Naloge iz biostatistike

Predmet: Temelji biostatistike

Pripravil
Lara Lusa

Junij 2014 - izdaja 2.2

Kazalo

Poglavje 1

Opisna statistika

1. Izvedli smo 5 meritev in dobili naslednje vrednosti: 4, 2, 3, 4 in 1. Izračunajte:

   (a) artimetično povprečje;

   (b) mediano;

   (c) modus;

   (d) varianco;

   (e) standardni odklon;

   (f) interkvartilni razmik;

   (g) razpon.

   Katere od naštetih so **mere razpršenosti** in katere so **mere središčnosti** (centralne tendence)?

   Rešitve
   Označimo z n velikost vzorca ($n = 5$) in z x_i i-to meritev ($x_1 = 4, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 1$).

   (a) artimetično **povprečje** je $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+4+4}{5} = 2.8$.

   (b) **mediana** je srednja vrednost: polovica vrednosti je manjših in polovica vrednosti je večjih od mediane; določamo jo tako, da uredimo podatke in izberemo srednjo vrednost. Urejeni podatki so: 1, 2, 3, 4, 4, torej mediana je Me=3.

   (c) **modus** je najpogostejša vrednost, Mo=4;

   (d) **varianca** je povprečna kvadrirana razlika podatkov iz aritmetičnega povprečja: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{(4-2.8)^2+(2-2.8)^2+(3-2.8)^2+(4-2.8)^2+(1-2.8)^2}{5-1} = 1.7$.

   (e) **standardni odklon** je koren variance $s = \sqrt{s^2} = \sqrt{1.7}$=1.3;

   (f) **interkvartilni razmik** je interval od 1. kvartila do 3. kvartila. 1. kvartil (ali 25. percentil) je mediana prve polovice podatkov (2) in nam pove katera je vrednost za katero velja, da je 25% podatkov manjših in 75% večjih. 3. kvartil (ali 75. percentil) je mediana druge polovice podatkov (4), torej je intervalni razmik interval od 2 do 4. Interkvartilni razmik nam pove, kje se nahaja srednjih 50% vrednosti.

   (g) **razpon** je interval od najmanjše do največje vrednosti: od 1 do 4.

Aritmetično povprečje, mediana in modus so mere središčnosti (oz. mere centralne tendence); varianca, standardni odklon, interkvartilni razmik in razpon so mere razpršenosti. Mere središčnosti opišejo *sredino* podatkov, medtem ko mere razpršenosti opišejo razpršenost podatkov okoli sredine.

2. V spodnji tabeli so podani rezultati izpitov iz biostatistike v akademskem letu 2009/2010 na Veterinarski fakulteti. Za vsakega študenta smo zabeležili najboljši rezultat.

	Frekvenca
ni opravil/a	8
zd(6)	13
db(7)	22
pd(8)	16
pd(9)	10
odl(10)	8

Vprašanja

(a) Kaj so statistične enote v tem primeru? Koliko jih je?

(b) Katero spremenljivko smo merili?

(c) Katere vrste je spremenljivka?

(d) Grafično prikažite podatke.

(e) Izpolnite tabeli.

	Frekvenca	Relativna frekvenca	Kumulativna relativna frekvenca
ni opravil/a	8		
zd(6)	13		
db(7)	22		
pd(8)	16		
pd(9)	10		
odl(10)	8		

	Izid izpita	Katere vrste je mera?
Modus		
Mediana		
Artimetično povprečje		
Razpon		
Interkvartilni razmik		
Standardni odklon		

Odgovori

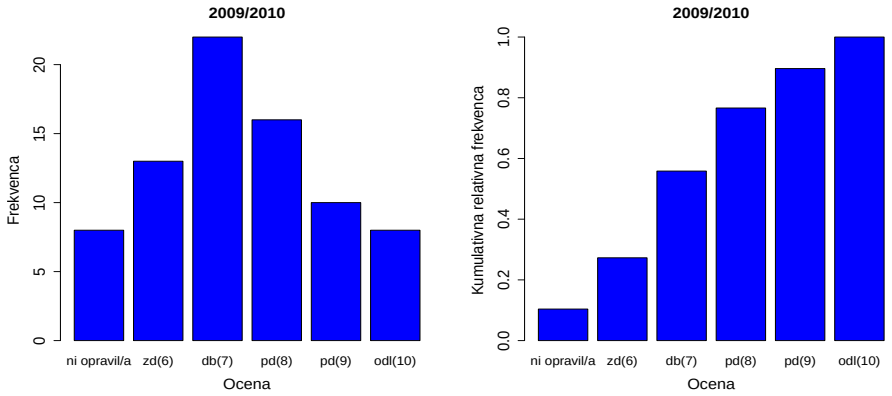
(a) **Statistične enote** so študentje. V vzorcu je 77 enot.

(b) Merili smo najboljši rezultat pri izpitu iz biostatistike.

(c) **Spremenljivka** je opisna urejenostna.

(d) Grafično lahko prikažemo podatke s stolpičnim diagramom, ki je prikazan na sliki 1.1 (levo).

(e)



Slika 1.1: *Grafična predstavitev podatkov.*

	Frekvenca	Relativna frekvenca	Kumulativna relativna frekvenca
ni opravil/a	8.00	0.10	0.10
zd(6)	13.00	0.17	0.27
db(7)	22.00	0.29	0.56
pd(8)	16.00	0.21	0.77
pd(9)	10.00	0.13	0.90
odl(10)	8.00	0.10	1.00

Razlaga

- Modus, mediana in aritmetično povprečje so mere središčnosti; razpon, interkvartilni razmik in standardni odklon so mere razpršenosti.
- Za opisne urejenostne spremenljivke ne moremo izračunati aritmetičnega povprečja in standardnega odklona.
- Modus je najpogostejša vrednost.
- Razpon je interval od najmanjše do največje vrednosti.
- Relativna frekvenca nekega rezultata je delež študentov, ki so dosegli ta rezultat. Izračunamo jo tako, da frekvenco rezultata delimo z velikostjo vzorca.
- Kumulativna relativna frekvenca nekega rezultata je delež študentov, ki so dosegli tak ali pa slabši rezultat. Izračunamo jo tako, da seštejemo relativne frekvence rezultatov, ki so enaki ali slabši od rezultata, ki nas zanima.
- Mediana je srednja vrednost podatkov (2. kvartil podatkov). Kvartile (in bolj nasplošno percentile) določamo s pomočjo kumulativnih relativnih frekvenc. Mediana je rezultat, za katerega velja, da je polovica študentov dosegla tak ali slabši rezultat, polovica študentov pa je dosegla tak ali boljši rezultat. Če pogledamo kumulativne relativne frekvence, vidimo, da je mediana danih podatkov rezultat db(7) (kumulativne relativne frekvence prvič presežejo vrednost 0.5 pri db(7)).
- Interkvartilni razmik je interval od 1. kvartila do 3. kvartila. Kvartile (in bolj nasplošno percentile) določimo s pomočjo kumulativnih relativnih frekvenc: 1.kvartil je zd(6) (kumulativne relativne frekvence prvič presežejo vrednost 0.25 za zd(6)) in 3. kvartil je pd(8) (tukaj gledamo vrednost 0.75 kumulativnih relativnih frekvenc).

	Izid izpita	Katere vrste je mera?
Modus	db(7)	Mera centralne tendence
Mediana	db(7)	Mera centralne tendence
Artimetično povprečje	Ne moremo izračunati	Mera centralne tendence
Razpon	ni opravil/a do odl(10)	Mera razpršenosti
Interkvartilni razmik	zd(6) do pd(8)	Mera razpršenosti
Standardni odklon	Ne moremo izračunati	Mera razpršenosti

3. Izpolnite naslednjo tabelo.

Spremenljivka	Vrsta spremenljivke	Mere središčnosti	Mere razpršenosti	Smiselen grafični prikaz
Spol				
Starost				
Ocena pri izpitu				
Vrsta raka				
Stadij raka				
Teža				
Višina				
Barva las				
Temperatura (v °C)				
Dohodek				
Inteligenčni kvocient				
Poštna številka				
Histologija tumorja				
Krvna skupina				
Stopnja izobrazbe				
Smer neba				
Znamka avtomobila				
Številka čevlja				
Število prijateljev				
Krvni pritisk (v mm Hg)				
Letnik rojstva				
Kajenje (Da ali ne)				
Število pokajenih cigaret				
pH vrednost				

Namig: 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 **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 nalogi 1, 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 stan-

dardni odklon, za asimetrično porazdeljene spremenljivke mediano in interkvartilni razmik, ker sta manj občutljivi na skrajne vrednosti. Ti dve meri se lahko izračunata tudi za opisne urejenostne spremenljivke. Za opisne imenske spremenljivke lahko (od mer iz naloge 1) izračunamo le modus.

Grafični prikazi, ki jih lahko uporabimo, za predstavitev podatkov so: **histogram** (primeren za številske spremenljivke, primer je na sliki 1.7), **graf škatla z brki** ali **okvir z ročaji** (primeren za številske spremenljivke, primer je na sliki 1.2), **stolpični diagram** (primeren za imenske spremenljivke, primer je na sliki 1.1).

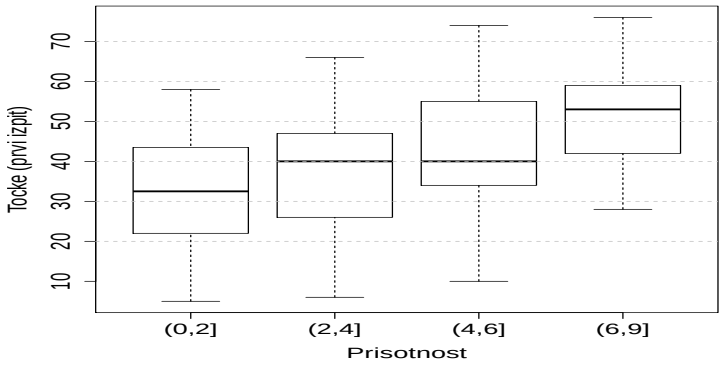
4. Poglejte grafikon.
Vprašanja

   (a) Kako se imenuje ta grafikon?

   (b) Kaj predstavlja črta v sredini škatle?

   (c) Kaj predstavljata spodnji in zgornji rob škatle?

   (d) Približno ocenite vrednosti, navedene v tabeli, za skupino, ki vsebuje študente, ki so obiskovali predavanja 5 ali 6 krat.



Slika 1.2: *Prisotnost in točke na izpitu.*

	<25%	25–49%	50–74%	≥75%
% študentov, ki niso opravili izpita (manj kot 40 točk)				
% študentov, ki niso dosegli več kot 50 točk				
% študentov, ki niso dosegli 30 točk				

   (e) Za skupini študentov, ki so obiskali predavanja največ dvakrat, oziroma 5 ali 6 krat, približno ocenite vrednosti, navedene v tabeli.

Rešitve in razlaga

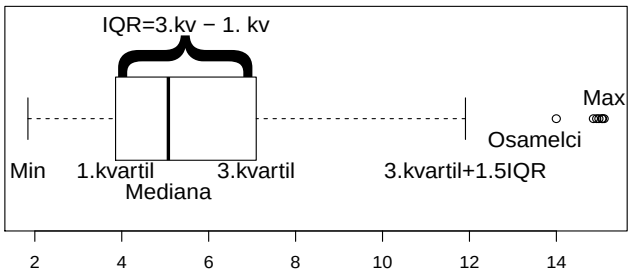
   (a) Grafikon se imenuje škatla z brki ali okvir z ročaji (*boxplot: box and whiskers plot*) (slika 1.3).

   (b) Odebeljena črta v sredini škatle predstavlja mediano.

	Največ dvakrat	5 ali 6 krat
Najmanjše število točk		
Največje število točk		
Povprečno število točk		
Mediansko število točk		
Standardni odklon		
Velikost vzorca		
Najpogostejši rezultat		
Razpon		
Interkvartilni razmik		

(c) Spodnji rob škatle predstavlja 1. kvartil (25. percentil) in zgornji rob predstavlja 3. kvartil (75. percentil). Višina škatle je torej interkvartilni razmik (IQR, *interquartile range*). Zgornja in spodnja črtica pri ročaju sta na sliki 1.2 najmanjša in največja vrednost.

V primeru, da je kakšna vrednost v vzorcu oddaljena od posameznega roba škatle za več kot 1.5 interkvartilnega razmika (slika 1.3), je vsaka taka vrednost prikazana s posamezno točko, črtica pri ročaju pa predstavlja to mejno vrednost (3. kvartil + $1.5 \times IQR$ za zgornjo črtico v primeru, da so *skrajne vrednosti* med največjimi, ali 1. kvartil - $1.5 \times IQR$ za spodnjo črtico v primeru, da so *skrajne vrednosti* med najmanjšimi). Pogosto se *skrajne vrednosti* imenujemo osamelci (*outliers*).



Slika 1.3: *Okvir z ročaji - razlaga.*

5. Stupica in soavtorji (Vector-Borne and Zoonotic Diseases, 2010) so merili 5 laboratorijskih spremenljivk pri pacientih z eritema eigrans (EM) (število levkocitov (WBC), število trombocitov (Platelets) in jetrne encime (Bilirubin, AST, ALT)).

Dobili so naslednje opisne statistike.

```
         Plt.0 WBC.0 ALT.0 AST.0 Bil.0
Min.    123.00  2.70  0.14  0.25  4.00
1st Qu. 216.80  5.08  0.28  0.34  7.95
Median  253.00  6.00  0.35  0.40 10.30
Mean    257.10  6.42  0.43  0.46 11.11
3rd Qu. 287.20  7.45  0.52  0.47 13.32
Max.    622.00 13.70  1.36  2.49 29.70
SD       66.95  2.00  0.22  0.27  4.87
```

(a) Kaj so vrednosti v tabeli?

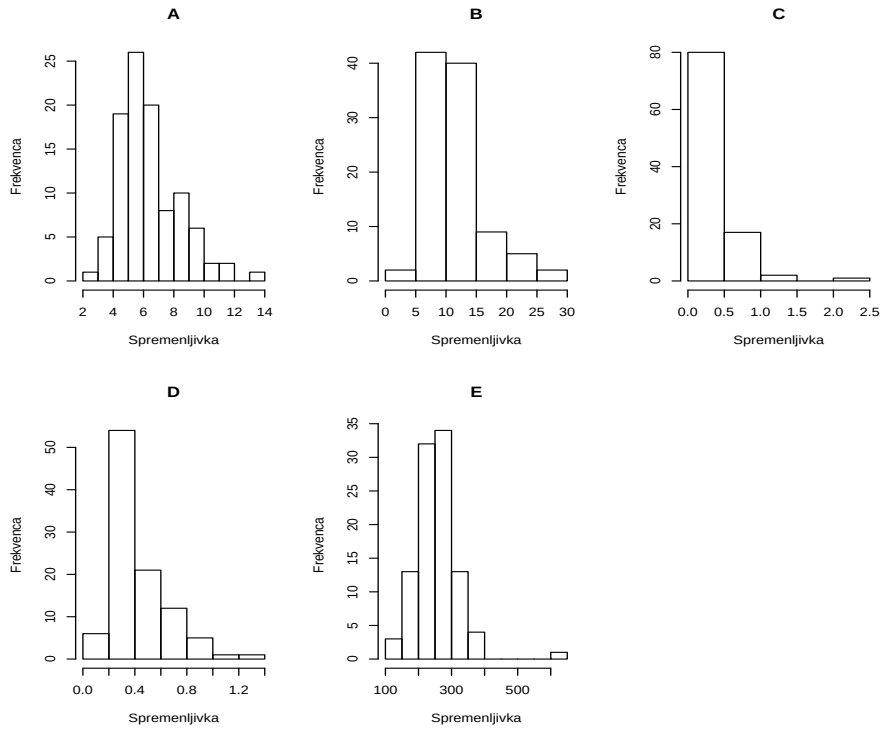
(b) Kakšno porazdelitev pričakujete za vsako spremenljivko? Prikažite grafično porazdelitve z okvirom z ročaji.

(c) Določite, kateri od spodnjih grafikonov prikazuje porazdelitev zgoraj navedenih spremenljvk (slika 1.4).

(d) Kaj predstavlja višina stolpcev histograma?

(e) Primerjajte grafični prikaz spremenljivke ALT na sliki 1.5. Kateri grafični prikaz se vam zdi bolj učinkovit? Katere informacije nam daje vsak graf?

(f) Primerjajte grafični prikaz spremenljivke ALT na sliki 1.6, kjer je porazdelitev prikazana posebej za ženke in za moške. Kateri grafični prikaz se vam zdi bolj učinkovit za neposredno primerjavo spremenljivke ALT glede spola?



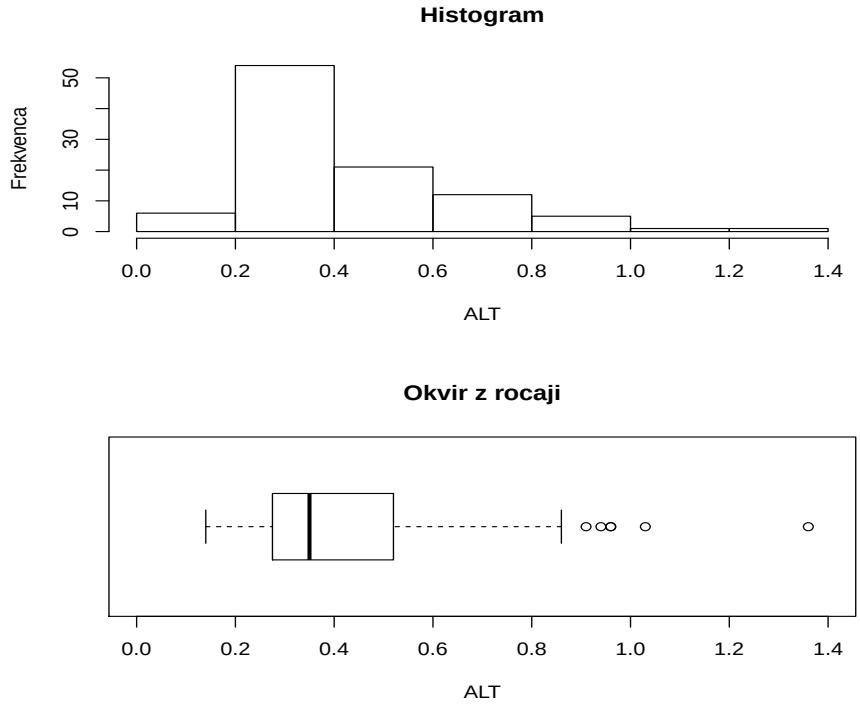
Slika 1.4: *Grafična predstavitev podatkov - histogram.*

6. V tabeli je podano število doseženih točk na izpitu iz biostatistike v akademskem letu 2009/2010 na Veterinarski fakluteti. Za vsakega študenta smo zabeležili samo rezultat prvega poizkusa. Grafično prikažite rezultate in na podlagi grafikona približno ocenite aritmetično sredino in standardni odklon za število točk doseženih na izpitu.

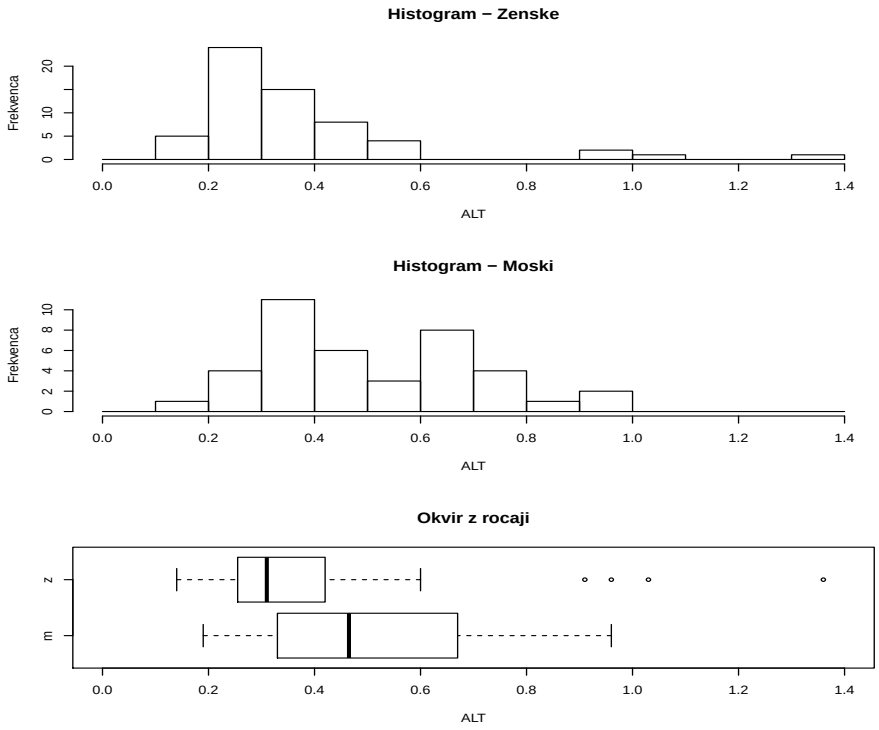
7. Zabeležili smo število doseženih točk na izpitu iz biostatistike, kjer je študent pozitivno opravil izpit (oziroma je dosegel vsaj 40 od 80 možnih točk) ter število dodatnih točk iz domačih nalog (od 0 do 20 točk).

   (a) Povprečno število točk je bilo 52.5, standardni odklon pa je bil 10.2 točk.

   (b) Povprečno število dodatnih točk iz domačih nalog je bilo 12.7, standardni odklon pa je bil 4.7 točk.



Slika 1.5: *Grafična predstavitev spremenljivke ALT - histogram in okvir z ročaji.*



Slika 1.6: *Grafična predstavitev spremenljivke ALT glede na spol- histogram in okvir z ročaji.*

Ali menite, da je na podlagi teh podatkov možno, da je število doseženih točk na pozitivnem izpitu, oziroma število dodatnih točk iz domačih nalog, normalno porazdeljeno? Utemeljite odgovor in skicirajte, kakšno porazdelitev

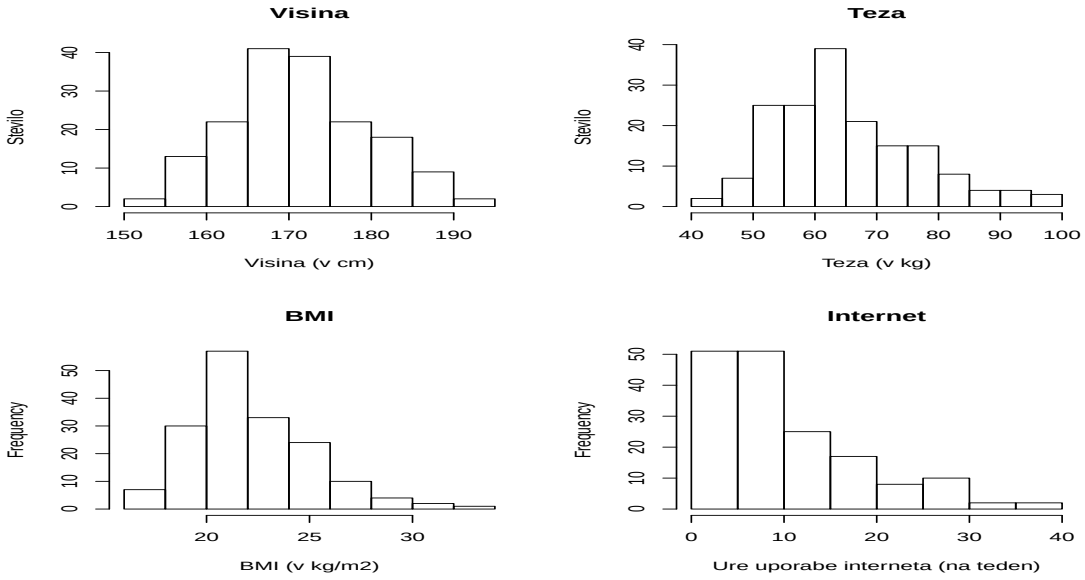
Razred	Frekvenca
$[0-10)$	6
$[10-20)$	3
$[20-30)$	9
$[30-40)$	18
$[40-50)$	14
$[50-60)$	13
$[60-70)$	6
$[70-80)$	4

bi pričakovali za vsako spremenljivko.

8. Izvedli smo pet meritev in dobili povprečje 30, mediano 25 in modus 20. Zapišite vrednosti posameznih meritev, ki se skladajo s temi opisnimi statistikami in skicirajte porazdelitveno funkcijo.

9. Izvedli smo 10 meritev in dobili naslednje podatke: 5, 7, 9, 9, 12, 13, 15, 20, 25, 30. Izračunajte eno (primerno) mero središčnosti in eno (primerno) mero razpršenosti, grafično predstavite podatke, navedite velikost vzorca in vrsto spremenljivke, ki smo jo merili.

