Preverjanje domnev o porazdelitvi spremenljivke

Do sedaj smo ocenjevali in preverjali domnevo o parametrih populacije kot μ, σ in π.

Sedaj pa bomo preverjali, če se spremenljivka porazdeljuje po določeni porazdelitvi.

Test je zasnovan na dejstvu, kako dobro se prilegajo empirične (eksperimentalne) frekvence vrednosti spremenljivke hipotetičnim (teoretičnim) frekvencam, ki so določene s predpostavljeno porazdelitvijo.

Preverjanje domneve o enakomerni porazdelitvi

Za primer vzemimo met kocke in za spremenljivko število pik pri metu kocke. Preizkusimo domnevo, da je kocka poštena, kar je enakovredno domnevi, da je porazdelitev spremenljivke enakomerna. Tedaj sta ničelna in osnovna domneva

H_0 : spremenljivka se porazdeljuje enakomerno,
H_1 : spremenljivka se ne porazdeljuje enakomerno.

Denimo, da smo 120-krat vrgli kocko ($n = 120$)
in štejemo kolikokrat smo vrgli posamezno število pik.

To so empirične ali opazovane frekvence, ki jih označimo s f_i.
Teoretično, če je kocka poštena, pričakujemo, da bomo dobili vsako vrednost z verjetnostjo 1/6 oziroma 20 krat.

To so teoretične ali pričakovane frekvence, ki jih označimo s f_i'.

Podatke zapišimo v naslednji tabeli

x_i	1	2	3	4	5	6
p_i	1/6	1/6	1/6	1/6	1/6	1/6
f_i'	20	20	20	20	20	20
f_i	20	22	17	18	19	24

S primerjavo empiričnih frekvenc z ustreznimi teoretičnim frekvencami se moramo odločiti, če so razlike posledica le vzorčnih učinkov in je kocka poštena ali pa sp razlike prevelike, kar kaže, da je kocka nepoštena. Statistika, ki meri prilagojenost empiričnih frekvenc teoretičnim je

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

ki se porazdeljuje po χ^2 porazdelitvi z $m = k-1$ prostostnimi stopnjami, ki so enake številu vrednosti spremenijivke ali celic (k) minus število količin dobljenih iz podatkov, ki so uporabljene za izračun teoretičnih frekvenc.

V našem primeru smo uporabili le eno količino in sicer skupno število metov kocke ($n = 120$). Torej število prostostnih stopenj je $m = k-1 = 6-1 = 5$.
Ničelna in osnovna domneva sta tedaj

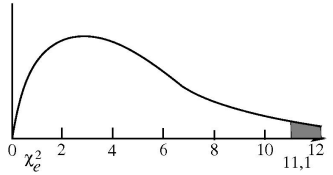
$$H_0 : \chi^2 = 0 \quad \text{in} \quad H_1 : \chi^2 > 0.$$

Doomnevo preverimo pri stopnji značilnosti $\alpha = 5\%$.
Ker gre za enostranski test, je kritična vrednost enaka

$$\chi^2_{1-\alpha}(k-1) = \chi^2_{0,95}(5) = 11,1.$$

Eksperimentalna vrednost statistike pa je

$$\chi^2_e = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = \frac{4+9+4+1+16}{20} = \frac{34}{20} = 1,7.$$



Ker ekperimentalna vrednost statistike ne pade v kritično območje, ničelne domneve ne moremo zavrniti. Empirične in teoretične frekvence niso statistično značilno različne med seboj.

Preverjanje domneve o normalni porazdelitvi

Omenjeni test najpogosteje uporabljamo za preverjanje ali se spremenjivka porazdeljuje normalno.

V tem primeru je izračun teoretičnih frekvenc potrebno vložiti malo več truda.

