

Zemljevod "statističnih testov"

Ena spremenljivka

- 1 opisna skupina; preizvemo Pologodje → klinični eksperimentalni test
- 1 številka → test za 1 vzorec

- Dve spremenljivki:
 - 2 opisni → test χ^2 → če je opisno z več kategorijami, tj. 3 skupine ali več → ANOVA, Ali Kruskal-Wallis
 - 1 številka in 1 opisna skupina (tj. 2 skupini) → test t za neodvisnost: Ali Mann-Whitneyev test
 - 2 številki:
 - zanima nas razlika → test t za parne meritve: Ali Wilcoxon test predstavlja rangov
 - zanima nas povezanost → linearna regresija

ni snov za izpit pri tem predmetu

Univerza v Ljubljani

Uporabna statistika, Uvod v statistiko

Inštitut za biostatistiko in medicinsko informatiko
5. vaja

1. del: Povezanost med dvema opisnima spremenljivkama

test χ^2

1. Zanima nas, ali v populaciji obstaja povezanost med lastništvom živali in izbrano fakulteto.

kontingenčna tabela

		Fakulteta		Vsota
		VF	MF	
Živali	DA	42	63	105
	NE	2	33	35
Vsota		44	96	140

- Verjetnost, da ima študent medicine domačo žival je:

$$p_{žival|MF} = \frac{63}{96} = 0,66$$

To primenjujemo

- Verjetnost, da ima študent veterine domačo žival je:

$$p_{žival|VF} = \frac{42}{44} = 0,95$$

- Zapišite ničelno domnevo testa χ^2 (hi-kvadrat):

$$H_0: \pi_{žival|MF} = \pi_{žival|VF} \quad (\text{populacija obvezna})$$

I. način:

Verjetnost lastništa živali je enaka v populaciji. Studentov MF in VF.

II. način:

Lastništvo živali ni fakulteta (MF/VF) v populaciji nista povezani.

- Izračunajte pričakovane frekvence pod ničelno domnevo:

$$\pi_{žival} = \frac{105}{140}$$

$$\pi_{žival|MF} = \pi_{žival|VF} = \frac{105}{140}$$

$$44 \cdot \pi_{žival} = \frac{44 \cdot 105}{140} = 33$$

		Fakulteta		Vsota
		VF	MF	
Živali	DA	33	72	105
	NE	11	24	35
Vsota		44	96	140

$$P_k = \frac{r_{obra 1} \cdot r_{obra 2}}{n}$$

- Izračunajte testno statistiko.

$$\chi^2 = \sum_k \frac{(O_k - P_k)^2}{P_k} = \frac{(42 - 33)^2}{33} + \frac{(63 - 72)^2}{72} + \frac{(2 - 11)^2}{11} + \frac{(33 - 24)^2}{24} = 14,32$$

Vejca: testna $\stackrel{H_0}{\sim} \chi^2_{df}$

$$df = (\text{st. vrstic} - 1) \cdot (\text{st. stolpcov} - 1) = 1 \cdot 1 = 1$$

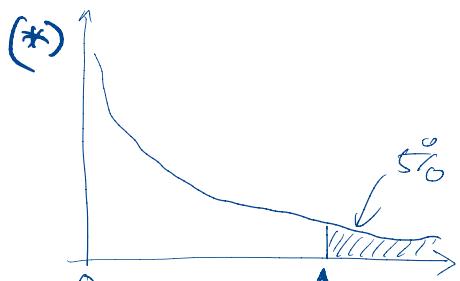
pričakovana porazdelitev

(+) Porazdelitev χ^2 ima različno obliko glede na df → glej npr. Wikipedia. Vse pa so pozitivne za slike ni bistveno, katero tečno narišete.

$$\alpha = 5\%$$

pod ničelno domnevo

- Narišite porazdelitev testne statistike, vrišite dobljeno vrednost, izračunajte in označite kritično vrednost ter vrednost p . Ali ničelno domnevo zavrnemo?