10. Zbrali smo podatke o študentih, ki so se udeležili predmeta biostatistika v obdobju 2008–2010. Oglejte si grafikone in povejte, katere spremenljivke so porazdeljene simetrično in katere so porazdeljene asimetrično. Za vsako spremenljivko približno ocenite aritmetično sredino in standardni odklon, ter povejte, ali pričakujete, da je mediana večja ali manjša od aritmentične sredine.



Slika 1.7: *Porazdelitev štirih spremenljivk.*

Povejte, katere trditve so pravilne in katere so napačne.

(a) Teža je porazdeljena simetrično.

(b) Višina je približno normalno porazdeljena.

(c) Povprečna in medianska višina sta si zelo podobni.

(d) Povprečni BMI je manjši kot medianski BMI.

(e) Porazdelitev števila ur tedenske uporabe interneta je negativno asimetrična.

(f) Standardni odklon višine je približno 20.

(g) Povprečje števila ur tedenske uporabe interneta je večje od mediane.

(h) Mediana števila ur tedenske uporabe interneta je približno 20 ur.

(i) Interkvartilni razmik števila ur tedenske uporabe interneta je približno od 15 do 35 ur.

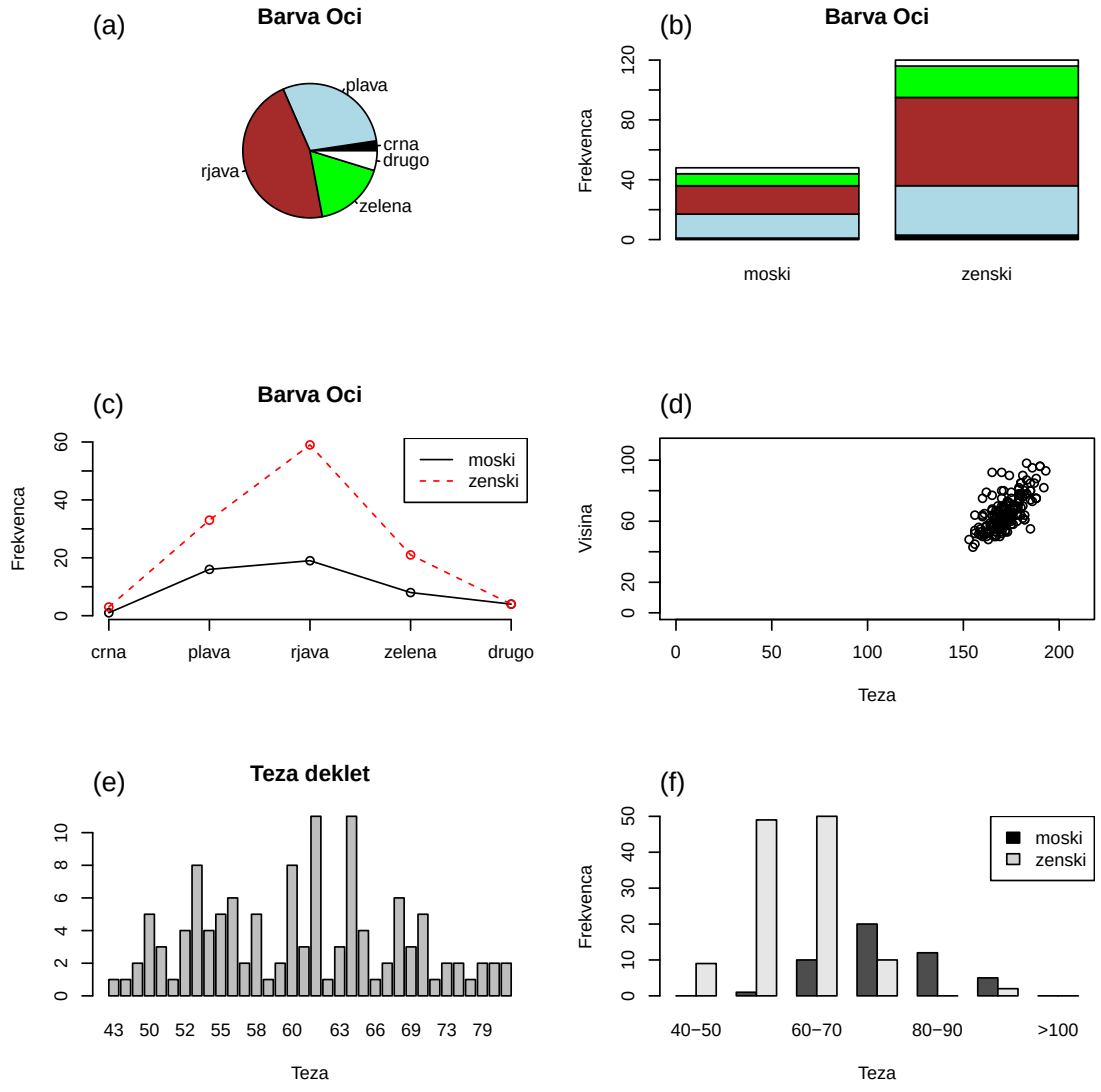
(j) Razpon višine je od 165 do 185 cm.

(k) Porazdelitev logaritemske transformacije števila ur tedenske uporabe interneta bi bila približno normalna.

(l) Približno 50 študentov je tehtalo med 50 in 60 kg.

(m) Manj kot 5 študentov je tehtalo več kot 80 kg.

11. Kako bi lahko izboljšali grafično predstavitev podatkov? Za vsak grafikon povejte, kaj je njegova slabost in kako bi ga izboljšali.



Slika 1.8: *Grafični prikazi podatkov.*

Poglavje 2

Verjetnost

2.1 Pogojna verjetnost, produkt in vsota dogodkov

1. Delež deklet med vsemi študenti veterine je 0.70. Trideset odstotkov deklet igra videoigrice, medtem ko to počne kar 60 odstotkov fantov.
Vprašanja

   (a) Kolikšna je verjetnost, da je naključno izbrana oseba z Veterinarske fakultete ženskega spola?

   (b) Kolikšna je verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice in je ženskega spola?

   (c) Kolikšna je verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice?

   (d) Kolikšna je verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice ali pa je ženskega spola?

   Rešitve
   $P(\text{Spol=M}) = 0.30$, $P(\text{Spol=F}) = 0.70$
   $P(\text{Igrice=Da}|\text{Spol=M}) = 0.60, P(\text{Igrice=Da}|\text{Spol=F}) = 0.30$

   (a) Verjetnost, da je naključno izbrana oseba z Veterinarske fakultete ženskega spola, je P(Spol=F)=0.70.

   (b) Verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice in ($\cap$) je ženskega spola, je:
   $P(\text{Spol=F}\cap\text{Igrice=Da}) = P(\text{Spol=F})P(\text{Igrice=Da}|\text{Spol=F}) = 0.70{\cdot}0.30 = 0.21$.

   (c) Verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice, je:
   $P(\text{Igrice=Da}) = P(\text{Igrice=Da}|\text{Spol=M})P(\text{Spol=M}) +$
   $P(\text{Igrice=Da}|\text{Spol=F})P(\text{Spol=F}) = 0.60 \cdot 0.30 + 0.30 \cdot 0.70 = 0.39$

   (d) Verjetnost, da naključno izbrana oseba z Veterinarske fakultete igra videoigrice ali ($\cup$) je ženskega spola, je:
   $P(\text{Spol=F} \cup \text{Igrice=Da}) = P(\text{Spol=F}) + P(\text{Igrice=Da}) - P(\text{Igrice=Da} \cap \text{Spol=F}) = 0.70 + 0.39 - 0.21 = 0.88$.

   Za rešitev smo uporabili sledeče teoretične rezultate:

- $P(E \cap H) = P(E|H)P(H)$; velja, da $P(E \cap H) = P(E)P(H)$, samo če sta E in H neodvisni, oziroma $P(E|H) = P(E)$.
- $P(E \cup H) = P(E) + P(H) - P(E \cap H)$; velja, da $P(E \cup H) = P(E) + P(H)$, samo če se E in H med sabo izključujeta, sta nezdružljiva ($P(E \cap H) = 0$).
- če sta H_1 in H_2 nezdružljiva in izčrpna dogodka (skupaj predstavljata vse mogoče dogodke v poskusu, $P(H_1 \cup H_2) = 1$), velja $P(E) = P(E \cap H_1) + P(E \cap H_2) = P(E|H_1)P(H_1) + P(E|H_2)P(H_2)$

Alternativna rešitev
Ta problem si lahko ogledamo tudi na morda bolj intuitiven način. Predpostavimo, da zajema populacija študentov veterine 1000 študentov. Lahko izračunamo število študentov, ki bi jih pričakovali v vsaki celici kontingenčne tabele, če veljajo naše predpostavke.

	Igrice=Da	Igrice=Ne	Vsota
Spol=F	a	b	a+b
Spol=M	c	d	c+d
Vsota	a+c	b+d	N=1000

Med 1000 študenti veterine, bi pričakovali 700 deklet (a+b=$1000 \cdot 0.7$) in 300 fantov (c+d=$1000 \cdot 0.3$). Med dekleti bi pričakovali 210 tistih, ki igrajo (a=$700 \cdot 0.3$) in 490 deklet, ki ne igrajo videoigric (b=700-120, oziroma $700 \cdot 0.70$). Med fanti pa bi pričakovali 180 igralcev (c=$300 \cdot 0.60$) in 120 neigralcev (d=300-120, oziroma $300 \cdot 0.40$). Tako bi imeli skupno 390 (a+c) igralcev in 610 (b+d) neigralcev. Izpolnjena kontingenčna tabela je:

	Igrice=Da	Igrice=Ne	Vsota
Spol=F	210	490	700
Spol=M	180	120	300
Vsota	390	610	N=1000

Iz te tabele lahko pridobimo vse verjetnosti, ki so nas zanimale.

- $P(\text{Spol=F}) = \frac{a+b}{N} = \frac{700}{1000} = 0.7$.
- $P(\text{Spol=F} \cap \text{Igrice=Da}) = \frac{a}{N} = \frac{210}{1000} = 0.21$.
- $P(\text{Igrice=Da}) = \frac{a+c}{N} = \frac{390}{1000} = 0.39$.
- $P(\text{Spol=F} \cup \text{Igrice=Da}) = \frac{a+b+c}{N} = \frac{210+490+180}{1000} = \frac{880}{1000} = 0.88$. (V tem primeru moramo sešteti vse celice, kjer je Spol=F ali Igrice=Da).

2. Prevalenca (pogostost) sladkorne bolezni tipa 2 je 10% v 60. letu starosti, medtem ko je prevalenca ishemične bolezni srca izmed diabetikov tipa 2 v 60. letu starosti 20%. V splošni populaciji 60-letnikov je ta prevalenca 5%.

   (a) Kolikšna je verjetnost, da ima oseba v 60. letu starosti sladkorno bolezen tipa 2 in srčno ishemijo?

   (b) Kolikšna je verjetnost, da ima oseba v 60. letu starosti sladkorno bolezen tipa 2 ali srčno ishemijo?

Rešitev

(a) $P(sl \cap is) = P(sl)P(is|sl) = 0.10 \cdot 0.20 = 0.02$

(b) $P(sl \cup is) = P(sl) + P(is) - P(is \cap sl) = 0.10 + 0.05 - 0.02 = 0.13$

NB: Če ne bi poznali pogojne verjetnosti $P(is|sl)$, ampak samo verjetnost ishemije $P(is)$, ne bi mogli izračunati verjetnosti, ki so nas zanimale ($P(sl \cap is)$ in $P(sl \cup is)$). Morali bi predpostavljati neodvisnost ishemične bolezni srca in diabetesa; ta predpostavka pa v tem primeru ne bi bila realistična.

3. Imamo tri različna zdravila (A, B in C). Verjetnost ozdravitve pacienta, zdravljenega z zdravilom A, je 0.7, za pacienta, zdravljenega z zdravilom B, je 0.6 ter za pacienta, zdravljenega z zdravilom C (placebo), je 0.2. V skupino A smo vključili 1/3 pacientov. Vsak pacient dobi eno zdravilo. Kakšen delež pacientov smo vključili v skupino B, če je skupna verjetnost ozdravitve enaka 0.38?

4. O Marsovcih vemo:

Marsovci	Verjetnost
So zeleni	0.85
So modri	0.10
So rumeni	0.05

Zeleni Marsovci	Verjetnost
Imajo zelene lase	0.90
Imajo modre lase	0.08
Imajo rumene lase	0.02

Modri Marsovci	Verjetnost
Imajo zelene lase	0.15
Imajo modre lase	0.80
Imajo rumene lase	0.05

Rumeni Marsovci	Verjetnost
Imajo zelene lase	0.25
Imajo modre lase	0.10
Imajo rumene lase	0.65

(a) Kolikšna je verjetnost, da je nakjučno izbran Marsovec rumene barve?

(b) Kolikšna je verjetnost, da je nakjučno izbran Marsovec rumene barve in ima rumene lase?

(c) Kolikšna je verjetnost, da ima nakjučno izbran Marsovec rumene Lasje?

(d) Kolikšna je verjetnost, da je nakjučno izbran Marsovec rumene barve ali ima rumene lase?

5. ** Raziskovalci so merili količino mačjega alergena Fel D1 v petih delih telesa; v raziskavo so vljučili šest mačk. Da bi določili, ali je povprečna količina Fel D1 različna med deli telesa so izvedli t-test za parne podatke za vsak par telesnih delov (na primer, primerjali so povprečni Fel D1 glave in hrbta, glave in trebuha, itd). Odločili so se, da bodo rezultati statistično značilni, če je P-vrednost manjša od 0.05.

(a) Kolikšna je verjetnost, da bo vsaj en test značilen, če imajo vsi deli telesa isto povprečno količino Fel D1 in če predpostavljamo, da rezultat vsakega testa ni odvisen od rezultatov drugih testov?

(b) Kako bi se ta verjetnost spreminjala, če bi se odločili, da bo rezultat testa statistično značilen, če je P-vrednost manjša od 0.005?

Rešitve

(a) Izvedli so 10 statističnih testov ($\binom{5}{2} = 10$).
$P(\text{vsaj en test značilen}|H_0) = 1 - P(\text{vsi testi neznačilni (NZ)}|H_0)$
Označimo s $P(T_{NZ}|H_0)$ verjetnost, da statistični test ni značilen, ko velja ničelna domneva.
$1 - P(T1_{NZ} \cap T2_{NZ} \cap ... \cap T10_{NZ}|H_0) =$
$1 - P(T1_{NZ}|H_0)P(T2_{NZ}|H_0)...P(T10_{NZ}|H_0) = 1 - 0.95 \cdot 0.95 \cdot \ldots 0.95 =$
$1 - 0.95^{10} = 1 - 0.60 = 0.40$

(b) Če bi se odločili, da bo rezultat testa statistično značilen, če je P-vrednost manjša od 0.005, potem bi za vsak test veljalo $P(T_{NZ}|H_0) = 0.995$ in z istim postopkom kot prej bi dobili $P(\text{vsaj en test značilen}|H_0)$=0.0489.

6. Crawder in soavtorji (JAVMA, št.9, 2005, strani 1503 − 1507) so preučevali uporabnost testa, ki temelji na polimerazni verižni reakciji PCR, *polymerase chain reaction* za diagnostiko mačjega virusa FIV-a FIV, *feline immunodeficiency virus infection*. V raziskavo so vključili 41 mačk s FIV-om in 83 mačk brez FIV-a. Diagnostični test je pravilno diagnosticiral 31 mačk s FIV-om in 81 mačk brez FIV-a (v članku: rezultati za PCR1). Ocenjen odstotek mačk s FIV-om je 2% (podatek velja za ZDA, vir: http://www.animalhealthchannel.com, predpostavljamo, da velja tudi v Sloveniji).

Vprašanja

(a) Koliko mačk so napačno diagnosticirali? Koliko lažno pozitivnih in koliko lažno negativnih mačk je bilo v tej raziskavi?

(b) Kolikšna je verjetnost, da ima mačka pozitivni test, če ima FIV? Kako imenujemo to verjetnost?

(c) Kolikšna je verjetnost, da ima mačka negativni test, če nima FIV-a? Kako imenujemo to verjetnost?

(d) Odločili smo se, da PCR test deluje dobro in ga začnemo uporabljati rutinsko; testiramo vsako mačko, ki pride na kotrolo. Kolikšna je verjetnost, da bo imela mačka s pozitivnim testom zares FIV? Kako imenujemo to verjetnost?

(e) Kolikšna pa je verjetnost, da mačka z negativnim testom nima FIV-a? Kako imenujemo to verjetnost?

(f) Lahko se odločimo, da bomo uporabljali PRC test samo za mačke s kliničnimi znaki drugih bolezni, za katere je pogostos FIV-a 15-odstotno (vir: http://www.animalhealthchannel.com). Katere izmed verjetnosti, ki ste jih izračunali pri prejšnjih točkah, bi se spremenile in kako?

(g) Ali je smiselno rutinsko testirati vse mačke, ali je bolj smiselno osredotočiti se samo na tiste, ki imajo klinične znake prisotnosti drugih bolezni?

Rešitve
Tudi v tem primeru lahko predstavimo podatke s kontingenčno tabelo. Označimo s T+ in T- pozitivne in negativne izide testa, s FIV+ in FIV- prisotnost in odsotnost FIV-a.

	FIV+	FIV-	Vsota
T+	31	2	33
T-	10	81	91
Vsota	41	83	N=124

(a) Napačno so diagnosticirali 10+2=12 mačk. 2 sta bila lažno pozitivna rezultata, 10 pa lažno negativnih rezultatov.

(b) Verjetnost, da ima mačka pozitivni test, če ima FIV je občutjivost testa (Sens, *sensitivity*). $Sens = P(T+|FIV+) = 31/41 = 0.76$.

(c) Verjetnost, da ima mačka negativni test, če nima FIV-a, je specifičnost testa (Spec, *specificity*). $Spec = P(T-|FIV-) = 81/83 = 0.98$.

(d) Ker zgleda, da ima PCR test dobre lastnosti, ga rutinsko uporabljamo za diagnozo FIV-a. Pogostost FIV-a v naši populaciji je 2% ($p = 0.02$). Verjetnost, da bo imela mačka s pozitivnim testom zares FIV je pozitivna napovedna vrednost testa (PPV, *positive predictive value*), $PPV = P(FIV+|T+)$.
Moramo se zavedati, da ta verjetnost ni enaka občutljivosti testa (P(T+|FIV+)) in se nanaša na informacijo, ki je najpomembnejša za (lastnika) mačko(e). Test je bil pozitiven; ali to pomeni, da ima mačka zares FIV?
Pogostost mačk s FIV-om v populaciji je zelo različna od pogostosti FIV-a na vzorcu, ki so ga uporabljali za ovrednotenje PCR testa. V vzorcu je imelo FIV 41/124=33.1 odstotkov mačk. Zato ni smiselno določati vrednost PPV-ja iz tabele, kjer bi dobili, da je PPV=31/33.
Lahko si pomagamo s kontingenčno tabelo, ki bi jo pričakovali med tisoč mačkami naše populacije.

	FIV+	FIV-	Vsota
T+	a	b	a+b
T-	c	d	c+d
Vsota	a+c	b+d	N=1000

Pričakovali bi 20 (a+c=1000·0.02) mačk s FIV-om in 980 mačk brez FIV-a (b+d=1000-20); izmed mačk s FIV-om, pričakujemo, da jih bo 15.1 imelo pozitivni test (a=$20 \cdot Sens$) in 4.9 negativni test (c=$20 - 20 \cdot Sens$). Izmed mačk, ki nimajo FIV-a, pričakujemo, da jih bomo pravilno diagnosticirali 956.4 mačk (d=$980 \cdot Spec$) in napačno 23.6 mačk (b=$980 - 980 \cdot Spec$). Torej skupno pričakujemo 38.7 pozitivnih in 961.3 negativnih testov. Podatki so prikazani v tabeli.

	FIV+	FIV-	Vsota
T+	15.1	23.6	38.7
T-	4.9	956.4	961.3
Vsota	20	980	N=1000

Zdaj lahko izračunamo PPV iz tabele, in dobimo PPV=15.1/38.7 = 0.39.

(e) Verjetnost, da mačka z negativnim testom nima FIV-a je negativna napovedna vrednost (NPV, *negative predictive value*): $NPV = P(FIV-|T-)$. Lahko jo izračunamo iz tabele, podobno kot PPV. NPV=956.4/ 961.3 = 0.995.

(f) Če bi rutinsko uporabljali PCR test samo za mačke s kliničnimi znaki drugih bolezni, se občutljivost in specifičnost testa ne bi spreminjali, saj nista odvisni od pogostosti bolezni; pozitivna in negativna napovedna vrednost pa bi se spremenili.

	FIV+	FIV-	Vsota
T+	113.4	20.5	133.9
T-	36.6	829.5	866.1
Vsota	150	850	N=1000

Torej bi dobili: PPV=113.4/133.9= 0.847 in NPV=829.5/736.1=0.958. Za mačke s kliničnimi znaki drugih bolezni ima pozitivni test bistveno drugčen pomen kot za asimptomatične mačke.

(g) Na podlagi prejšnjih podatkov o PPV, izgleda, da rutinsko testiranje asimptomatičnih mačk ni smiselno, saj bi bilo več kot 60 odstotkov pozitivnih testov napačnih. PPV v splošni populaciji bi lahko povečali , če bi imel PCR test še večjo specifičnost.

Rešitve s formulami
Pomagamo si z Bayesovim izrekom:

$$PPV = P(FIV+|T+) = \frac{P(T+|FIV+)P(FIV+)}{P(T+)}$$

$P(T+) = P(T+|FIV+)P(FIV+) + P(T+|FIV-)P(FIV-) =$
$= P(T+|FIV+)P(FIV+) + (1 - P(T-|FIV-))(1 - P(FIV+)) =$
$= Sens \cdot p + (1 - Spec)(1 - p)$

$$PPV = P(FIV+|T+) = \frac{Sens \cdot p}{Sens \cdot p + (1 - Spec) \cdot (1 - p)}$$

$$NPV = P(FIV-|T-) = \frac{P(T-|FIV-)P(FIV-)}{P(T-)}$$

$$P(T-) = 1-P(T+) = 1-(Sens \cdot p+(1-Spec) \cdot (1-p)) = (1-Sens) \cdot p+Spec \cdot (1-p)$$

$$NPV = P(FIV-|T-) = \frac{Spec \cdot (1 - p)}{(1 - Sens) \cdot p + Spec \cdot (1 - p)}$$

Z uporabo enačb nam ni treba vsakič izračunati tabele in lahko določamo PPV in NPV, če poznamo občutljivost in specifičnost testa, ter pogostost bolezni v populaciji, za katero želimo uporabljati test.

2.2 Normalna (gaussova) porazdelitev

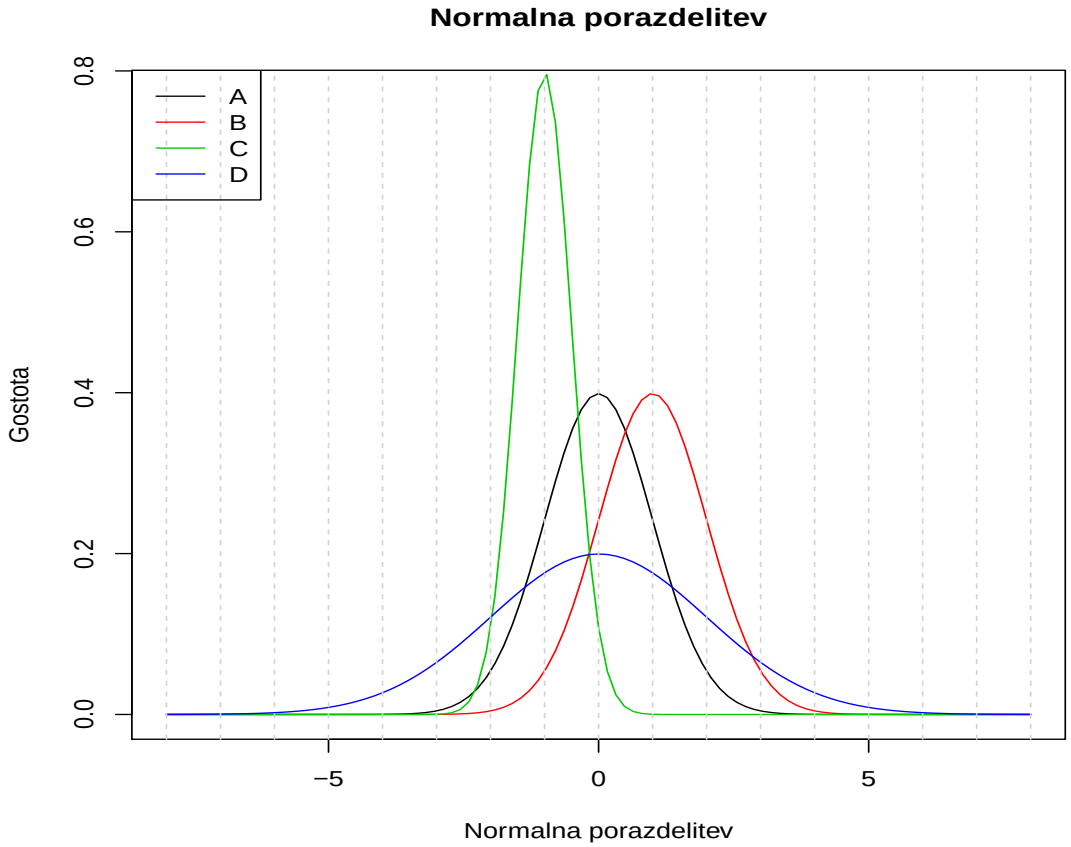
7. Na sliki 2.1 so prikazane štiri normalne porazdelitve. Za vsako približno ocenite povprečje in standardni odklon.