Primer: Preizkusimo domnevo, da se spremenljivka telesna višina porazdeljuje normalno $N(177, 10)$. Domnevo preverimo pri 5% stopnji značilnosti.
Podatki za 100 slučajno izbranih oseb so urejeni v frekvenčni porazdelitvi takole:

	f_i
nad 150-160	2
nad 160-170	20
nad 170-180	40
nad 180-190	30
nad 190-200	8
	100

Ničelna in osnovna domneva sta tedaj

$$H_0 : \chi^2 = 0 \quad \text{in} \quad H_1 : \chi^2 \neq 0.$$

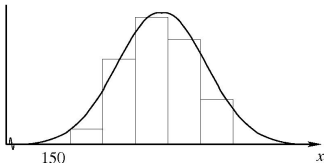
Za test uporabimo statistiko

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

ki se porazdeljuje po χ^2 porazdelitvi z $m = 5 - 1$ prostostnimi stopnjami.
Kritična vrednost je

$$\chi^2_{0,95}(4) = 9,49.$$

V naslednjem koraku je potrebno izračunati teoretične frekvence.
Najprej je potrebno za vsak razred izračunati verjetnost p_i, da spremenljivka zavzame vrednosti določenega intervala, če se porazdeljuje normalno.
To lahko prikažemo na sliki:



Tako je na primer verjetnost, da je višina med 150 in 160 cm:

$$P(150 < X < 160) = P\left(\frac{150 - 177}{10} < Z < \frac{160 - 177}{10}\right) =$$

$$= P(-2,7 < Z < -1,7) = H(2,7) - H(1,7) = 0,4965 - 0,4554 = 0,0411.$$

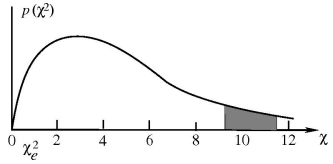
Podobno lahko izračunamo ostale verjetnosti. Teoretične frekvence so $f_i' = n \times p_i$. Izračunane verjetnosti p_i in teoretične frekvence f_i' so

	f_i'	p_i	f_i'
nad 150-160	2	0,0411	4,11
nad 160-170	20	0,1974	19,74
nad 170-180	40	0,3759	37,59
nad 180-190	30	0,2853	28,53
nad 190-200	8	0,0861	8,61
	100		98,58

Eksperimentalna vrednost statistike je tedaj

$$\chi^2_e = \frac{(2 - 4.11)^2}{4,11} + \frac{(20 - 19,74)^2}{19,74} + \frac{(40 - 37,59)^2}{37,59}$$

$$+ \frac{(30 - 28,53)^2}{28,53} + \frac{(8 - 8,61)^2}{8,61} \approx 1$$



Ker ekperimentalna vrednost ne pade v kritično območje, ne moremo zavrniti ničelne domneve, da je spremenljivka normalno porazdeljena.

Obstajajo tudi drugi testi za preverjanje porazdelitve spremenljivke, npr. Kolmogorov-Smirnov test.

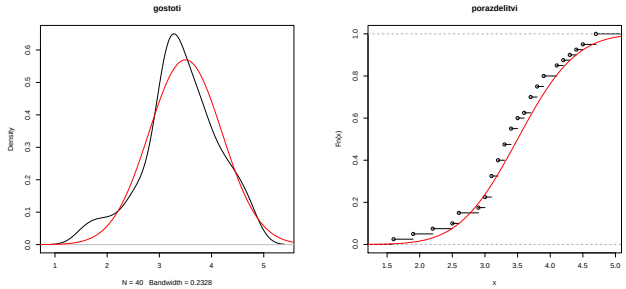
Pri objavi anketiranih rezultatov je potrebno navesti:

1. naročnika in izvajalca,
2. populacijo in vzorčni okvir,
3. opis vzorca,
4. velikost vzorca in velikost realiziranega vzorca (stopnja odgovorov)
5. čas, kraj in način anketiranja,
6. anketno vprašanje,
7. vzorčno napako.