$$\chi^2_{df=1}$$

• testna $> \chi^2_{kritična} \Rightarrow H_0$ zavrnemo

• $p < \alpha \Rightarrow H_0$ zavrnemo

R koda

$$\begin{aligned} \text{reduct } p &= P(\chi^2_1 > 14.32) = \text{pchisq}(14.32, df=1) = \\ &\quad \text{LE EN REP} \\ &0.0002 \\ \chi^2_{kritična} &= \chi^2_{0.95; df=1} = \text{qchisq}(0.95, df=1) = 3.84 \\ &\quad \text{R koda} \end{aligned}$$

- Zapišite vsebinski sklep.
- Iz za razliko med delžema → OTS

II. Lastništvo živali na fakultata sta v populaciji prevezema.
I. Delož studentov z živaljo je na VF vegji kot na MF.

POPULACIJA

- Kaj so bile predpostavke tega testa? Kaj lahko storimo v primeru, ko predpostavke testa niso izpolnjene?

vsaj 80% prisotnosti frekvenc je ≥ 5

(če je testna pod H_0 približno porazdeljena kot χ^2_{df})

→ za naš primer so izpolnjene

če niso izpolnjene:

- Povečemo naloge, če je to je mogoče.
 - Zdržimo kategorije spremenjivke, če je to mogoče.
 - Uporabimo Fisherjev eksaktni test (če pa nimajo te predpostavke).
- ↓ enako H_0

χ^2 test v R:

```
dd <- read.table("Ankete1011.txt", header = T, dec = ",", sep = "\t", fill = T)
dd$zival <- ifelse(dd$Domace_zivali == "Ne", "Ne", "Da")
```

```
test <- chisq.test(table(dd$zival, dd$Fakulteta), cor = F) ## hi2 test brez Yatesovega popravka
chisq.test(table(dd$zival, dd$Fakulteta), cor = T) ## hi2 test z Yatesovim popravkom
fisher.test(table(dd$zival, dd$Fakulteta)) ## Fisherjev eksaktni test
```

test\$expected → prisotnosti frekvence

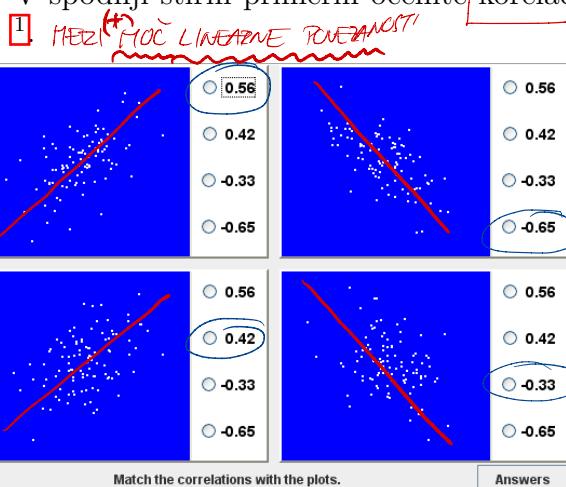
$$\chi^2 = \sum_k \frac{(O_k - E_k - 0.5)^2}{E_k}$$

NI SNOV ZA IZPIT

2. del: Linearna regresija in korelacija

Oznaka:

1. V spodnji štirih primerih ocenite korelacijski koeficient na podlagi razsevnega diagrama



Največja možna vrednost korelacijskega koeficiente je: 1

(*) **Kdaj so bliži so točke premici?**
Kdaj so razpršeni?

Najmanjša možna vrednost korelacijskega koeficiente je: -1

NAKLON NI POMEMBEN

Če je korelacijski koeficient 0 pomeni, da:

NI linearne povezanosti

$r=0$
obe stevilci

$r=0$
ni linearne povezanosti,
ni pa linearne

Izračun korelacije v R:

`x <- rnorm(100)` *generira 2 vrstega*

`y <- rnorm(100)+x`

`cor(x, y)` → Pearsonov korelacijski koeficient = moyed osrednja, ta pri tem produkt
`cor.test(x, y)` → $H_0: \text{mesta povezani}$ oz. pop. korelacijski koef. = 0] To je isto
 D linearno

2. Zaženite program R in odprite datoteko [Ankete1011.txt](#).

```
dd <- read.table("Ankete1011.txt", header = T, dec = ",", sep = "\t", fill = T)
```

