar{x}$ =38.6 °C.
- Standardni odklon na vzorcu: s=0.32 °C.

- 95% *referenčni razpon*: od 38 do 39.3 °C.
- 95% *interval zaupanja* za populacijsko povprečno temperaturo (μ): od 38.58 do 38.71 °C.
- Povprečna temperatura zdravih psov v populaciji ni 38 °C ($P < 0.001$, *test t za en vzorec*, H0: $\mu = 38$ °C)
- Predpostavke: neodvisnost psov, gaussova porazdelitev temperature v populaciji.

7.1

Primerjava z bolnimi psi

Raziskava

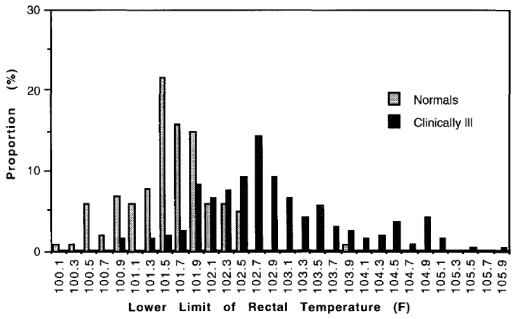


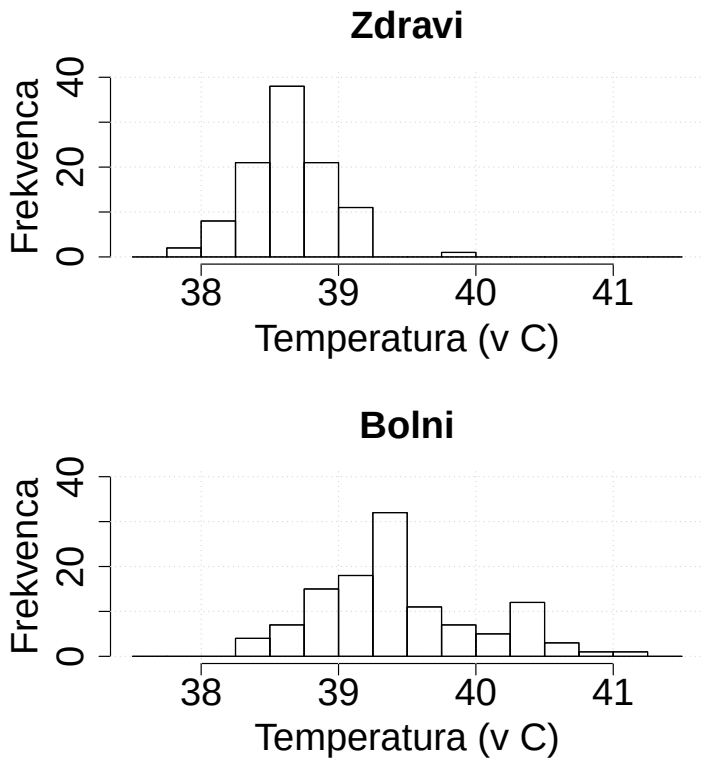
Figure 2.6 Frequency distribution of rectal temperatures for clinically normal and abnormal dogs.

- Klinično bolni psi: imajo znake okužbe dihal ali prebavil.
- *Kolikšna je povprečna razlika* temperature med zdravimi in bolnimi?
- *Ali je razlika* dovolj velika, da lahko sklepamo, da je *prisotna tudi v populaciji*?

