Anuška Ferligoj, Katja Lozar Manfreda, Aleš Žiberna:

OSNOVE STATISTIKE NA PROSOJNICAH

Študijsko gradivo pri predmetu Statistika. Fakulteta za družbene vede, Univerza v Ljubljani

Ljubljana, 2011

5 BIVARIATNA ANALIZA

5.1 UVOD: BIVARIATNA ANALIZA

Bivariatna analiza: analiza dveh spremenljivk.

Primer: Povezanost med številom ur učenja in doseženim številom točk na izpitu. Preverjamo domnevo, da so tisti, ki so se učili več ur, dosegli tudi višjo oceno (domneva o Pearsonovem koeficientu korelacije, ki meri linearno povezanost med spremenljivkama).

X $\longleftrightarrow$ Y Povezanost

X $\longrightarrow$ Y Odvisnost

Mere povezanosti in odvisnosti ločimo glede na tip spremenljivk. Med njimi se bomo učili naslednjih:

1. NOMINALNI tip para spremenljivk (ena od spremenljivk je nominalna): χ^2, kontingenčni koeficienti (povezanost).
2. ORDINALNI tip para spremenljivk (ena spremenljivka je ordinalna, druga ordinalna ali boljša): Spearmanov koeficient korelacije rangov (povezanost).
3. INTERVALNI/RAZMERNOSTNI tip para spremenljivk (obe spremenljivki sta intervalni/razmernostni): koeficient korelacije (povezanost), regresijski koeficient (odvisnost).

5.2 POVEZANOST ZA NOMINALNI TIP PARA SPREMENLJIVK

5.2.1 Ureditev podatkov v kontingenčno tabelo

Primer:

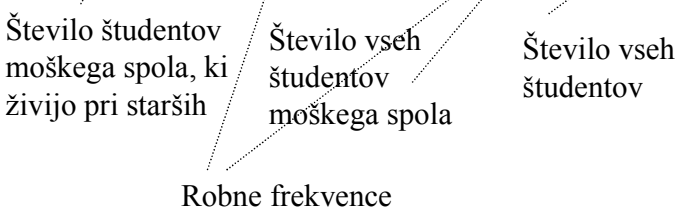
- ENOTA: dodiplomski študent neke fakultete v letošnjem študijskem letu;
- VZOREC: slučajni vzorec 200 študentov;
- 1. SPREMENLJIVKA: spol;
- 2. SPREMENLJIVKA: stanovanje v času študija

Zanima nas, ali študentke v času študija drugače stanujejo kot študenti oz. ali sta spol in stanovanje v času študija povezana. V ta namen podatke študentov pri obeh spremenljivkah uredimo v dvodimenzionalno frekvenčno porazdelitev. To tabelo imenujemo **kontingenčna tabela**.

Denimo, da so podatki za vzorec urejeni v naslednji kontingenči tabeli:

Tabela: Kontingenčna tabela dimenzije 2x3 z vpisanimi empiričnimi frekvencami

	Starši	Št. dom	Zasebno	Skupaj
Moški	16	40	24	80
Ženske	48	36	36	120
Skupaj	64	76	60	200



Robne frekvence

Ker nas zanima, ali študentke drugače stanujejo v času študija kot študenti, moramo porazdelitev stanovanja študentk primerjati s porazdelitvijo študentov. Ker je število študentk različno od števila študentov, moramo zaradi primerjave izračunati relativne frekvence.

Relativne frekvence – strukturni % po vrsticah (po spolu):

	Starši	Št. dom	Zasebno	Skupaj
Moški	20	50	30	100
Ženske	40	30	30	100
Skupaj	32	38	30	100

Interpetacija:

32% vseh študentov živi pri starših, 38% v študentskem domu in 30% zasebno.

Med študenti moškega spola jih 20% živi pri starših, 50% v študentskem domu in 30% zasebno.

Med študentkami jih 40% živi pri starših, 30% v študentskem domu in 30% zasebno.

Vidimo lahko, da glede bivanja zasebno med študenti in študentkami ni razlik. Prihaja pa do razlik pri drugih dveh tipih stanovanja: med študenti jih več živi v študentskem domu, medtem ko jih med študentkami več živi pri starših.

Relativne frekvence – strukturni % po stolpcih (po stanovanju v času študija):

	Starši	Št. dom	Zasebno	Skupaj
Moški	25	57	40	40
Ženske	75	43	60	60
Skupaj	100	100	100	100

Interpetacija:

Med vsemi študenti je 40% moških in 60% žensk.

Med tistimi, ki živijo pri starših, je 25% moških in 75% žensk.

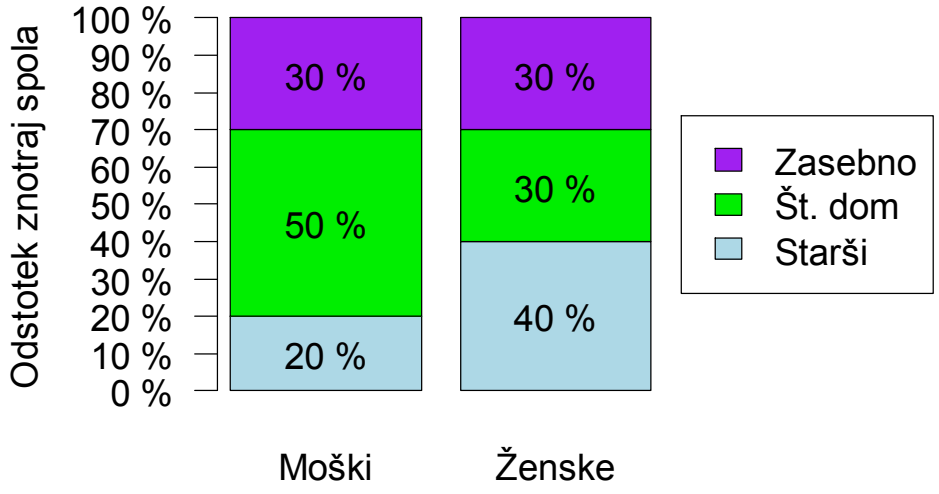
Med tistimi, ki živijo v študentskem domu, je 57% moških in 43% žensk.

Med tistimi, ki živijo zasebno, je 40% moških in 60% žensk.

(Nepravilno, nejasno: Pri starših živi 25% moških in 75% žensk.)

Vidimo lahko, da za ta primer interpretacija po stolpcih ni najbolj primerna. Na osnovi teh podatkov bi namreč lahko napačno zaključili, da pri zasebnikih pogosteje bivajo študentke kot študenti. Vendar dejansko med njimi ni razlik in je porazdelitev pri tej kategoriji popolnoma enaka porazdelitvi po spolu v celotnem vzorcu.

Grafični prikaz primernih strukturnih odstotkov (v našem primeru po vrsticah)



Iz grafa lahko razberemo, da študentke pogosteje kot študenti bivajo pri starših, medtem ko študenti pogosteje bivajo v študentskem domu. Glede bivanja zasebno med njimi ni razlik.

5.2.2 Test povezanosti

Relativni odstotki po spolu (po vrsticah):

	Starši	Št. dom	Zasebno	Skupaj
Moški	20	50	30	100
Ženske	40	30	30	100
Skupaj	32	38	30	100

Če med spoloma ne bi bilo razlik, bi bile v zgornji tabeli obe porazdelitvi (za moške in ženske) enaki porazdelitvi pod "skupaj". Naš primer kaže, da se odstotki razlikujejo: npr. le 20% študentov in kar 40% študentk živi med študijem pri starših. Odstotki v študentskih domovih pa so ravno obratni. Zasebno pa stanuje enak odstotek deklet in fantov. Že pregled relativnih frekvenc kaže, da sta spremenljivki povezani med seboj.

Vprašanje: Ali je do povezanosti na vzorcu prišlo zato, ker na populaciji res obstaja povezanost (res obstaja razlika med spoloma)? Ali pa je to zgolj posledica slučaja izbire enot v vzorec?

Postopek pri testu o povezanosti za nominalni tip para spremenljivk (χ^2 test)

Kontingenčna tabela kaže podatke za slučajni vzorec. Zato nas zanima, ali so razlike v porazdelitvi tipa stanovanja v času študija po spolu statistično značilne (torej obstaja povezanost med spremenljivkama na populaciji) in ne le učinek vzorca (torej slučajne).

V ta namen postavimo naslednji dve domnevi o povezanosti za nominalni tip para spremenljivk na populaciji:

H_o: Spremenljivki na populaciji nista povezani.

H_1: Spremenljivki na populaciji sta povezani.

Za preverjanje teh domnev (oz. ničelne domneve) na osnovi vzorčnih podatkov, podanih v kontingenčni tabeli (dvo-razsežni frekvenčni porazdelitvi), lahko uporabimo **χ^2 test** (hi-kvadrat test, angl. *chi square*). Ta test sloni na primerjavi **empiričnih** (dejanskih) **frekvenc** s frekvencami, kakršne bi bile v primeru nepovezanosti (**teoretične frekvence**) (torej, če bi bili porazdelitvi stanovanja v času študija deklet in fantov enaki). Če se bodo empirične frekvence zadosti razlikovale od teoretičnih frekvenc, bomo sklepali, da na populaciji povezanost obstaja.

Izračun teoretičnih frekvenc

Empirične frekvence (frekvence iz vzorca):

	Starši	Št. dom	Zasebno	Skupaj
Moški	16	40	24	80 $= n_1$
Ženske	48	36	36	120 $= n_2$
Skupaj	64 = m_1	76 = m_2	60 = m_3	200 = n

Teoretična frekvenca za celico *ij* (celica s frekvenco na presečišču *i*-te vrstice in *j*-tega stolpca):

$$f_{ij}' = \frac{n_i \cdot m_j}{n}$$

n_i – robna frekvenca *i*-te vrstice

m_j – robna frekvenca *j*-tega stolpca

n – skupno število enot

Teoretične frekvence (izračunane po zgornjem obrazcu za vsako celico):

	Starši	Št. dom	Zasebno	Skupaj
Moški	25.6	30.4	24	80
Ženske	38.4	45.6	36	120
Skupaj	64	76	60	200

Teoretične frekvence so frekvence, kakršne bi dobili, če med spremenljivkama ne bi bilo povezanosti. V tem primeru bi bile obe porazdelitvi (za "moške" in za "ženske") enaki porazdelitvi pod "skupaj". Torej bi bile relativne teoretične frekvence v vseh treh vrsticah enake.

Relativne teoretične frekvence:

	Starši	Št. dom	Zasebno	Skupaj
Moški	32	38	30	100
Ženske	32	38	30	100
Skupaj	32	38	30	100

χ^2 statistika

χ^2 statistika, ki primerja empirične (dejanske) in teoretične (hipotetične) frekvence, je naslednja:

$\chi^2 = \sum_{i=1}^{v}\sum_{j=1}^{s}\frac{(f_{ij} - f_{ij}')^2}{f_{ij}'}$ ali krajše $\chi^2 = \sum_{i=1}^{k}\frac{(f_i - f_i')^2}{f_i'}$, kjer je *i* oznaka za *i*-to celico med *k* celicami kontingenčne tabele.

χ^2 meri razlike med empiričnimi in teoretičnimi frekvencami. Ko so empirične frekvence enake ali zelo podobne frekvencam, kakršne bi bile v primeru nepovezanosti (torej teoretičnim frekvencam), takrat je $\chi^2 \cong 0$. V tem primeru lahko zaključimo, da vzorčni podatki ne kažejo na povezanost med spremenljivkama.

Če pa se empirične frekvence razlikujejo od teoretičnih frekvenc (takrat je $\chi^2 > 0$), potem lahko zaključimo, da vzorčni podatki kažejo na povezanost med spremenljivkama.

Domnevi

Torej lahko domnevi zapišemo na naslednji način:

H_0: Spremenljivki nista povezani.

$\chi^2 = 0$ (V obliki $\gamma = \gamma_H$.) Ni razlik med empiričnimi in teoretičnimi frekvencami.

H_1: Spremenljivki sta povezani.

$\chi^2 > 0$ So razlike med empiričnimi in teoretičnimi frekvencami.

(Enostranski test.)

Izračunajmo eksperimentalno vrednosti statistike χ^2 za naš primer.