Izpolnite tabelo

Porazdelitev	Povprečje	Standardni odklon
A		
B		
C		
D		

Namig
Pomagajte si z interaktivnimi statističnimi tabelami (`http://ibmi.mf.uni-lj.si/sl/centri/biostatisticni-center/interaktivno/statisticne-tabele`, izberite porazdelitev Distribution type: Normal in spremenite povprečje (*mean*) in standardni odklon (*standard deviation*)).

Povrečje normalne porazdelitve (μ) je v sredini porazdelitve (je enako mediani in modusu). Standardni odklon (σ) približno ocenimo tako, da določimo, v katerem simetričnem intervalu okrog povprečja je približno 95% opazovanj. Spodnji in zgornji limit tega intervala sta približno: $\mu - 2\sigma$ in $\mu + 2\sigma$ (bolj



Slika 2.1: *Normalne porazdelitve.*

natančno: morali bi množiti σ z 1.96).

Na podlagi ocenjenih povrečij in standardnih odklonov, za vsako spremenljivko ocenite, kolikšna je verjetnost, da ima naključno izbrana enota pozitivno vrednost (P(A>0), ..., P(D>0)).

8. Holesterol (v mg/dl) je normalno porazdeljen, ima povprečje 210 in standardni odklon 20.

   (a) Narišite porazdelitev holesterola.

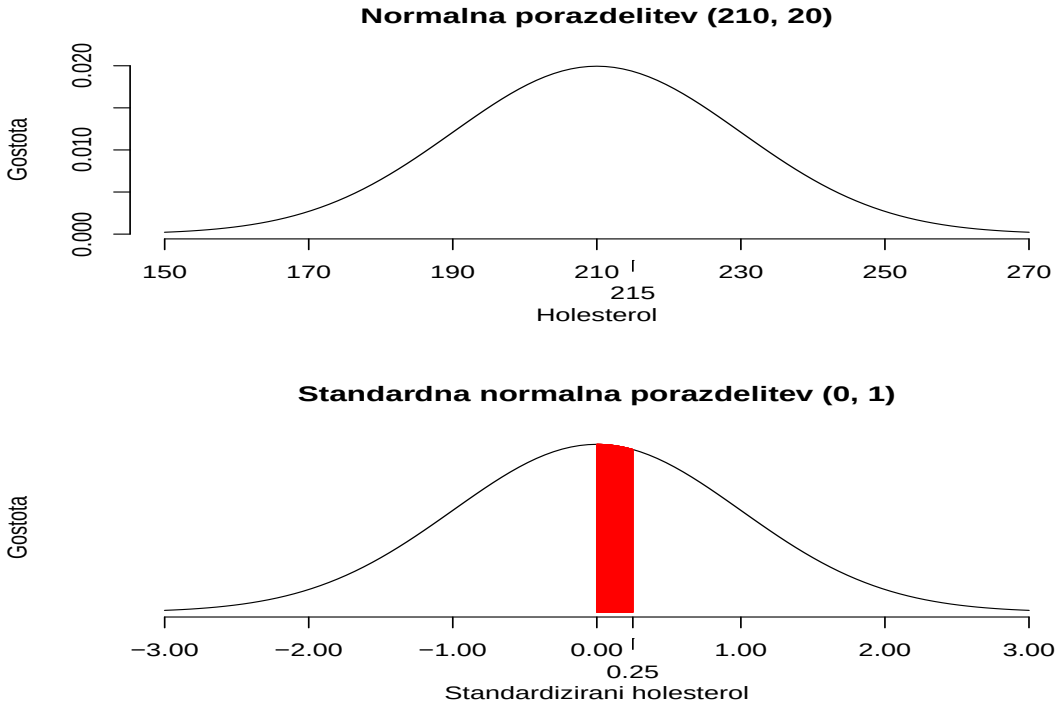
   (b) Izračunajte kolikšna je verjetnost, da ima oseba holesterol med 210 in 215 mg/dl.

   (c) Izračunajte kolikšna je verjetnost, da ima oseba holesterol več kot 215 mg/dl.

   (d) Izračunajte 97.5-ti percentil holesterola;

Rešitve
Namig
Za spremenljivko (X), ki je porazdeljena normalno s povprečjem μ in standardnim odklonom σ ($X \sim N(\mu, \sigma)$) velja, da jo lahko standardiziramo z naslednjo formulo

$$Z = \frac{X - \mu}{\sigma}.$$



Slika 2.2: *Porazdelitev holesterola (normalna s povprečjem 210 in standardnim odkolonom 20 in standardizirana).*

Nova spremenljivka Z ima še vedno normalno porazdelitev, s povprečjem 0 in standardnim odklonom 1 ($Z \sim N(\mu = 1, \sigma = 0)$); taki spremenljivki pravimo standardizirana normalna porazdelitev. Standardna normalna porazdelitev je uporabna, ker je njena porazdelitvena funkcija tabelirana (oziroma, v vsaki statistični knjigi lahko najdete, kolikšna je verjetnost, da je Z večji ali manjši od neke dane vrednosti, na primer: $P(Z < -1,96) = 0.025, P(Z > 0) = 0.5, P(Z > 2.58) = 0.005$, itd).

Uporaba statističnih tabel
Iz statistične tabele standardne normalne porazdelitve (glej tabelo B.1 na koncu knjige) odčitamo vrednosti za izračun P(Z<0.25) in P(Z<0). Vemo, da je standardna normalna porazdelitev simetrično porazdeljena okrog ničle; oziroma velja, da za katerokoli pozitivno vrednost a, $P(Z < -a) = P(Z > a)$. Spomnimo se tudi, da $P(Z < a) = 1 - P(Z \geq a) (= P(Z > a)$, ker je verjetnost vsake posamezne vrednosti 0, P(Z=a)=0)).
$P(Z < 0.25) = 1 - P(Z \geq 0.25)$; $P(Z \geq 0.25)$ odčitamo iz tabele standardne normalne porazdelitve: izberemo vrstico **0.2** in stolpec **0.05**, vrednost na presečišču je $P(Z \geq 0.25) = 0.401$. Podobno najdemo $P(Z \geq 0) = 0.5$ (vrstica **0** in stolpec **0**).

(a) Za spremenljivko holesterol (Hol) vemo, da $Hol \sim N(\mu = 210, \sigma = 20)$. Porazdelitev holesterola je prikazana na sliki 2.2.

(b) Verjetnost, da ima oseba holesterol med 210 in 215 mg/dl: $Z = \frac{Hol-210}{20} \sim N(\mu = 0, \sigma = 1)$.
$P(210 \leq Hol \leq 215) = P(\frac{210-210}{20} \leq \frac{Hol-210}{20} \leq \frac{215-210}{20}) = P(0 \leq Z \leq 0,25)$

$= P(Z \leq 0,25) - P(Z \leq 0) = 0.60 - 0.50 = 0.10$

(c) Verjetnost, da ima oseba holesterol med 210 in 215 mg/dl: $z1 = (210 - 210)/20 = 0$
$z2 = (215 - 210)/20 = 0,25$
$P(Hol > 215) = P(Z > 0.25) = 1 - P(Z \leq 0,25) = 1 - 0.60 = 0.40$

(d) 97.5-ti percentil holesterola je vrednost holesterola H, za katero velja P(Hol<H)=0.975. Tudi v tem primeru si lahko pomagamo z standardno normalno porazdelitev. Za Z $\sim N(0,1)$ že vemo, da je P(Z<1.96)=0.975 (glejte tabelo B.1). Zaenkrat smo uporabljali to enačbo: $\frac{Hol-\mu}{\sigma} = Z$, torej lahko izrazimo tudi Hol kot funkcijo od Z. $Hol = Z \cdot \sigma + \mu$. V našem primeru je $\mu = 210, \sigma = 20$, in vrednost, ki nas zanima je $z = 1.96$. Torej $x = 1.96 \cdot 20 + 210$=249.2 mg/dl.
V praksi bi pomenilo, da ima 97.5% oseb vrednost holesterola, ki je manjša od 249.2 mg/dl.

9. Holesterol se lahko meri v mmol/l ali pa v mg/dL. Formula, s katero transformiramo meritve iz mg/d v mmol/l, je:

$$mmol/l = mg/dl \cdot 0.02586$$

Narišite porazdelitev holesterola iz prejšnje naloge (Holesterol (mg/dL) $\sim N(\mu = 210, \sigma = 20)$) za meritve v mmol/l.

10. Na podlagi podatkov, zbranih za vsa slovenska dekleta, vemo, da je višina deklet približno normalno porazdeljena s povprečjem 168 cm in s standardnim odklonom 6 cm. Na podlagi teh podatkov ocenite

(a) simetrični interval glede povprečja, za katerega pričakujemo, da bo vseboval višino 95% slovenskih deklet;

(b) simetrični interval glede povprečja, za katerega pričakujemo, da bo vseboval višino 99% slovenskih deklet;

(c) 97.5-ti percentil višine slovenskih deklet;

(d) peti percentil višine slovenskih deklet;

(e) delež slovenskih deklet z višino, ki je 168 ali večja;

(f) * verjetnost, da bo slovensko dekle imelo višino, ki je točno enaka 168;

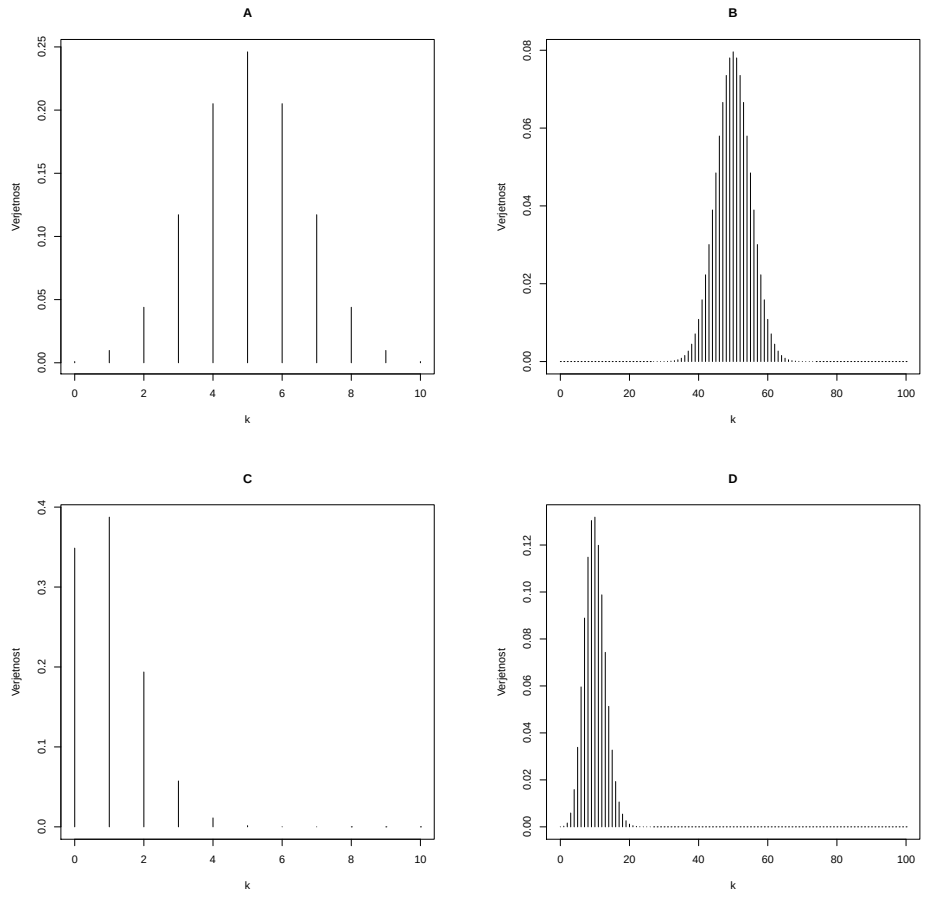
(g) delež slovenskih deklet z višino, ki je 156 ali večja.

2.3 Binomska porazdelitev

11. Na sliki 2.3 so prikazane štiri binomske porazdelitve. Za vsako ocenite parametra π in n. V vsakem grafikonu so prikazani vsi možni izidi (vse možne vrednosti binomske slučajne spremenljivke). Možni odgovori so $\pi = 0.1$ ali $\pi = 0.5$ in $n = 10$ ali $n = 100$.

Izpolnite tabelo

Porazdelitev	π	n
A		
B		
C		
D		



Slika 2.3: *Binomske porazdelitve.*

Namig

Pomagajte si z interaktivnimi statističnimi tabelami (`http://ibmi.mf.uni-lj.si/sl/centri/biostatisticni-center/interaktivno/statisticne-tabele`, izberite porazdelitev Distribution type: Binomial in spremenite število enot (poskusov) (*Number of trials*) in verjetnost uspeha (*Probability of success*)).

Povprečje binomske porazdelitve (μ) je $n\pi$ in standardni odklon je $\sqrt{\frac{\pi(1-\pi)}{n}}$. Binomsko porazdelitev je mogoče aproksimirati z normalno porazdelitvijo; aproksimacija je dobra, če sta $n\pi$ in $n(1-\pi)$ večja od 5.

12. Izpit iz statistike je sestavljen iz 10 vprašanj; vsako vprašanje ima 5 možnih odgovorov, izmed katerih študent izbere en pravilen odgovor.

    (a) Kolikšna je verjetnost, da bo izpit pozitivno opravil študent, ki se ni učil za izpit in ki naključno izbere odgovore (oziroma bo pravilno odgovoril na vsaj 5 vprašanj)?

    (b) Na izpitni rok se prijavi 15 študentov, ki se niso učili za izpit in ki naključno odgovarjajo na vprašanja. Kolikšna je verjetnost, da niti eden izmed študentov ne opravi izpita?

(a) $n = 10$: število poskusov
$\pi = 0.20$: verjetnost uspeha pri posameznem poskusu

K: število uspehov
$P(K \geq 5|n,\pi) = 1 - (P(K=0|n,\pi) + P(K=1|n,\pi) + P(K=2|n,\pi) + P(K=3|n,\pi) + P(K=4|n,\pi))$
$P(K=k|n,\pi) = \binom{n}{k}\pi^k(1-\pi)^{(n-k)}$
$P(K=0|n=10,\pi=0.2) = \binom{10}{0}0.2^0(1-0.2)^{(10-0)}$=0.11
$P(K=1|n=10,\pi=0.2) = \binom{10}{1}0.2^1(1-0.2)^{(10-1)}$=0.27
$P(K=2|n=10,\pi=0.2) = \binom{10}{2}0.2^2(1-0.2)^{(10-2)}$=0.3
$P(K=3|n=10,\pi=0.2) = \binom{10}{3}0.2^3(1-0.2)^{(10-3)}$=0.2
$P(K=4|n=10,\pi=0.2) = \binom{10}{4}0.2^4(1-0.2)^{(10-4)}$=0.09
$P(K \geq 5|n,\pi) = 1 - 0.97 = 0.03$

(b) $n = 15$: število študentov (poskusov)
$\pi = 0.03$: verjetnost uspeha pri posameznem izpitu
$k = 0$: število uspehov
$P(K=0|n=15,\pi=0.03) = \binom{15}{0}0.03^0(1-0.03)^{(15-0)} = 0.63$

13. Verjetnost, da gensko modicirana miš preživi vsaj en mesec je 0.20. Naredili smo raziskavo z desetimi mišmi.

    (a) Kolikšna je verjetnost, da po enem mesecu živi natanko ena?

    (b) Kolikšna je verjetnost, da po enem mesecu ne preživi niti ena?

    (c) Kolikšna je verjetnost, da po enem mesecu pre?ivi vsaj ena?

2.4 Druge porazdelitve

14. Na sliki 2.4 so prikazane štiri t porazditve. Za vsako ocenite stopinje prostosti (*degrees of freedom, df*) (možni odgovori so: 2, 8, 18 in 98). Za vsako porazdelitev določite tudi (kritično) vrednost $t_{df,0.025}$, oziroma vrednost t porazdelitve z df stopinjami prostosti, za katero velja, da $P(t_{df} \geq t_{df,0.025}) = 0.025$.

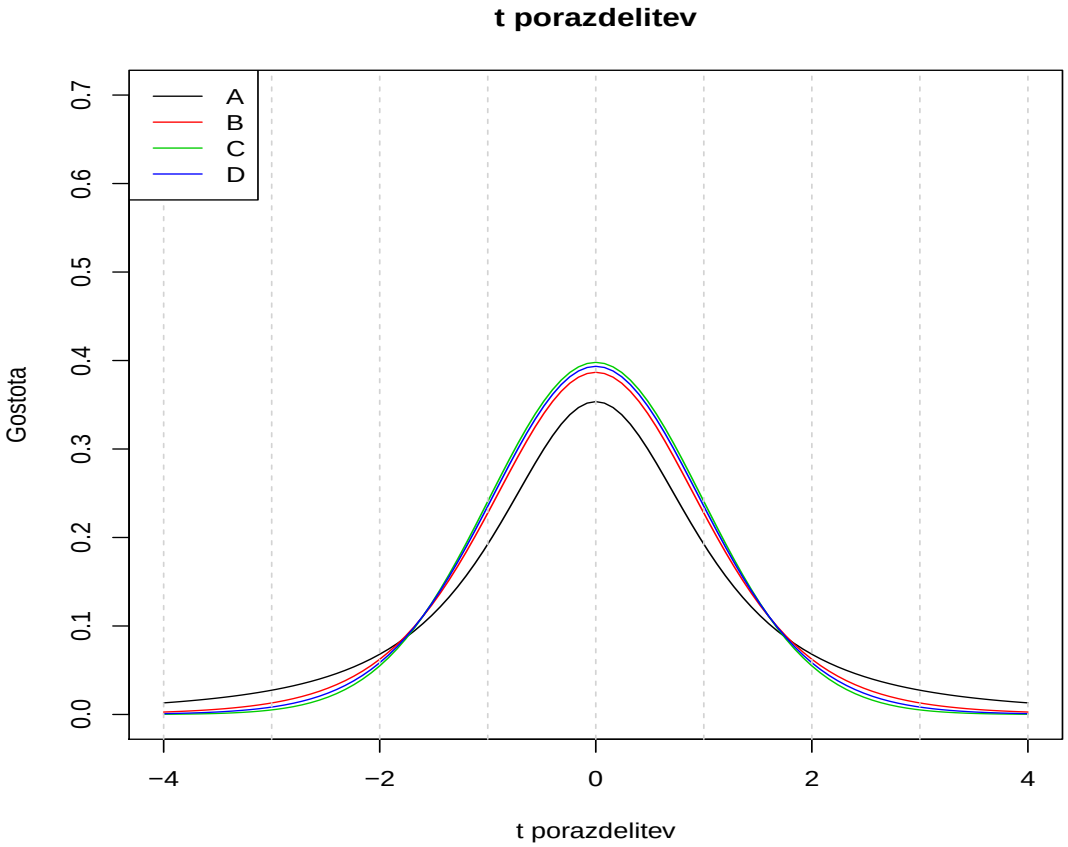
Izpolnite tabelo

Porazdelitev	Stopinje prostosti	$t_{0.025}$
A		
B		
C		
D		

Namig

Pomagajte si z interaktivnimi statističnimi tabelami (`http://ibmi.mf.uni-lj.si/sl/centri/biostatisticni-center/interaktivno/statisticne-tabele`, izberite porazdelitev Distribution type: t in spremenite število stopinj prostosti (*Degrees of freedom*).

Stopinje prostosti je parameter t porazdelitve. t porazdelitev je zelo podobna standardni normalni porazdelitvi, če ima veliko stopinj prostosti (df>50). t porazdelitev je bolj razpršena, ko ima manj stopinj prostosti.
$t_{df=8,0.025}$ je vrednost za katero velja, da $P(t_{df} \geq t_{df=8,0.025}) = 0.025$; odčitamo jo iz statistične tabele t porazdelitve. Izberemo vrstico $df = 8$ in stolpec $\alpha = 0.025$ in dobimo vrednost 2.306 ($P(t_{df=8} \geq 2.306) = 0.025$).



Slika 2.4: *t porazdelitve.*

15. Na sliki 2.5 so prikazane štiri hi-kvadrat (χ^2) porazditve. Za vsako ocenite stopinje prostosti (*degrees of freedom, df*) (možni odgovori so: 1, 2, 3 in 8). Za vsako porazdelitev določite tudi (kritično) vrednost $\chi^2_{df,0.05}$, oziroma vrednost χ^2 porazdelitve z *df* stopinjami prostosti, za katero velja, da $P(\chi^2_{df} \geq \chi^2_{df,0.05}) = 0.05$.

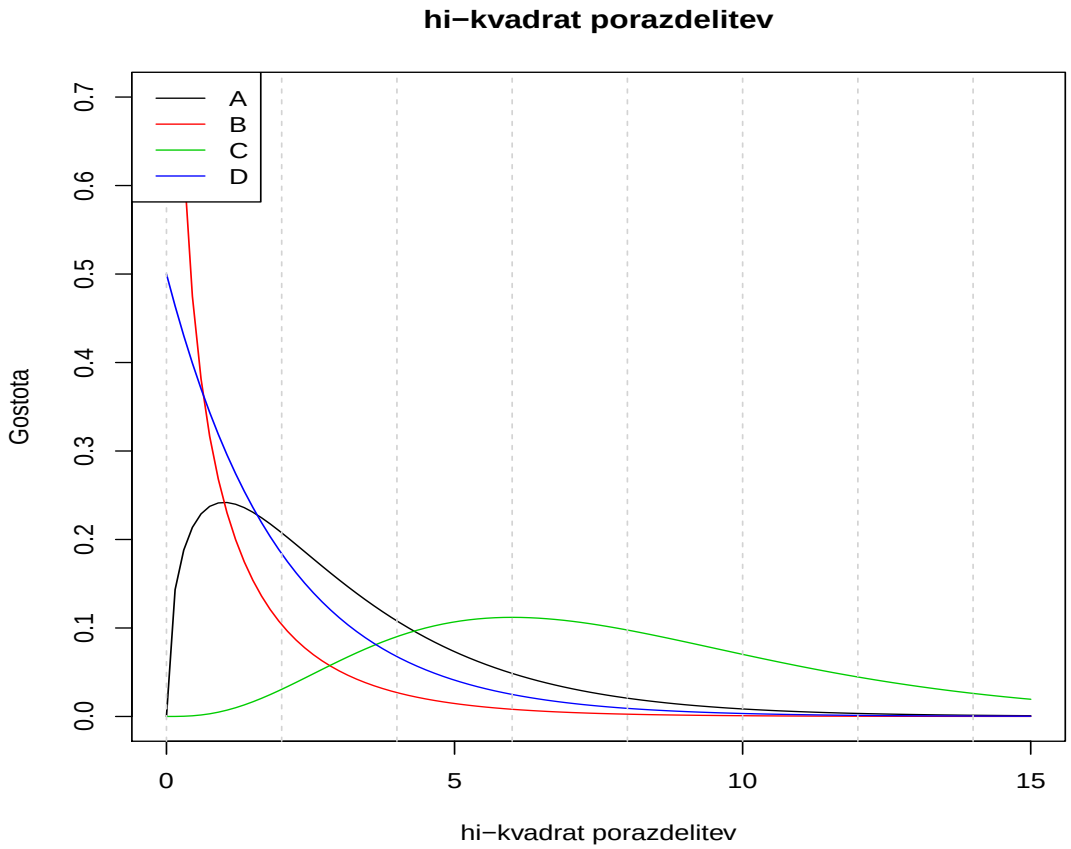
Izpolnite tabelo

Porazdelitev	Stopinje prostosti	$\chi^2_{0.05}$
A		
B		
C		
D		

Namig
Stopinje prostosti je parameter χ^2 porazdelitve. χ^2 porazdelitev je bolj asimetrična v desno, ko ima malo stopinj prostosti.
$\chi^2_{df=1,0.05}$ je vrednost za katero velja, da $P(\chi^2_{df=1} \geq \chi^2_{df=1,0.05}) = 0.05$; odčitamo jo iz statistične tabele χ^2 porazdelitve. Izberemo vrstico $df = 1$ in stolpec $\alpha = 0.05$ in dobimo vrednost 3.841 ($P(\chi^2_{df=1} \geq 3.841) = 0.05$).