- Narišite razsevni diagram, pri čemer naj bo odvisna spremenljivka teža (na ordinatni osi), neodvisna spremenljivka pa višina (na abscisni osi).

```
plot(dd$Visina, dd$Teza, xlab = "Visina", ylab = "Teza")
```

- V diagram vrišite regresijsko premico.

```
abline(a = lm(Teza ~ Visina, data = dd)$coef[1],  
       b = lm(Teza ~ Visina, data = dd)$coef[2])
```

- Geometrijsko ocenite približno vrednost regresijske konstante in regresijskega koeficiente ter presodite, kolikšna se vam zdi razpršenost točk okoli regresijske premice.



- Za iste podatke izvedite linearno regresijsko analizo ter dobljeni vrednosti regresijske konstante in regresijskega koeficiente primerjajte s približno ocenjenima.

```
fit <- lm(Teza ~ Visina, data = dd)
```

¹Kratek opis korelacijskega koeficiente najdete na <http://en.wikipedia.org/wiki/Correlation>.

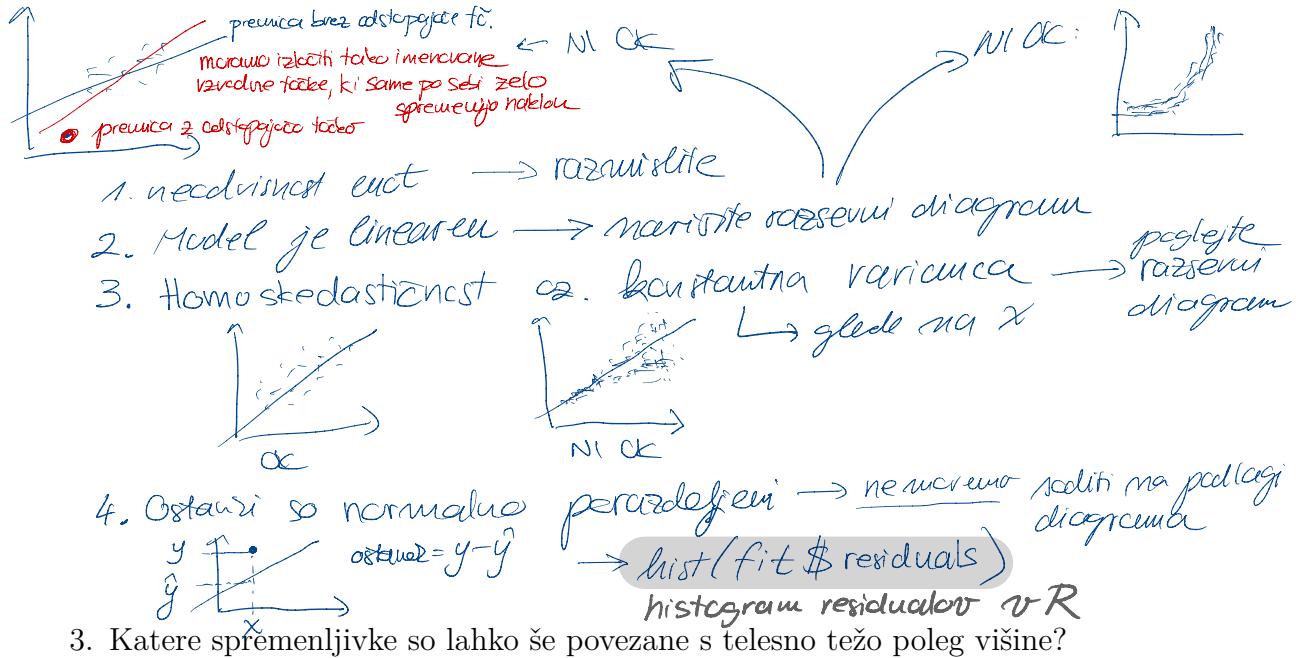
- Ali je smiselno interpretirati koeficient pri konstanti?

$\text{O mi v razsevu podatkov} \Rightarrow \text{Ne.}$

- Kaj lahko na podlagi modela poveste o teži 100 cm velikega otroka?