Preizkus Kolmogorov-Smirnova v R-ju

```
> t <- c(1.6,1.9,2.2,2.5,2.6,2.6,2.9,3.0,3.0,3.1,3.1,3.1,
+ 3.1,3.2,3.2,3.2,3.3,3.3,3.3,3.4,3.4,3.4,3.5,3.5,3.6,3.7,
+ 3.7,3.7,3.8,3.8,3.9,3.9,4.1,4.1,4.2,4.3,4.4,4.5,4.7,4.7)
> z <- sample(t)
> z
 [1] 4.1 2.2 4.7 3.8 4.7 3.5 3.3 3.6 4.4 2.9 4.3 4.2 3.9 3.1
[15] 3.1 2.5 4.5 3.0 2.6 3.4 1.9 3.5 3.2 3.1 3.7 2.6 3.1 3.2
[29] 3.4 3.7 3.4 3.3 4.1 1.6 3.9 3.3 3.0 3.7 3.2 3.8
> "lot(density(z),main="gostoti")
> "urve(dnorm(x,mean=3.5,sd=0.7),add=TRUE,col="red")
> "lot(ecdf(z),main="porazdelitvi")
> curve(pnorm(x,mean=3.5,sd=0.7),add=TRUE,col="red")
> ks.test(z,"pnorm",mean=3.5,sd=0.7)

        One-sample Kolmogorov-Smirnov test

data:  z
D = 0.1068, p-value = 0.7516
alternative hypothesis: two.sided

Warning message:
cannot compute correct p-values with ties in:
  ks.test(z, "pnorm", mean = 3.5, sd = 0.7)
```


Preizkus Kolmogorov-Smirnova v R-ju



V R-ju so pri preizkusih izpisane vrednosti `p-value` $= \Pi$
(preizkusna statistika ima pri veljavnosti osnovne domneve vrednost vsaj tako ekstremno, kot je zračunana).

$[0, 0.001]$ – izjemno značilno (***);
$(0.001, 0.01]$ – zelo značilno (**);
$(0.01, 0.05]$ – statistično značilno (*);
$(0.05, 0.1]$ – morda značilno;
$(0.1, 1]$ – neznačilno.

Osnovno domnevo zavrnemo, če je `p-value` pod izbrano stopnjo značilnosti.

Preizkus Kolmogorov-Smirnova v R-ju

```
> p <- pnorm(t,mean=3.5,sd=0.7)
> s <- (1:40)/40; d <- abs(s-p)
> options(digits=3)
> cbind(t,p,s,d)
       t     p     s     d        t     p     s     d
 [1,] 1.6 0.00332 0.025 0.02168  [21,] 3.4 0.44320 0.525 0.08180
 [2,] 1.9 0.01114 0.050 0.03886  [22,] 3.4 0.44320 0.550 0.10680
 [3,] 2.2 0.03165 0.075 0.04335  [23,] 3.5 0.50000 0.575 0.07500
 [4,] 2.5 0.07656 0.100 0.02344  [24,] 3.5 0.50000 0.600 0.10000
 [5,] 2.6 0.09927 0.125 0.02573  [25,] 3.6 0.55680 0.625 0.06820
 [6,] 2.6 0.09927 0.150 0.05073  [26,] 3.7 0.61245 0.650 0.03755
 [7,] 2.9 0.19568 0.175 0.02068  [27,] 3.7 0.61245 0.675 0.06255
 [8,] 3.0 0.23753 0.200 0.03753  [28,] 3.7 0.61245 0.700 0.08755
 [9,] 3.0 0.23753 0.225 0.01253  [29,] 3.8 0.66588 0.725 0.05912
[10,] 3.1 0.28385 0.250 0.03385  [30,] 3.8 0.66588 0.750 0.08412
[11,] 3.1 0.28385 0.275 0.00885  [31,] 3.9 0.71615 0.775 0.05885
[12,] 3.1 0.28385 0.300 0.01615  [32,] 3.9 0.71615 0.800 0.08385
[13,] 3.1 0.28385 0.325 0.04115  [33,] 4.1 0.80432 0.825 0.02068
[14,] 3.2 0.33412 0.350 0.01588  [34,] 4.1 0.80432 0.850 0.04568
[15,] 3.2 0.33412 0.375 0.04088  [35,] 4.2 0.84134 0.875 0.03366
[16,] 3.2 0.33412 0.400 0.06588  [36,] 4.3 0.87345 0.900 0.02655
[17,] 3.3 0.38755 0.425 0.03745  [37,] 4.4 0.90073 0.925 0.02427
[18,] 3.3 0.38755 0.450 0.06245  [38,] 4.5 0.92344 0.950 0.02656
[19,] 3.3 0.38755 0.475 0.08745  [39,] 4.7 0.95676 0.975 0.01824
[20,] 3.4 0.44320 0.500 0.05680  [40,] 4.7 0.95676 1.000 0.04324
> options(digits=7)
> max(d)
[1] 0.1067985
```