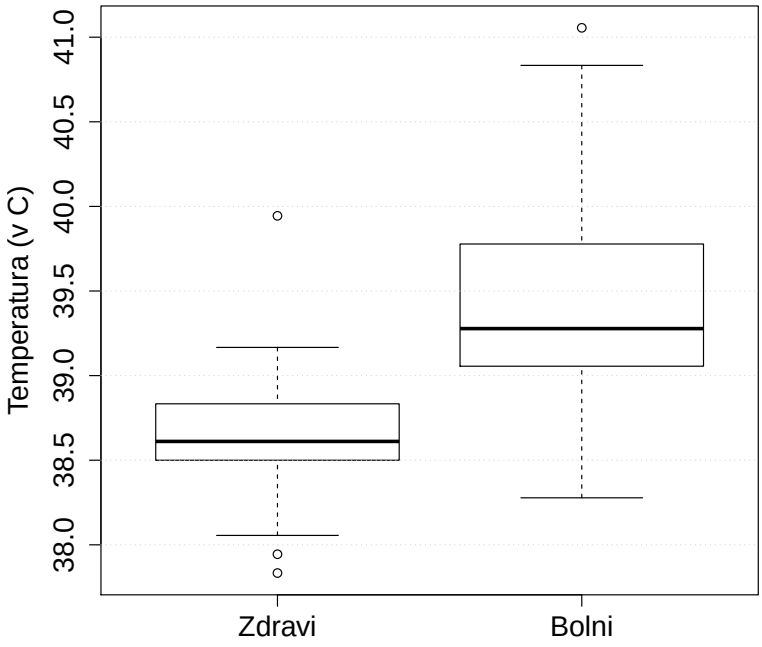
7.2

Grafični prikaz

Histogram



Boxplot



	Zdravi	Bolni
n	102	116
$\bar{x}$	38.6	39.4
s	0.3	0.6
Mediana	38.6	39.3
Največja vrednost	39.9	41.1

7.3

Ocena razlike povprečiji

Razlika v povprečni temeperaturi

- Razlika na vzorcu: $\bar{x}_{bolni} - \bar{x}_{zdravi}$ = 39.4 - 38.6 = 0.8 °C.
- 95% interval zaupanja za razliko populacijskih povprečij ($\mu_{bolni} - \mu_{zdravi}$): 0.67 do 0.93 °C.
- Interpretacija?
- Kako smo ga izračunali?

$$\bar{x}_{bolni} - \bar{x}_{zdravi} \pm t_{0.975,n-2} \cdot \hat{SE}$$

- Za povprečje ene populacije smo uporabili formulo:

$$\bar{x}_{zdravi} \pm t_{0.975,n-1} \cdot \hat{SE}; \hat{SE} = \frac{s}{\sqrt{n}}$$

- Formula za standardno napako za ralizko dveh povprečij je bolj zakomplicirana...

7.4

Standardna napaka za razliko povprečij

Formule

1. Skupna varianca

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

2. Skupni standardni odklon $s_p = \sqrt{s_p^2}$
3. Standardna napaka:

$$\hat{SE} = s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

	Zdravi	Bolni
n	102	116
$\bar{x}$	38.6	39.4
s	0.3	0.6
s^2	0.1	0.3

Izračun

1. Skupna varianca

$$s_p^2 = \frac{101 * 0.1 + 115 * 0.35}{102 + 116 - 2} = 0.23$$

2. Skupni standardni odklon

$$s_p = \sqrt{0.23} = 0.48$$

3. Standardna napaka

$$\hat{SE} = 0.48 \sqrt{\left(\frac{1}{102} + \frac{1}{116}\right)} = 0.07$$

7.5

Interval zaupanja za razliko povprečij

Formule

1. 95% interval zaupanja za $\mu_{bolni} - \mu_{zdravi}$

$$\bar{x}_{bolni} - \bar{x}_{zdravi} \pm t_{0.975,n-2} \cdot \hat{SE}$$

2. 99% interval zaupanja za $\mu_{bolni} - \mu_{zdravi}$

$$\bar{x}_{bolni} - \bar{x}_{zdravi} \pm t_{0.995,n-2} \cdot \hat{SE}$$

	Vrednost
$\bar{x}_{bolni} - \bar{x}_{zdravi}$	39.4-38.6=0.8 °C
$\hat{SE}$	0.07
n	102+116=218
$t_{0.975,218-2}$	1.97
$t_{0.995,218-2}$	2.6

Izračun

1. 95% interval zaupanja za $\mu_{bolni} - \mu_{zdravi}$

$$(39.4 - 38.6) \pm 0.07 \cdot 1.97$$

od 0.67 do 0.93.

2. 99% interval zaupanja za $\mu_{bolni} - \mu_{zdravi}$

$$(39.4 - 38.6) \pm 0.07 \cdot 2.6$$

od 0.63 do 0.97.