$\text{Micesar, ker } 100 \text{ cm v razsevu podatkov}$

- Kaj so predpostavke za uporabo linearne regresije? Na podlagi razsevnega diagrama komentirajte njihovo izpolnjeno.



Oglejte si primer, kjer smo vključili spol kot neodvisno spremenljivko. Kodiranje spola: 1: moški, 2: ženski.

Linearna regresija: $y = \text{funkcija}$, X je opredeljeno

Coefficients^a

regresija: $y = \text{funkcija}$, X je opredeljeno

Izpis SPSS

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.
	B	Std. Error			
1 (Constant)	89,018	3,305		26,932	,000
Spol	-14,599	1,884	-,551	-7,750	,000

a. Dependent Variable: Teza

$$y = a + bx$$

- Kakšno težo model napove za moškega in kakšno za žensko?

$\text{kodiruje spola v SPSS}$
 $M: 89 - 14,6 \cdot 1 = 74,4$

$\Sigma: 89 - 14,6 \cdot 2 = 55,8$

- Kako interpretiramo regresijski koeficient? $-14,6$

Ženske so v povprečju lažje za $14,6$ kg. (vzorec)

significance
 $= \text{vrednost } P$

drugična 1. vrstica tabele - razlog je različno kodiranje spola
 \rightarrow končni rezultat je sedem enak

$\sqrt{Z-u}: M: 74,4 - 14,6 \cdot 0 = 74,4$

$\text{Intercept} = 74,4 \quad \Sigma: 74,4 - 14,6 \cdot 1 = 55,8$

$\text{kodiruje spola v R}$

- Zapišite ničelno domnevo.

Spol in teza v pop. niste lin. povezana.

- Kaj zaključimo na podlagi rezultatov? $p < 0,001 \Rightarrow$ H0 zavrnemo

$\text{H1: Spol in teza v pop. lin. povezana.}$

Ni del testa t , te je test s
 $H_0: \sigma_1^2 = \sigma_2^2$ 4. Primerjajte dobljene rezultate z rezultati testa t za dva neodvisna vzorca.
 $P_{\text{Levene}} < \alpha \Rightarrow \sigma_1^2 \neq \sigma_2^2$
 \Rightarrow nujno 2 vrstical
 $P_{\text{Levene}} > \alpha \Rightarrow$ matematska lastnost
test v obeh vrsticah
To ni bistreno → glej komendar pri Vajil 4, str. 9.

Group Statistics				
Spol	N	Mean	Std. Deviation	Std. Error Mean
Teza moski	43	74,419	11,2594	1,7170
zenski	97	59,820	9,8246	,9975

Independent Samples Test

Teza	Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
	2,330	,129	7,750	138	,000	14,5990	1,8838	10,8741	18,3239
Equal variances assumed			7,352	71,569	,000	14,5990	1,9858	10,6401	18,5580
Equal variances not assumed									

- Zapišite ničelno domnevo za test t . Primerjajte z ničelno domnevo pri linearni regresiji.

$$H_0: \mu_M = \mu_Z$$

- Ali so rezultati obeh testov podobni? Da, sicer.

- Primerjajte napovedani teži iz regresijskega modela s povprečnima težama v vsaki skupini.

izlokat povprečni teži

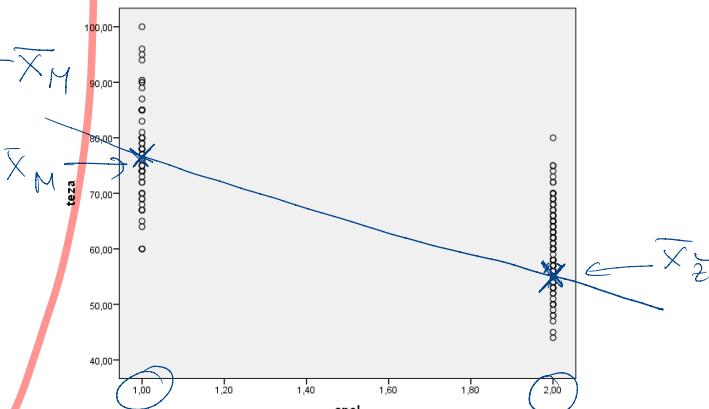
$$\bar{\mu}_M = 74,4 \\ \bar{\mu}_Z = 55,8$$

- Vrišite premico v spodnji graf. Skozi kateri dve točki poteka?

$$\beta = \frac{\Delta Y}{\Delta X} = \frac{\bar{X}_Z - \bar{X}_M}{2-1} = \bar{X}_Z - \bar{X}_M$$

Ipd:
 $\beta = \mu_Z - \mu_M$

LR: $H_0: \beta = 0$



- Ali testa odgovarjata na enako vprašanje? Ali odgovarjata enako?