Preizkus χ^2 v R-ju

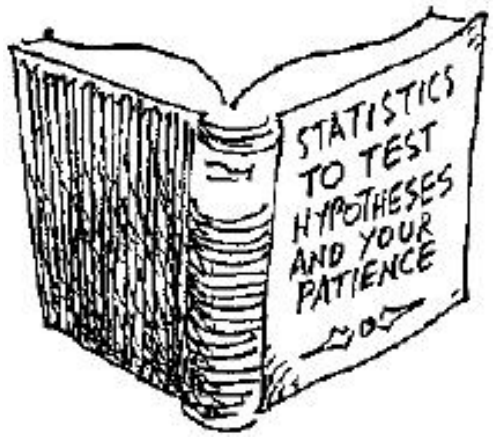
```
> a <- rbind(c(80,5,15),c(40,20,20),c(20,30,20))
> rownames(a) <- c("za","proti","neodlocen")
> colnames(a) <- c("do 25","25-50","nad 50")
> a
          do 25 25-50 nad 50
za           80     5     15
proti        40    20     20
neodlocen    20    30     20
> chisq.test(a)

        Pearson's Chi-squared test

data:  a
X-squared = 51.4378, df = 4, p-value = 1.808e-10
```

Poleg `ks.test` in `chisq.test` obstaja v R-ju še več drugih preizkusov: `prop.test`, `binom.test`, `t.test`, `wilcox.test`, `var.test`, `shapiro.test`, `cor.test`, `fisher.test`, `kruskal.test`.

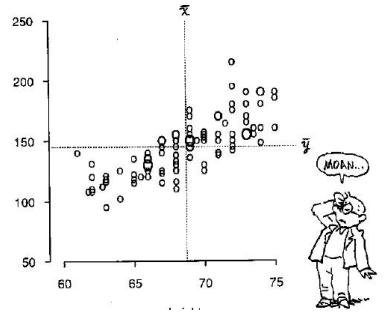
Opise posameznega preizkusa dobimo z zahtevo `help` (*preizkus*) .

Spearmanov preizkus povezanosti v R-ju

```
> slo <- c(5,3,1,2,4)
> mat <- c(5,2,3,1,4)
> slo
[1] 5 3 1 2 4
> mat
[1] 5 2 3 1 4
> cor.test(slo,mat,method="spearm")

        Spearman's rank correlation rho

data:  slo and mat
S = 6, p-value = 0.2333
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7
```

II.6. Bivariantna analiza in regresija



Bivariantna analiza

$X \longleftrightarrow Y$ povezanost

$X \longrightarrow Y$ odvisnost

Mere povezanosti ločimo glede na tip spremenljivk:

1. NOMINALNI tip para spremenljivk (ena od spremenljivk je nominalna): χ^2, kontingenčni koeficienti, koeficienti asociacije;
2. ORDINALNI tip para spremenljivk (ena spremenljivka je ordinalna druga ordinalna ali boljša) koeficient korelacije rangov;
3. ŠTEVILSKI tip para spremenljivk (obe spremenljivki sta številski): koeficient korelacije.

Preverjanje domneve o povezanosti dveh nominalnih spremenljivk

Vzemimo primer:

- ENOTA: dodiplomski študent neke fakultete v letu 1993/94;
- VZOREC: slučajni vzorec 200 študentov;
- 1. SPREMENLJIVKA: spol;
- 2. SPREMENLJIVKA: stanovanje v času študija.

Zanima nas ali študentke drugače stanujejo kot študentje oziroma: ali sta spol in stanovanje v času študija povezana. V ta namen podatke študentov po obeh spremenljivkah uredimo v dvodimenzionalno frekvenčno porazdelitev. To tabelo imenujemo kontingenčna tabela.