Empirične frekvence f_i:

	Starši	Št. dom	Zasebno	Skupaj
Moški	16	40	24	80
Ženske	48	36	36	120
Skupaj	64	76	60	200

Teoretične frekvence f_i':

	Starši	Št. dom	Zasebno	Skupaj
Moški	25.6	30.4	24	80
Ženske	38.4	45.6	36	120
Skupaj	64	76	60	200

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'} =$$

$$= \frac{(16-25.6)^2}{25.6} + \frac{(40-30.4)^2}{30.4} + \frac{(24-24)^2}{24} + \frac{(48-38.4)^2}{38.4} + \frac{(36-45.6)^2}{45.6} + \frac{(36-36)^2}{36} = 11.05$$

Preverjanje domneve in χ^2 porazdelitev

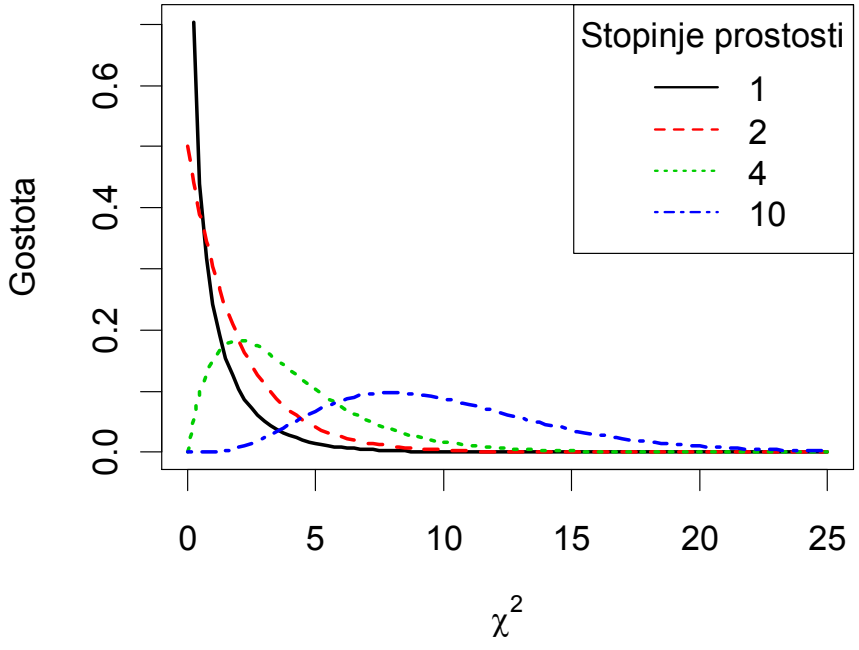
Kdaj je χ^2 zadosti večji od 0, da zavrnemo ničelno domnevo in rečemo, da sta spremenljivki povezani? → Določimo **kritično območje**.

Statitika χ^2 (če jo izračunamo iz kontingenčne tabele na vseh možnih vzorcih velikosti n) se porazdeljuje po **χ^2 porazdelitvi**.

χ^2 porazdelitev:

- Definirana le za pozitivne vrednosti slučajne spremenljivke.
- Unimodalna, asimetrična v desno.
- Odvisna od prostostnih stopenj. Pri manjšem številu prostostnih stopenj je bolj asimetrična v desno, z večanjem prostostnih stopenj pa se približuje normalni porazdelitvi.
- V primeru χ^2 testa za preverjanje domneve o povezanosti za nominalni tip para spremenljivk se statistika χ^2 porazdelujejo po χ^2 porazdelitvi z $m=(s-1)(v-1)$ prostostnimi stopnjami, pri čemer je s število stolpcev, v pa število vrstic v kontingenčni tabeli.

Slika: χ^2 porazdelitve pri različnih prostostnih stopnjah



Kritično območje

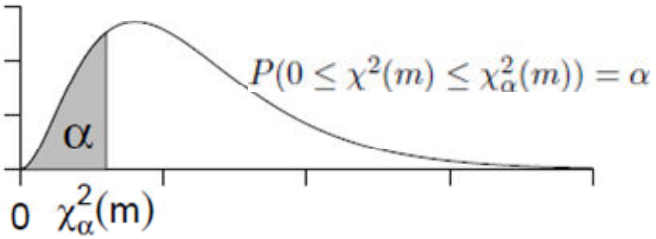
Kako določimo kritično območje?

Iz tabele porazdelitve standardizirane spremenljivke χ^2 razberemo tisto vrednost, ki loči $\alpha\%$ ekstremnih vrednosti (torej loči $\alpha\%$ vzorcev, ki so malo verjetni, od bolj verjetnih vzorcev). Pri tem iščemo vrednosti χ^2 pri verjetnosti $(1-\alpha)$ in prostostnimi stopnjami $[(s-1)(v-1)]$, pri čemer je s št. stolpcev, v pa število vrstic v kontingenčni tabeli. Na ta način določimo kritično območje, t.j. območje zavračanja ničelne domneve.

Poiščimo kritično vrednost za naš primer pri 5% stopnji značilnosti:

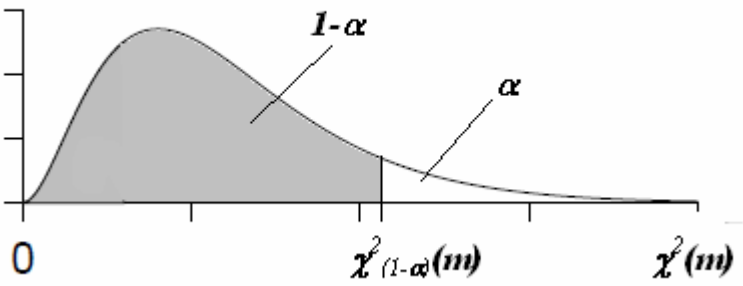
$$\chi^2_{1-\alpha}[(s-1)(v-1)] = \chi^2_{1-0.05}[(3-1)(2-1)] = \chi^2_{0.95}(2) = 5.99$$

Tabela za χ^2 porazdelitev (del tabele):



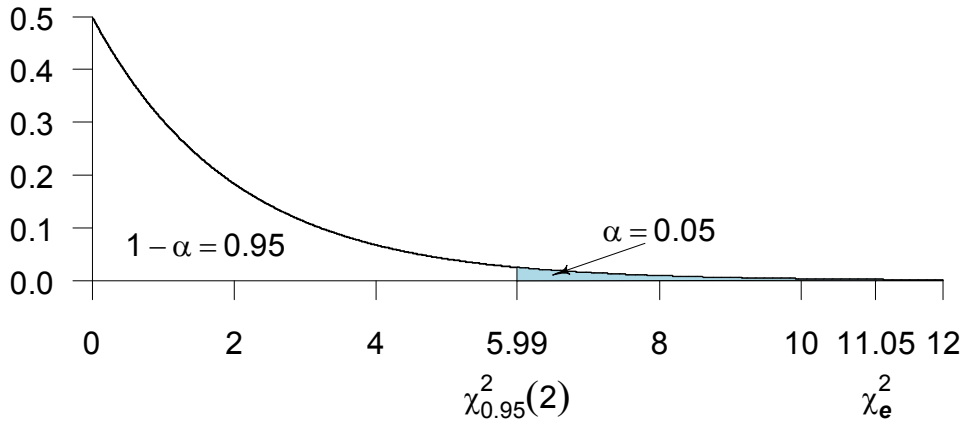
m \ α	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84

V primeru povezanosti za nominalni tip para spremenljivk iščemo $\chi^2_{(1-\alpha)}[(v-1)(s-1)]$.



Sklep

$\chi^2_e = 11.05 > \chi^2_{(1-\alpha)} = 5.99$



Ker eksperimentalna vrednost pade v kritično območje, ničelno domnevo lahko zavrnemo. Pri 5% stopnji značilnosti lahko zaključimo, da sta spremenljivki statistično značilno povezani med seboj, torej da se stanovanje v času študija po spolu razlikuje.

(Izpis iz stat. programa: "sign =.004", kar beremo kot "*p* vrednost je 0.004", kar je manj ko 0.05 in ničelno domnevo zavrnemo).

Tako velika razlika med empiričnimi in teoretičnimi frekvencami, merjena s χ^2 statistiko, kot smo jo dobili na našem vzorcu, se lahko zgodi le pri manj kot 0.5% vseh vzorcih pri domnevi, da ni povezanosti. Torej lahko zavrnemo domnevo o nepovezanosti.

Ponovimo: Postopek pri testu o povezanosti za nominalni tip para spremenljivk

1. Postavimo domnevi:

H_0: Spremenljivki nista povezani.

$\chi^2=0$ Ni razlik med empiričnimi in teoretičnimi frekvencami.

H_1: Spremenljivki sta povezani.

$\chi^2>0$ So razlike med empiričnimi in teoretičnimi frekvencami.

2. Izračunamo teoretične frekvence:

$$f_{il}' = \frac{n_i \cdot m_j}{n} = f_i'$$

3. Izračunamo eksperimentalno vrednost testne statistike:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - f_i')^2}{f_i'}$$

, kjer je *i* oznaka za *i*-to celico med *k* celicami kontingenčne tabele.

4. Poiščemo kritično vrednost $\chi^2_{1-\alpha}[(s-1)(v-1)]$.

5. Sklep: primerjamo χ^2_e in $\chi^2_{(1-\alpha)}$:

$\chi^2_e < \chi^2_{(1-\alpha)}$ H_0 ne zavrnemo. (Spremenljivki nista povezani.)

Obrazložitev: Majhna razlika med empiričnimi in teoretičnimi frekvencami, ki smo jo dobili na vzorcu, je zelo verjetna v primeru, da spremenljivki na populaciji nista povezani. Sklepamo torej, da spremenljivki nista povezani.

$\chi^2_e > \chi^2_{(1-\alpha)}$ H_0 zavrnemo ob stopnji značilnosti α, kot pravilna nam ostane H_1. (Spremenljivki sta povezani).

Obrazložitev: Tako velika razlika med empiričnimi in teoretičnimi frekvencami, kot smo jo dobili na vzorcu, je zelo malo verjetna (le $\alpha\%$ verjetnosti) v primeru, da spremenljivki na populaciji nista povezanosti. Sklepamo torej, da sta spremenljivki povezani.

Lastnosti χ^2 statistike

- Statistika χ^2 je lahko le pozitivna (gre namreč za kvadrate razlik med empiričnimi in teoretičnimi frekvencami). Zato je tudi test preverjanja domneve enostranski ($H_1: \chi^2>0$).
- Zavzame lahko vrednosti na intervalu $[0, \chi^2_{max}]$, kjer je $\chi^2_{max}=n(h-1)$, pri čemer je n število enot v vzorcu ter $h=min(v,s)$ (najmanjše število izmed števila vrstic in stolpcev kontingenčne tabele).
- χ^2 test pove, ali na populaciji obstaja povezanost med obavnavanima nominalnima spremenljivkama. Ne pove pa, kako močna je ta povezanost. χ^2 statistika namreč ni primerljiva med različnimi kontingenčnimi tabelami, ker je njena vrednost odvisna od:
  a) n, števila enot v vzorcu,
  b) števila celic v tabeli (števila vrednosti obravnavanih spremenljivk).

5.2.3 Moč povezanosti: kontingenčni koeficienti

Ker χ^2 statistika ni primerljiva med različnimi tabelami, nam ne podaja informacije o moči povezanosti. Slednjo merimo s kontingenčnimi koeficienti, ki so izpeljani iz χ^2 ter normirani (imajo določen interval možnih vrednosti in so zato primerljivi med različnimi tabelami).

Cramerjev α

$$\alpha = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n(h-1)}}$$, kjer je $h = min(v,s)$. $\alpha \in [0, 1]$

Interpretiramo na naslednji način (za dano skupino enot, npr. vzorec):

$0.05 < \alpha < 0.3$ šibka povezanost

$0.3 < \alpha < 0.6$ srednje močna povezanost

$0.6 < \alpha < 1$ močna povezanost

Pearsonov koeficient kontingence *C*

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Pearsonov koeficient kontigence C je definiran na intervalu $[0, C_{max}]$, pri čemer je

$$C_{\max} = \sqrt{\frac{h-1}{h}}$$, kjer je $h = min(v,s)$.

Kot tak ni primerljiv med različnimi tabelami, zato izračunamo **popravljeni Pearsonov koeficient kontingence C_{pop}:**

$$C_{pop} = \frac{C}{C_{\max}}$$ $C_{pop} \in [0, 1]$

Interpretiramo na naslednji način (za dano skupino enot, npr. vzorec):

$0.1 < C_{pop} < 0.3$ šibka povezanost

$0.3 < C_{pop} < 0.6$ srednje močna povezanost

$0.6 < C_{pop} < 1$ močna povezanost

Naš primer:

Cramejerev α $\alpha = \sqrt{\frac{\chi^2}{n(h-1)}} = \sqrt{\frac{11.05}{200(2-1)}} = 0.23$

... šibka povezanost.

Popravljeni Pearsonov koeficient kontigence C_{pop}:

$$C_{pop} = \frac{C}{C_{\max}} = \frac{\sqrt{\frac{\chi^2}{\chi^2 + n}}}{\sqrt{\frac{h-1}{h}}} = \frac{\sqrt{\frac{11.05}{11.05 + 200}}}{\sqrt{\frac{2-1}{2}}} = 0.32$$

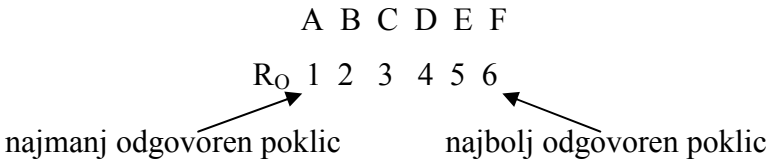
... srednje močna povezanost.

5.3 POVEZANOST ZA ORDINALNI TIP PARA SPREMENLJIVK

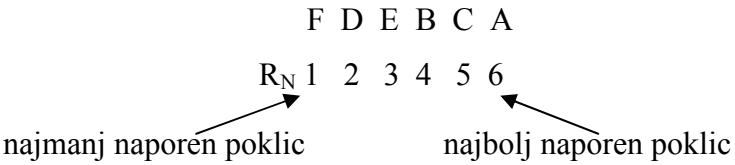
Primer

Vzemimo slučajni vzorec šestih poklicev in ocenimo, koliko so odgovorni (O) in koliko so fizično naporni (N). Poklice nato uredimo od najmanj odgovornega do najbolj odgovornega in podobno, od najmanj fizično napornega do najbolj napornega. Poklice, ki so naše enote analize, torej uredimo v ranžirno vrsto in jim pripišemo rang (t.j. zaporedno mesto v ranžirni vrsti).