2.5 Intervali zaupanja

16. Želeli ste oceniti povprečno število ovac, ki so v oskrbi posameznega pastirja na slovenskih planinah. V ta namen ste pri 8 pastirjih zbrali podatke o številu



Slika 2.5: χ^2 *porazdelitve.*

ovac v njegovi oskrbi. Dobili ste spodnje podatke:

$$23, 34, 64, 129, 45, 75, 153, 98.$$

Vprašanja

(a) Iz preteklih raziskav veste, da je varianca števila ovac, ki so v oskrbi enega pastirja 123. Izračunajte 95% interval zaupanja za povprečno število ovac v oskrbi enega pastirja.

(b) Ne poznate variance števila ovac, ki so v oskrbi enega pastirja, ker prejšnjih raziskav na tem področju ni. Izračunajte 95% interval zaupanja za povprečno število ovac v oskrbi enega pastirja. (Ta situacija je bolj realistična, ker ponavadi ne poznamo populacijske variance, ampak jo moramo oceniti na podlagi podatkov.)

(c) Kaj bi se spremenilo pri izračunu, če bi želeli podati tudi 99% intervala zaupanja (za oba primera)?

Rešitev Velikost vzorca: n=8 (število pastirjev).
Povprečje vzorca: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{23+34+\ldots+98}{8} = \frac{621}{8} = 77.62$.
Želimo oceniti 95% interval zaupanja za μ, povprečno število ovac v oskrbi enega pastirja v populaciji slovenskih pastirjev.

(a) Primer, ko je populacijska varianca znana. Varianca števila ovac: $\sigma^2 = 123$, $\sigma = \sqrt{123} = 11.09$.
Standardna napaka: $\hat{SE} = \frac{\sigma}{\sqrt{n}} = \frac{11.09}{\sqrt{8}} = 3.92$.
Vzorčno povprečje je normalno porazdeljeno, s povprečjem μ in standardnim odklonom $\hat{SE}$.
Spodnja meja za 95% IZ za μ: $\bar{x} - z_{0.025} * \hat{SE} = 77.62 - 1.96 * 3.92 = 69.94$.
Zgornja meja za 95% IZ za μ: $\bar{x} + z_{0.025} * \hat{SE} = 77.62 + 1.96 * 3.92 = 85.3$.
Vrednost $z_{0.025}$ je vrednost, za katero velja, da $P(Z \geq z_{0.025}) = 0.025$; $z_{0.025} = 1.96$: odčitamo jo iz statistične tabele standardne normalne porazdelitve (glej nalogo o porazdelitvi holesterola za razlago).

(b) Populacijska varianca ni znana, torej jo moramo oceniti na podlagi podatkov.
Ocenjena varianca števila ovac:
$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(23-77.62)^2+(34-77.62)^2+...+(98-77.62)^2}{8-1} = 2125.7$, $s = \sqrt{s^2} = 46.11$.
Standardna napaka: $\hat{SE} = \frac{s}{\sqrt{n}} = \frac{46.11}{\sqrt{8}} = 16.3$.
Standardizirano vzorčno povprečje ima t porazdelitev s $df = n - 1$=7 stopinjami prostosti.
Spodnja meja za 95% IZ za μ: $\bar{x} - t_{df=7,0.025} * \hat{SE} = 77.62 - 2.36 * 16.3 = 39.08$.
Zgornja meja za 95% IZ za μ: $\bar{x} + t_{df=7,0.025} * \hat{SE} = 77.62 + 2.36 * 16.3 = 116.16$.
$t_{df=7}$ je t porazdelitev s 7 stopinjami prostosti. Vrednost $t_{df=7,0.025}$ je vrednost za katero velja, da $P(t_{df=7} \geq t_{df=7,0.025}) = 0.025$; odčitamo jo iz statistične tabele t porazdelitve. Izberemo vrstico $df = 7$ in stolpec $\alpha = 0.025$ in dobimo vrednost 2.365 ($P(t_{df=7} \geq 2.365) = 0.025$).

(c) Postopek bi ostal isti, spremenili bi samo vrednost $z_{0.005}$ oziroma $t_{df=7,0.005}$.

17. Na podlagi podatkov, ki smo jih zbrali za študente moškega spola, ki so obiskovali predmet biostatistika med letoma 2008 in 2010, smo ocenili, da je višina študentov približno normalno porazdeljena ter da je povprečna višina študentov 182 cm in standardni odklon 6 cm. Vzorec je vseboval 48 študentov.

    Na podlagi teh podatkov ocenite

    (a) 95% interval zaupanja za povprečno višino slovenskih študentov moškega spola;

    (b) 99% interval zaupanja za povprečno višino slovenskih študentov moškega spola;

    (c) ali lahko na podlagi naših podatkov trdimo, da je povprečna višina slovenskih študentov moškega spola različna od 187 cm;

18. Kdaj je ocena populacijske aritmetične sredine bolj natančna?

    (a) pri manjšem vzorcu

    (b) pri večjem vzorcu

    (c) natančnost ni odvisna od velikosti vzorca.

19. Kako bomo z večjo verjetnostjo zajeli populacijsko povprečje teže?

(a) Če merimo 10000 oseb in izračunamo 95% interval zaupanja.

(b) Če merimo 100 oseb in izračunamo 95% interval zaupanja.

(c) Verjetnost ni odvisna od velikosti vzorca.

20. Pri raziskavi smo merili spremenljivko X in za povprečje X-a (μ) ocenili 95% interval zaupanja. Rezultat je 1.5 do 3.5. Odgovorite na spodnji vprašanji

(a) Kakšno je bilo povprečje našega vzorca?

(b) Standardni odklon populacije je bil 1.613. Koliko statističnih enot je bilo vključenih v raziskavo?

21. Kolikokrat se povečata/zmanjšata standardna napaka in interval zaupanja, če vzorec povečamo 9x? (Predpostavljamo, da je populacijski standardni odklon znan.)

(a) sprememba standardne napake

(b) sprememba intervala zaupanja.

2.6 Statistične napake

22. Katere trditve so pravilne in katere so napačne za napako 1. oziroma napako 2. vrste?

(a) Odvisna je od velikosti vzorca.

(b) Odvisna je od variabilnosti spremenljivke.

(c) Odvisna je od druge napake.

(d) Manjša je pri večjih raziskavah.

(e) Manjša je pri bolj variabilnih spremenljivkah.

(f) Posledica te napake je, da dobimo lažno pozitivni rezultat.

(g) Posledica te napake je, da dobimo lažno negativni rezultat.

(h) Naredimo jo, če zavrnemo pravilno ničelno domnevo.

(i) Naredimo jo, če ne zavrnemo napačne ničelne domneve.

(j) Enaka je stopnji značilnosti.

(k) Odvisna je od moči testa.

(l) Odvisna je od velikosti učinka, ki ga želimo zaznati.

(m) **Pri primerjavi dveh povprečij je odvisna od populacijskega povprečja.

(n) **Pri primerjavi dveh deležev je odvisna od populacijskih deležev.

Poglavje 3

Primerjava skupin - številske spremenljivke

Pri izračunu p vrednosti si pomagajte s tabelami, vaš rezultat bo zato le približen in odvisen od natančnosti vaših tabel. Pomagajte si tudi z interaktivnimi statističnimi tabelami `http://ibmi.mf.uni-lj.si/sl/centri/biostatisticni-center/interaktivno/statisticne-tabele`. Rezultat, ki ga najdete v rešitvah, je ponavadi izračunan z računalnikom.

1. Želeli ste oceniti razliko v povprečni širini dlak med psi in mačkami. Zmerili ste širino dlake pri 10 psih in 10 mačkah ter dobili naslednje podatke.

   - Ali je povprečna širina dlake pri psih in mačkah razlikuje?
   - Izračunajte 95% interval zaupanja za razliko v povprečni širini dlak med psi in mačkami ter ga interpretirajte. Lahko predpostavimo, da sta varianci enaki.

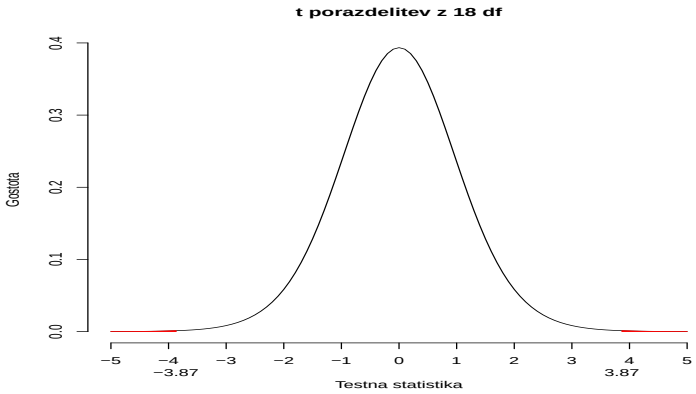
Širina (mm)	Povprečje ($\bar{x}$)	Standardni odklon (s)	Število (n)
Psi (1)	103.2	25.0	10
Mačke (2)	68.2	13.9	10

Rešitve

- Statistična metoda: **t-test za neodvisna vzorca**.
- Skupni standardni odklon: $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$
- Standardna napaka: $\hat{SE} = s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}$
- Testna statistika: $t = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{\bar{x}_1-\bar{x}_2}{\hat{SE}}$
- Stopinje prostosti: $df = n_1 + n_2 - 2$
- Porazdelitev testne statistike pod ničelno domnevo: t porazdelitev z $n_1 + n_2 - 2$ stopinjami prostosti ($t_{n_1+n_2-2}$) (prikazana na sliki 3.1).

$s_p = \sqrt{\frac{(10-1)25^2+(10-1)13,9^2}{10+10-2}} = 20.23$ mm.
$\hat{SE} = 20.23\sqrt{(\frac{1}{10}+\frac{1}{10})} = 9.05$ mm

$\bar{x}_1 - \bar{x}_2 = 103.2 - 68.2 = 35$ mm.
$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{SE}} = \frac{35}{9.05} = 3.87$
$p = P(t_{18} < -3.87|H_0) + P(t_{18} > 3.87|H_0) = 2P(t_{18} > 3.87|H_0) = 2 * 0.0006 = 0.0011$



Slika 3.1: *t porazdelitev (18 stopinj prostosti).*

Uporaba statističnih tabel
P-vrednost lahko odčitamo iz statistične tabele za porazdelitev t (tabela 1.2). V tem primeru nas zanima porazdelitev t z 18 stopinjami prostosti, torej izberemo vrstico kjer je df=18. V tej vrstici iščemo vrednost testne statistike (t=3.87). Največja vrednost v tej vrstici je 3.922 (za $\alpha = 0.0005$), druga največja pa 3.610 (za $\alpha = 0.001$). To pomeni, da $P(t_{18} > 3.610) = 0.001$ in $P(t_{18} > 3.922) = 0.0005$. Ker $3.610 < 3.87 < 3.922$ vemo, da velja tudi, da $P(t_{18} > 3.87) < 0.001$ in $P(t_{18} > 3.87) > 0.0005$. Za izračun p-vrednosti rabimo tudi $P(t_{18} < -3.87)$; porazdelitev t je simetrična, torej velja tudi, da $P(t_{18} < -3.87) < 0.001$. $p < 2 \cdot 0.001 = 0.0002$. Na podlagi te tabele ne moremo priti do bolj natančnega rezultata.

Če bi želeli samo preveriti, ali je p-vrednost manjša ali večja od 0.05 potem bi morali primerjati vrednost izračunane testne statistike s kritično vrednostjo $t_{0.025} : P(t_{18} > t_{0.025}) = 0.025$, ki je 2.101. Ker je (absolutna vrednost) testne statistike večja od 2.101, potem lahko sklepamo, da je p-vrednost< 0.05.

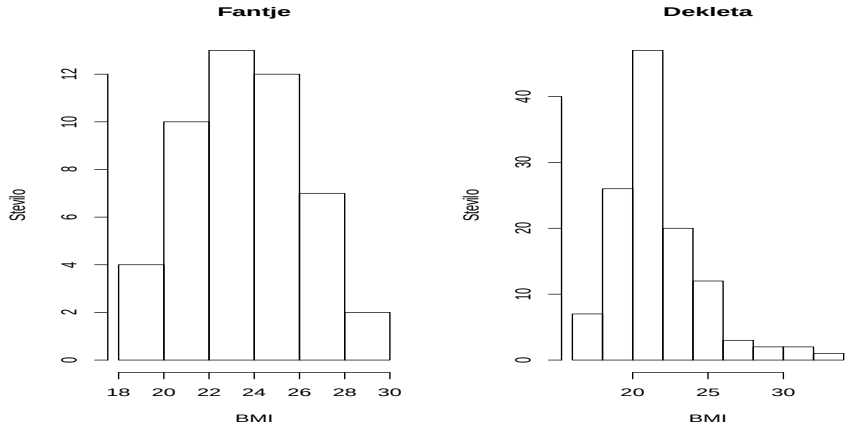
95% interval zaupanja za $\mu_{psi} - \mu_{macke}$: od $(\bar{x}_1 - \bar{x}_2) - \hat{SE}t_{18,0.025}$ do $(\bar{x}_1 - \bar{x}_2) + \hat{SE}t_{18,0.025}$= od $35 - 9.05 \cdot 2.1$ do $35 + 9.05 \cdot 2.1 =$ od 16 do 54.

2. Želimo primerjati povprečno višino slovenskih fantov in deklet. Zbrali smo naslednje podatke med letoma 2008 in 2010.

	Povprečje	Standardni odklon	Število
Fantje	181.50	5.63	48.00
Dekleta	168.32	6.03	120.00

- Ugotovite, ali se povprečna višina fantov statistično značilno razlikuje od povprečne višine deklet.
- Izračunajte tudi 95% interval zaupanja za populacijsko razliko povprečij (fantje-dekleta) in ga interpretirajte.

3. Primerjati želimo povprečni indeks telesne mase (BMI, *body mass index*) slovenskih fantov in deklet. Zbrali smo podatke med letoma 2008 in 2010 in dobili naslednje rezultate. Porazdelitev BMI-ja v skupini fantov in deklet je prikazana na sliki 3.2.



Slika 3.2: *Porazdelitev BMI-ja.*

	Povprečje	Standardni odklon	Število
Fantje	23.60	2.55	48.00
Dekleta	21.62	2.93	120.00

```
        Two Sample t-test

data:  BMI by Spol
t = -4.086, df = 166, p-value = 0.00006817
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.927462 -1.020040
sample estimates:
mean in group Dekleta  mean in group Fantje
             21.62420              23.59795
```

(a) Ali se povprečni BMI fantov statistično značilno razlikuje od povprečni BMI deklet?

(b) Kateri statistični test smo uporabili, da bi določili statistično značilnost razlik v BMI-ju?

(c) Izračunajte 95% interval zaupanja za populacijsko razliko povprečij (fantje-dekleta) in ga interpretirajte?

(d) Na sliki 3.2 sta prikazani vzorčni porazdelitvi BMI-ja pri fantih in dekletih. Razmislite o izpolnjenosti zahtevanih predpostavk za korektno izvedbo uporabljenega statističnega testa. Katere predpostavke bi bile v tem primeru problematične. Utemeljite.

(e) Kako lahko bolj učinkovito grafično predstavimo razlike v BMI med fanti in dekleti?

Rešitve
Glej dodatek za točke (a)-(c).

(d) Predpostavke na katerih temelji t-test za dva neodvisna vzorca z enakimi varianci so

- statistične enote so neodvisne;
- populacijski varianci sta v obeh skupinah enaki;
- številska spremenljivka je v obeh skupinah porazdeljena normalno.

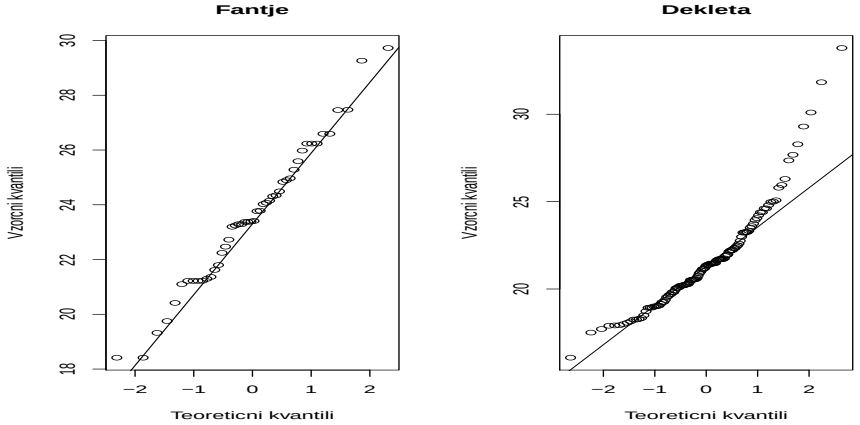
Enakost varianc lahko preverimo s testom F; ničelna domneva pravi, da je razmerje med populacijskima varijancima enako 1 ($H_0 : \sigma_F^2/\sigma_D^2 = 1$).

```
        F test to compare two variances

data:  my.data.2$BMI by my.data.2$Spol
F = 0.7589, num df = 47, denom df = 119, p-value = 0.2846
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4806524 1.2607176
sample estimates:
ratio of variances
         0.7588931
```

Na podlagi tega rezultata ne moremo zavrniti ničelne domneve o enakost varianc pri dekletih in fantih (P=0.285, 95% IZ za σ_F^2/σ_D^2: od 0.48 do 1.26).

BMI je pri dekletih porazdeljen nekoliko asimetrično v desno (glej sliko 3.2). Odmiki od normalnosti so še bolj razvidni, če narišemo graf normalnih kvantilov (*normal quantile-quantile plot, normal QQ-plot*). Z grafikonom primerjamo teoretične kvantile standardne normalne porazdelitve (os x) s kvantili našega vzorca (os y). Vrišemo tudi premico enakosti, kjer bi pričakovali točke v primeru normalne porazdelitve danih podatkov. V primeru, da so točke bistveno oddaljene od premice, lahko (*po občutku*) sklepamo, da vzorec ne izhaja iz normalno porazdeljene spremenljivke. Graf kvantilov je podan na sliki 3.3, kjer opazimo, da so pri dekletih kvantili višjih vrednosti večji od pričakovanih pri normalno porazdeljeni spremenljivki.



Slika 3.3: *Graf kvantilov za BMI.*

V primeru, da v populaciji niso spremenljivke normalno porazdeljene, lahko uporabimo **Mann-Whitneyev test**, ki ne predpostavlja normalne porazdelitve številskih spremenljivk. Mann-Whitneyev test je znan tudi kot **Wilcoxonov test vsote rangov** (*Wilcoxon rank sum test*). Ničelna domneva pravi, da je porazdelitev spremenljivke enaka ne glede na skupino.

Mann-Whitneyev test uporabi range namesto dejanskih podatkov. Podatke uredimo ne glede ne skupino: oseba z najmanjšim BMI pridobi rang r=1, tista z drugim najmanjšim BMI rang r=2, itd. Če imata dve osebi isto vrednost BMI pridobita povprečni rang. Za vsako skupino izračunamo vsoto rangov. Primer izračuna rangov in vsote rangov v vsaki skupini je podan v tabeli.

BMI	Spol	Rang (r)
19.5	F	1
20.1	F	2
22.1	M	3
23.2	F	4
23.5	F	5.5
23.5	M	5.5
24.0	M	7

Vsota rangov za moške je 3+5.5+7=15.5, vsota rangov za ženske je 1+2+4+5.5=12.5. Znan rezultat je, da je vsota naravnih števil od 1 do n: $n(n+1)/2$. Torej je vsota rangov za primer, ki je podan v tabeli: $\frac{n(n+1)}{2}=\frac{7\times 8}{2} = 28$.

Za primerjavo BMI-ja smo izvedli Mann-Whitneyev test s statističnem programom R in smo pridobili naslednje rezultate.

```
        Wilcoxon rank sum test

data:  my.data.2$BMI by my.data.2$Spol
W = 4185, p-value = 0.000004605
alternative hypothesis: true location shift is not equal to 0
```

Testna statistika U (označena z W v statističnem programu R) je izračunana z uporabo podatkov za najmanjšo skupino (moški, $n_1 = 48$)

$$U = \sum_{i=1}^{n_1} r_i - \frac{n_1(n_1+1)}{2},$$

$\sum_{i=1}^{n_1} r_i$ je vsota rangov za moške. Vrednost na našem vzorcu je U=4185.

Lahko bi izračunali eksatno p-vrednost s pomočjo verjetnostnega računa (ali s statističnim programom), ampak za velike vzorce ponavadi uporabimo normalni približek. Če velja ničelna domneva, U je normalno porazdeljen s povprečjem $\frac{n_1\times n_2}{2}$ in s standardnim odklonom $\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$. Standardiziramo testno statistiko U in dobimo vrednost

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}},$$

ki izhaja iz standardne normalne porazdelitve, če velja ničelna domneva.

$$\frac{4185 - \frac{48\times 120}{2}}{\sqrt{\frac{48\times 120(48+120+1)}{12}}} = 4.58.$$

P-vrednost je <0.001, torej lahko zavrnemo ničelno domnevo, da imata fante in dekleta isto porazdelitev BMI-ja. Lahko tudi rečemo, da če naključno izberemo fanta in dekle iz populacije, imamo $p = U/(n_1 n_2) = 0.73$ verjetnost, da bo vrednost BMI-ja večja za fanta kot za dekle iz tega para.

4. Primerjati želimo povprečno višino študentov, ki igrajo videoigrice s popvrečno višino študentov, ki ne igrajo videoigric. Zbrali smo podatke med letoma 2008 in 2010 ter jih analizirali s t-testom za neodvisna vzorca ob predpostavki enakih varianc. Dobili naslednje rezultate.

```
        Two Sample t-test

data:  visina by igrice
t = 2.6677, df = 166, p-value = 0.008394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9084266 6.0819580
sample estimates:
   mean in group Igra mean in group Ne igra
             174.2500              170.7548
```

   (a) Koliko oseb smo zajeli v vzorec?

   (b) Interpretirajte rezultate.

   (c) Kako bi lahko razložili ta rezultat?

5. Raziskava, ki proučuje sistolični krvni tlak, primerja novo zdravilo (Trt) s placebom tako, da je vsak pacient dobil placebo, mesec kasneje pa še novo zdravilo. Rezultati so naslednji (v mmHg):

Zdravilo	152	145	136	156	116	95	126	116	152	140
Placebo	150	180	140	157	120	132	135	126	170	136
Razlika (D)	2	-35	-4	-1	-4	-37	-9	-10	-18	4

Vprašanja

- Kateri statistični test lahko uporabite, da bi ugotovili ali se učinkovitost novega zdravila in placeba razlikujeta?
- Natančno zapišite ničelno (in alternativno) domnevo, ki jo pri tem testu testiramo.
- Ali je učinkovitost novega zdravila statistično značilno različna od placeba?
- Ali 95% interval zaupanja za razliko povprečnega sistoličnega tlaka med meritvami, ki so bile narejene po zdravljenju z novim zdravilom, in meritvami, ki so bile narejene po zdravljenu s placebom, vsebuje vrednost 0?

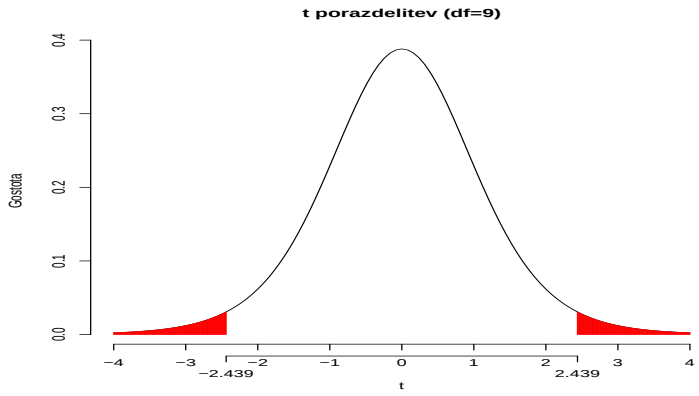
Rešitve

- Da bi ugotovili, ali je učinkovitost novega zdravila različna od placeba, lahko uporabimo **t-test za odvisne podatke**. T-testa za neodvisna vzorca ne moremo uporabiti, saj smo merili vsako osebo dvakrat. Predpostavka o neodvisnosti enot bi bila v tem primeru kršena.