7.6

test t za neodvisna vzorca

test t za neodvisna vzorca; n: skupna velikost vzorca $(n_1 + n_2)$

- $\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{SE}} \sim t_{n-2}$
- *Ničelna domneva*: H0: $\mu_1 = \mu_2$
- *Testna statistika*:

$$t = \frac{\bar{x} - 0}{\hat{SE}},$$

  izhaja iz t_{n-2}, če drži ničelna domneva.
- vrednost p: verjetnost, da opazimo testno statistiko, ki smo jo opazili na vzorcu ali vrednost, ki je bolj skrajna.

7.7

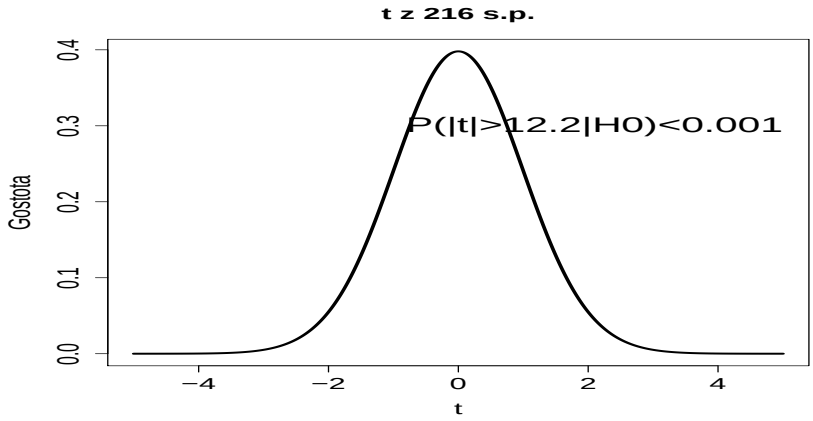
test t za neodvisna vzorca

Temperatura zdravih in bolnih psov

- Ničelna domneva: H0: $\mu_{zdravi} = \mu_{bolni}$. V populaciji je povprečna temperatura zdravih in bolnih psov enaka.
- Testna statistika

$$t = \frac{39.4 - 38.6}{0.07} = 12.22$$

- Porazdelitev testne statistike, če drži ničelna domneva: t_{218-2}.



-
- $P < 0.001$. Zavrnemo ničelno domnevo. Povprečna temperatura zdravih in bolnih psov ni enaka.

7.8

Analiza s pomočjo računalnika (program R)

```
  t.test(temp.C.ill, temp.C.normal, var.eq=TRUE)


Two Sample t-test

data:  temp.C.ill and temp.C.normal
t = 12.22, df = 216, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6712 0.9293
sample estimates:
mean of x mean of y
    39.45     38.65
```

7.9

test t za neodvisna vzorca: Predpostavke

Predpostavke

- Enote so neodvisne. Enote=psi
- Spremenljivka je v populaciji normalno porazdeljena. Spremenljivka=temperatura
- Varianci obeh skupin sta v populaciji enaki.$\sigma^2_{zdravi} = \sigma^2_{bolni}$

7.10

Analiza s pomočjo računalnika (program R) - brez predpostavke o enakosti varianc

```
  t.test(temp.C.ill, temp.C.normal, var.eq=FALSE)


Welch Two Sample t-test

data:  temp.C.ill and temp.C.normal
t = 12.66, df = 182.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6755 0.9250
sample estimates:
mean of x mean of y
    39.45     38.65
```

V čem so razlike? Stopinje prostosti ocenimo na podlagi podatkov.

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

7.11

test t za dva vzorca: Kršitev predpostavk

Kako bi analizirali...?

- Razliko v povprečni temperaturi zdravih in bolnih psov, če istega psa merite v obeh *stanjih*?
- Katero predpostavko testa t za dva neodvisna vzorca bi kršili?

Kako lahko analiziramo take podatke?