Ureditev poklicev po odgovornosti:



Ureditev poklicev po fizični napornosti:



5.3.1 Ureditev podatkov

Podatke zapišemo v tabelo, kjer v prvi stolpec vpišemo enote, v drugi in tretji stolpec pa zaporedno mesto enote v ranžirni vrsti glede na obravnavani lastnosti. Obravnavani spremenljivki sta torej rang glede na odgovornost (R_O) in rang glede na fizično napornost (R_N). Gre za ordinalni spremenljivki, saj smo enote uredili glede na obravnavana kriterija.

Poklic	R_O	R_N
A	1	6
B	2	4
C	3	5
D	4	2
E	5	3
F	6	1

5.3.2 Izračun povezanosti med dvema ordinalnima spremenljivkama

Ko imamo podatke ustrezno urejene v tabelo, lahko povezanost med dvema ordinalnima spremenljivkama merimo s *Spearmanovim koeficientom korelacije rangov*, ki meri povezanost med rangi obravnavanih spremenljivk. Izračunamo ga po obrazcu:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

Oznake v tem obrazcu pomenijo naslednje:
r_s ... oznaka za *Spearmanov koeficient korelacije rangov*, ki ga izračunamo na vzorcu
d_i ... razlika med rangoma za i-to enoto
n ... število enot v vzorcu

Izračun povezanosti za naš primer:

Poklic	R_O	R_N	d_i	d_i^2
A	1	6	-5	25
B	2	4	-2	4
C	3	5	-2	4
D	4	2	2	4
E	5	3	2	4
F	6	1	5	25
Vsota	21	21	0	66

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \cdot 66}{6\left(6^2 - 1\right)} = -0.88$$

Spearmanov koeficient korelacije rangov lahko zavzame vrednosti na intervalu [-1,1]. Pomen ključnih vrednosti v tem intervalu je:

r_s = 1 ... Popolna pozitivna povezanost (za primer glej spodaj).

Če se z večanjem rangov po prvi spremenljivki večajo tudi rangi po drugi spremenljivki, gre za pozitivno povezanost. Takrat je koeficient pozitiven in blizu 1.

r_s = -1 ... Popolna negativna povezanost (za primer glej spodaj).

Če se z večanjem rangov po prvi spremenljivki manjšajo rangi po drugi spremenljivki, gre za negativno povezanost. Takrat je koeficient negativen in blizu -1.

r_s = 0 ... Ni povezanosti med spremenljivkama.

Če ne obstaja nobena povezanost med rangi po prvi in rangi po drugi spremenljivki, takrat je koeficient blizu 0.

Ekstremne vrednosti koeficienta korelacije rangov se v praksi seveda redko zgodijo. V našem konkretnem primeru smo dobili koeficient korelacije rangov r_s = -0.88, kar je blizu -1. To pomeni,

da na izbranem vzorcu poklicev obstaja močna negativna povezanost med spremenljivkama: tisti poklici, ki so bolj odgovorni, so v povprečju fizično manj naporni.

Primer: $r_s = 1$... Bolj kot je poklic odgovoren, bolj je tudi fizično naporen

Poklic	R_O	R_N
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5
F	6	6

Primer: $r_s = -1$... Bolj kot je poklic odgovoren, manj je fizično naporen

Poklic	R_O	R_N
A	1	6
B	2	5
C	3	4
D	4	3
E	5	2
F	6	1

5.3.3 Sklepanje iz vzorca na populacijo – Preverjanje domneve

Na vzorcu poklicev smo izračunali Spearmanov koeficient korelacije rangov, s katerim smo ugotovili, kako močna je povezanost med odgovornostjo in fizično napornostjo teh poklicev na danem vzorcu poklicev.

Zanima nas, ali smo takšno povezanost na vzorcu dobili zato, ker na populaciji poklicev res obstaja povezanost, torej ali lahko posplošimo na vse poklice. Ali pa je povezanost na vzorcu le posledica slučajne izbire poklicev v vzorec?

Na zastavljeno vprašanje odgovorimo tako, da testiramo domnevo o nepovezanosti ordinalnih spremenljivk pri izbrani stopnji značilnosti. Postopek testiranja domnev je običajen, kot ga že poznamo.

1. Postavimo ničelno in alternativno domnevo

H_0: $\rho_s = 0$

Spremenljivki na populaciji nista povezani, torej je Spearmanov koeficient korelacije rangov na populaciji[1] ρ_s enak 0 oz. ni statistično značilno različen od 0.

H_1: $\rho_s \neq 0$

Spremenljivki na populaciji sta povezani, torej je Spearmanov koeficient korelacije rangov na populaciji ρ_s statistično značilno različen od 0[2].

2. Izberemo ustrezno testno statistiko in jo izračunamo

Testna statistika, s katero preverjamo domnevo o povezanosti dveh ordinalnih spremenljivk, je statistika *t*, ki se porazdeljuje približno po (nam že poznani) Studentovi *t* porazdelitvi z (*n* - 2) prostostnimi stopnjami. Porazdelitev testne statistike je torej odvisna od števila enot v vzorcu (število enot v vzorcu – 2).

Testno statistiko t, ki je standardizirani koeficient korelacije rangov, izračunamo po naslednjem obrazcu:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

Oznake v tem obrazcu pomenijo naslednje:

r_s ... Spearmanov koeficient korelacije rangov, izračun na vzorcu

n ... število enot v vzorcu

Za naš primer je izračun eksperimentalne vrednosti testne statistike *t* naslednji:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{-0.88\sqrt{6-2}}{\sqrt{1-(-0.88)^2}} = -3.71$$

3. Določimo kritično vrednost testne statistike pri izbrani stopnji značilnosti

Za naš primer izračunamo vrednost testne statistike pri 5% stopnji značilnosti. Ker gre za dvostranski test, iščemo dve kritični vrednosti, ki sta po absolutni vrednosti enaki. Iščemo torej $\pm t_{\alpha/2}(n - 2)$. Za naš konkreten primer iz tabele Studentove porazdelitve razberemo ustrezno vrednost $\pm t_{\alpha/2}(n - 2) = \pm t_{0.05/2}(6 - 2) = \pm t_{0.025}(4) = \pm 2.776$.

[1] Spearmanov koeficient korelacije rangov na populaciji označujemo z grško črko ρ_s (izgovarjamo "ro") z indeksom "s". Na vzorcu smo ga označili r_s.

[2] V tem primeru imamo dvostranski test. Lahko bi testirali tudi domnevo, da je koeficient statistično značilno večji ali manjši od 0 ($\rho_s > 0$ oz. $\rho_s < 0$), torej bi uporabili enostranski test.

4. Sklep: primerjava kritične in eksperimentalne vrednosti testne statistike *t*

$t_e = -3.71$

$t_{0.025}(4) = \pm 2.776$

$|t_e| > |t_{\alpha/2}|$

Eksperimentalna vrednost pade v kritično območje, torej v območje zavračanja ničelne domneve (eksperimentalna vrednost je po absolutni vrednosti večja od kritične vrednosti). Zavrnemo ničelno domnevo in s 5% stopnjo značilnosti ugotavljamo, da sta na populaciji spremenljivki negativno povezani med seboj. Za vse poklice lahko posplošimo, da so tisti poklici, ki so bolj odgovorni, manj fizično naporni.

5.3.4 Kaj storimo, če imamo spremenljivke različnih merskih lestvic (npr.: eno ordinalno in eno intervalno ali razmernostno spremenljivko)?

V tem primeru spremenljivko, ki ima boljšo mersko lestvico, obravnavamo, kot da ima tako mersko lestvico, kot spremenljivka s slabšo mersko lestvico. Če bi preverjali domnevo o povezanosti med nominalno in ordinalno spremenljivko, bi obe obravnavali kot nominalni in torej uporabili χ^2 test.

Če pa imamo eno ordinalno in eno intervalno ali razmernostno spremenljivko, tak par spremenljivk obravnavamo kot ordinalni tip para spremenljivk. Pred analizo pa moramo enotam glede na vrednost intervalne/razmernostne spremenljivke pripisati rang in ta rang obravnavati kot ordinalno spremenljivko.

Npr. ugotoviti želimo, ali so poklici, ki so bolj odgovorni, tudi bolje plačani. Poleg že znanih podatkov o odgovornosti vzorca šestih poklicev imamo še podatek o povprečnem mesečnem osebnem dohodku za ta vzorec poklicev. Poklice uredimo v ranžirno vrsto glede na velikost povprečnega mesečnega osebnega dohodka, npr. od tistega z najmanjšim do tistega z največjim povprečnim osebnim dohodkom. Poklicem nato pripišemo ustrezni rang za spremenljivko R_D (rang glede na dohodek).

Poklic	*Dohodek*	R_D
A	800	1
C	1000	2
B	1500	3
D	1800	4
F	2000	5
E	2500	6

Spremenljivki (rang glede na fizično odgovornost in rang glede na dohodek) uredimo v ustrezno tabelo. Nato nadaljujemo s postopkom, kot je opisano zgoraj.

Poklic	R_O	R_D
A	1	1
B	2	3
C	3	2
D	4	4
E	5	6
F	6	5

5.3.5 Kaj storimo, če so kakšne vrednosti enake?

Včasih se zgodi, da imata dve ali več enot glede na obravnavano ordinalno spremenljivko enake vrednosti. V tem primeru je potrebno izračunati povprečne pripadajoče range.

Npr. v vzorcu 8 poklicev so poklici A, B in C najmanj in enako odgovorni. Po odgovornosti sledi poklic D, nato E, nato pa F in G, ki sta spet enako odgovorna. Najbolj odgovoren pa je poklic H. Poklicem, ki so enako odgovorni, je potrebno pripisati enak rang, in sicer povprečni rang med siceršnjimi zaporednimi številkami v ranžirni vrsti. Nato nadaljujemo s postopkom, kot je opisano zgoraj.

A = B = C < D < E < F = G < H

Napačno R_O	1	2	3	4	5	6	7	8
Pravilno R_O	2	2	2	4	5	6.5	6.5	8

2 kot povprečje (1+2+3)/3=2

6.5 kot povprečje (6+7)/2=6.5

5.4 POVEZANOST ZA INTERVALNI/RAZMERNOSTNI TIP PARA SPREMENLJIVK

5.4.1 Linearna povezanost dveh intervalnih/razmernostnih spremenljivk

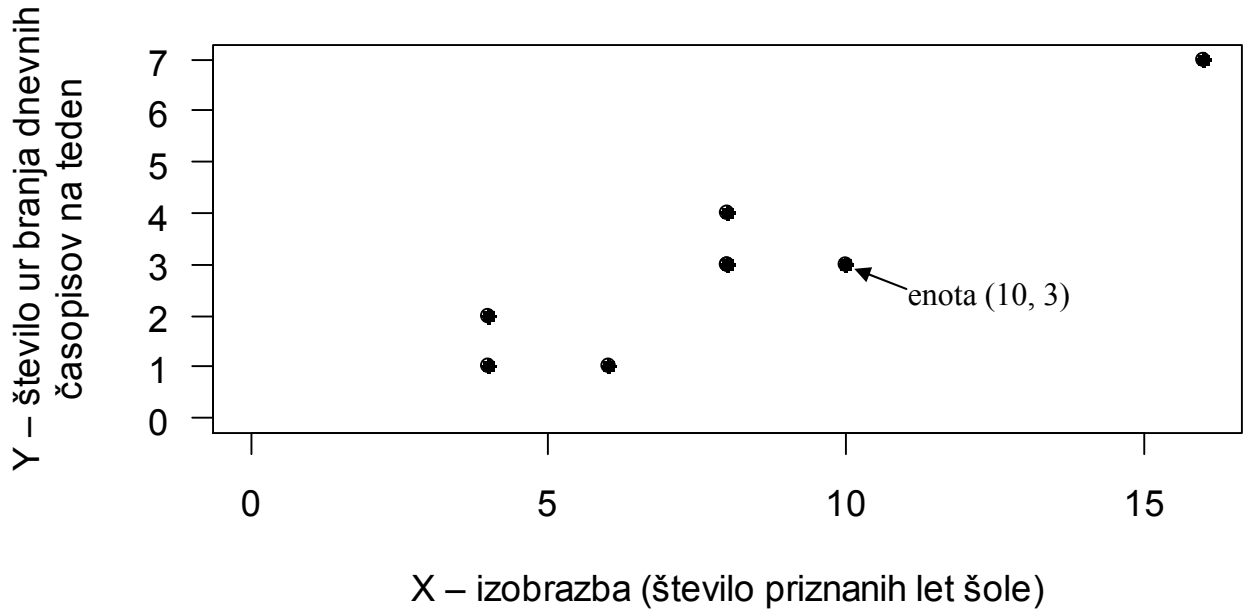
Vzemimo primer dveh razmernostnih spremenljivk:

- *X – izobrazba (število priznanih let šole)*
- *Y – število ur branja dnevnih časopisov na teden*

Podatki za 8 slučajno izbranih oseb so:

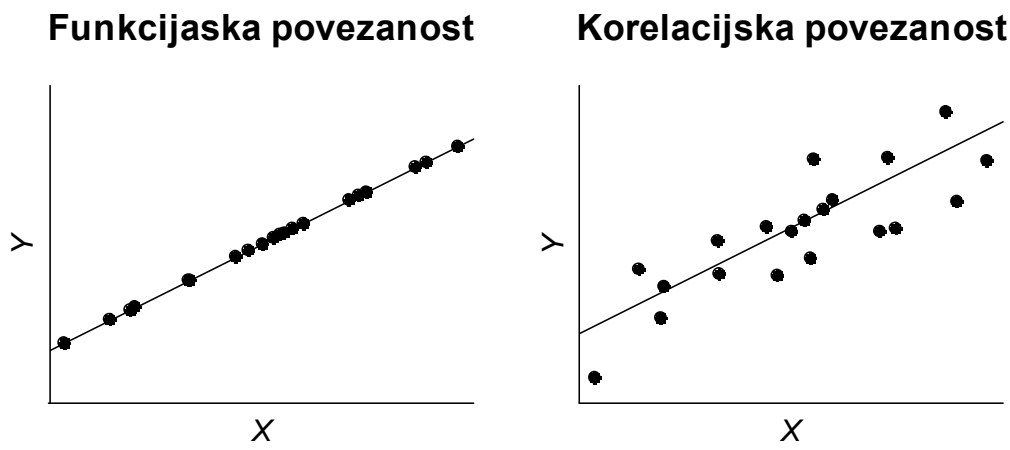
X	*Y*
10	3
8	4
16	7
8	3
6	1
4	2
8	3
4	1

Grafično lahko ponazorimo povezanost med dvema intervalnima/razmernostnima spremenljivkama z **razsevnim grafikonom**. To je graf, v katerem v koordinatni sistem, kjer koordinati predstavljata spremenljivki, vrišemo enote s pari vrednosti obeh spremenljivk. V našem primeru je razsevni grafikon:

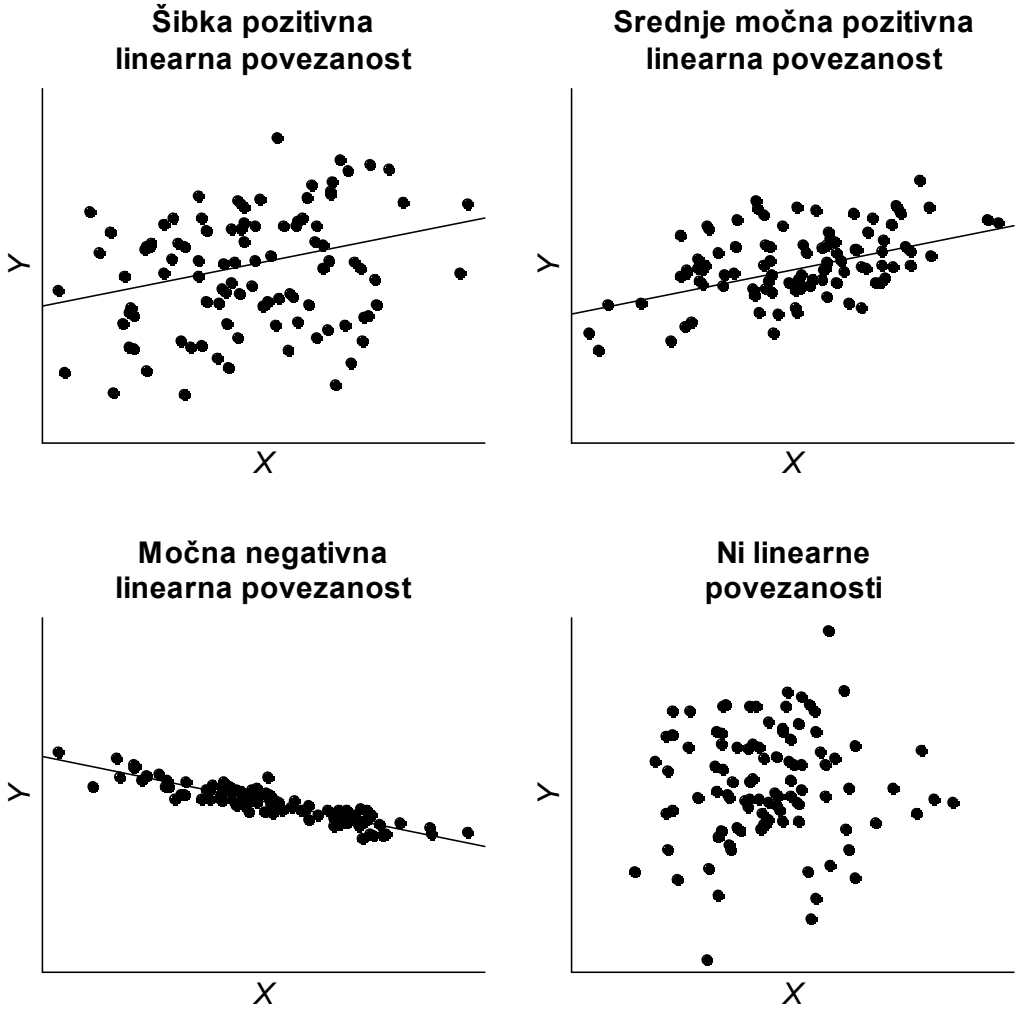


Iz razsevnega grafikona lahko razberemo **tipe povezanosti** med spremenljivkama:

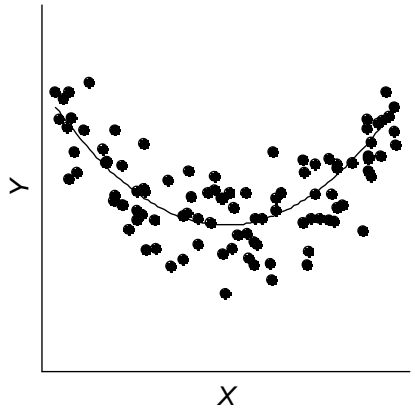
- **funkcijska** povezanost: vse točke ležijo na krivulji;
- **korelacijska** (stohastična) povezanost: točke se od krivulje bolj ali manj odklanjajo (večja ali manjša povezanost).



Zelo pogosta je **linearna povezanost** med dvema spremenljivkama (točke se gostijo okoli premice – linearne funkcije). Tipični primeri linearne povezanosti spremenljivk:



Možna je tudi **nelinearna povezanost** med spremenljivkama. Primer nelinearne povezanosti:



Npr. za avtomobile: povezanost med starostjo avtomobila in vrednostjo avtomobila.

5.4.2 Kovarianca in Pearsonov koeficient korelacije

Povezanost za intervalni/razmernostni tip para spremenljivk lahko merimo s **kovarianco**, ki meri **linearno povezanost** med spremenljivkama.

Populacija

$$C_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

Vzorec

$$C_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

$(x_i-\mu_X)$ ... Razlika (razdalja) med vrednostjo spremenljivke X za i-to enoto in aritmetično sredino spremenljivke X (odklon enote od povprečja za X).

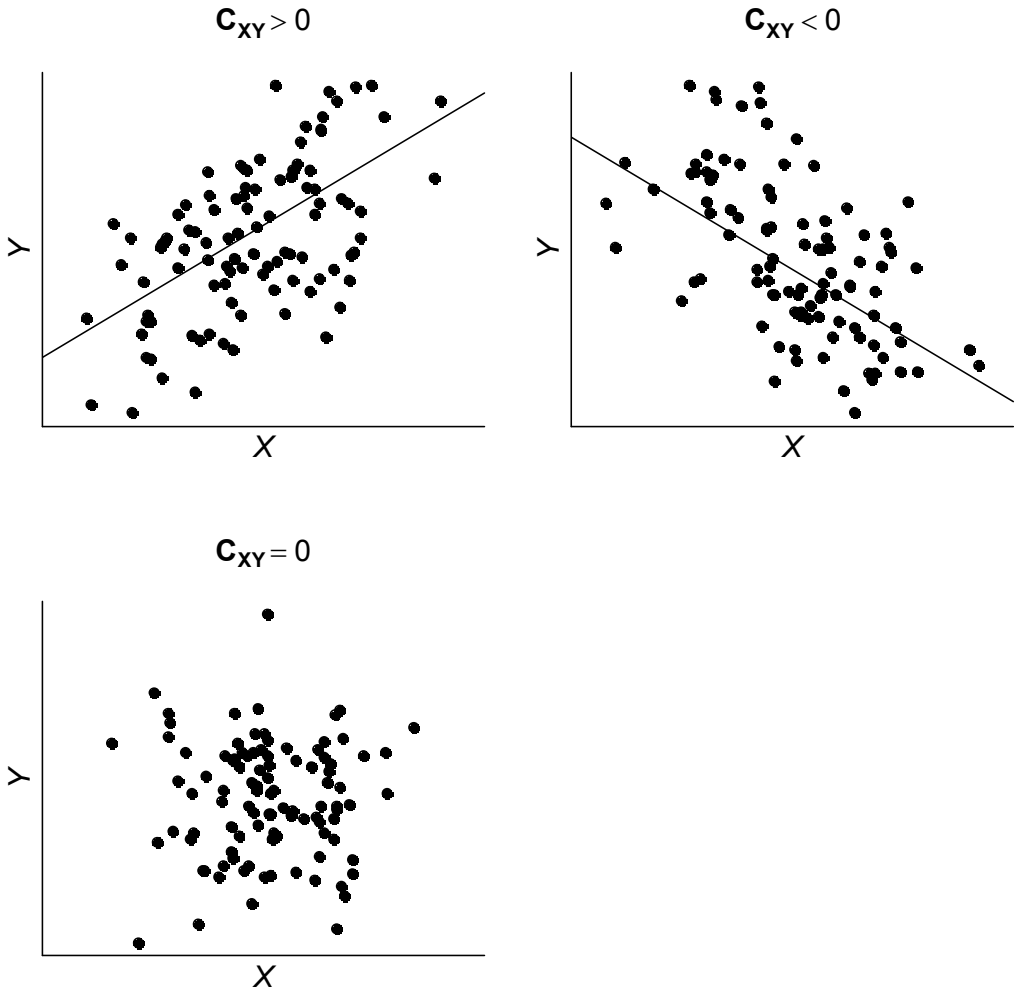
$(y_i-\mu_Y)$ ... Razlika (razdalja) med vrednostjo spremenljivke Y za i-to enoto in aritmetično sredino spremenljivke Y (odklon enote od povprečja za Y).

Interpretiramo: $C_{XY} > 0$ pozitivna linearna povezanost

$C_{XY} = 0$ ni linearne povezanosti

$C_{XY} < 0$ negativna linearna povezanost

Primeri linearne (ne)povezanosti in kovarianca:



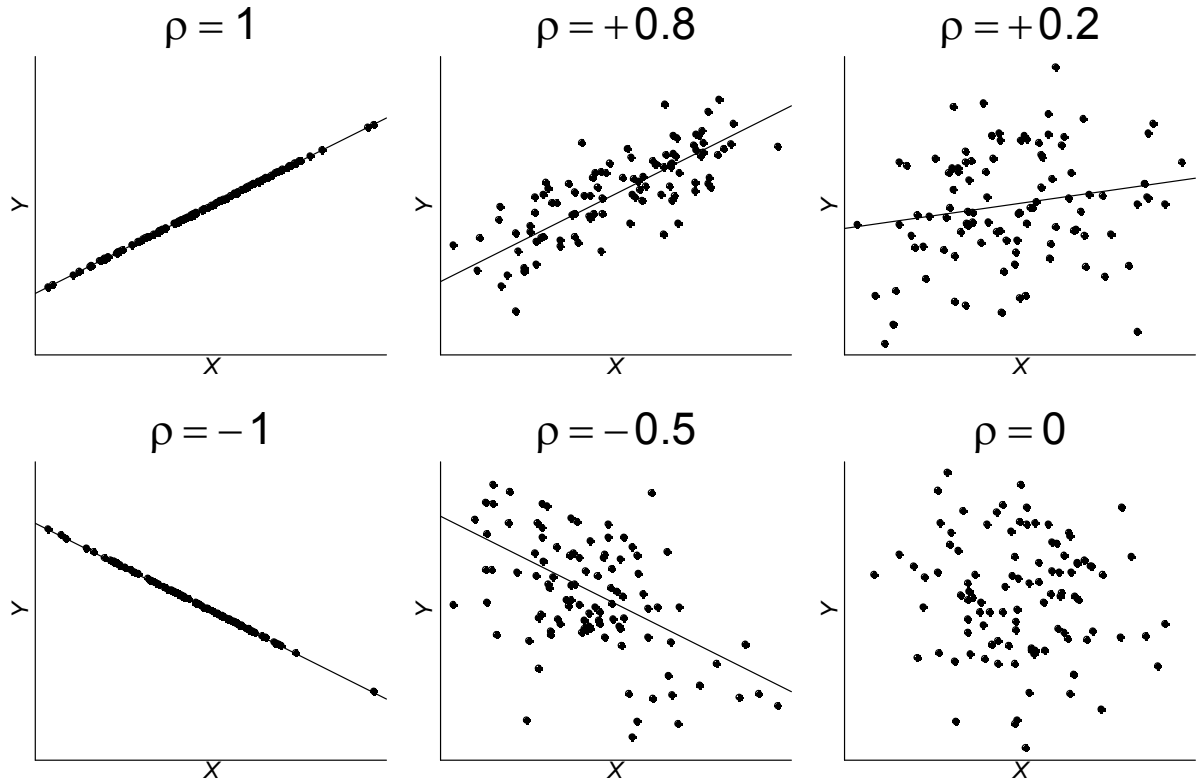
Ker kovarianca ni primerljiva, računamo **Pearsonov koeficient korelacije** kot mero linearne povezanosti med dvema intervalnima/razmernostnima spremenljivkama.

$$\rho_{XY} = \frac{C_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \cdot \sum_{i=1}^{N}(y_i - \mu_Y)^2}}$$

Koeficient korelacije lahko zavzame vrednosti v intervalu [-1, 1].