LR je identičen testu + nekoliko bolj vs. 1

predpostavilo enake lastnosti varianc

H_0 sta enazi

5. Odprite datoteko **teza.txt**, ki vsebuje podatke o teži žensk pred in po nosečnosti. Tež pred in po porodu primerjajte najprej s parnim testom t , nato pa še z linearno regresijo. Pri vsaki analizi zapišite ničelno domnevo ter interpretirajte rezultate.

- Parni test t :

- Ničelna domneva:

Priporočna teža pred

porodcu je enaka povez. teži po porodu
v populaciji nekonic.

- Vrednost p in statistični sklep:

$$p = 0,035 < 5\% \Rightarrow \text{Ho zavrnemo}$$

- Interpretacija rezultatov:

— je različna od —

- Linearna regresija:

- Ničelna domneva:

Teža pred porodcu ni enačena s težo po porodcu.

- Vrednost p in statistični sklep:

$$p < 2 \cdot 10^{-76} \Rightarrow \text{Ho zavrnemo}$$

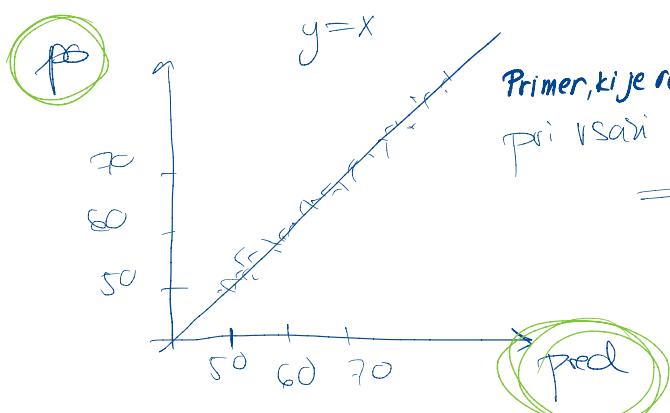
- Interpretacija rezultatov:

— je ena. —

- Ali si lako zamislite primer, ko bi eden izmed testov dal značilen rezultat, drugi pa ne?

LR : Ho zavrnemo, torej sta lin. povezani

test t : Ho ne zavrnemo, torej ne moremo trditi, da sta priporočeni razlicni (naj si bosta na vzorec)



Primer, ki je realističen:

pri vsoti \bar{x} : teža pred = teža po

$$\Rightarrow \bar{x}_{\text{pred}} = \bar{x}_{\text{po}} \Rightarrow \text{test } t \text{ ne zavrne}$$

$$b = 1 \neq 0 \Rightarrow \text{LR } \text{Ho zavrsi.}$$

Pri povezovanju med spremenljivkami, ki ne imajo linearne povezave, ne moremo uporabiti linearne regresije in preverjati povezavnost!

Podatki pred & po: seveda sta meritvi med seboji povezani, Ho linearne regresije ni posebej zavrnuta.
dejansko mora biti enak, ali se teži pred in po razlikujejo. Tu je resnično smiseln test t za parne meritve!

- y odvisna spremenljivka oz. izid (an. outcome)
- 1 neodvisna spremenljivka (x) } umiravatna analiza
- (pojasnjevalna)
(an. Predictor)

Lahko imamo več neodvisnih sprem. → naslednja naloga.

Multipla linearma regresija → več neodvisnih spremenljivk

↳ uporablja se tudi izraz multivariatna linearma regresija,
čeprav le-ta ni pravilen za ta okvir → pri multivariatni višini več izdelov (y), pri multipli pa več x

6. Vrnite se k podatkom iz ankete (Ankete1011.txt). Ocenimo hkrati povezanost višine in spola s težo: v 1m uporabite formulo **Teza ~ Visina + Spol**. Rezultate tega modela bomo primerjali s prejšnjim modelom **Teza ~ Visina**.