- Ničelna domneva je, da je v populaciji povprečje krvnega tlaka enako ne glede na zdravilo, oziroma $H_0 : \mu_{\text{Zdravilo}} = \mu_{\text{Placebo}}$ (oziroma $\delta = \mu_{\text{Zdravilo}} - \mu_{\text{Placebo}} = 0$), alternativna domneva pa je $H_a : \mu_{\text{Zdravilo}} \neq \mu_{\text{Placebo}}$ (oziroma $\delta = \mu_{\text{Zdravilo}} - \mu_{\text{Placebo}} \neq 0$)
- Z D smo označili spremenljivko *Razlika (D)*, ki za vsakega pacienta meri razliko med sistoličnim krvnim tlakom po zdravljenju z novim zdravilom in po zdravljenju s placebom. Izračunamo razliko za vsakega pacienta, vzorčno povrečje ($\bar{D}$) in vzorčni standardni odklon te spremenljivke (s_D).

	Povrečje	SD	n
Zdravilo	133.40	19.78	10.00
Placebo	144.60	19.38	10.00
Razlika (D)	-11.20	14.52	10.00

NB: Povprečna razlika ($\bar{D}$) je enaka razliki povprečij ($\bar{x}_{\text{Zdravilo}} - \bar{x}_{\text{Placebo}}$). Za izračun standardnega odklona razlik (s_D) pa rabimo podatke posameznikov (izračun je možen neposredno samo če sta spremenljivki neodvisni). Velikost vzorca (n=10) je število neodvisnih enot, ki smo jih vključili v raziskavo, ne pa število meritev ($2n = 2 * 10$).



Slika 3.4: *t porazdelitev (9 stopinj prostosti).*

Testna statistika se izračuna s formulo

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} = \frac{\bar{D}}{\hat{SE}};$$

s_D je standardni odklon razlik, $s_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}}$, $\hat{SE} = \frac{s_D}{\sqrt{n}}$ je standardna napaka. Če velja ničelna domneva, je testna statistika porazdeljena po t_{n-1} porazdelitvi (t porazdelitev z n-1 stopinjami prostosti; pazite, da je df=10-1=9 (in ne 19 ali 18). V tem primeru je standardna napaka $\hat{SE} = \frac{s_D}{\sqrt{n}} = \frac{14.52}{3.16} = 4.592$ in vrednost testne statistike je:
$t = \frac{\bar{D}}{\hat{SE}} = \frac{-11.2}{4.592} = -2.439$.
t porazdelitev z 9 stopinjami prostosti je prikazana na sliki 3.4 in p-vrednost predstavlja pobarvano površino. P-vrednost izračunamo podobno kot pri t-testu za neodvisna vzorca, $p = P(t_{df} \leq -|t||H_0) +$

$P(t_{df} \geq |t||H_0) = 2P(t_{df} \geq |t||H_0)$, kar v našem primeru pomeni, da je $p = 2 * P(t_9 \geq 2.439) = 2 * 0.019 = 0.038$

Za ta primer je izpis iz statističnega programa R naslednji:

```
        Paired t-test

data:  trt and placebo
t = -2.4391, df = 9, p-value = 0.03742
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.5873282  -0.8126718
sample estimates:
mean of the differences
                  -11.2
```

- 95% interval zaupanja za razliko povprečnega sistoličnega tlaka med meritvami (95% IZ za $\bar{D}$) ne vsebuje vrednosti 0, saj je rezultat statistično značilen pri stopnji značilnosti 5%. 99% IZ pa bi zajel vrednost 0, ker je $p > 0.01$.

6. Drugi raziskovalec želi odgovoriti na vprašanje, ali je novo zdravilo bolj učinkovito od placeba, tako da meri vrednosti tlaka pri 20 pacientih, ki jih naključno razdeli v 2 skupini. Ena skupina dobi placebo, druga pa novo zdravilo. Rezultati so numerično popolnoma enaki tistim, ki so opisani v prejšnji nalogi.

   (a) Ali lahko ta raziskovalec pride do istega zaključka kot prvi raziskovalec, ne da bi ponovil statistični test?

   (b) Kateri statistični test lahko uporabite, da bi ugotovili ali je učinkovitost novega zdravila različna od placeba?

   (c) Izračunajte

   - Vzorčno razliko povprečij
   - Standardno napako
   - Testno statistiko
   - p-vredost
   - 95% interval zaupanja

   (d) Kaj lahko zaključite iz teh rezultatov ?

7. S testom t za odvisna vzorca smo na vzorcu 30 bolnikov z ankilozirajočim spondilitisom primerjali sedimentacijo eritrocitov (v mm/h) pred in po zdravljenju z nesteroidnimi protivnetnimi zdravili. Dobili smo spodnji rezultat.

	Mean	$\hat{SE}$	95% Conf Int		t	df	Sig.(2-tailed)
Sed. pred - Sed. po zdr	8,167	2,114	3,843	12,490	3,863	29	,002

   (a) Natančno navedite ničelno domnevo, ki smo jo testirali.

   (b) Ali jo zavrnemo ali sprejmemo ($\alpha = 0.05$)? (Za utemeljitev odločitve obkrožite ustrezno vrednost v izpisu)

   (c) Kako se porazdeli testna statistika, če velja ničelna domneva?

   (d) Kolikšen je bil standardni odklon razlik eritrocitov pred in po zdravljenju?

(e) Zakaj nismo uporabili testa t za neodvisna vzorca?

(f) Naštejte katere vrednosti v zgornji tabeli bi se spremenile, če bi ga? Kako?

8. Merili ste število obiskov pri veterinarju v desetih (istih) klinikah v letu 2008 in 2009. Dobili ste, da je bilo v povprečju leta 2008 90 obiskov, leta 2009 pa 100. Standardni odlkon razlik je bil 7 obiskov. Iz prejšnjih raziskav veste, da je smiselno predpostavljati, da se razlika obiskov porazdeli normalno.

   (a) Kateri statistični test bi lahko uporabili, da bi sklepali, ali je prišlo do razlike v povprečnem številu obiskov med letoma 2008 in 2009?

   (b) Izračunajte testno statistiko in p-vrednost. Kaj na podlagi statističnega testa na naših podatkih lahko sklepate o razliki v povprečnem številu obiskov v veterinarskih klinikah? Natančno navedite ničelno domnevo, ki jo testiramo.

   (c) Kako interpretiramo p-vrednost?

   (d) Ali lahko na podlagi rezultatov pridobljenih s statističnim testiranjem skepamo, ali bo vrednost 0 vključena v 95% intervalu zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008)?

   (e) Izračunajte 95% interval zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008).

   (f) Kako interpretiramo ta interval zaupanja?

   Drugi raziskovalec se je lotil iste raziskave, ampak malo drugače. Leta 2008 in 2009 je vključil v svojo študijo različne veterinarske klinike (10 vsako leto). Dobil je natanko ista povprečja, standardna odklona števila obiskov pa sta bila 5 v letu 2008 in 7 v letu 2009. Iz prejšnjih raziskav vemo, da je smiselno predpostavljati, da se variabilnost števila obiskov ne spreminja skozi čas, in da je število obiskov v populaciji normalno porazdeljeno.

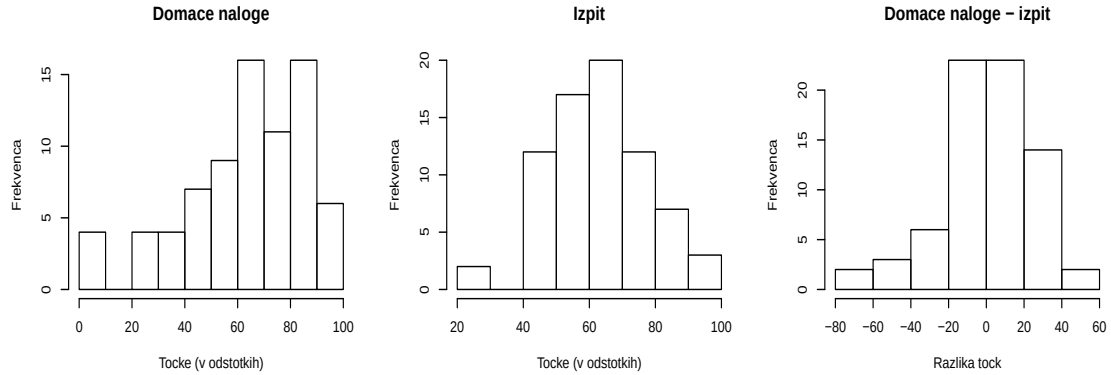
   (a) Izračunajte 95% interval zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008). Ali je potrebno ponoviti izračun, ali lahko uporabite rezultate statističnega sklepanja iz prejšnje naloge, zato ker je raziskovalec dobil natanko iste rezultate?

   (b) Kateri statistični test bi lahko uporabil ta raziskovalec, da bi sklepal, ali je prišlo do razlike v povprečnem številu obiskov med letoma 2008 in 2009?

   (c) Ali lahko na podlagi rezultatov pridobljenih s 95% intervalom zaupanja sklepamo, ali bo p-vrednost pridobljena s statističnim testiranjem večja ali manjša kot 0.05?

   (d) Izračunajte testno statistiko in p-vrednost ter sklepajte, ali je na podlagi naših podatkov med letoma 2008 in 2009 prišlo do razlike v povprečnem številu obiskov v veterinarskih klinikah. Natančno navedite ničelno domnevo, ki jo testiramo.

9. Primerjati želimo število doseženih točk na izpitu biostatistike z dodatnimi točkami iz domačih nalog. Oba podatka imamo za 73 študentov. Ker je maksimalno število točk različno (20 za domače naloge in 80 na izpitu), smo se odločili, da bomo uporabili za obe spremenljivki uporabili odstotek doseženih možnih točk. Na sliki 3.5 je prikazana s histogramom porazdelitev spremenljivk ter porazdelitev njune razlike.



Slika 3.5: *Porazdelitve spremenljivk.*

S katerim statističnim testom bi preverili, ali sta rezultat na izpitu in dodatne točke iz domačih nalog povezane?

Rešitev
Podatki so parni, ker smo pridobili dve vrednosti za vsakega študenta. V tem primeru bi lahko primerjali povprečje spremenljivk s t-testom za dva odvisna vzorca (parni t-test). Predpostavke za uporabo t-testa za dva odvisna vzorca so:

- neodvisnost statističnih enot;
- normalnost porazdelitve razlike v populaciji.

Lahko predpostavljamo, da so rezultati študentov neodvisni, ampak je razvidno s slike, da razlika odstotnih točk ni normalno porazdeljena. Statistični test za primerjavo dveh povezanih spremenljivk, ki ne predpostavlja normalnost razlik, je **Wilcoxonov test predznačenih rangov** *Wilcoxon signed rank test*. Podobno kot Mann-Whitneyev test v primeru neodvisnih vzorcev, tudi ta test uporablja range namesto dejanskih podatkov. V tem primeru rangiramo absolutne vrednosti razlik, oziroma razlike ne glede na predznak. Primer izračuna rangov je podan v naslednji tabeli.

Domače naloge	Izpit	Razlika	Rang (r)
15	12	3	1
30	34	-4	2
100	95	5	3
90	100	-10	4
70	100	-30	5
70	30	40	6
80	80	0	–

Iz vzorca odstranimo enote, za katere je razlika 0. Vsota rangov pozitivnih razlik je 1+3+6=10, vsota rangov negativnih razlik je 2+4+5=11. Ničelna domneva Wilcoxonovega testa predznačenih rangov pravi, da je populacijska mediana razlik 0. V praksi to pomeni, da testiramo, ali je verjetnost pozitivnih in negativnih razlik enaka. Rezultat

Za primerjavo izpitnih in dodatnih točk smo izvedli Wilcoxonov test predznačenih rangov s statističnim programom R in smo pridobili naslednje rezultate.

```
        Wilcoxon signed rank test

data:  DN and max.points
V = 1488.5, p-value = 0.448
alternative hypothesis: true location shift is not equal to 0
```

Imeli smo podatke za 73 študentov. Za nobenega študenta nismo dobili iste vrednosti pri obeh spremenljivkah (n=velikost vzorca - število vezanih podatkov (kjer je razlika=0)). Pozitivnih razlik je bilo 39, njihova vsota rangov je bila 1488.5; negativnih razlik je bilo 34, njihova vsota rangov je bila 1212.5. Testna statistika V je največja vsota rangov.

Tudi v tem primeru bi lahko izračunali eksatno p-vrednost s pomočjo verjetnostnega računa (ali s statističnim programom), ampak za velike vzorce ponavadi uporabimo normalni približek. Če velja ničelna domneva, V je normalno porazdeljen s povprečjem $\frac{n(n+1)}{4}$ in s standardnim odklonom $\sqrt{\frac{n(n+1)(2n+1)}{24}}$.

Standardiziramo testno statistiko V in dobimo vrednost

$$\frac{V - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2*n+1)}{24}}},$$

ki izhaja iz standardne normalne porazdelitve, če velja ničelna domneva.

$$\frac{1488.5 - \frac{73\times 74}{4}}{\sqrt{\frac{73\times 74\times 147}{24}}} = 0.76.$$

P-vrednost je 0.4481, torej ne moremo zavrniti ničelne domneve, da je mediana razlik dodatnih in izpitnih točk 0. Ta rezultat je bil pričakovan, saj je bila opažena razlika median na vzorcu zelo majhna (mediana dodatnih točk =66.5, mediana izpitnih točk = 63.75).

Poglavje 4

Primerjava skupin - opisne spremenljivke

1. Opazovali smo 100 mačk (50 samcev in 50 samic). Levkemijo mačk je imelo 25 samcev in 16 samic. Testirajte povezanost med spolom in prisotnostjo levkemije.

   Vprašanja

   - Kateri test bi uporabili v tem primeru?
   - Navedite ničelno domnevo, ki jo testiramo.
   - Izračunajte testno statistiko.
   - Kako se porazdeli testna statistika, če velja ničelna domneva?
   - Izračunajte p-vrednost in sklepajte, ali lahko na podlagi naših podatkov zavrnemo ničelno domnevo.
   - Ali bi pričakovali manjšo ali večjo p-vrednost, če bi opazili, da ima 27 samcev levkemijo in 10 samic? (izračuna ni potrebno ponoviti)

   Rešitve

   Spodaj je podana kontingenčna tabela z opazovanimi frekvencami.

	Samec	Samica	Vsota
Levkemija	25	16	41
Ni levkemije	25	34	59
Vsota	50	50	100

   Ponavadi opazovane frekvence označimo s črko O (iz angleščine *observed frequencies*). $N(=100)$ je velikost vzorca.

	Samec	Samica	Vsota
Levkemija	$O_{\text{levkemija, samec}}$	$O_{\text{levkemija, samica}}$	$O_{\text{levkemija}}$
Ni levkemije	$O_{\text{ni levkemije, samec}}$	$O_{\text{ni levkemije, samica}}$	$O_{\text{ni levkemije}}$
Vsota	O_{samec}	O_{samica}	N

   Statistični test, ki ga lahko uprabimo za preverjanje povezanosti med spolom in mačjo levkemijo, je test hi-kvadrat. Ničelna domneva je, da v populaciji mačk ni povezanosti med spolom in mačjo levkemijo, oziroma, da delež mačk z levkemijo ni odvisen od spola.

Izračunamo pričakovane frekvence ob veljavni ničelni domnevi. V tem primeru so pričakovane frekvence (E_{ij}) enake številu mačk, ki bi jih pričakovali v vsaki kategoriji, če bi bila spol in mačja levkemija neodvisna. Označimo jih s črko E (iz angleščine, *expected frequencies*).

Kontigenčna tabela pričakovanih frekvenc (E) je:

	Samec	Samica	Vsota
Levkemija	$E_{\text{levkemija, samec}}$	$E_{\text{levkemija, samica}}$	$O_{\text{levkemija}}$
Ni levkemije	$E_{\text{ni levkemije, samec}}$	$E_{\text{ni levkemije, samica}}$	$O_{\text{ni levkemije}}$
Vsota	O_{samec}	O_{samica}	N

Predpostavljamo, da sta robna vrstica in robni stolpec pričakovanih frekvenc enaka opazovanim.

$E_{\text{levkemija,samec}} = N * P(\text{levkemija} \cap \text{samec}|H_0) = N\frac{^{N}\text{levkemija}}{N}\frac{^{N}\text{samec}}{N} =$
$N * P(\text{levkemija}) * P(\text{samec}) = N * \frac{^{O}\text{levkemija}}{N} * \frac{^{O}\text{samec}}{N} =$
$100\frac{41}{100}\frac{50}{100} = 20.5.$

Druge pričakovane frekevence lahko izračunamo z istim postopkom, ali bolj enostavno: odštejemo že izračunane pričakovane vrednosti od robnih vrednosti:
$E_{\text{levkemija,samica}} = O_{\text{levkemija}} - E_{\text{levkemija,samec}}$=41-20.5=20.5
$E_{\text{ni levkemije,samec}} = O_{\text{samec}} - E_{\text{levkemija,samec}}$=50-20.5=29.5
$E_{\text{ni levkemije,samica}} = O_{\text{samica}} - E_{\text{levkemija,samica}}$=50-20.5=29.5

Izračunane pričakovane frekvence so:

	Samec	Samica	Vsota
Levkemija	20.5	20.5	41
Ni levkemije	29.5	29.5	59
Vsota	50	50	100

Testna statistika hi-kvadrat se izračuna po formuli:

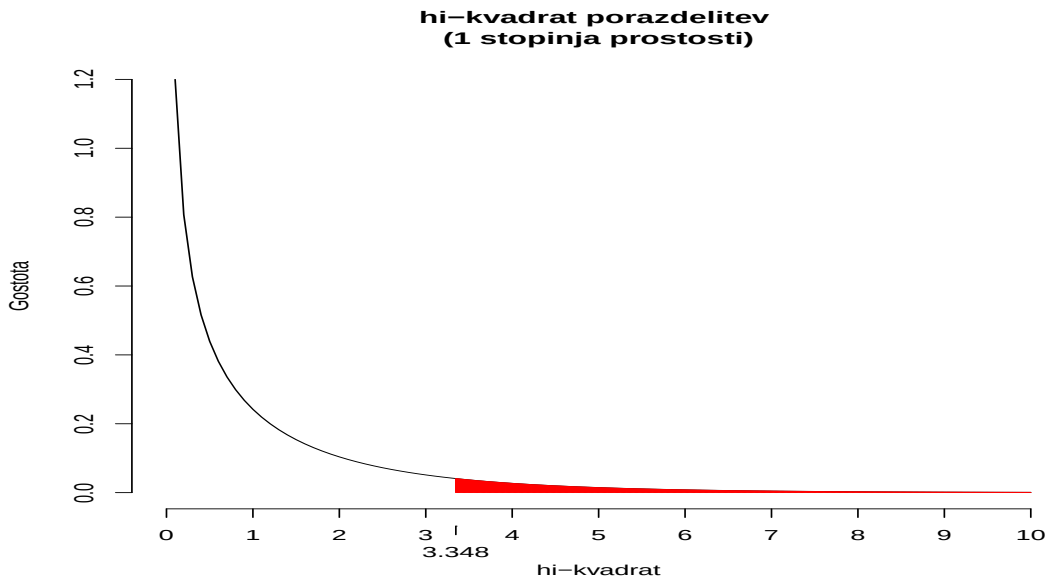
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}};$$

indeksi i in j določajo vrstice (r) oziroma stolpce (c) kontingenčne tabele.

V našem primeru je vrednost testne statistike
$\chi^2 = \frac{(25-20.5)^2}{20.5} + \frac{(16-20.5)^2}{20.5} + \frac{(25-29.5)^2}{29.5} + \frac{(34-29.5)^2}{29.5} =$
$\frac{20.25}{20.5} + \frac{20.25}{20.5} + \frac{20.25}{29.5} + \frac{20.25}{29.5} = 3.348$

Testna statistika je ob veljavni ničelni domnevi porazdeljena po hi-kvadrat porazdelitevi z eno stopinjo prostosti (df, *degrees of freedom*). Stopinje prostosti določamo glede na število vrstic (r) in stolpcev (c) kontingenčne tabele. Ker ima v našem primeru kontingenčna tabela 2 vrstici in 2 stolpca ($r = c = 2$), so stopinje prostosti $df = (r-1)(c-1) = (2-1)(2-1) = 1$. Hi-kvadrat porazdelitev z eno stopinjo prostosti, X_1^2, je prikazana na sliki 4.1. Na sliki 4.1 sta prikazani tudi vrednost izračunane testne statistike in p-vrednost (grafično, pobarvana površina).

P-vrednost je verjetnost, da ob veljavni ničelni domnevi dobimo vrednost testne statistike, ki je večja ali enaka tisti, ki smo jo opazili(p=P($X_1^2 \geq \chi^2|H_0$)). Vrednost določamo s pomočjo statističnega programa oziroma s pomočjo statističnih tabel. V našem primeru je p=0.067. Kritična vrednost za hi-kvadrat



Slika 4.1: *Hi-kvadrat porazdelitev z eno stopinjo prostosti, vrednost izračunane testne statistike in p-vrednost (površina označena z rdečo barvo).*

porazdelitev z eno stopinjo prostosti pri stopnji tveganja $\alpha = 0.05$ je 3.84; če je vrednost testne statistike večja od 3.84 zavrnemo ničelno domnevo. Kritično vrednost odčitamo iz statistične tabele za hi-kvadrat: izberemo vrstico **df=1** in stolpec $\alpha = 0.05$) in najdemo kritično vrednost na presečišču.
NB: v tem primeru smo s tem postopkom že upoštevali, da je test *dvostranski*, ker so vsa odstopanja testne statistike od 0 pozitivna, torej gledamo samo en (desni) rep porazdelitve.

Izpis iz programa R:

```
        Pearson's Chi-squared test

data:  my.table.q1p
X-squared = 3.3485, df = 1, p-value = 0.06727
```

Če bi opazili, 27 (namesto 25) samcev z levkemijo in 10 (namesto 16) samic z levkemijo, bi pričakovali, da bi bila p-vrednost manjša Izračuna ni potrebno ponoviti, ker bi ob isti velikosti vzorca imeli trdnejši dokaz, da obstaja povezanost med spolom in levkemijo.

2. Stupica in soavtorji (Vector-Borne and Zoonotic Diseases, 2010) so raziskovali povezanost med pozitivno borelijsko kulturo kože (osamitev *Borrelia burgdorferi sensu lato* iz kože) in drugimi dejavniki pri pacientih z eritema migrans (EM). V raziskavo so vključili 252 zaporednih pacientov z EM in iz vzorcev kože skušali osamiti borelije. Med pacienti je bilo 60 žensk. Pri 69 bolnikih so iz kože osamili borelije (pozitivni bolniki), pri 31 bolniku borelij niso osamili (negativni bolniki). Med pozitivnimi pacienti je bilo 42 žensk. Za ugotavljanje povezanosti med spolom in pozitivnostjo borelijske kulture kože so uporabili test hi-kvadrat z Yatesovim popravkom za zveznost. Dobili so naslednje rezultate (izpis iz statističnega programa R).