$\rho > 0$ oz. blizu 1	Pozitivna linearna povezanost - če se z večanjem vrednosti prve spremenljivke večajo vrednosti tudi druge spremenljivke.
$\rho < 0$ oz. blizu -1	Negativna linearna povezanost - če se z večanjem vrednosti prve spremenljivke vrednosti druge spremenljivke manjšajo.
$\rho \sim 0$ (blizu 0)	Ne gre niti za pozitivno, niti za negativno linearno povezanost, spremenljivki nista linearno povezani.

Primeri linearne (ne)povezanosti in koeficient korelacije:



5.4.3 Sklepanje iz vzorca na populacijo: test povezanosti

Če imamo podatke za vzorec, izračunamo vzorčni koeficient korelacije (interpretiramo ga na enak način kot za populacijo):

$$r_{XY} = \frac{C_{YX}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

Na osnovi vzorčnega koeficienta korelacije želimo oceniti, ali obstaja povezanost na populaciji ali pa je njegova vrednost zgolj posledica slučaja izbire enot v vzorec.

Statistično sklepanje o korelacijski povezanosti – test o povezanosti za intervalni/razmernostni tip para spremenljivk

Postavimo **ničelno in alternativno domnevo**:

H_0: $\rho = 0$ Na populaciji spremenljivki nista linearno povezani.

H_1: $\rho \neq 0$ Na populaciji sta spremenljivki linearno povezani.

(Alternativna domneva je lahko tudi:

H_1: $\rho > 0$ Na populaciji sta spremenljivki pozitivno linearno povezani.

H_1: $\rho < 0$ Na populaciji sta spremenljivki negativno linearno povezani.)

Testna statistika (standardizirani vzorčni koeficient korelacije), s katero testiramo domnevo, je

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Testa statistika se porazdeluje po Studentovi *t* porazdelitvi z *m=(n-2)* prostostnimi stopnjami.

Primer

Preverimo domnevo, da sta izobrazba (merjena s številom priznanih let šole) in število ur branja dnevnih časopisov na teden povezana med seboj pri 5% stopnji značilnosti.

1. Izračunamo vzorčni koeficient korelacije.

x_i	y_i	$x_i-\overline{X}$	$y_i-\overline{Y}$	$(x_i-\overline{X})^2$	$(y_i-\overline{Y})^2$	$(x_i-\overline{X})(y_i-\overline{Y})$
10	3	2	0	4	0	0
8	4	0	1	0	1	0
16	7	8	4	64	16	32
8	3	0	0	0	0	0
6	1	-2	-2	4	4	4
4	2	-4	-1	16	1	4
8	3	0	0	0	0	0
4	1	-4	-2	16	4	8
64	24	0	0	104	26	48

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}\cdot 64 = 8 \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{8}\cdot 24 = 3$$

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{X})^2 \cdot \sum_{i=1}^{n}(y_i-\overline{Y})^2}} = \frac{\sum_{i=1}^{8}(x_i-8)(y_i-3)}{\sqrt{\sum_{i=1}^{8}(x_i-8)^2 \cdot \sum_{i=1}^{8}(y_i-3)^2}} =$$

$$= \frac{48}{\sqrt{104\cdot 26}} = 0.92$$

Pearsonov koeficient korelacije na vzorcu je velik (blizu 1) in pozitiven, torej na vzorcu obstaja močna pozitivna povezanost med spremenljivkama. Tiste osebe, ki imajo več priznanih let šole, tudi pogosteje berejo dnevne časopise.

Ali se je povezanost na vzorcu zgodila, ker dejansko na populaciji obstaja povezanost med spremenljivkama, ali pa je to zgolj slučaj?

2. Postavimo ničelno in alternativno domnevo.

H_0: $\rho = 0$ Na populaciji spremenljivki nista linearno povezani. (Izobrazba in pogostost branja dnevnih časopisov nista linearno povezana.)

H_1: $\rho \neq 0$ Na populaciji sta spremenljivki linearno povezani. (Izobrazba in pogostost branja dnevnih časopisov sta linearno povezana).

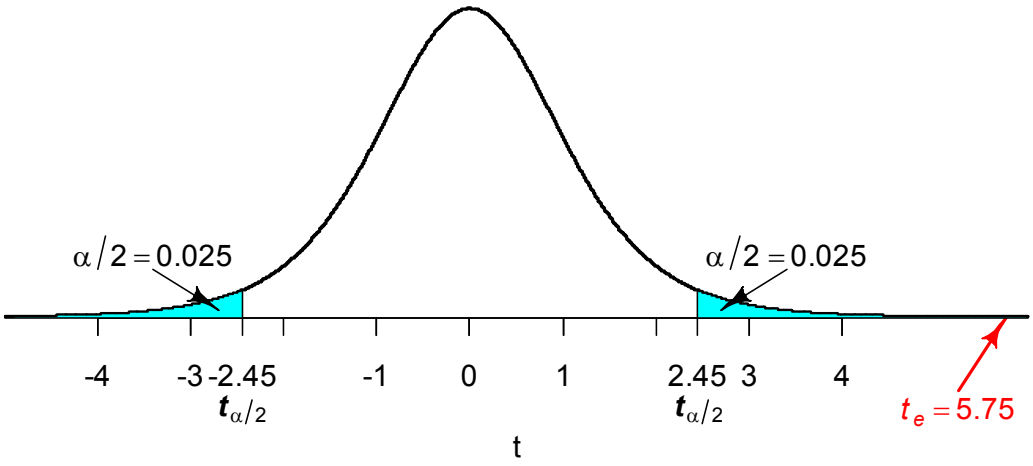
3. Izračunamo eksperimentalno vrednost testne statistike.

$$t_e = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.92\sqrt{8-2}}{\sqrt{1-0.92^2}} = 5.75$$

4. Določimo kritično območje.
Ker gre za dvostranski test, je kritično območje določeno s kritičnima vrednostima:

$$\pm t_{\alpha/2}(n\text{-}2) = \pm t_{0.05/2}(8\text{-}2) = \pm t_{0.025}(6) = \pm 2.45$$

5. Sklep.



$5.75 > 2.45$ oz. $|t_e| > |t_{\alpha/2}|$ H_0 zavrnemo, spremenljivki sta linearno povezani.

Ob 5% stopnji značilnosti ugotavljamo, da sta izobrazba in število ur branja dnevnih časopisov na teden na populaciji statistično značilno povezani med seboj.

(Izpis iz stat. programa: $p = .001$, kar je $P(t > 5.75)+P(t < -5.75)$. Ker je p manjši od 0.05, H_0 zavrnemo. Verjetnost, da dobimo na vzorcu koeficient korelacije po absolutni vrednosti 0.92 ali več oz. t-statistiko po absolutni vrednosti 5.75 ali več, ob predpostavki, da na populaciji ni linearne povezanosti, je zgolj 0.1%. Zato podvomimo v pravilnost ničelne domneve o nepovezanosti.)

5.5 ODVISNOST ZA INTERVALNI/RAZMERNOSTNI TIP PARA SPREMENLJIVK: REGRESIJA

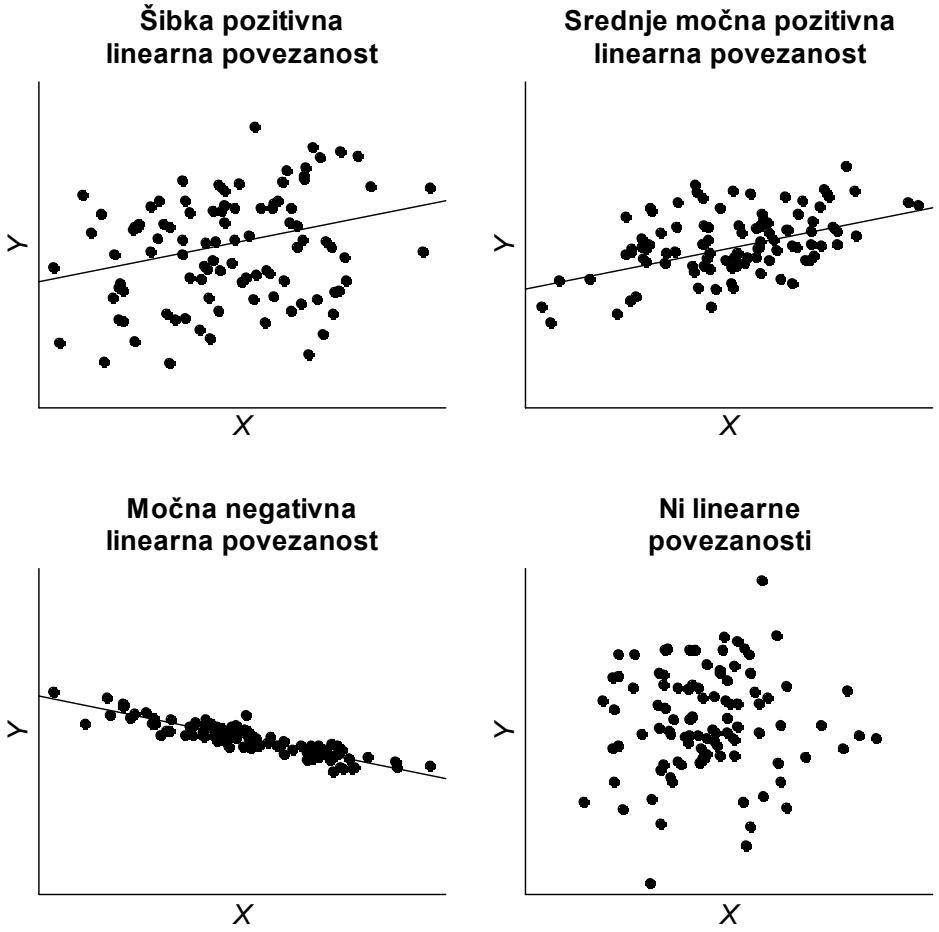
X	⟷	Y	Povezanost
X	⟶	Y	Odvisnost

Za intervalni/razmernostni tip para spremenljivk merilo lahko:

1. povezanost s Pearsonovim koeficientom korelacije,
2. odvisnost: z regresijsko analizo.

5.5.1 Regresijska analiza - uvod

Problem regresijske analize: Kako med točke v razsevnem grafikonu vrisati krivuljo (npr. premico), da bo najbolje ponazarjala odvisnost med spremenljivkama?



- Regresijska analiza = metoda za analizo odvisnosti med intervalnimi/razmernostnimi spremenljivkami. Natančneje, metoda za pojasnjevanje in napovedovanje vrednosti odvisnih spremenljivk s pomočjo vrednosti neodvisnih spremenljivk.
- Tipi regresijske analize glede na število spremenljivk:
  - Bivariatna regresija: ena neodvisna, ena odvisna spremenljivka.
  - Multipla regresija: več neodvisnih, ena odvisna spremenljivka.
- Tipi regresijske analize glede na obliko regresijske funkcije (odvisnosti):
  - Linearna regresija: regresijska funkcija je linearna.
  - Nelinearna regresija: regresijska funkcija ni linearna.
- V okviru predmeta Statistika bomo obravnavali **bivariatno linearno regresijo**.

5.5.2 Postopek regresijske analize

1. S pregledom grafične predstavitve odnosa med dvema spremenljivkama v razsevnem grafikonu ugotovimo, ali se med spremenljivkama nakazuje linearna povezanost.
2. Če lahko spremenljivki vsebinsko definiramo kot odvisno (Y) in neodvisno (X) spremenljivko, je smiselno izračunati regresijsko premico. S premico bomo pojasnili odvisnost odvisne spremenljivke Y od neodvisne spremenljivke X. Če imamo vzorec, pri tem preverjamo tudi statistično značilnost regresijskega koeficienta, ki nastopa v regresijski premici.
3. S pomočjo premice lahko tudi napovedujemo vrednosti odvisne spremenljivke Y s pomočjo vrednosti neodvisne spremenljivke X.
4. Kakovost pojasnjevanja odvisnosti ocenimo z determinacijskim koeficientom, kakovost napovedovanja vrednosti pa s standardno napako ocene.

5.5.3 Regresijska premica

Kako določimo **linearno regresijsko funkcijo – regresijsko premico**?

Grafično: Regresijska premica je med točke v razsevnem grafikonu postavljena tako, da so odkloni točk od premice čimmanjši. Na ta način najbolje ponazarja odvisnost med spremenljivkama.

Matematično: Regresijska premica je določena z **metodo najmanjših kvadratov**, pri kateri iščemo takšne vrednosti napovedane odvisne spremenljivke (točke na premici), da bodo kvadrati odklonov pravih vrednosti (dejanskih točk) od teh napovedanih vrednosti (točk na premici) čimmanjši.

Definirajmo ....

Regresijska funkcija $Y' = f(X)$ kaže, kakšen bi bil vpliv spremenljivke X na Y, če razen vpliva spremenljivke X ne bi bilo drugih vplivov na spremenljivko Y. Ker pa so ponavadi še drugi vplivi na proučevano spremenljivko Y, se točke, ki predstavljajo enote v razsevnem grafikonu, odklanjajo od idealne krivulje

$$Y = Y' + E = f(X) + E$$

kjer X imenujemo **neodvisna spremenljivka**, Y **odvisna spremenljivka** in E **člen napake** (ali motnja, disturbanca).