- Kaj se je zgodilo z deležem pojasnjene variabilnosti glede na prejšnji model? Komentirajte.

regresijska enačba: $\hat{teza} = a + b_1 \cdot visina + b_2 \cdot spol$

$$R^2 = 0,4732 \text{ se je povečal } (R^2 \text{ ni več } (1)^2) \\ (R^2 \text{ prej } 0,4632) \rightarrow R^2 \text{ se redno poveča ob dodajanjih spremenljivk}$$

→ sledi, da se poveča tudi ob dodajanjih spremenljivk, ki niso povezane z izdelom
→ zato se nanaša na R^2 uporabljajo **Adjusted R²** (glej izpls)

- Interpretirajte regresijska koeficienta in pripadajoči vrednosti p. Primerjajte z interpretacijo v prejšnjem modelu.

• $b_{visina} = 0,7984$: Če priverjam dva študenta/ki enakega spola, potem bo imel 1 cm višji težo \hat{teza} 0,7984 kg več težo.

• $p_{visina} < 0,001$: Višina je v populaciji povezana s težo neodvisno od spola.

• $b_{spol} = -3,7467$: Ženske so v povprečju za 3,7467 kg lažje kot moški ob enaki višini.

• $p_{spol} = 0,1081$: Ne moremo trditi, da je spol povezan s težo neodvisno od višine.

Interpretacija: b in p kot v modelu z eno sprem., vendar s pomembnimi dodatkovimi – priverjam neodvisno od preostalih sprem. v modelu, tj. enote izvračimo glede na celotne sprem.

- Napovejte teži za 170 cm velikega moškega in za 170 cm veliko žensko. Ali sta različni ali enaki? Primerjajte s prejšnjim modelom.

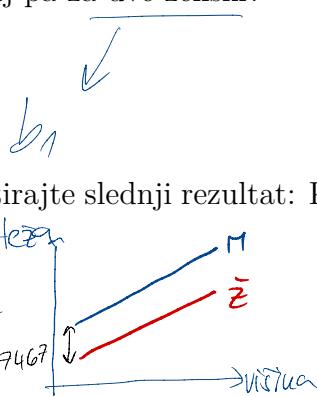
$$64,7322$$

$$60,9855$$

62,689, nglede na spol

V tem modelu teži oblikoma različno oceno teže za M in Ž.

- Kakšna je pričakovana razlika v teži za dva moška, katerih višina se razlikuje za 1 cm? Kaj pa za dve ženski?



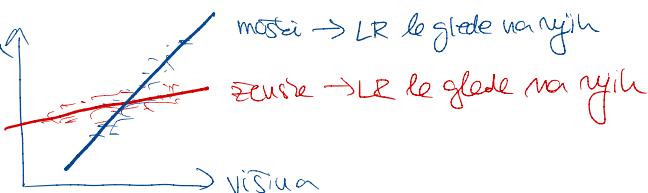
V tem modelu teži ne dovolimo razlike spremenljivke v teži za M in Ž
→ M in Ž imajo enak način za težo glede na višino
→ TO JE DODATNA IMPLICITNA PREDP. TEGA MODELA!

- Komentirajte slednji rezultat: Kaj to pomeni? Ali je to smiselno? Kdaj je smiselno?

Grafično predstavljanje teži predstavljajoča prejšnje modela

→ dovoljujemo različen intercept glede na spol, ne pa tudi naklona

Narisemo si podatke: 100x
Dovim, da razgledajo težo.



več o tem pri predmetu Linearni modeli

V tačnem primeru zgoraji model ni smiseln.

→ Dodati nacrtati f.i. interakcijo med spolom in višino, tj. prilagodljivo model:

višina + spol + višina * spol
→ ta model dodeljuje različna naklona za M in Ž
→ prav tako dodeljuje različen intercept za M in Ž, to pa je dovoljno že prejšnji model
V tem modelu lahko tudi pogledamo statistično značilnost določene interdecije.