- Za vsako *enoto* računamo razliko Pes 1: $razlika_1 = temperatura_{1,bolan} - temperatura_{1,zdrav}$ Pes 2: $razlika_2 = temperatura_{2,bolan} - temperatura_{2,zdrav}$ ...
- Ocenimo vzorčno povrečno razliko temperature in 95% interval zaupanja za populacijsko povprečno razliko (δ).
- S testom t za en vzorec lahko sklepamo, ali je populacijska povprečna razlika različna od neke dane številke Tipično nas zanima, ali sta povprečij različni (H0: $\delta = 0$)

7.12

Ali imajo mačke več Fel d1 alergena na hrbtu ali na vratu?

Kako bi izvedli raziskavo?

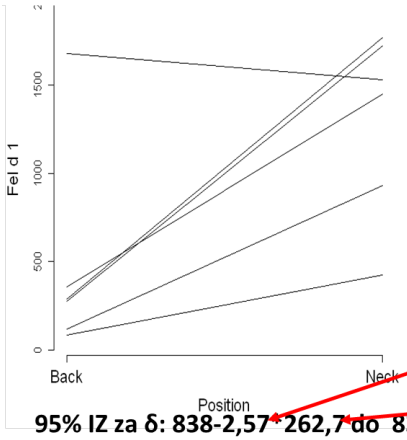
Parni t-test (t-test za dva odvisna vzorca)

TABLE II. Distribution of allergen on cats

Cat	Chest*	Stomach†	Neck	Back†	Back haunches*
1	1042	703	1720	275	540
2	653	1480	1450	356	355
3	487	281	1530	1680	376
4	82	40	426	83	114
5	82	136	931	118	152
6	298	347	1770	288	366
Mean	441	498	1300	464	318

Hair samples were obtained by using a grooming glove and assayed for Fel d 1 content by ELISA. Values are expressed as micrograms of Fel d 1 per gram of hair. Values from the neck region were compared with results for other anatomic sites by using Student's paired *t* test.
*$p < 0.002$.
†$p < 0.05$.

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$



	Razlika d (Neck-Back)
Cat1	1445
Cat2	1094
Cat3	-150
Cat4	343
Cat5	813
Cat6	1482
Povprečje	838
SD	643,5
SE=SD/√n	262,7

$$\frac{\overline{D} - \delta}{\frac{s_d}{\sqrt{n}}} \sim t_{n-1}$$

H0: δ=0

Testna statistika

t=838/262,7=3,19, df=5

p= 0,024

$t_{5,\ 1-0,05/2}$=2,57

95% IZ za δ: 838-2,57*262,7 do 838+2,57*262,7 =162,8 do 1513,1 Fel d 1 za gram dlake

7.13

t-test za en vzorec

t-test; n: velikost vzorca

- $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$
- 95% interval zaupanja za μ : $\bar{x} \pm t_{0.975,n-1} \cdot \hat{SE}; \hat{SE} = s/\sqrt{n}$.
- *Ničelna domneva*: H0: $\mu = \mu_0$ (dana vrednost).
- *Testna statistika*:

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

, izhaja iz t_{n-1}, če drži ničelna domneva.

- P vrednost: verjetnost, da opazimo testno statistiko, ki smo jo opazili na vzorcu ali vrednost, ki je bolj skrajna.

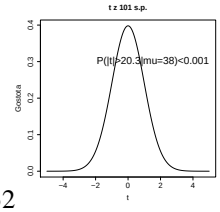
7.14

t-test za en vzorec

Temperatura zdravih psov

- n=102; $\bar{x}$=38.65 °C. ; s=0.32 °C.; $\hat{SE}$=0.03 ; $t_{0.975,101} = 1.98$
- 95% interval zaupanja: $38.65 \pm 1.98 \cdot 0.03$: 38.58 do 38.71 °C.
- Ničelna domneva: H0: $\mu = 38$ °C.
- Vrednost testne statistike

$$t = \frac{38.65 - 38}{0.03} = 20.32$$



- P<0.001. Zavrnemo ničelno domnevo. Povprečna temperatura zdravih psov ni 38 °C.

7.15

Analiza s pomočjo računalnika (program R)

```
  t.test(temp.C.normal, mu=38)


One Sample t-test

data:  temp.C.normal
t = 20.32, df = 101, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 38
95 percent confidence interval:
 38.58 38.71
sample estimates:
mean of x
    38.65
```

7.16

Kazalo