Če je regresijska funkcija linearna, zapišemo:

$$Y' = f(X) = a + bX$$

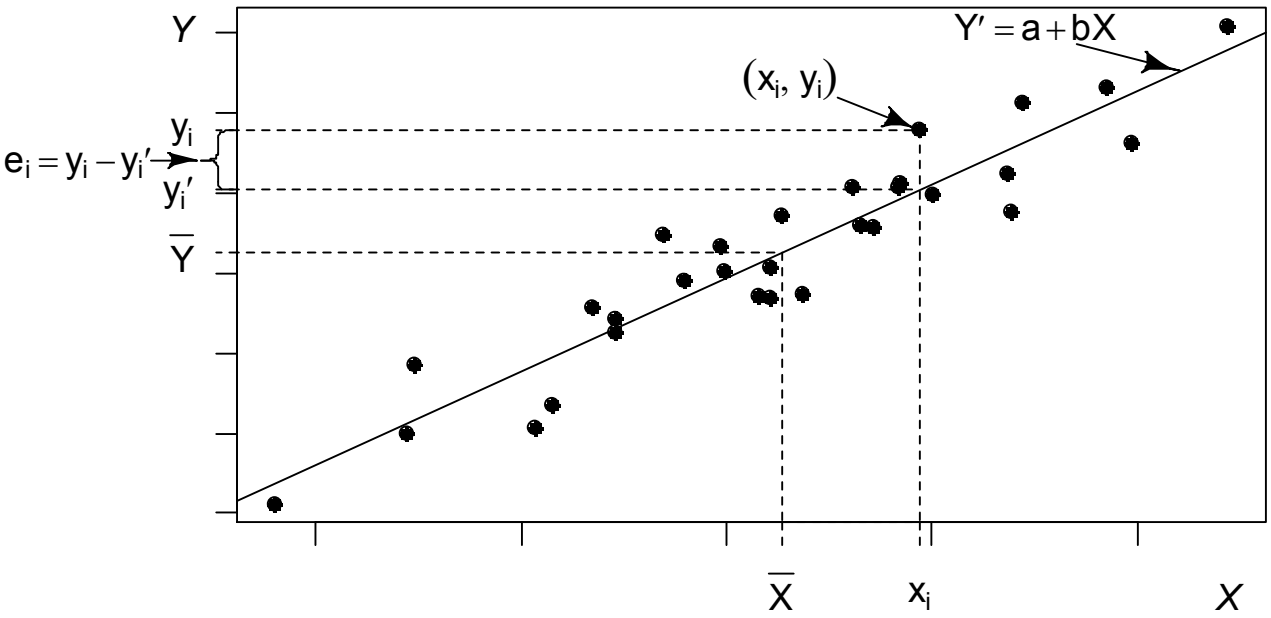
V tem primeru je regresijska odvisnost:

$$Y = Y' + E = a + bX + E$$

oz. za *i*-to enoto

$$y_i = y_i' + e_i = a + bx_i + e_i$$

Regresijsko odvisnost si lahko nazorno predstavimo v razsevnem grafikonu.



Linearna regresijska funkcija je le ena od možnih regresijskih funkcij. Sicer pa regresijsko funkcijo lahko v splošnem zapišemo

$$Y' = f(X, a, b, ...)$$

kjer so *a*, *b*, ... parametri funkcije. Ponavadi se moramo na osnovi pregleda razsevnega grafikona odločiti za tip regresijske funkcije in nato oceniti parametre funkcije, tako da se regresijska krivulja kar se da dobro prilega točkam v razsevnem grafikonu. Kot merilo prilagojenosti krivulje točkam vzamemo

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i')^2 = \min$$

$e_i = y_i - y_i'$ ... razlika med dejansko vrednostjo y_i in vrednostjo y_i', napovedano z regresijsko funkcijo (npr. premico) – odklon točke od krivulje (npr. premice) (napaka, rezidual).

To metodo ocenjevanja parametrov regresijske funkcije imenujemo **metoda najmanjših kvadratov**. Želimo namreč, da je krivulja (funkcija) postavljena tako, da so kvadrati odklonov točk od premice čim manjši.

Ocenjevanje parametrov linearne regresijske funkcije – linearna regresija

Y'= f(X, a, b, ...) ... kakršnakoli funkcija (krivulja)

Y'= a+bX ... primer regresijske funkcije – linearna funkcija (krivulja je premica) → linearna regresijska funkcija

V primeru linearne regresijske funkcije lahko ocenimo parametra *a* in *b* po metodi najmanjših kvadratov takole:

$$F = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i')^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2 = \min$$

Minimum funkcije F lahko določimo tako, da parcialno odvajamo po obeh parametrih

$$\frac{\partial F}{\partial a} = 0; \qquad \frac{\partial F}{\partial b} = 0$$

Dobimo sistem dveh linearnih enačb, iz katerih lahko izračunamo neznana parametra *a* in *b:*

$$a = \mu_Y - \frac{C_{XY}}{\sigma_X^2}\mu_X \qquad b = \frac{C_{XY}}{\sigma_X^2}$$

Če izračunana parametra *a* in *b* vstavimo v regresijsko funkcijo *Y' = a + bX*, dobimo

$$Y' = \mu_Y + \frac{C_{XY}}{\sigma_X^2}(X - \mu_X)$$

To funkcijo imenujemo tudi **prva regresijska funkcija**.

Podobno bi lahko ocenili linearno regresijsko funkcijo, kjer je *X* odvisna in *Y* neodvisna spremenljivka:

$$X' = a^* + b^* Y$$

Če z metodo najmanjših kvadratov podobno ocenimo parametra *a** in *b**, dobimo

$$X' = \mu_X + \frac{C_{XY}}{\sigma_Y^2}(Y - \mu_Y)$$

To funkcijo imenujemo **druga regresijska funkcija**.

Za vsak par spremenljivk je mogoče izračunati obe regresijski funkciji, **le ena od njih pa je običajno vsebinsko smiselna**.

$$Y' = \mu_Y + \frac{C_{XY}}{\sigma_X^2}(X - \mu_X)$$

Y' ... odvisna spremenljivka
X ... neodvisna spremenljivka
μ_Y ... aritmetična sredina odvisne spremenljivke *Y*
μ_X ... aritmetična sredina neodvisne spremenljivke *X*
σ_X^2 ... varianca neodvisne spremenljivke *X*
C_{XY} ... kovarianca med spremenljivkama *X* in *Y*

Izračun regresijske premice, če imamo podatke za **populacijo**:

$$Y' = \mu_Y + \frac{C_{XY}}{\sigma_X^2}(X - \mu_X)$$

Izračun regresijske premice, če imamo podatke za **vzorec**:

$$Y' = \overline{Y} + \frac{C_{XY}}{s_X^2}(X - \overline{X}) \quad \text{, kjer sta}$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \quad \text{in} \quad C_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})$$

Lahko pa enako zapišemo/izračunamo tudi takole (če imamo podatke za **vzorec)**:

$$Y' = a + bX$$

$$b = \frac{C_{XY}}{s_X^2}, \quad a = \overline{Y} - b\overline{X}$$

Primer
Vzemimo primer 8 oseb, ki smo ga obravnavali v poglavju o povezanosti dveh intervalnih/razmernostnih spremenljivk. Spremenljivki sta bili:
X – izobrazba (število priznanih let šole)
Y – število ur branja dnevnih časopisov na teden

Spomnimo se podatkov za teh 8 slučajno izbranih oseb:

X	Y
10	3
8	4
16	7
8	3
6	1
4	2
8	3
4	1

Zanje izračunajmo obe regresijski premici in ju vrišimo v razsevni grafikon.

Ko smo računali koeficient korelacije, smo že izračunali aritmetični sredini:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}\cdot 64 = 8 \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{8}\cdot 24 = 3$$

Izračunali smo tudi vsoti kvadratov odklonov od aritmetične sredine za obe spremenljivki:

$$\sum_{i=1}^{n}(x_i - \overline{X})^2 = 104 \quad , \qquad \sum_{i=1}^{n}(y_i - \overline{Y})^2 = 26 \quad ,$$

in vsoto produktov odklonov od obeh aritmetičnih sredin:

$$\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = 48$$

Prva regresijska premica je potem:

$$Y' = \overline{Y} + \frac{C_{XY}}{s_X^2}(X - \overline{X}) = \overline{Y} + \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}(X - \overline{X}) =$$

$$= 3 + \frac{48}{104}(X - 8) = -0.68 + 0.46X$$

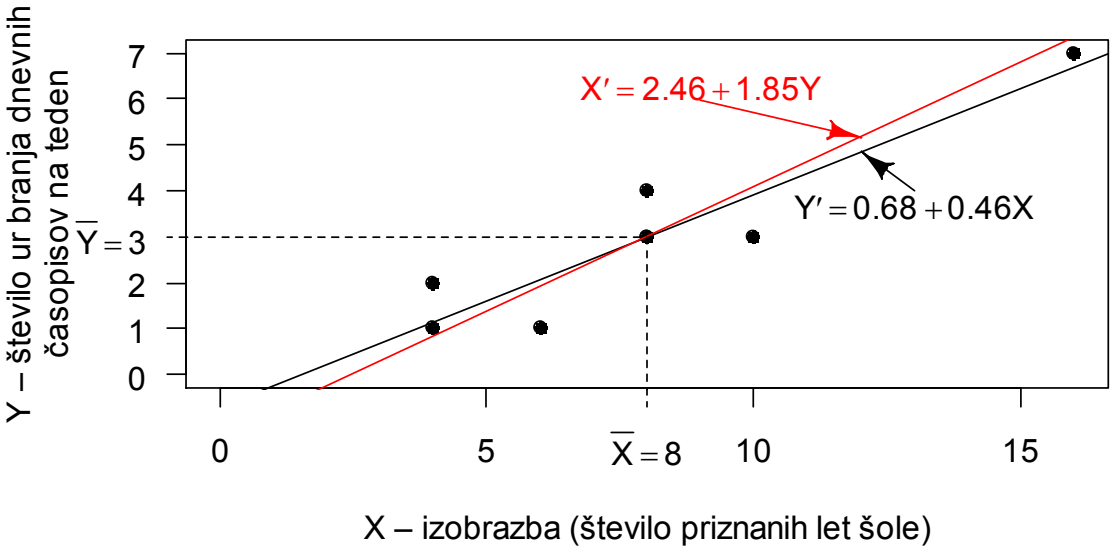
Druga regresijska premica pa je:

$$X' = \overline{X} + \frac{C_{XY}}{s_Y^2}(Y - \overline{Y}) = \overline{X} + \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(y_i - \overline{Y})^2}(Y - \overline{Y}) =$$

$$= 8 + \frac{48}{26}(Y - 3) = 2.46 + 1.85Y$$

Obe regresijski premici lahko vrišemo v razsevni grafikon in preverimo, če se res najbolje prilegata točkam v grafikonu.

Vsebinsko smiselna je v tem primeru prva regresijska premica, ki prikazuje odvisnost števila ur branja dnevnih časopisov na teden (odvisna spremenljivka *Y*) od števila priznanih let šolanja (neodvisna spremenljivka *X*).



Regresijski premici se sekata v točki $(\overline{X}, \overline{Y})$, (8, 3), določeni z aritmetičnima sredinama spremenljivk *X* in *Y*.

Kaj je razvidno iz regresijske premice?

$$Y' = a + bX$$

Odnos med *Y* in *X* je odvisen od dveh parametrov:

a ... določa, kje regresijska premica seka ordinato, *a=Y'(0)*, torej *Y'* ima vrednost *a*, ko ima *X* vrednost *0*.

b ... določa naklon premice (pozitivna ali negativna povezanost in moč povezanosti)

b imenujemo REGRESIJSKI KOEFICIENT. Pove, za koliko se spremeni vrednost Y, če se X spremeni za eno enoto. Če je $b=0$, potem Y ni odvisna od X (spremenljivka Y je konstanta, za katerokoli vrednost spremenljivke X ima isto vrednost, t.j. $Y' = a$).

S pomočjo regresijske premice lahko napovedujemo vrednosti odvisne spremenljivke: izračunamo vrednost spremenljivke Y pri dani vrednosti spremenljivke X.

$$Y'(x_i)=y_i'=a+bx_i$$

Primer
Primer 8 oseb, za katere imamo podatek o X – izobrazbi (število priznanih let šole) in Y - številu ur branja dnevnih časopisov na teden.
Izračunali smo regresijsko premico $Y' = -0.68 + 0.46X$, ki pojasnjuje, kako je število ur branja dnevnih časopisov (odvisna spremenljivka Y) odvisno od izobrazbe (neodvisna spremenljivka X).

Kaj lahko razberemo iz te regresijske premice?

Napovejmo vrednost, npr. $X = 10 \rightarrow Y' = -0.68 + 0.46*10 = 3.92$.
Če bi bila pogostost branja dnevnih časopisov na teden odvisna le od izobrazbe in nobenega drugega dejavnika, potem bi za osebo, ki ima 10 let šolanja, napovedali, da bere dnevne časopise približno 4 ure na teden.

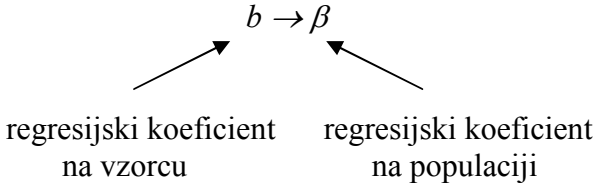
$b = 0.46$ ... Pozitivna odvisnost: tisti, ki imajo več let šolanja, tudi pogosteje berejo dnevne časopise.
Za vsako dodatno leto šolanje se število ur branja poveča za 0.46 ure. Npr. $X = 10 \rightarrow Y' = 3.92$, $X = 11 \rightarrow Y' = 4.38$, razlika med njima je ravno 0.46 ure.