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

	starši	št. dom	zasebno	skupaj
moški	16	40	24	80
ženske	48	36	36	120
skupaj	64	76	60	200

Ker nas zanima ali študentke drugače stanujejo v času študija kot študentje, moramo porazdelitev stanovanja študentk primerjati s porazdelitvijo študentov.

Ker je število študentk različno od števila študentov, moramo zaradi primerjave izračunati relativne frekvence:

	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 bili obe porazedelitvi (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 (po vrsticah) kaze, da sta spremenljivki povezani med seboj.

Relativne frekvence lahko računamo tudi po stolpcih:

	starši	št. dom	zasebno	skupaj
moški	25	56,6	40	40
ženske	75	43,4	60	60
skupaj	100	100	100	100

Relativno frekvenco lahko prikažemo s stolpci ali krogi.

Kontingenčna tabela kaže podatke za slučajni vzorec.

Zato nas zanima, ali so razlike v porazdelitvi tipa stanovanja v času študija po spolu statistčno značilne in ne le učinek vzorca.

H_0: spremenljivki nista povezani
H_1: spremenljivki sta povezani

Za preverjanje domneve o povezanosti med dvema nominalnima spremenljivkama na osnovi vzorčnih podatkov, podanih v dvo-razsežni frekvenčni porazdelitvi, lahko uporabimo χ^2 test.

Ta test sloni na primerjavi empiričnih (dejanskih) frekvenc s teoretičnimi frekvencami, ki so v tem primeru frekvence, ki bi bile v kontingenčni tabeli, če spremenljivki ne bi bili povezani med seboj.

To pomeni, da bi bili porazdelitvi stanovanja v času študija deklet in fantov enaki.

Če spremenljivki nista povezani med seboj, so verjetnosti hkratne zgoditve posameznih vrednosti prve in druge spremenljivke enake produktu verjetnosti posameznih vrednosti. Npr., če označimo moške z M in stanovanje pri starših s S, je:

$$P(M) = \frac{80}{200} = 0,40;$$

$$P(S) = \frac{64}{200} = 0,32;$$

$$P(M \cap S) = P(M) \cdot P(S) = \frac{80}{200} \cdot \frac{64}{200} = 0,128.$$

Teoretična frekvenca je verjetnost $P(M \cap S)$ pomnožena s številom enot v vzorcu:

$$f'(M \cap S) = n \cdot P(M \cap S) = 200 \cdot \frac{80}{200} \cdot \frac{64}{200} = 25,6.$$

Podobno izračunamo teoretične frekvence tudi za druge celice kontingenčne tabele.

Če teoretične frekvence zaokrožimo na cela števila, je tabela izračunanih teoretičnih frekvenc f'_i naslednja:

	starši	št. dom	zasebno	skupaj
moški	26	30	24	80
ženske	38	46	36	120
skupaj	64	76	60	200

Spomnimo se tabel empiričnih (dejanskih) frekvenc f_i:

χ^2 statistika, ki primerja dejanske in teoretične frekvence je

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

kjer je k število celic v kontingenčni tabeli. Statistika χ^2 se porazdeljuje po χ^2 porazdelitvi s $(s-1)(v-1)$ prostostnimi stopnjami, kjer je s število vrstic v kontingenčni tabeli in s število stolpcev.

Ničelna in osnovna domneva sta v primeru tega testa

H_0: $\chi^2 = 0$ (spremenljivki nista povezani)
H_1: $\chi^2 > 0$ (spremenljivki sta povezani)

Iz tabele za porazdelitev χ^2 lahko razberemo kritične vrednost te statistike pri 5% stopnji značilnosti:

$$\chi^2_{1-\alpha}[(s-1)(v-1)] = \chi^2_{0,95}(2) = 5,99.$$

Eksperimentalna vrednost statistike χ^2 pa je:

$$\chi^2_e = \frac{(16-26)^2}{26} + \frac{(40-30)^2}{30} + \frac{(24-24)^2}{24}$$

$$+\frac{(48-38)^2}{38} + \frac{(36-46)^2}{46} + \frac{(36-36)^2}{36} = 12.$$

Ker je ekperimentalna vrednost večja od kritične vrednosti, pomeni, da pade v kritično območje.