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

data:  Borrelia and Sex
X-squared = 0.0019, df = 1, p-value = 0.9648
```

(a) Kako se razlikuje hi-kvadrat test z Yatesovim popravkom od navadnega testa hi-kvadrata? Zakaj ga uporabljamo?

(b) *Ali pričakujete, da bo p-vrednost večja ali manjša od tiste, ki bi jo dobili s testom hi-kvadrat?

(c) Napišite kontingenčno tabelo z opazovanimi frekvencami.

(d) Interpretirajte rezultate.

(e) Katere mere lahko izračunamo, da bi ovrednotili povezanost med spolom in pozitivno borelijsko kulturo? Izračunajte tri različne mere povezanosti (s 95% intervalom zaupanja).

(f) ** Avtorji niso posebno poudarili tega rezultata v članku (povezanost med spolom in pozitivno borelijsko kulturo); zaključili so celo, da so pozitivni in negativni pacienti primerljivi glede vseh značilnosti, ki so jih merili ob vključitvi v raziskavo. Kaj bi lahko bil razlog za tak zaključek?

(g) * Katero mero povezanosti bi izračunali, če bi avtorji vključili v raziskavo 100 pacientov s pozitivno borellijsko kulturo in 100 pacientov z negativno borelijsko kulturo, ter si ogledali, ali sta spol in pozitivnost povezani?

(h) * Kako sta povezana razmerje obetov in relativno tveganje?

Rešitve

(a) **Hi-kvadrat test z Yatesovim popravkom** se razlikuje od navadnega testa hi-kvadrata pri izračunu testne statistike. Testna statistika je

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_i};$$

ničelna domneva in porazdelitev testne statistike sta enaki kot pri hi-kvadratu testu. Navadni hi-kvadrat test ne da točnih rezultatov, ko so pričakovane frekevence majhne; p-vrednosti so lahko premajhne. Z Yatesovim popravkom izboljšamo zanesljivost rezultatov tudi pri majhnih vzorcih.

(b) * p-vrednost je večja če uporabimo Yatesov popravek, ker je testna statistika manjša (pri izračunu odštejemo 0.5 od vsake vrednosti). Razlika je bolj izrazita pri manjših vzorcih.

(c) Spodaj je podana kontingenčna tabela z opazovanimi frekvencami.

Borrelia	Ženski	Moški	Vsota
Pozitivni	a=42	b=27	a+b=69
Negativni	c=18	d=13	c+d=31
Vsota	a+c=60	b+d=40	N=100

(d) Zavrnemo ničelno domnevo, da v populaciji bolnikov z EM ni povezanosti med spolom in pozitivno borelijsko kulturo; opazili smo, da imajo ženske večjo verjetnost, da so pozitivne (P_F^+=0.7, za moške pa P_M^+=0.68).

(e) Poleg p-vrednosti moramo podati tudi neko mero povezanosti med spremenljivkama, s katero bomo ovrednotili povezanost med spolom in pozitivno borelijsko kulturo. To je zelo pomembno, ker tako lahko sporočimo ne samo, ali obstaja neka značilna povezanost med spremenljivkama, ampak tudi *kolikšna* je povezanost. Možne mere, ki jih lahko podamo so:

- razlika deležev;
- relativno tveganje (*RR, relative risk*);
- razmerje obetov (*OR, odds ratio*).

$$p = P(Borrelia = +) = \frac{a+b}{N} = \frac{69}{100} = 0.69$$

.

$$p_F = P(Borrelia = +|Spol = F) = \frac{a}{a+c} = \frac{42}{42+18} = 0.7$$

$$p_M = P(Borrelia = +|Spol = M) = \frac{b}{b+d} = \frac{27}{27+13} = 0.68$$

$$n_M = b + d = 27 + 13 = 40, n_F = 42 + 18 = 60$$

Razlika deležev
Ocena: $p_F - p_M$. **Standardna napaka:**

$$\hat{SE}(p_F - p_M) = \sqrt{var(p_F) + var(p_M)} = \sqrt{\frac{p_F(1-p_F)}{n_F} + \frac{p_M(1-p_M)}{n_M}}.$$

95% interval zaupanja za razliko deležev: od $(p_F - p_M) - 1.96 \times \hat{SE}(p_F - p_M)$ do $(p_F - p_M) + 1.96 \times \hat{SE}(p_F - p_M)$, ker je razlika deležev normalno porazdeljena (velja, če je vzorec *velik*, oziroma če $N(p_F - p_M)$ in $N(1 - (p_F - p_M)) > 5$).
V našem primeru je razlika deležev:

$$p_F - p_M = 0.7 - 0.68 = 0.02$$

in standardna napaka je

$$\hat{SE}(p_F - p_M) = \sqrt{\frac{0.7 \times 0.3}{60} + \frac{0.68 \times 0.32}{40}} = 0.09.$$

95% interval zaupanja za razliko deležev pozitivnih pacientov glede spola je: $0.02 - 1.96 \cdot 0.09$ do $0.02 + 1.96 \cdot 0.09$, oziroma -0.16 do 0.21.

Relativno tveganje
Ocena:

$$RR = \frac{p_F}{p_M}.$$

Naravni logaritem od RR ($\ln(RR)$) je normalno porazdeljen in **standardna napaka od** $\ln(RR)$ **je:**

$$\hat{SE}(ln(\frac{p_F}{p_M})) = \sqrt{\frac{1}{a} - \frac{1}{a+c} + \frac{1}{b} - \frac{1}{b+d}}.$$

95% interval zaupanja za $\ln(RR)$ je:

$$\ln(RR) - 1.96\hat{SE}(\ln(RR)) \text{do} \ln(RR) + 1.96\hat{SE}(\ln(RR)).$$

95% interval zaupanja za RR je:

$$e^{\ln(RR)-1.96\hat{SE}(\ln(RR))} \text{ do } e^{\ln(RR)+1.9\hat{SE}(\ln(RR))},$$

$e = 2.718$ je osnova naravnega logaritma, Eulerjevo številko.

V našem primeru je relativno tveganje:

$$RR = \frac{p_F}{p_M} = \frac{0.7}{0.68} = 1.04,$$

in $\ln(RR) = 0.04$. Standardna napaka od $\ln(RR)$ je

$$\hat{SE}(ln(\frac{p_F}{p_M})) = \sqrt{\frac{1}{42} - \frac{1}{42+18} + \frac{1}{27} - \frac{1}{27+13}} = 0.14.$$

95% interval zaupanja za $\ln(RR)$ je od $0.04-1.96\cdot0.14$ do $0.04+1.96\cdot0.14$, oziroma od -0.24 do 0.31. 95% interval zaupanja za RR je: $e^{-0.24}$ do $e^{0.31}$, oziroma od 0.79 do 1.36.

Razmerje obetov
Ocena:

$$OR = \frac{\frac{p_F}{1-p_F}}{\frac{p_M}{1-p_M}}.$$

Naravni logaritem od OR ($\ln(OR)$) je normalno porazdeljen in **standardna napaka** $\ln(OR)$ **je:**

$$\hat{SE}(ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{d} + \frac{1}{c} + \frac{1}{d}}.$$

95% interval zaupanja za $\ln(OR)$ je:

$$\ln(OR) - 1.96\hat{SE}(\ln(OR)) \text{ do } \ln(OR) + 1.96\hat{SE}(\ln(OR)).$$

95% interval zaupanja za OR je:

$$e^{\ln(OR)-1.96\hat{SE}(\ln(OR))} \text{ do } e^{\ln(OR)+1.96\hat{SE}(\ln(OR))}.$$

V našem primeru je razmerje obetov:

$$OR = \frac{\frac{p_F}{1-p_F}}{\frac{p_M}{1-p_M}} = \frac{\frac{0.7}{1-0.7}}{\frac{0.68}{1-0.68}} = 1.12 = \frac{ad}{bc},$$

in $\ln(OR) = 0.12$. Standardna napaka od $\ln(OR)$ je

$$\hat{SE}(lnOR) = \sqrt{\frac{1}{42} + \frac{1}{27} + \frac{1}{18} + \frac{1}{13}} = 0.44.$$

95% interval zaupanja za $\ln(OR)$ je od $0.12-1.96\cdot0.44$ do $0.12+1.96\cdot0.44$, oziroma od -0.75 do 0.98. 95% interval zaupanja za OR je: $e^{-0.75}$ do $e^{0.98}$, oziroma od 0.47 do 2.66.

(f) Poleg spola so avtorji testirali povezanost med Borrelio in 24 drugimi spremenljivkami (starost, prisotnost drugih bolezni, prisotnost simptomov, itd, glej tabelo 1 v članku). Verjetnost, da bi vsaj ena spremenljivka imela p-vrednost<0.05 v primeru, da ne bi bilo nobene povezanosti, je bila 0.72 (glej nalogo 5 v poglavju Verjetnost). P-vrednosti vseh ostalih testiranih spremenljivk so bile večje od 0.05. Zato avtorji niso dali posebnega pomena temu rezultatu in so trdili, da nobena spremenljivka ni značilno povezana s pozitivno borelijsko kulturo.

(g) * V tem primeru bi avtorji izvedli raziskavo primerov in kontrol (*case control study*). Za tovrstne raziskave je edina veljavna mera povezanosti razmerje obetov, saj ta mera ohranja vrednost kljub spreminjanju razmerja med številom primerov in kontrol. Kot primer, si lahko ogledate naslednji tabeli, kjer smo v drugem primeru (tabela (b)) vzeli 10 krat več pozitivnih pacientov. OR se ne spremeni, medtem ko se RR in razlika deležev spremenita.

Tabela (a)	Ženski	Moški	Vsota
Poz	a	b	a+b
Neg	c	d	c+d
Vsota	a+c	b+d	

Tablela (b)	Ženski	Moški	Vsota
Poz	$10 \cdot a$	$10 \cdot b$	$10 \cdot a + 10 \cdot b$
Neg	c	d	$c + d$
Vsota	$10 \cdot a + c$	$10 \cdot b + d$	

$p_F^{(b)} = \frac{10a}{10a+c} \neq \frac{a}{a+c} = p_F^{(a)}$;
$p_M^{(b)} = \frac{10b}{10b+d} \neq \frac{b}{b+d} = p_M^{(a)}$;
$RR^{(b)} = p_F^{(b)}/p_M^{(b)} = \frac{10a/(10a+c)}{10b/(10b+d)} \neq RR^{(a)} = \frac{a/(a+c)}{b/(b+d)}$
$OR^{(b)} = \frac{10a \times d}{10b \times c} = \frac{ad}{bc} = OR^{(a)}$.

(h) * Kako sta povezana razmerje obetov in relativno tveganje?

$$RR = \frac{p_F}{p_M} = \frac{\frac{p_F(1-p_F)}{1-p_F}}{\frac{p_M(1-p_M)}{1-p_M}} = OR\frac{1-p_F}{1-p_M}$$

RR in OR sta si med sabo zelo podobni, če sta p_F in p_M zelo majhni.

3. Raziskovali smo povezanost med načinom uživanja tobaka oz. alkohola in lokacijo raka ustne votline. Dobili smo naslednje podatke

	Jezik	Drugo
Žvečenje	146	267
Kajenje	71	166
Alkohol	51	71

(a) Kateri statistični test lahko uporabite, da bi ugotovili ali obstaja povezanost?

(b) Ali sta spremenljivki povezani?

Namig: pazite na stopinje prostosti ($df = (r-1)(c-1)$) in na kritično vrednost.

4. Med študenti, ki so med letoma 2008 in 2010 obiskovali predmet biostatistika, je na anketo odgovorilo 120 deklet in 48 fantov. Videoigrice je igralo 30% deklet in 58.33% fantov. Ali sta spol in igranje videoigric povezana (oziroma, ali sta populacijska deleža deklet in fantov, ki igrajo igrice, enaka?)?

5. Ogledali smo si povezanost med kajenjem (da ali ne) in lastništvom domačih živali. Dobili smo naslednje rezultate

	Živali - Ne	Živali - Da
Kajenje - Ne	24	125
Kajenje - Da	1	18

Ali lahko v tem primeru uporabimo hi-kvadrat test, da bi ugotovili, ali obstaja povezanost med kajenjem in lastništvom domačih živali? Odgovor utemeljite.

Rešitev

Predpostavka za uporabo χ^2 testa ni izpolnjena, saj je ena pričakovana frekevenca manjša od 5. Uporabimo **Fisherjev eksatni test**.

	Živali - Ne	Živali - Da
Kajenje - Ne	22.17	126.83
Kajenje - Da	2.83	16.17

Tabela 4.1: Pričakovane frekvence

```
        Fisher's Exact Test for Count Data

data:  my.table.o
p-value = 0.3133
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.494003 149.822531
sample estimates:
odds ratio
  3.438398
```

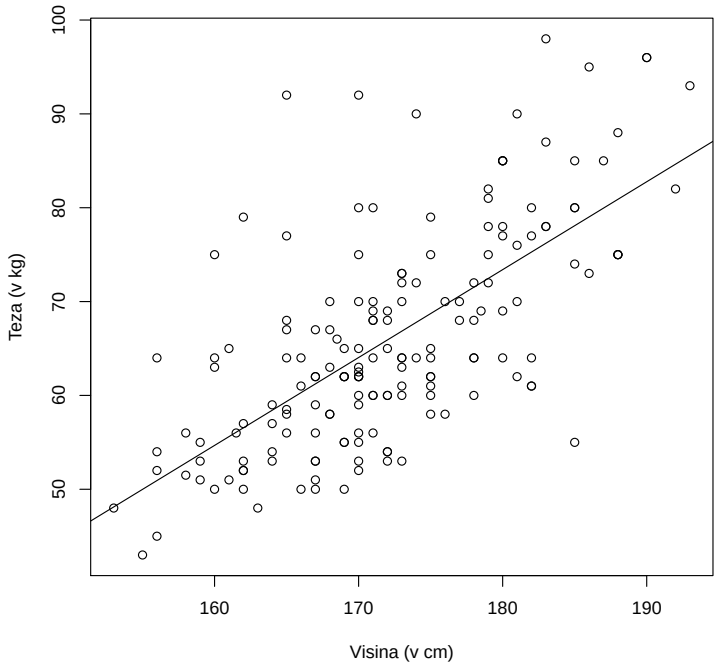
Na podlagi Fisherjevega eksaktnega testa ne moremo zavrniti ničelne domneve, da ni povezanosti med kajenjem in lastištvom domačih živali (P=0.313).

Poglavje 5

Povezanost med številskimi spremenljivkami

1. Proučevali smo povezanost med višino (v centimetrih) in težo (v kg). V vzorec smo zajeli 168 ljudi; podatki so prikazani na sliki 5.1. Rezultati statistične obdelave podatkov so prikazani v tabeli.

   Vrednost determinacijskega koeficienta je ($R^2 = 0.45$).



Slika 5.1: *Grafična predstavitev podatkov.*

	Ocena	Standardna napaka	Testna statistika t	p-vrednost
konstanta	-95.17	13.82	-6.89	<0,0001
Višina	0.94	0.08	?	?

Vprašanja

(a) Katero statistično metodo smo uporabili, da bi preučili povezanost med težo in višino?

(b) Katero spremenljivko smo izbrali kot odvisno in katero kot neodvisno?

(c) Napišite enačbo ocenjenega modela in skicirajte ocenjeno premico.

(d) Kakšna je vrednost testne statistike t za regresijski koeficient za višino?

(e) Kakšna je porazdelitev testne statistike za regresijski koeficient za višino, če velja ničelna domneva?

(f) Kakšna je p-vrednost za regresijski koeficient za višino?

(g) Ali je teža statistično značilno povezana z višino?

(h) Katere vrste povezanosti testiramo s tem modelom?

(i) Kako interpretiramo vrednost regresijskega koeficienta za višino?

(j) Kako interpretiramo regresijsko konstanto?

(k) Izračunajte 95% interval zaupanja za regresijski koeficient za višino.

(l) Kakšno težo pričakujemo za osebo, ki je visoka 175 cm? (Oziroma, kolikšna je ocenjena povprečna teža oseb, ki so visoke 175 cm?)

(m) Kakšno razliko v teži pričakujemo pri dveh osebah, katerih višina se razlikuje za 8 cm?

(n) Koliko variabilnosti teže smo pojasnili s tem modelom? Kaj to pomeni?

(o) Ocenite Pearsonovo korelacijo med težo in višino.

(p) Kakšno vrednost bi dobili za oceno regresijskega koeficienta za višino, če bi višino namesto v centimetrih merili v metrih?

Rešitve

(a) Uporabili smo univariatno linearno regresijo.

(b) Odvisna spremenljivka je teža in neodvisna spremenljivka je višina.

(c) Enačba ocenjenega modela je: $Tea = a + b \cdot Viina$, če označimo z a ocenjeno vrednost konstante in z b ocenjeno vrednost regresijskega koeficienta za višino. V tem primeru je enačba ocenjenega modela: $Tea = -95.17 + 0.94 \cdot Viina$. Ocenjena premica je prikazana na sliki 5.1.

(d) Testno statistiko t za regresijski koeficient za višino izračunamo po formuli $t = \frac{\text{Ocena}}{\text{Standardna napaka}} = \frac{0.94}{0.08} = 11.75$.

(e) Testna statistika za regresijski koeficient za višino je ob predpostavljeni ničelni domnevi porazdeljena po t porazdelitvi z df=168-2=166 stopinjami prostosti. Ta porazdelitev je zelo podobna standardni normalni porazdelitvi, ker je veliko stopinj prostosti.

(f) P-vrednost je verjetnost, da ob veljavni ničelni domnevi dobimo vrednost testne statistike, ki je od 0 oddaljena bolj od izračunane vzorčne vrednosti. Ponavadi za izračun p-vrednosti uporabljamo statistične programe. Lahko jo izračunamo tudi sami s pomočjo statistčnih tabel. $p = P(t_{df} < -|t| \,|H_0) + P(t_{df} > |t| \,|H_0) = 2 \times P(t_{df} > |t| \,|H_0)$.
V našem primeru je $p = 2 \times P(t_{166} > 11.75) < 0,001$
Kritična vrednost pri stopnji tveganja 0.05 je vrednost $t_{df,\frac{0.05}{2}}$ iz t_{df} porazdelitve za katero velja: $p = 2 \times P(t_{df} > t_{df,\frac{0.05}{2}}) = 0.05$.
V našem primeru je kritična vrednost $t_{df,\frac{0.05}{2}}$=1.97, ker velja, da $p =$

$2P(t_{166} > 1.97) = 0.05$. Če poznamo kritično vrednost, lahko preko primerjave izračunane testne statistike s kritično vrednostijo ugotivimo, ali je p-vrednost manjša ali večja od stopnje tveganja. Na primer, če izberemo stopnjo tveganja $\alpha = 0.05$ in je absolutna vrednost testne statistike večja od kritične vrednosti $t_{df,\frac{0.05}{2}}$, potem bo p-vrednost manjša od 0.05 in obratno. Kritično vrednost uporabljamo tudi pri izračunu intervala zaupanja.

(g) Višina je statistično značilno povezana s težo, ker je p-vrednost za regresijski koeficient višine manjši od 0.05.

(h) S tem modelom testiramo samo linearno povezanost.

(i) Vrednost regresijskega koeficienta za višino je naklon ocenjene premice. V praksi to pomeni, da če primerjamo dve osebi, ki imata 1 cm razlike v višini, pričakujemo, da bo ocenjena razlika v teži 0.94 kg.

(j) Pomen konstante je v tem primeru povpečna teža oseb, ki imajo 0 cm višine. V tem primeru ni smiselno interpretirati vrednosti konstante, ker takih oseb ni (in jih tudi nismo vključili v vzorec).

(k) 95% interval zaupanja za regresijski koeficient za višino v populaciji izračunamo po tej formuli:
od $b - t_{df,\frac{0.05}{2}} * \hat{SE}$ do $b - t_{df,\frac{0.05}{2}} * \hat{SE}$.
V našem primeru je 95% interval zaupanja za regresijski koeficient za višino v populaciji od 0.94-1.97*0.08 do 0.94+1.97*0.08, oziroma od 0.78 do 1.1. Imamo torej 95% zaupanje, da je neznani regresijski koeficient višine vsebovan v intervalu od 0.78 in 1.1.

(l) Ocenjeno povprečno težo oseb, ki so visoke 175 cm, določimo tako, da vstavimo vrednost 175 v enačbo ocenjenega modela. $Tea = a + b * 175$, v našem primeru je $-95.17 + 0.94 * 175 = 69.33$ kg.

(m) Pri dveh osebah, katerih višina se razlikuje za 8 cm, pričakujemo $8 * b = 8 * 0.94 = 7.52$ kg razlike v teži.

(n) S tem modelom smo pojasnili 45% variabilnosti teže, ker je determinacijski koeficient $R^2 = 0.45$.

(o) Pri univariatni linearni regresiji je Pearsonova korelacija med odvisno in neodvisno spremenljivko (težo in višino v našem primeru) koren determinacijskega koeficienta. Torej v tem primeru $r = \sqrt{R^2} = \sqrt{0.45} = 0.67$.

(p) Če bi višino namesto v centimetrih merili v metrih bi bil ocenjeni model:

	Ocena	Standardna napaka	Testna statistika t	p-vrednost
konstanta	-95.17	13.82	-6.89	<0,0001
Višina	93.65	8.02	11.68	<0,0001

Vidimo, da se ocenjene vrednosti za konstanto ne spremenijo, medtem ko za regresijski koeficient in za njegovo standardno napako dobimo vrednosti, ki so 100 krat večje od tistih iz modela, kjer je bila višina izmerjena v centerimetrih (1m=100cm); testna statistika in p-vrednost se ne spreminjata (ker $t = \frac{100*b}{100*\hat{SE}} = \frac{b}{\hat{SE}}$). Seveda se tudi determinacijski koeficient ne spreminja.

2. V poskusu smo želeli primerjati učinek različnih doz zdravila (od 10 do 25 mg) pri zdravljenju revmatoidnega artritisa. Izid našega poskusa je povzet v indeksu učinkovitosti, ki meri stopnjo ublažitve različnih simptomov (večja vrednost indeksa pomeni bolj učinkovito zdravljenje). V poskus smo zajeli 80

bolnikov, z modelom linearne regresije smo dobili naslednje rezultate:

Parameter	B	Std. napaka	Wald	p
konstanta	-6.3	4.30	-1.5	0.148
doza	3.6	0.08	45.3	<0.001

(a) Natančno zapišite ničelno domnevo, ki jo preverjamo.

(b) Skicirajte odnos med dozo zdravila in učinkovitostjo zdravljenja.

(c) Ali je vpliv doze statistično značilen pri stopnji tveganja 0.05? Obkrožite ustrezni podatek v zgornji tabeli, ki to pove.

(d) Izračunajte pričakovano razliko med indeksoma učinkovitosti, če je pacient A prejemal 5mg večjo dozo kot pacient B. Pri katerem pacientu pričakujete večji indeks učinkovitosti?

(e) Izračunajte pričakovani indeks učnkovitosti za pacienta, ki je prejemal dozo 15mg.

(f) Ali lahko rezultate posplošimo na situacijo, ko zdravila ne dajemo (doza = 0 mg)?

(g) Izračunajte 95% interval zaupanja za koeficient za dozo.

3. Napovedati želimo oceno izpita iz statistike na podlagi števila tedenskih ur posvečenih učenju. Grafično predstavite (izmišljene) podatke in na podlagi le-teh izpolnite tabelo.

	Ocena	Standardna napaka	Testna statistika t	p-vrednost
konstanta				
Ure				

4. Na sliki 5.2 so prikazani podatki, ki smo jih opazili pri 4 različih vzorcih. Za vsak vzorec določite, kateri rezultat smo dobili z uporabo linearne regresije.

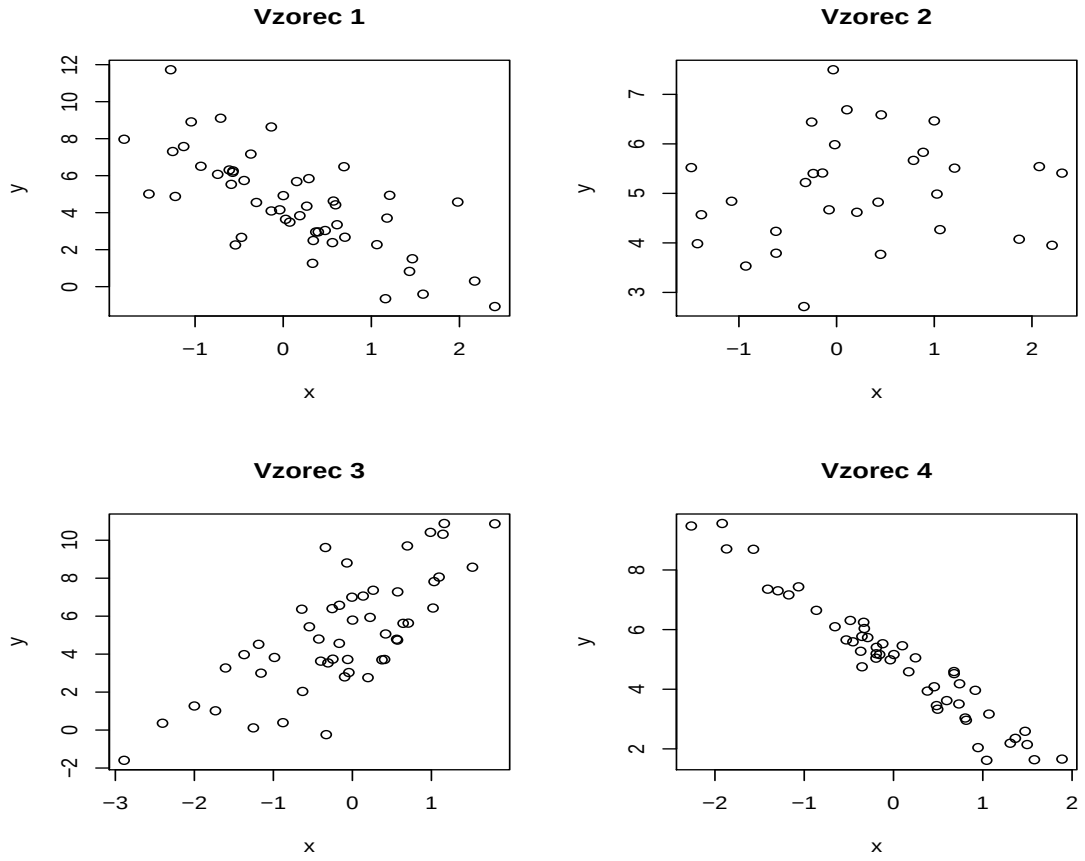
Vzorec ?				
Parameter	Ocena	Std. napaka	t	p-vrednost
(Konstanta)	5.399	0.286	18.859	<0.001
x	2.327	0.289	8.059	<0.001

Vzorec ?				
Parameter	Ocena	Std. napaka	t	p-vrednost
(Konstanta)	5.034	0.205	24.568	<0.001
x	0.137	0.193	0.71	0.484

Vzorec ?				
Parameter	Ocena	Std. napaka	t	p-vrednost
(Konstanta)	4.999	0.07	71.171	<0.001
x	-2.001	0.073	-27.413	<0.001

Vzorec ?				
Parameter	Ocena	Std. napaka	t	p-vrednost
(Konstanta)	4.694	0.259	18.122	<0.001
x	-1.993	0.268	-7.436	<0.001

5. Primerjati želimo povprečno višino študentov, ki igrajo videoigrice s povrečno višino študentov, ki ne igrajo videoigric. Zbrali smo podatke med leti 2008



Slika 5.2: *Razsevni diagram podatkov opaženih pri 4 vzorcih).*

ter 2010 in analizirali podatke s t-testom za neodvisna vzorca ob predpostavki enakih varianc. Dobili smo naslednje rezultate.