(a = -0.68 ... Pomenilo bi, koliko ur tedensko berejo dnevne časopise osebe, ki nimajo nobenega priznanega leta šole. V tem primeru vsebinsko nesmiselno.)

5.5.4 Sklepanje iz vzorca na populacijo – statistično sklepanje o regresijskem koeficientu

Na vzorcu smo odvisnost spremenljivke Y od spremenljivke X opisali z linearno funkcijo $Y' = a + bX$ (**regresijska premica na vzorcu**).

Zanima nas, kakšna odvisnost med spremenljivkama velja za populacijo, torej kakšna je $Y' = \alpha + \beta X$ (**regresijska premica na populaciji**).



Testiramo **domnevo o regresijskem koeficientu**, najpogosteje domnevo, da je $\beta = 0$, torej da Y ni odvisen od X.

Testiranje domneve o regresijskem koeficientu β

Postavimo domnevi: $H_0: \beta = \beta_H$ $\quad$ $H_1: \beta \neq \beta_H$

Nepristranska cenilka za populacijski regresijski koeficient β je **vzorčni regresijski koeficient:**

$$b = \frac{C_{XY}}{s_X^2}$$

Če ga izračunamo na vseh možnih vzorcih, se porazdeljuje po Studentovi *t* porazdelitvi z $m = (n-2)$ prostostnimi stopnjami in

$$E(b) = \beta \quad \text{in} \quad SE(b) = \frac{s_y\sqrt{1-r^2}}{s_x\sqrt{n-2}}$$

Testna statistika (standardizirani vzorčni regresijski koeficient) je tedaj

$$t = \frac{b-\beta_H}{SE(b)} = \frac{s_x\sqrt{n-2}}{s_y\sqrt{1-r^2}}(b-\beta_H)$$

Primer

Primer 8 oseb, za katere imamo podatek o *X* – izobrazbi (število priznanih let šole) in *Y*- številu ur branja dnevnih časopisov na teden. Za ta vzorec smo z regresijsko premico izračunali odvisnost števila ur branja dnevnih časopisov od izobrazbe:

$$Y' = -0.68 + 0.46X$$

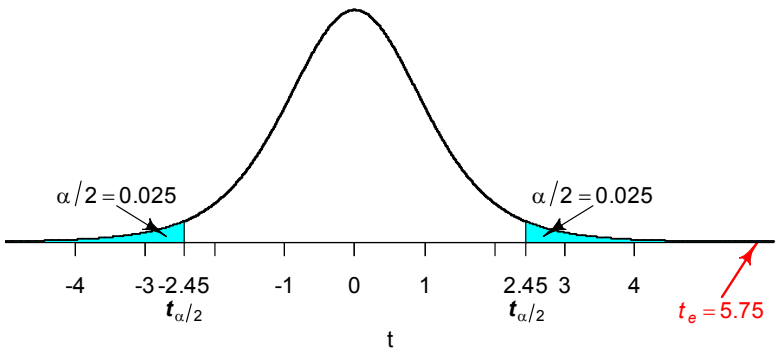
Preverimo domnevo, da je regresijski koeficient različen od 0 pri 5% stopnji značilnosti. Preverimo torej domnevo, da na populaciji *X* res vpliva na *Y* (da izobrazba vpliva na branje dnevnih časopisov).

1. Postavimo ničelno in alternativno domnevo:
   $H_0: \beta = 0$ *X* ne vpliva na *Y*. Izobrazba ne vpliva na pogostost branja.
   $H_1: \beta \neq 0$ *X* vpliva na *Y*. Izobrazba vpliva na pogostost branja.

2. Ker gre za dvostranski test, je ob 5% stopnji značilnosti kritično območje določeno s kritičnima vrednostima:
   $$\pm t_{\alpha/2}(n\text{-}2) = \pm t_{0.05/2}(8\text{-}2) = \pm t_{0.025}(6) = \pm 2.45$$

3. Eksperimentalna vrednost testne statistike je:
   $$t_e = \frac{s_x\sqrt{n-2}}{s_y\sqrt{1-r^2}}(b-\beta_H) = \frac{\sqrt{(x_i-\bar{X})^2}\sqrt{n-2}}{\sqrt{(y_i-\bar{Y})^2}\sqrt{1-r^2}}(b-\beta_H) =$$
   $$= \frac{\sqrt{104\cdot(8-2)}}{\sqrt{26\cdot(1-0.92^2)}}(0.46-0) = 5.75$$

4. Sklep: $5.75 > 2.45$ oz. $|t_e| > |t_{\alpha/2}|$, H_0 zavrnemo, *Y* je odvisna od *X*.



(Izpis iz stat. programa, $p < .001$, kar je $P(t > 5.75) + P(t < -5.75)$. Ker je manjše od 0.05, *H0* zavrnemo.)

5. Interpretacija
Število ur branja dnevnih časopisov na teden je linearno odvisno od priznanih let šole pri 5% stopnji značilnosti. (Verjetnost, da ob domnevi, da na populaciji odvisnosti ni (torej, da je $\beta = 0$), na vzorcu dobimo vzorčni regresijski koeficient z absolutno vrednostjo 0.46 ali več oz. *t*-statistiko po absolutni vrednosti 5.75 ali večji, je tako majhna (manjša od 0.1%), da lahko upravičeno podvomimo v pravilnost domneve o linearni neodvisnosti.)

5.5.5 Kvaliteta regresijskega modela

Regresijska funkcija – regresijski model.

Z regresijskim modelom (regresijsko funkcijo):

- **opisujemo odvisnost** odvisne spremenljivke (*Y*) od neodvisne spremenljivke (*X*) in
- **napovedujemo** vrednosti odvisne spremenljivke (*Y*) na osnovi vrednosti neodvisne spremenljivke (*X*).

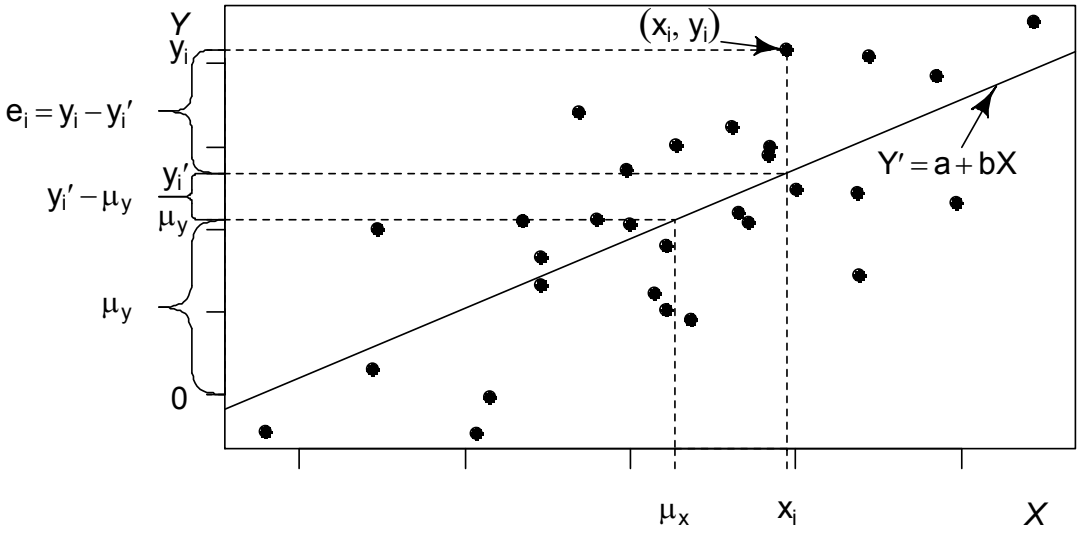
Kako dober je regresijski model (regresijska funkcija)? → 2 kazalca:

1. **Determinacijski koeficient** (**delež pojasnjene variance**) = kazalec kvalitete opisa **odvisnosti** med spremenljivkama z regresijsko premico.
2. **Standardna napaka ocene** = kazalec kvalitete **napovedovanja** vrednosti odvisne spremenljivke s pomočjo regresijske premice.

Definirajmo ...

Vrednost odvisne spremenljivke y_i lahko razstavimo na tri komponente:

$$y_i = \mu_Y + (y_i' - \mu_Y) + (y_i - y_i')$$



Enačbo lahko preoblikujemo v:

$$y_i - \mu_Y = (y_i' - \mu_Y) + (y_i - y_i')$$

μ_Y ... rezultat splošnih vplivov (povprečje)
$y_i' - \mu_Y$ ... rezultat vpliva spremenljivke X
$y_i - y_i'$ ... rezultat vpliva drugih dejavnikov

Zgornjo enačbo lahko preoblikujemo tako, da enakost najprej na obeh straneh enačaja kvadriramo, nato seštejemo po vseh enotah in nato delimo s številom enot N. Dobimo:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_Y)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i' - \mu_Y)^2 + \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2$$

To lahko zapišemo takole: $\sigma_Y^2 = \sigma_{Y'}^2 + \sigma_e^2$

kjer posamezni členi pomenijo:

σ_Y^2 .... celotna varianca spremenljivke Y
$\sigma_{Y'}^2$ .... pojasnjena varianca spremenljivke Y (pojasnjena, odvisna od spremenljivke X)
σ_e^2 ... nepojasnjena varianca spremenljivke Y (odvisna od drugih vplivov)

Determinacijski koeficient

Delež pojasnjene variance spremenljivke Y s spremenljivko X v celotni varianci imenujemo **determinacijski koeficient** in definiran je na intervalu [0, 1].

$$R = \frac{\sigma_{y'}^2}{\sigma_y^2} = \frac{\sum_{i=1}^{N}(y_i' - \mu_Y)^2}{\sum_{i=1}^{N}(y_i - \mu_Y)^2}$$

Pokazati se da, da je v primeru bivariatne linearne regresijske odvisnosti determinacijski koeficient enak kvadratu Pearsonovega koeficienta korelacije:

$$R = \rho^2$$

Če determinacijski koeficient v primeru bivariatne linearne regresije računamo na vzorcu, je obrazec naslednji:

$$R = \frac{s_{y'}^2}{s_y^2} = \frac{\sum_{i=1}^{n}(y_i' - \overline{Y})^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2} = r^2$$

Determinacijski koeficient oz. delež pojasnjene variance nam pove, koliko (kakšen delež) variabilnosti v vrednostih odvisne spremenljivke (Y) lahko pripišemo vrednostim neodvisne spremenljivke (X) in ne kakšnim drugim vplivom.

Determinacijski koeficient oz. delež pojasnjene variance je tako nek kazalec kvalitete regresijske funkcije. Pove nam, kako dobro ocenjena regresijska funkcija/regresijski model (ocenjena parametra *a* in *b*) opisuje odvisnost odvisne spremenljivke (*Y*) od neodvisne spremenljivke (*X*), ki je vključena v analizo. Želimo, da je delež pojasnjene variance oz. determinacijski koeficient čim večji.

Delež pojasnjene variance ima seveda vrednost na intervalu [0, 1].

Standardna napaka ocene

Kvadratni koren iz nepojasnjene variance σ_e^2 imenujemo **standardna napaka regresijske ocene** in meri razpršenost točk okoli regresijske krivulje. V primeru linearne regresijske odvisnosti je standardna napaka enaka:

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} = \sigma_Y\sqrt{1-\rho^2}$$

Če standardno napako ocene računamo na vzorcu, je obrazec naslednji:

$$s_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - y_i')^2} = s_Y\sqrt{1-r^2}$$

Tudi standardna napaka regresijske ocene meri kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo/regresijskim modelom. Pove nam, kakšen je standardni odklon dejanskih vrednosti y_i od napovedanih vrednosti y_i' (napovedanih z regresijsko funkcijo). Pove torej, kako dobro ocenjena regresijska premica (ocenjena parametra *a* in *b*) napoveduje vrednosti odvisne spremenljivke (*Y*) s pomočjo vrednosti neodvisne spremenljivke (*X*), ki je vključena v analizo.

Želimo, da je standardna napaka čim manjša, torej da so odkloni od regresijske krivulje (razlike med dejanskimi in napovedanimi vrednostmi) čim manjši.

Primer

Ponovno vzemimo primer 8 oseb, za katere imamo podatke o dveh spremenljivkah:
X – izobrazba (število priznanih let šole)
Y – število ur branja dnevnih časopisov na teden

Ocenimo, kako dobro izračunana regresijska funkcija $Y' = -0.68 + 0.46X$ opisuje odvisnost pogostosti branja dnevnih časopisov na teden od izobrazbe (torej determinacijski koeficient) in kako dobro lahko na osnovi izobrazbe napovedujemo pogostost branja dnevnih časopisov (torej standardno napako ocene).