```
        Two Sample t-test

data:  visina by igrice
t = 2.6677, df = 166, p-value = 0.008394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9084266 6.0819580
sample estimates:
   mean in group Igra mean in group Ne igra
             174.2500              170.7548
```

- Koliko oseb smo vključili v vzorec?
- Interpretirajte rezultate.
- Napišite enačbo ocenjenega regresijskega modela.
- Kako bi lahko lahko razložili ta rezultat?

** Analizo ponovimo z uporabo linearne regresije in dobimo iste rezultate. V tem primeru smo kodirali igranje videoigrica z 1 in ne-igranje z 0. Izid je višina, neodvisna spremenljivka pa je igranje z videoigricami. Primerjajte rezultate

t-testa in linearne regresije. Kako lahko interpretiramo konstanto? Kako pa koeficient za igrice?

```
Call:
lm(formula = my.data.2$Visina ~ igrice)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2500  -4.7548  -0.7548   5.2452  21.2452

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 170.7548     0.8087 211.157 < 0.0000000000000002 ***
igrice        3.4952     1.3102   2.668              0.00839 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.247 on 166 degrees of freedom
Multiple R-squared:  0.04111,        Adjusted R-squared:  0.03533
F-statistic: 7.117 on 1 and 166 DF,  p-value: 0.008394
```

V model smo vključili poleg igranja tudi spol. Kodirali smo moški spol z 1 in ženski spol z 0.

```
Call:
lm(formula = visina ~ igrice + sex)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2358  -3.2358   0.7148   3.7642  16.7642

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) 168.2358     0.6162  273.00 <0.0000000000000002 ***
igrice        0.2835     0.9778    0.29               0.772
sex          13.0989     1.0511   12.46 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.937 on 165 degrees of freedom
Multiple R-squared:  0.5061,        Adjusted R-squared:  0.5001
F-statistic: 84.52 on 2 and 165 DF,  p-value: < 0.00000000000000022
```

Komentirajte rezultate modela. Kako lahko interpretiramo konstanto v tem primeru? Kako koeficient za igrice? In koeficient za spol? Kateri model pojasni več variabilnosti?

Oglejte si še naslednji model, kjer smo vključili tudi težo in ga interpretirajte.

```
Call:
lm(formula = visina ~ igrice + spol + teza)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.7888  -3.2818   0.5107   3.8433  18.2243

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 152.21033    2.92491  52.039 < 0.0000000000000002 ***
igrice       -0.37035    0.90656  -0.409                0.683
spol          8.89353    1.22510   7.259      0.0000000000148 ***
teza          0.26482    0.04742   5.585      0.0000000952966 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.458 on 164 degrees of freedom
Multiple R-squared:  0.585,        Adjusted R-squared:  0.5774
F-statistic: 77.06 on 3 and 164 DF,  p-value: < 0.00000000000000022
```

Dodatek A

Rešitve

A.1 Opisna statistika

4.

(a) Grafikon se imenuje škatla z brki (*boxplot* ali okvir z ročaji ali *box and whiskers plot*).

(b) Odebeljena črta v sredini škatle predstavlja mediano.

(c) Spodnji rob škatle predstavlja 1. kvartil (25. percentil) in zgornji rob predstavlja 3. kvartil (75. percentil). Višina škatle je torej interkvartilni razmik (IQR, *interquartile range*). Zgornja in spodnja črtica pri ročaju sta v tem primeru najmanjša in največja vrednost.

V primeru, da je kakšna vrednost v vzorcu oddaljena od posameznega roba škatle za več kot 1.5 interkvartilnega razmika, je vsaka taka vrednost prikazana s posamezno točko, črtica pri ročaju pa predstavlja to mejno vrednost (3. kvartil + $1.5 \times IQR$ za zgornjo črtico v primeru, da so *skrajne vrednosti* med največjimi, ali 1. kvartil - $1.5 \times IQR$ za spodnjo črtico v primeru, da so *skrajne vrednosti* med najmanjšimi). Pogosto se *skrajne vrednosti* imenujemo osamelci (*outliers*).

(d) Porazdelitev podatkov za študente, ki so obiskovali predavanja dvakrat ali manj, je podana s prvim grafikonom, tista za študente, ki so obiskovali predavanja 5 ali 6 krat pa s tretjim.

	<25%	25–49%	50–74%	≥75%
% študentov, ki niso opravili izpita (manj kot 40 točk)		X		
% študentov, ki niso dosegli več kot 50 točk			X	
% študentov, ki niso dosegli 30 točk	X			

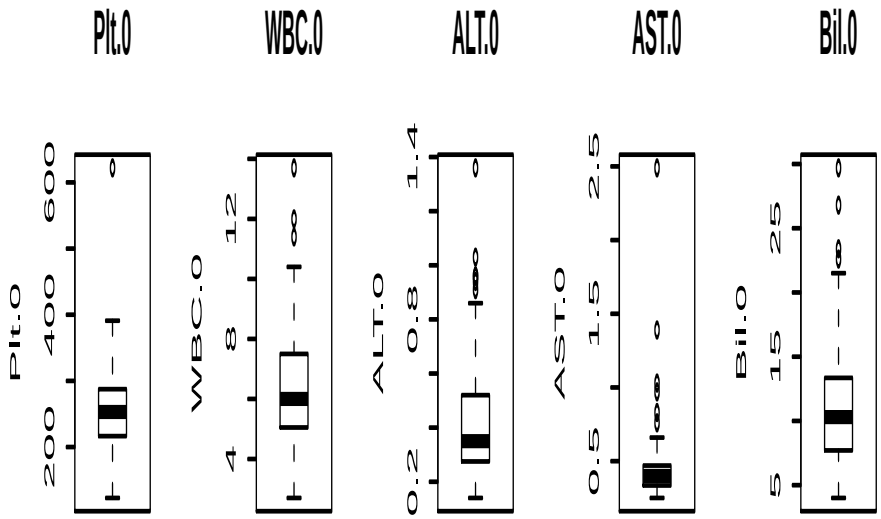
5.

(a) Min je najmanjša vrednost; 1st Qu. je prvi kvartil; Median je mediana; Mean je aritmetično povprečje; 3rd Qu. je tretji kvartil; Max. je največja vrednost; NAs je število manjkajočih podatkov (*Not Available*); SD je standardni odklon (*Standard Deviation*).

(b) Grafična predstavitev z okvirom z ročaji.

(c) Vrstni red je:

	Največ dvakrat	5 ali 6 krat
Najmanjše število točk	5	10
Največje število točk	58	74
Povprečno število točk	Ne moremo določiti	Ne moremo določiti
Mediansko število točk	32	40
Standardni odklon	Ne moremo določiti	Ne moremo določiti
Velikost vzorca	Ne moremo določiti	Ne moremo določiti
Najpogostejši rezultat	Ne moremo določiti	Ne moremo določiti
Razpon	5 do 58	10 do 74
Interkvartilni razmik	20 do 42	34 do 55



Slika A.1: *Grafična predstavitev podatkov - okvir z ročaji.*

```
[1] "WBC.0" "Bil.0" "AST.0" "ALT.0" "Plt.0"
```

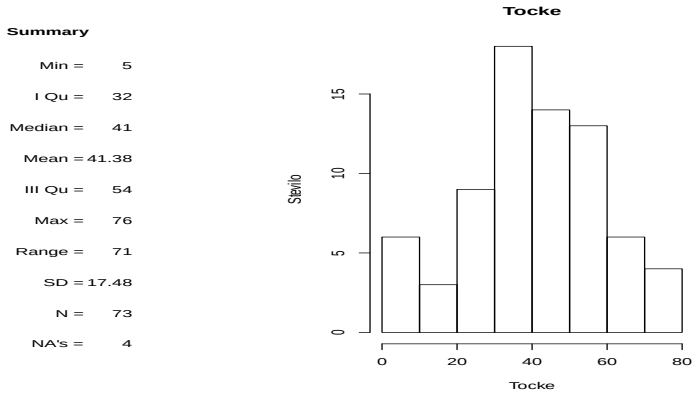
(d) Višina stolpca histograma pove, koliko statističnih enot ima vrednost spremenljivke, ki je vključena v intervalu, na katerem se stolpec nanaša. Na primer, 54 oseb ima vrednost ALT med 0.2 in 0.4 (slika 1.5). (Ponavadi je spodnja vrednost intervala vključena in zgornja vrednost izključena; v tem primeru gledamo interval $[0.2, 0.4)$).

(e) Histogram in okvir z ročaji po

Primerjajte grafični prikaz spremenljivke ALT na sliki 1.5. Kateri grafični prikaz se vam zdi bolj učinkovit? Katere informacije nam daje vsak graf?

(f) Primerjajte grafični prikaz spremenljivke ALT na sliki 1.6, kjer je porazdelitev prikazana posebej za ženke in za moške. Kateri grafični prikaz se vam zdi bolj učinkovit za neposredno primerjavo spremenljivke ALT glede spola?

6. Podatke lahko grafično prikažemo s histogramom (slika A.2). V tem primeru ne uporabimo stolpičnega diagrama kot v nalogi 2, saj je število doseženih točk na izpitu številska (razmernostna) spremenljivka.



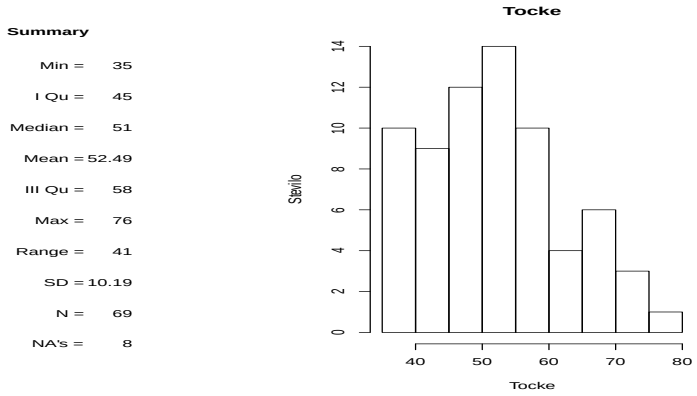
Slika A.2: *Dosežene točke na izpitu.*

Porazdelitev je približno simetrična in normalno porazdeljena. Povrečje normalne porazdelitve (μ) je v sredini porazdelitve (je enako mediani in modusu). Za normalno porazdeljene spremenljivke standardni odklon (σ) približno ocenimo tako, da določimo, v katerem simetričnem intervalu okrog povprečja je približno 95% opazovanj. Spodnja in zgornja meja tega intervala sta približno: $\mu - 2\sigma$ in $\mu + 2\sigma$ (bolj natančno: morali bi množiti σ z 1.96). Na podlagi grafikona približno ocenimo, da je aritmetično sredina za število točk doseženih na izpitu 40 in standardni odklon približno 15 točk, ker je večina oprazovanj med 10 in 70 točkami. Dejanske vrednosti so: aritmetično povprečje, $\bar{x}$ =41.38, standardni odklon: s =17.48. Mediana je Me=41.

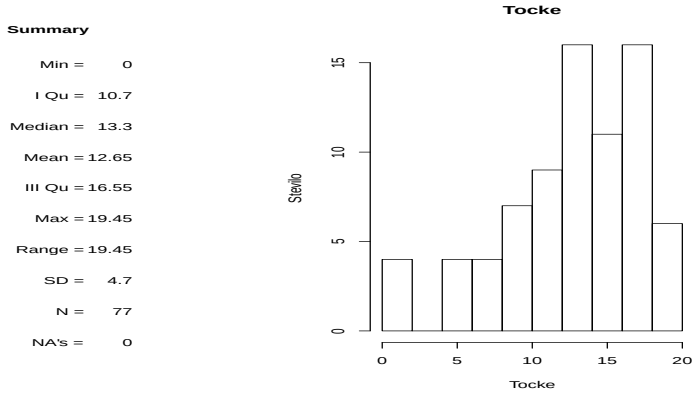
7.

(a) Število točk na izpitu. V tem primeru so bile vse vrednosti večje ali enake od 40 točk, ker smo vključili v vzorec samo pozitivno opravljene izpite, $\bar{x}$ =52.5 in s =10.2 točk. Če bi bili podatki normalno porazdeljeni bi pričakovali, da je približno 95% vseh študentov doseglo rezultat med $52.5 - 2 \cdot 10.2$=32.1 in $52.5 + 2 \cdot 10.2$=72.9 točk, kar ni smiselno, ker je vrednost spodnje meje intervala manjša kot 40 točk in ne streza razponu vzorčnih podatkov. Porazdelitev števila točk doseženih na izpitu biostatistike pri uspešnih študentih, je prikazana na sliki A.3 s histogramom. Kot pričakovano je porazdelitev asimetrična v desno ($Me < \bar{x}$).

(b) Število dodatnih točk iz domačih nalog. V tem primeru so vrednosti manjše od 20; $\bar{x}$ =12.7 in s =4.7 točk. Če bi bili podatki normalno porazdeljeni bi pričakovali, da je približno 95% vseh študentov doseglo med $12.7 - 2 \cdot 4.7$=3.3 in $12.7 + 2 \cdot 4.7$=22.1 dodatnih točk, kar ni smiselno, ker je vrednost zgornje meje intervala večja kot 20 točk. Porazdelitev števila točk doseženih na izpitu biostatistike, kjer je študent pozitivno opravil izpit, je prikazana na sliki A.4 s histogramom. Kot pričakovano je porazdelitev negativno asimetrična ($Me > \bar{x}$).



Slika A.3: *Dosežene točke na pozitivnem izpitu.*



Slika A.4: *Dodatne točke iz domačih nalog.*

8. n=5, $\bar{x}$=30, Me=25, Mo=20. Tri najmanjše vrednosti morajo biti: $x_1 = 20, x_2 = 20, x_3 = 25$; x_4 in x_5 morata biti taka, da $\sum_{i=1}^{n} x_i = 30 \cdot 5 = 150$ (ker je $\bar{x}$=30), oziroma $x_4+x_5=150-(20+20+25) = 85$; vrednosti morata biti večji od 25 (ker Me=25) in med sabo različni (ker Mo=20). Na primer lahko izberemo $x_4 = 35$ in $x_5 = 50$.

9. Porazdelitev podatkov (5, 7, 9, 9, 12, 13, 15, 20, 25, 30) je grafično prikazana na sliki A.5.

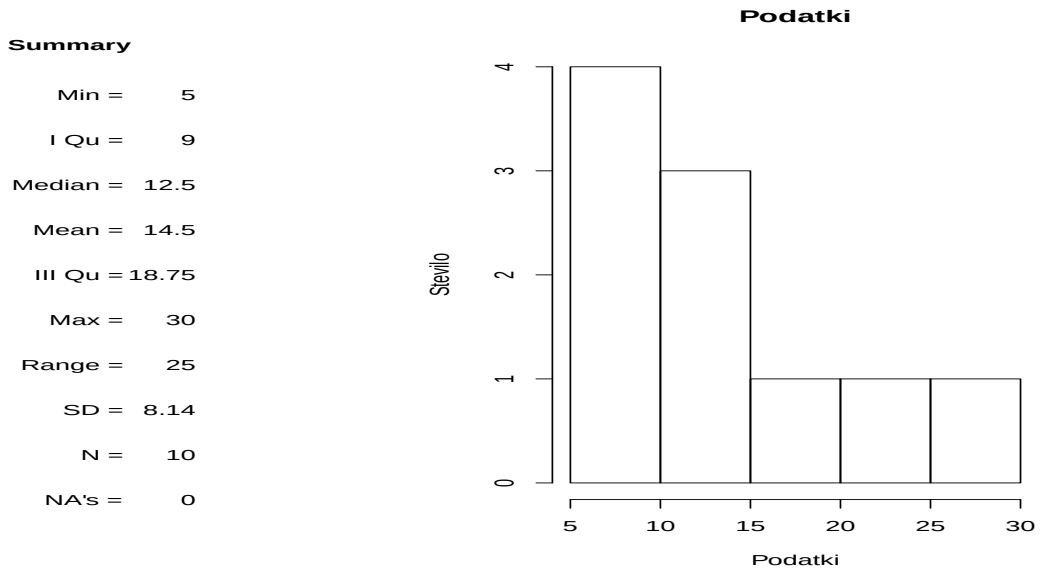
Ker je porazdelitev asimetrična v desno, izberemo mediano kot mero središčnosti in interkvartilni razmik kot mero razpršenosti. Me=12.5 (število opazovanj je soda številka, torej mediano dobimo tako, da izračunamo aritmetično povprečje dveh srednjih vrednosti, 12 in 13 v tem primeru), interkvartilni razmik= od 8 (1. kvartil), do 17.5 (3. kvartil). Lahko bi tudi rekli, da je mediana vključena med 12 in 13. Ta definicija mediane je uporabna tudi za opisne urejenostne spremenljivke, kjer ne moremo izračunati aritmetičnega povprečja.

10. Porazdelitev spremenljivk je opisana v tabeli A.1. Višina je približno simetrično in normalno porazdeljena, spremenljivki teža in indeks telesne mase sta rahko pozitivno asimetrično porazdeljeni, uporaba interneta je porazdeljena izrazito pozitivno asimetrično.

Trditve (pravilne (T) ali napačne (F)).

(a) Teža je porazdeljena simetrično. F

(b) Višina je približno normalno porazdeljena. T



Slika A.5: *Grafična predstavitev podatkov in opisne statistike.*

	Višina	Teža	BMI	internet
Min.	153.00	43.00	16.07	0.00
1st Qu.	167.00	57.75	20.20	5.00
Median	171.00	64.00	21.68	10.00
Mean	172.10	65.99	22.19	11.78
3rd Qu.	178.10	73.25	23.82	15.00
Max.	193.00	98.00	33.79	40.00
SD	8.40	11.71	2.96	8.60

Tabela A.1: Opisne statistike

(c) Povprečna in medianska višina sta si zelo podobni. T

(d) Povprečni BMI je manjši kot medianski BMI. F (porazdelitev je rahko pozitivno asimetrična, torej $Me < \bar{x}$)

(e) Porazdelitev števila ur tedenske uporabe interneta je negativno asimetrična. F

(f) Standardni odklon višine je približno 20 cm. F (je približno 10 cm)

(g) Povprečje števila ur tedenske uporabe interneta je večje od mediane. T

(h) Mediana števila ur tedenske uporabe interneta je približno 20 ur. F

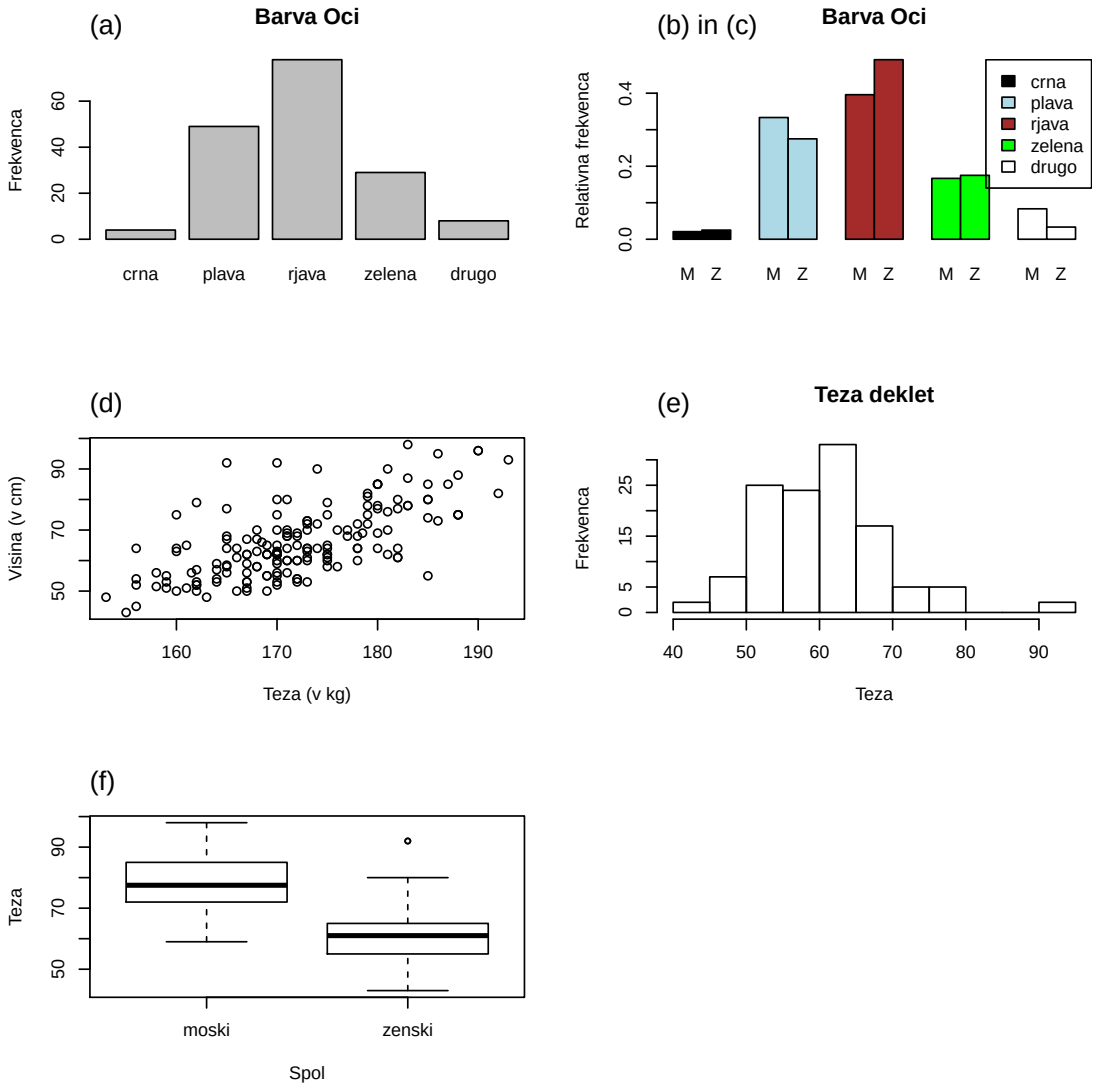
(i) Interkvartilni razmik števila ur tedenske uporabe interneta je približno od 15 do 35 ur. F

(j) Razpon višine je od 165 do 185 cm. F

(k) Porazdelitev logaritmične transformacije števila ur tedenske uporabe interneta bi bila približno normalna. T

(l) Približno 50 študentov je tehtalo med 50 in 60 kg. T

(m) Manj kot 5 študentov je tehtalo več kot 80 kg. F



Slika A.6: *Grafični prikazi podatkov.*

11. Na sliki A.6 so prikazani bolj primerni grafični prikazi podatkov.

(a) Kolač (ali strukturni krog) ni najboljša izbira za grafično predstavitev opisnih spremenljivk, ker človeško oko težko primerja površine. Stolpični diagram frekvenc ali relativnih frekvenc je boljša izbira.

(b) Grafikon predstavlja frekvenčno porazdelitev barve oči glede na spol. Če nas zanima, ali so si moški in ženske podobne glede na barvo oči, s tem grafikonom težko primerjamo spola, ker je število v skupinah različno in so prikazane absolutne frekvence. Rajši prikažemo stolpični diagram relativnih frekvenc. Dodatna pomankljivost je, da manjka legenda - kaj pomeni posamezna barva.

(c) Tudi v tem grafikonu so prikazane absolutne frekvence, kot v prejšnjem primeru. Dodatna napaka je, da so vrednosti povezane, kar poda napačni vtis, da so barve oči urejene.

(d) V tem razsevnem diagramu je začetna vrednost osi 0 (nepotrebno), kar zavzame velik del grafikona. Povezanost med spremenljivkama je zato manj razvidna.

(e) Porazdelitev teže je podana s stolpičnim diagramom, ki ni primerni grafični prikaz za številske spremenljivke. Prikazana je frekvenca posameznih vredosti, na osi x niso prikazane vse možne vrednosti, ampak samo tiste, ki so bile opazovane na vzorcu - zato vrednosti niso enakomerno porazdeljene. Težko lahko dojamemo obliko porazdelitve, če ne združimo številskih vrednosti v razrede. Manjka oznaka na osi y (gre za absolutne frekvence). Rajši uporabimo histogram.

(f) Radi bi primerjali težo med spoloma. Z uporabo absolutnih frekvenc je težko primerjati skupini različnih velikosti. Tudi uporaba stolpičnega diagrama je zavajajoča. Za neposredno primerjavo je primernejša škatla z brki, ali pa dva ločena histograma relativnih frekvenc, ki imata isti osi x.

A.2 Verjetnost

3. $P(Z) = P(Z|A)P(A) + P(Z|B)P(B) + P(Z|C)P(C) = 1/3 * 0.7 + X * 0.6 + (2/3 - X) * 0.2 = 0.38$, ker so A, B in C izčrpni in nezdružljivi dogodki. $X = \frac{0.38 - 1/3 * .7 - 2/3 * .2}{(0.6 - 0.2)} = 0.0333$

4. Marsovci.

(a) $P(Barva = R) = 0.05$ (iz tabele)

(b) $P(Barva = R \cap Lasje = R) = P(Barva = R)P(Lasje = R|Barva = R) = 0.05 * 0.65 = 0.0325$

(c) $P(Lasje = R) = P(Lasje = R|Barva = R)P(Barva = R) + P(Lasje = R|Barva = Z)P(Barva = Z) + P(Lasje = R|Barva = M)P(Barva = M) = 0.65 * 0.05 + 0.05 * 0.10 + 0.02 * 0.85 = 0.0545$

(d) $P(Barva = R \cup Lasje = R) = P(Barva = R) + P(Lasje = R) - P(Barva = RinLasje = R) = 0.05 + 0.0545 - 0.0325 = 0.072$

7.

Porazdelitev	Povprečje	Standardni odklon	Verjenost pozitivne vrednosti
A	0	1	0.5
B	1	1	0.84
C	-1	0.5	0.02
D	0	2	0.5

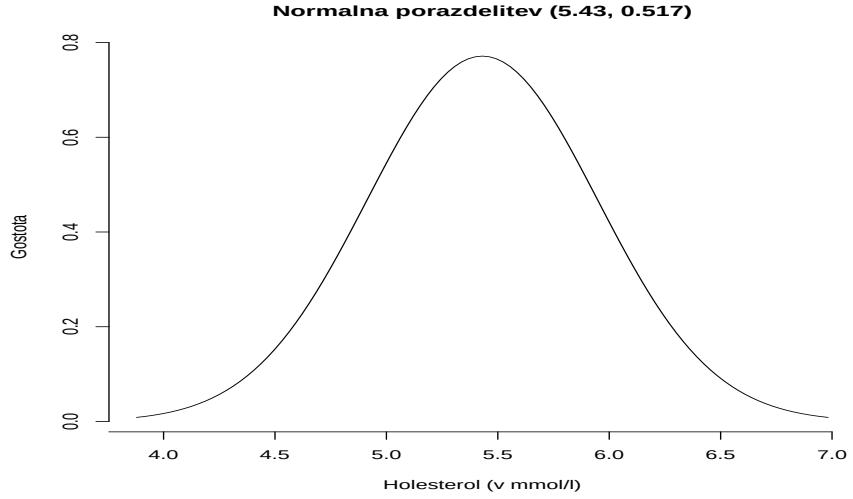
9. Holesterol (mg/dl) $\sim N(\mu = 210, \sigma = 20))$. Formula, s katero transformiramo meritve iz mg/dl v mmol/l, je:

$$mmol/l = mg/dl * 0.02586$$

Za normalno porazdeljene spremenljivke velja, da so tudi njihove linearne transformacije normalno porazdeljene, oziroma $X \sim N(\mu, \sigma)$, $Y = a + bX \sim N(a + b\mu, b\sigma)$.

V tem primeru je $a = 0$ in $b = 0.02586$, torej Holesterol (v mmol/l) $\sim N(210 * 0.02586, 20 * 0.02586)$=N(5.43, 0.52).

Porazdelitev holesterola v mmol/l je grafično prikazana na sliki **??**.



Slika A.7: *Grafični prikazi podatkov.*

10. Višina $\sim N(168cm, 6cm)$.

(a) od 168-1.96*6=156.24 cm do 168+1.96*6=179.76 cm; $1.96 = z_{0.05/2}$.

(b) od 168-2.58*6=152.52 cm do 168+2.58*6=183.48 cm; $2.58 = z_{0.01/2}$

(c) 168+1.96*6=179.76 cm;

(d) 168-1.64*6=158.16; $-1.64 = z_{0.95}$

(e) $P(Visina \geq 168cm) = 0.5$.

(f) * $P(Visina = 168cm) = 0$; verjetnost posamezne vrednosti je 0, ker je višina zvezna spremenljivka.

(g) $P(Visina \geq 156cm) = P(Z > -2) = 0.977$.

11. Binomska porazdelitev

Porazdelitev	π	n
A	0.5	10
B	0.5	100
C	0.1	10
D	0.1	100

13. $\pi = 0.20$, $n = 10$.

(a) $P(K = 1) = n * \pi * (1 - \pi)^9 = 0.27$

(b) $P(K = 0) = (1 - \pi)^{10} = 0.11$

(c) $P(K > 0) = 1 - P(K = 0) = 1 - (1 - \pi)^{10} = 0.89$

17. Višina $\sim N(\mu, \sigma)$; $\bar{x} = 182$ cm, $s = 6$) cm, $n = 48$ študentov. μ je povprečna višina študentov v populaciji in σ je populacijski standardni odklon. Standardna napaka $\hat{SE} = \frac{s}{\sqrt{n}} = \frac{6}{\sqrt{48}} = 0.87$ in $\frac{\bar{X}-\mu}{\hat{SE}} \sim t_{df=47}$.
$t_{df=47,\alpha=0.025} = 2.01$ $t_{df=47,\alpha=0.005} = 2.68$

(a) 95% IZ za μ: $182 - 2.01 \times 0.87 = 180.2513$ cm do $182 + 2.01 \times 0.87 = 183.7487$ cm.

(b) 99% IZ za μ: 182 - 2.68×0.87=179.6684 cm do $182 + 2.68 \times 0.87 = 184.3316$ cm.

(c) Vrednost 187 ni vključena v 95% (oziroma 99%) IZ. Na podlagi naših rezultatov bi izključili, da je povprečna višina slovenskih študentov 187 cm (P<0.01 za $H_0 : \mu = 187$ cm).

14. t porazdelitev.

Porazdelitev	Stopinje prostosti	$t_{0.025}$
A	2	4.3
B	8	2.31
C	98	1.98
D	18	2.1

11. χ^2 porazdelitev.

Porazdelitev	Stopinje prostosti	$\chi^2_{0.05}$
A	3	7.81
B	1	3.84
C	8	15.51
D	2	5.99

18. Ocena populacijske aritmetične sredine je bolj natančna pri večjem vzorcu.

19. Verjetnost, da bomo zajeli populacijsko povprečje teže, ni odvisna od velikosti vzorca (pri večjem vzorcu imamo večjo natančnost pri oceni).

20. 95% interval zaupanja za μ: od 1.5 do 3.5, σ=1.613.

(a) $\bar{x} = 2.5$ (srednja točka intervala)

(b) $3.5 - 2.5 = 1 = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96 \times 1.613/\sqrt{n}$; $n = 10$

21. n, $n_2 = 9n$. Kolikokrat se povečata/zmanjšata standardna napaka in interval zaupanja, če vzorec povečamo 9x?

(a) sprememba standardne napake: 3x; $\hat{SE} = \frac{\sigma}{\sqrt{n}}; \hat{SE}_2 = \frac{\sigma}{\sqrt{n_2}} = \frac{\sigma}{\sqrt{9n}} = 1/3 \times \hat{SE}$

(b) sprememba intervala zaupanja: 3x; širina intervala : $2 \times 1.96 * \frac{\sigma}{\sqrt{n}}$, nova širina intervala: $2 \times 1.96 \times \frac{\sigma}{\sqrt{9n}} = 1/3 \times 2 \times 1.96 \times \frac{\sigma}{\sqrt{n}}$.

22. Navedena je napaka za katero je trditev pravilna.

(a) 2.

(b) 2.

(c) 2.

(d) 2.

(e) nobena, 2. je večja pri bolj variabilnih spremenljivkah.

(f) 1.

(g) 2.

(h) 1.

(i) 2.

(j) 1.

(k) 2., moč testa je 1-napaka 2. vrste

(l) 2.

(m) 2.

(n) 2.

A.3 Primerjava skupin - številske spremenljivke

2. Primerjava povprečne višine slovenskih fantov in deklet.

```
        Two Sample t-test

data:  my.data.2[, 4] by my.data.2[, 3]
t = 13.0346, df = 166, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.18291 15.17543
sample estimates:
mean in group 1 mean in group 2
       181.5000        168.3208
```

(a) Povprečna višina fantov se statistično razlikuje od povprečne višine deklet, P<0.001.

(b) 95% IZ za $\mu_M - \mu_F = 11.2$ do 15.2 cm; s 95% zaupanjem lahko trdimo, da je razlika med povprečno višino slovenskih fantov in deklet vključena v intervalu od 11.2 do 15.2 cm.

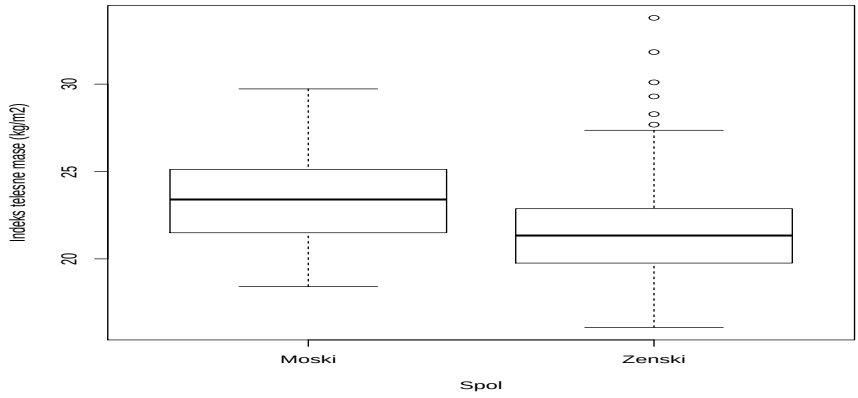
3. Primerjava indeks telesne mase slovenskih fantov in deklet.

(a) Da, P<0.001.

(b) t-test za dva neodvisna vzorca z enakima variancama.

(c) 95% interval zaupanja za populacijsko razliko povprečij (fantje-dekleta) je od 1.02 do 2.93 kg/m^2. Imamo 95% zaupanje, da je populacijska razlika povprečij vključena v intervalu.

(d) Razlike v BMI med fanti in dekleti bolj učinkovito grafično predstavimo z uporabo škatke z brki (slika A.8).



Slika A.8: *Graf škatla z brki za BMI glede na spol.*

4. Primerjava povprečne višine študentov, ki igrajo videoigrice s povrečno višino študentov, ki ne igrajo videoigric.

(a) V vzorec smo zajeli $n = df + 2 = 168$ oseb.

(b) Povprečna višina študentov, ki igrajo videoigrice je statistično značilno večja od povprečne višine študentov, ki ne igrajo videoigric (P=0.0084, razlika=3.5 cm, 95% interval zaupanja za populacijsko razliko od 0.91 do 6.08 cm).

(c) Glej nalogo 1 v poglavju Verjetnost.

6. Raziskovalec naključno razdeli paciente v 2 skupini. Prva skupina dobi placebo, druga pa novo zdravilo.

(a) Raziskovalec mora ponoviti analizo podatkov, ker ima neodvisne podatke.

(b) t-test za dva neodvisna vzorca

(c) Izračuni

- Vzorčna razlika povprečij: -144.6+133.4=-11.2
- Standardna napaka= 8.76
- Testna statistika= - 1.279
- p-vredost=0,217
- 95% interval zaupanja: -29.6 do 7.20

(d) Na podlagi teh rezultatov ne moremo zavrniti ničelne domneve

Izpis iz statističnega programa R je naslednji:

```
        Two Sample t-test

data:  trt and placebo
t = -1.2791, df = 18, p-value = 0.2171
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -29.596635   7.196635
sample estimates:
mean of x mean of y
    133.4     144.6
```

7. Primerjava sedimentacije eritrocitov.

(a) Ničelna domneva je $\mu_{pred} = \mu_{po}$; μ_{pred} je populacijska povprečna sedimentacija eritrocitov pred zdravljenjem in μ_{po} je povprečna sedimentacija eritrocitov po zdravljenju, v populaciji bolnikov z ankilozirajočim spondilitisom.

(b) Zavrnemo ničelno domnevo ($p = 0.002$, označeno s Sig. (2-tailed) v tabeli)

(c) $t_{df=29}$

(d) $\hat{SE} = \frac{s}{\sqrt{n}}$, s je standardni odklon razlik eritrocitov pred in po zdravljenju in $n = df + 1 = 30$ je velikost vzorca; $s = \hat{SE} * \sqrt{n} = 2.114 * \sqrt{30}$=11.58.

(e) Nismo uporabili testa t za neodvisna vzorca, ker smo merili istega pacienta dvakrat; podatki niso bili neodvisni.

(f) Samo povprečje razlike se ne bi spremenilo. $df = 58$, načeloma bi pričakovali večjo standardno napako, širši interval zaupanja in večjo p-vrednost.

8. Število obiskov pri veterinarju v desetih (istih) klinikah v letu 2008 in 2009. $\bar{x}_{2008} = 90$, $\bar{x}_{2009} = 100$; $s = 7$: standardni odlkon razlik; razlika se porazdeli normalno.

(a) t-test za odvisna vzorca, ker smo merili vsako kliniko dvakrat.

(b) $H_0 : \mu_{2009} = \mu_{2008}$; $\hat{SE} = 7/\sqrt{10}$ =2.21; $\bar{x}_{2009} - \bar{x}_{2008}$=10. $t = \frac{\bar{x}_{2009} - \bar{x}_{2008}}{\hat{SE}} = 4.518$; df=9; p=0.0015.

(c) Zavrmemo H_0. Med letoma 2008 in 2009 je prišlo do statistično značilne razlike v povprečnem številu obiskov v veterinarskih klinikah.

(d) Na podlagi rezultatov pridobljenih s statističnim testiranjem lahko skepamo, da vrednost 0 ne bo vključena v 95% intervalu zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008), saj je bila P-vrednost manjša od 0.05.

(e) 95% interval zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008): $t_{df=9,0.025} = 2.26$ od $\bar{x}_{2009} - \bar{x}_{2008} - t_{df=9,0.025} \times \hat{SE}$ do $\bar{x}_{2009} - \bar{x}_{2008} + t_{df=9,0.025} \times \hat{SE}$, torej od 4.99 do 15.01.

(f) Imamo 95% zaupanje, da je povprečna razlika v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih vključena med 4.99 in 15.01 obiski.

Drugi raziskovalec: $s_{2008} = 5, s_{2009} = 7$, $\bar{x}_{2008} = 90$, $\bar{x}_{2009} = 100$, $n_{2008} = 10$, $n_{2009} = 10$.

(a) Potrebno je ponoviti izračun, ker je drugi raziskovalec vsako leto vključil različne klinike; podatki so torej neodvisni.

Skupni standardni odklon

$$s_s = \sqrt{\frac{s_{2009} \times (n_{2009} - 1) + s_{2008} \times (n_{2008} - 1)}{n_{2009} + n_{2008} - 2}} = 6.08;$$

standardna napaka $\hat{SE} = s_s\sqrt{1/n_{2009} + 1/n_{2008}}$ =2.72. $t_{df=18,0.025}$=2.1.

95% interval zaupanja za povprečno razliko v številu obiskov v slovenskih veterinarskih klinikah v teh dveh letih (2009-2008): od $\bar{x}_{2009} - \bar{x}_{2008} - t_{df=18,0.025} \times \hat{SE}$ do $\bar{x}_{2009} - \bar{x}_{2008} + t_{df=18,0.025} \times \hat{SE}$, torej od 4.28 do 15.72.

(b) t-test za dva neovdisna vzorca.

(c) Na podlagi rezultatov pridobljenih s 95% intervalom zaupanja lahko sklepamo, da bo p-vrednost pridobljena s statističnim testiranjem s t-testom za dva neodvisna vzorca manjša kot 0.05, ker interval zaupanja ne vsebuje 0.

(d) $H_0 : \mu_{2009} = \mu_{2008}$; $t = \frac{\bar{x}_{2009} - \bar{x}_{2008}}{\hat{SE}} = 3.676$; df=18; p=0.0017.

A.4 Primerjava skupin - opisne spremenljivke

3. Povezanost med načinom uživanja tobaka oz. alkohola in lokacijo raka ustne votline.

(a) χ^2 test.

(b) Izpis is statističnega programa R.

```
        Pearson's Chi-squared test

data:  tmp.data
X-squared = 5.1449, df = 2, p-value = 0.07635
```

Spremenljivki nista statistično značilno povezani (P=0.076). Pričakovane frekvence so podane v tabeli A.2.

	Jezik	Drugo
Žvečenje	143.37	269.63
Kajenje	82.27	154.73
Alkohol	42.35	79.65

Tabela A.2: Pričakovane frekvence

Testna statistika je ob veljavni ničelni domnevi porazdeljena po χ^2 porazdelitev z $df = 2$.

4. Povezanost med igranjem videoigric in spolom.

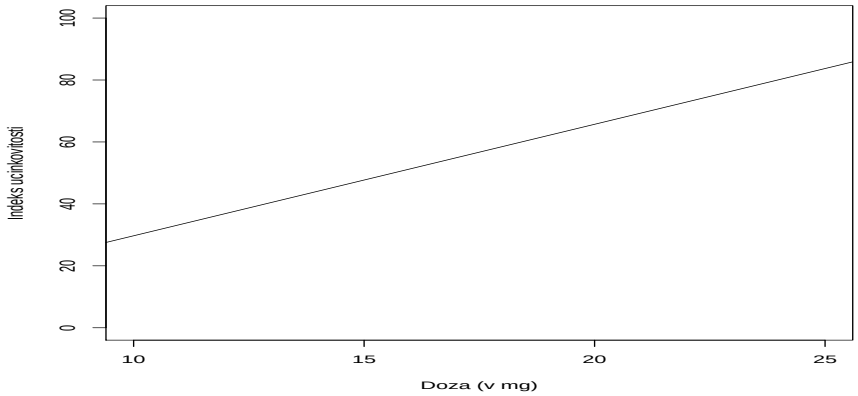
```
        Pearson's Chi-squared test

data:  table(my.data.2[, 3], my.data.2[, 16])
X-squared = 11.6712, df = 1, p-value = 0.0006348
```

Igranje videoigric in spol sta povezani (P=0.0006).

A.5 Povezanost med številskimi spremenljivkami

2. Povezanost med dozo in indeksom učinkovitosti.



Slika A.9: *Ocenjena povezanost med dozo in indeksom učinkovitosti.*

(a) $H_0 : \beta = 0$, v populaciji doza in indeks učinkovitosti nista linearno povezani.

(b) Ocenjen odnos med dozo zdravila in učinkovitostjo zdravljenja je prikazan na sliki A.9.

(c) Da, P<0.001 za regresijski koeficient doze.

(d) Pričakujemo večji indeks učinkovitosti pri pacientu A, ker je regresijski koeficient za dozo pozitiven. Pričakovana razlika je $5 \times 3.6 = 18$.

(e) Pričakovani indeks učnkovitosti za pacienta, ki je prejemal dozo 15mg: $-6.3 + 3.6 \times 15 = 47.7$.

(f) Rezultatov ne moremo posplošiti na situacijo, ko zdravila ne dajemo (doza = 0 mg), saj v raziskavo nismo vključili pacientov, ki niso bili zdravljeni (doza je bila med 10 in 25 mg).

(g) 95% interval zaupanja za koeficient za dozo (β): $3.6 - 1.99 \times 0.08$=3.44 do $3.6 + 1.99 \times 0.08$=3.76.

3. Grafična predstavitev (izmišljenih) podatkov je podana na sliki A.10. Na podlagi teh podatkov smo ocenili model, ki je podan v tabeli.

Slika A.10: *Povezanost med ure učenja in rezultatom na izpitu.*

Ocenjen model linearne regresije				
Parameter	Ocena	Std. napaka	t	p-vrednost
(Konstanta)	3.136	1.476	2.125	0.036
Ure	7.37	0.217	33.949	<0.001

4. Vrstni red: vzorec 3, vzorec 2, vzorec 4, vzorec 1.

Dodatek B

Statistične tabele

B.1 Standardna normalna porazdelitev

V tabeli je vrednost $P(Z \geq z)$. z je vrednost navedena na začetku vrstice sešteta z vrednostjo navedeno na vrhu stolpca. Na primer, če nas zanima vrednost 1.96, izberemo vrstico "1.9"in stolpec "0.06". Vrednost na presečišču (??) je verjetnost $P(Z \geq 1.96) = 0.025$.

	0	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**
0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464	0.460
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425	0.421
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386	0.382
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348	0.345
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312	0.309
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278	0.274
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245	0.242
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215	0.212
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187	0.184
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161	0.159
1	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138	0.136
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117	0.115
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099	0.097
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082	0.081
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068	0.067
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056	0.055
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046	0.045
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037	0.036
1.8	0.036	0.035	0.034	0.034	0.033	0.032	0.031	0.031	0.030	0.029	0.029
1.9	0.029	0.028	0.027	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023
2	0.023	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018
2.1	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014	0.014
2.2	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011	0.011
2.3	0.011	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008
2.4	0.008	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006	0.006
2.5	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
2.6	0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003
2.7	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
2.8	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
2.9	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001
3	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.2	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
3.3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

B.2 t porazdelitev

V tabeli je vrednost t_{alpha} za katero velja: $P(t_{df} \geq t_{\alpha}) = \alpha$, t_{df} je t porazdelitev s df stopinjami prostosti. (stopinje prostosti po vrsticah, α po stolpcih).
Primer: za $\alpha = 0.025$ in $df = 5$ odčitamo $t_{\alpha} = 2.571$.

	0.1	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
∞	1.282	1.645	1.960	2.326	2.576	3.090

B.3 χ^2 porazdelitev

V tabeli je vrednost χ^2_{alpha} za katero velja: $P(X^2_{df} \geq \chi_\alpha) = \alpha$, X^2_{df} je χ^2 porazdelitev s *df* stopinjami prostosti. (stopinje prostosti po vrsticah, α po stolpcih).
Primer: za $\alpha = 0.05$ in $df = 1$ odčitamo $\chi^2_\alpha = 3.841$.

	0.1	**0.05**	**0.025**	**0.01**	**0.001**	**5e-04**
1	2.706	3.841	5.024	6.635	10.828	12.116
2	4.605	5.991	7.378	9.210	13.816	15.202
3	6.251	7.815	9.348	11.345	16.266	17.730
4	7.779	9.488	11.143	13.277	18.467	19.997
5	9.236	11.070	12.833	15.086	20.515	22.105
6	10.645	12.592	14.449	16.812	22.458	24.103
7	12.017	14.067	16.013	18.475	24.322	26.018
8	13.362	15.507	17.535	20.090	26.124	27.868
9	14.684	16.919	19.023	21.666	27.877	29.666
10	15.987	18.307	20.483	23.209	29.588	31.420
11	17.275	19.675	21.920	24.725	31.264	33.137
12	18.549	21.026	23.337	26.217	32.909	34.821
13	19.812	22.362	24.736	27.688	34.528	36.478
14	21.064	23.685	26.119	29.141	36.123	38.109
15	22.307	24.996	27.488	30.578	37.697	39.719
16	23.542	26.296	28.845	32.000	39.252	41.308
17	24.769	27.587	30.191	33.409	40.790	42.879
18	25.989	28.869	31.526	34.805	42.312	44.434
19	27.204	30.144	32.852	36.191	43.820	45.973
20	28.412	31.410	34.170	37.566	45.315	47.498
21	29.615	32.671	35.479	38.932	46.797	49.011
22	30.813	33.924	36.781	40.289	48.268	50.511
23	32.007	35.172	38.076	41.638	49.728	52.000
24	33.196	36.415	39.364	42.980	51.179	53.479
25	34.382	37.652	40.646	44.314	52.620	54.947
26	35.563	38.885	41.923	45.642	54.052	56.407
27	36.741	40.113	43.195	46.963	55.476	57.858
28	37.916	41.337	44.461	48.278	56.892	59.300
29	39.087	42.557	45.722	49.588	58.301	60.735
30	40.256	43.773	46.979	50.892	59.703	62.162
40	51.805	55.758	59.342	63.691	73.402	76.095
60	74.397	79.082	83.298	88.379	99.607	102.695