Z regresijsko premico najprej izračunajmo **napovedane vrednosti** za vseh 8 oseb v vzorcu. Vstavljamo vrednosti xi v regresijsko funkcijo

$$y_i' = -0.68 + 0.46\,x_i$$

x_i	y_i	y_i'
10	3	3.92
8	4	3
16	7	6.68
8	3	3
6	1	2.08
4	2	1.16
8	3	3
4	1	1.16

Determinacijski koeficient ali **delež pojasnjene variance** izračunajmo (1) s pomočjo odklonov napovedanih vrednosti od povprečja $\overline{Y}$ = 3 ali (2) s Pearsonovim koeficientom korelacije r = 0.92.

x_i	y_i	y_i'	$(y_i - \overline{Y})^2$	$(y_i' - \overline{Y})^2$
10	3	3.92	0	0.8464
8	4	3	1	0
16	7	6.68	16	13.5424
8	3	3	0	0
6	1	2.08	4	0.8464
4	2	1.16	1	3.3856
8	3	3	0	0
4	1	1.16	4	3.3856
			26	22.0064

$$R = \frac{s_{y'}^2}{s_y^2} = \frac{\sum_{i=1}^{n}(y_i' - \overline{Y})^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2} = \frac{22.0064}{26} = 0.8464$$

$$R = r^2 = 0.92^2 = 0.8464$$

Interpretacija: Delež pojasnjene variance oz. determinacijski koeficient je 0.8464, kar pomeni, da z izobrazbo (t.j. s priznanim številom let šole) lahko pojasnimo 85% variance (variabilnosti, razlik) v pogostosti branja dnevnih časopisov (v številu ur branja dnevnih časopisov na teden). Ostale razlike (15%) so rezultat nekih drugih vplivov, ki jih nismo zajeli v naš regresijski model.

Standardno napako ocene izračunajmo (1) s pomočjo odklonov napovedanih vrednosti od dejanskih vrednosti (rezidualov, ostankov) ali (2) s pomočjo standardnega odklona spremenljivke in Pearsonovega koeficienta korelacije.

x_i	y_i	y_i'	$(y_i - \overline{Y})^2$	$(y_i - y_i')^2$
10	3	3.92	0	0.8464
8	4	3	1	1
16	7	6.68	16	0.1024
8	3	3	0	0
6	1	2.08	4	1.1664
4	2	1.16	1	0.7056
8	3	3	0	0
4	1	1.16	4	0.0256
			26	3.8464

$$s_e = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - y_i^{'})^2} = \sqrt{\frac{1}{7}\cdot 3.8464} = 0.75$$

$$s_e = s_Y\sqrt{1-r^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{Y})^2}\sqrt{1-r^2} = \sqrt{\frac{1}{8-1}\cdot 26}\cdot\sqrt{1-0.92^2} = 0.75$$

Interpretacija: Standardna napaka ocene je 0.75 ure. Pri napovedanju pogostosti branja dnevnih časopisov (v številu ur branja dnevnih časopisov na teden) s pomočjo izobrazbe (s številom priznanih let šole) se standardno zmotimo za 0.75 ure.

5.6 VAJE

1. Ferligoj: Naloge iz statistike.
   - Nominalni tip para spremenljivk: 9.1 do 9.7.
   - Ordinalni tip para spremenljivk: 9.8 do 9.11.
   - Intervalni/razmernostni tip para spremenljivk: 9.12 – 9.19, 9.21 - 9.23, 9.25, 9.26.

2. V neki anketi leta 2001 so anketiranci odgovarjali na vprašanje "*Ali uporabljate internet za opravljanje bančnih/borznih transakcij?*". Podatki so naslednji:

		Spol		Skupaj
		Moški	Ženski	
Ali uporabljate internet za opravljanje bančnih/borznih transakcij?	Da, že uporabljam	176	39	215
	Še ne uporabljam, a me zanima	160	42	202
	Ne in me ne zanima	43	24	67
Skupaj		379	105	484

   a) Kaj je enota analize, kaj sta spremenljivki? Kolikšna je velikost vzorca?
   b) Izračunajte smiselne strukturne odstotke, grafično jih predstavite ter interpretirajte. Ali na danem vzorcu obstaja razlika med moškimi in ženskami glede uporabe interneta za opravljanje bančnih/borznih transakcij?
   c) Ali lahko na osnovi danih podatkov za populacijo ugotovimo, da obstaja povezanost med spolom in uporabo interneta za opravljanje bančnih/borznih transakcij? Domnevo preverite pri 10% stopnji značilnosti.
   d) Kakšna je moč povezanosti?
   e) Preverite domnevo, da je na populaciji delež žensk, ki že uporabljajo internet za opravljanje borznih/bančnih transakcij, manjši od deleža moških. Domnevo preverite pri 5% stopnji značilnosti.

3. Leta 2002 sta istočasno dve organizaciji izvedli raziskavo o obiskanosti spletnih mest. Slučajni vzorec 10 spletnih mest je bil v vsaki od obeh raziskav rangiran nekoliko drugače glede na izmerjeno pogostost obiskovanja. Rangiranje glede na ti dve raziskavi je predstavljeno v spodnji tabeli, pri čemer nižji rang predstavlja višjo obiskanost (višje mesto spletne strani na lestvici), višji rang pa nižjo obiskanost (nižje mesto spletne strani na lestvici).

Spletno mesto	Rang glede na 1. raziskavo	Rang glede na 2. raziskavo
Markurja.com	1	2
Najdi.si	2	1
Mobitel	3	7
Email.si	4	4
24ur.com	5	6
SiOL	6	3
Pinkponk	7	8
Delo	8	10
RTV Slovenija	9	9
TIS Telekom	10	5

a) Kaj je enota analize in kaj sta spremenljivki?

b) Kako močna je povezanost med rangiranjema dveh raziskav na danem vzorcu spletnih mest?

c) Ali lahko za celotno populacijo spletnih mest rečemo, da obe raziskavi dajeta podobne rezultate? (Preverite domnevo, da med rangiranjem spletnih mest v dveh raziskavah obstaja statistično značilna pozitivna povezanost pri 1% stopnji značilnosti.)

4. Spodnja tabela prikazuje podatke o številu priseljenih prebivalcev in številu razvez na 100 prebivalcev v nekem koledarskem letu za slučajni vzorec 10 slovenskih občin.

Občina	Število razvez/100 preb.	Število priseljenih/100 preb.
Borovnica	1	4
Cerklje	5	9
Gorišnica	2	5
Grosuplje	19	26
Miklavž	7	18
Pivka	3	12
Prevalje	3	5
Selnica	2	8
Tišina	5	1
Vransko	1	2

a) Kaj je enota analize in kaj sta spremenljivki?

b) Preverite, ali na populaciji vseh slovenskih občin obstaja statistično značilna linearna povezanost med številom priseljenih in številom razvez, pri 5% stopnji značilnosti.

5. Na osnovi vzorca 10-ih enot iz neke ankete o politični participaciji analizirajte vpliv posameznikovega občutka o možnosti vplivanja na družbene dogodke na pogostost izražanja mnenja v javnosti. Možnost vpliva je merjena na lestvici od 1 do 5, kjer 1 pomeni "*nikakršne možnosti nimam za vplivanje*", 5 pa "*velike možnosti imam za vplivanje"*. Pogostost izražanja mnenja v javnosti pa je merjeno s število izražanj mnenja glede družbenih in političnih problemov v javnih situacijah v zadnjih 4 letih. Podani so tudi podatki o spolu anketiranca (1 – *moški*, 2 – *ženski*).

Oseba	Občutek vplivanja	Pogostost izračanja mnenja	Spol
1	4	6	1
2	2	1	2
3	3	2	1
4	5	9	1
5	3	1	2
6	2	0	2
7	5	5	2
8	4	4	1
9	2	7	2
10	4	5	1

a) Kaj je enota analize, koliko in katere spremenljivke obravnavamo?
b) Grafično ponazorite odnos med spremenljivkama "*možnost vpliva na družbene dogodke*" in "*pogostost izražanja mnenja v javnosti*" ter interpretirajte.
c) Izračunajte Pearsonov koeficient korelacije med njima in preverite, ali na populaciji obstaja linearna povezanost med spremenljivkama (upoštevajte 10% stopnjo značilnosti). Interpretirajte.
d) Izračunajte tudi ustrezno regresijsko premico in jo vrišite v razsevni grafikon. Interpretirajte.
e) Preverite domnevo, ali za populacijo velja, da občutek o možnosti vplivanja na družbene dogodke pozitivno vpliva na pogostost izražanja mnenja v javnosti (upoštevajte 10% stopnjo značilnosti).
f) Kolikšen del variabilnosti (variance) v pogostosti izražanja mnenja v javnosti ne uspemo pojasniti z občutkom vplivanja na družbene dogodke? Kolikšen del variance pa uspemo pojasniti?
g) Ocenite, kako pogosto bi izražal mnenje nekdo, ki je najmanj prepričan, da lahko vpliva na družbene dogodke (ima torej vrednost 1).
h) Kakšna je standardna napaka ocene pri takšni napovedi?
i) Izračunajte standardno napako še kot razliko med dejanskimi in napovedanimi vrednostmi iz regresijske premice. Rezultata se morata ujemati.
j) Ali se pogostost izražanja mnenja v javnosti na populaciji med moškimi in ženskami statistično značilno razlikuje? Upoštevajte 5% stopnjo značilnosti. (Predpostavljamo, da je variabilnost v pogostosti izražanja mnenja enaka za populacijo moških in žensk.)

6. Spodnja tabela prikazuje za 6 slučajno izbranih oseb podatke o številu ur gledanja televizije na teden in številu obiskov kinopredstav na mesec.

Oseba	Št. ur gledanja TV/teden	Št. obiskov kinopredstav/mesec
1	10	2
2	15	1
3	6	2
4	7	4
5	20	1
6	8	2

   a) Kaj se enote in kaj sta spremenljivki? Kakšna je merska lestvica spremenljivk?
   b) Podatke grafično predstavite z razsevnim grafikonom.
   c) Z linearno regresijsko funkcijo ocenite, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden televizijo. Premico vrišite v razsevni grafikon.
   d) Kolikšna je pri takem ocenjevanju standardna napaka ocene?
   e) Kolikšen delež variance obiska kinopredstav lahko pojasnimo z gledanjem televizije?

7. Primeri izpitnih vprašanj.

Domnevo o povezanosti spremenljivk *spol* in *smer študija* preverjamo

a) s χ^2 testom
b) s Pearsonovim koeficientom korelacije
c) s Spearmanovim koeficientom korelacije rangov
d) z intervalom zaupanja

V kontingenčni tabeli, ki prikazuje vrednosti dveh nominalnih spremenljivk na vzorcu 400 enot in ki ima 4 vrstice in 6 stolpcev, izračunamo vrednost $\chi^2 = 108$. Kolikšen je tedaj Cramerjev koeficient?

a) $C = 0{,}3$
b) $C = 0{,}46$
c) $\alpha = 0{,}33$
d) $\alpha = 0{,}3$

Kontingenčna tabela kaže povezanost med spremenljivkama *spol* in *zanimanje za šport* (z vrednostmi *majhno*, *srednje*, *veliko*). Vrednost χ^2 statistike je 23,4 in stopnja tveganja 5%. Ugotovimo lahko naslednje:

a) $\chi^2_{(1-\alpha)} = 5{,}99$; domneve, da sta spremenljivki nepovezani, ne moremo zavrniti
b) $\chi^2_{\alpha} = 0{,}10$; zavrnemo domnevo, da sta spremenljivki nepovezani
c) $\chi^2_{(1-\alpha)} = 5{,}99$; zavrnemo domnevo, da sta spremenljivki povezani
d) $\chi^2_{(1-\alpha)} = 5{,}99$; zavrnemo domnevo, da sta spremenljivki nepovezani

Na slučajnem vzorcu 20-ih dijakov smo izračunali Spearmanov koeficient korelacije rangov med *razvrstitvijio dijiakov glede na učni uspeh* (od tistih z nezadostnim uspehom do tistih z odličnim uspehom) ter *glede na sodelovanje v obšolskih dejavnostih* (od tistih, ki nikoli ne sodelujejo, do tistih, ki redno sodelujejo). Vrednost koeficienta je 0,72. Ali lahko za populacijo dijakov trdimo, da obstaja pozitivna povezanost med učnim uspehom in sodelovanjem v obšolskih dejavnostih, ob 5% stopnji značilnosti?

a) $H_0: \rho_s = 0$, $H_1: \rho_s > 0$, $z_e = 4.4$, $z_\alpha = 1.65$; ne, pozitivna povezanost ne obstaja
b) $H_0: \rho_s = 0$, $H_1: \rho_s \neq 0$, $t_e = 4.4$, $t_{\alpha/2} = \pm 2.10$; da, pozitivna povezanost obstaja
c) $H_0: \rho_s = 0$, $H_1: \rho_s \neq 0$, $z_e = 4.4$, $z_{\alpha/2} = \pm 1.96$; ne, pozitivna povezanost ne obstaja
d) $H_0: \rho_s = 0$, $H_1: \rho_s > 0$, $t_e = 4.4$, $t_\alpha = 1.73$; da, pozitivna povezanost obstaja

Pearsonov koeficient korelacije med dvema spremenljivkama je -0.82. Spremenljivki sta

a) močno pozitivno linearno povezani
b) močno negativno linearno povezani
c) močno eksponentno povezani
d) nepovezani

Če je pri bivariatni linearni regresijski analizi Pearsonov koeficient korelacije 0.50, je delež pojasnjene variance

a) 0,75
b) 0,25
c) 0
d) 1

Med spremenljivkama X in Y velja linearna regresijska zveza $Y' = 5 + 2X$. Napovejte vrednost spremenljivke Y', če je vrednost spremenljivke X enaka 15.

a) 35
b) 5
c) 3,5
d) 7