To pomeni, da ničelno domnevo zavrnemo.

Pri 5% stopnji značilnosti lahko sprejmemo osnovno domnevo, da sta spremenljivki statistično značilno povezani med seboj.

Statistika χ^2 je lahko le pozitivna. Zavzame lahko vrednosti v intervalu $[0, \chi^2_{\max}]$, kjer je $\chi^2_{\max} = n(k-1)$, če je $k = \min(v, s)$.

χ^2 statistika v splošnem ni primerljiva. Zato je definiranih več **kontingenčnih koeficientov**, ki so bolj ali manj primerni. Omenimo naslednje:

1. **Pearsonov koeficient**:

$$\Phi = \frac{\chi^2}{n},$$

ki ima zgornjo mejo $\Phi^2_{\max} = k - 1$.

2. **Cramerjev koeficient**:

$$\alpha = \sqrt{\frac{\Phi^2}{k-1}} = \sqrt{\frac{\chi^2}{n(k-1)}},$$

ki je definiran na intervalu $[0, 1]$.

3. **Kontingenčni koeficient**:

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

ki je definiran na intervalu $[0, C_{\max}]$, kjer je $C_{\max} = \sqrt{k/(k-1)}$.

Koeficienti asociacije

Denimo, da imamo dve nominalni spremenljivki, ki imata le po dve vrednosti (sta dihotomni). Povezanost med njima lahko računamo poleg kontingenčnih koeficientov s **koeficienti asociacije** na osnovi frekvenc iz kontingenčne tabele 2×2:

$Y \backslash X$	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	d	$c+d$
	$a+c$	$b+d$	N

kjer je $N = a + b + c + d$. Na osnovi štirih frekvenc v tabeli je definirnih več koeficientov asociacije:

- **Yulov koeficient asociacije**:

$$Q = \frac{ad - bc}{ad + bc} \in [-1, 1].$$

- **Sokal Michenerjev koeficient**:

$$S = \frac{a+d}{a+b+c+d} = \frac{a+d}{N} \in [0, 1].$$

- **Pearsonov koeficient**:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1, 1].$$

Velja

$$\chi^2 = N \cdot \phi^2.$$

- **Jaccardov koeficient**:

$$J = \frac{a}{a+b+c} \in [0, 1],$$

- in še več drugih.

Primer:
Vzemimo primer, ki kaže povezanost mod kaznivimi dejanji in alkoholizmom. Tabela kaže podatke za $N = 10.750$ ljudi

alk. \ kaz. d.	DA	NE	skupaj
DA	50	500	550
NE	200	10.000	10.200
skupaj	250	10.500	10.750

Izračunajmo koeficiente asociacije:

$$Q = \frac{50 \times 10000 - 200 \times 500}{50 \times 10000 + 200 \times 500} = 0,67;$$

$$S = \frac{10050}{10750} = 0,93;$$

$$J = \frac{50}{50 + 500 + 200} = 0,066.$$

Izračunani koeficienti so precej različni. Yulov in Sokal Michenerjev koeficient kažeta na zelo močno povezanost med kaznjivimi dejanji in alkoholizmom, medtem kot Jaccardov koeficient kaže, da med spremenljivkama ni povezanosti. Pri prvih dveh koeficientih povezanost povzroča dejstvo, da večina alkoholiziranih oseb ni naredila kaznivih dejanj in niso alkoholiki (frekvenca d). Ker Jaccardov koeficient upošteva le DA DA ujemanje, je lažji za interpretacijo. V našem primeru pomeni, da oseba, ki je naredila kaznivo dejanje, sploh ni nujno alkoholik.

Preverjanje domneve o povezanosti dveh ordinalnih spremenljivk

V tem primeru gre za študij povezanosti med dvema spremenljivkama, ki sta vsaj ordinalnega značaja.

Primer:

Vzemimo slučajni vzorec šestih poklicev in ocenimo, koliko so odgovorni (O) in koliko fizično naporni (N).

V tem primeru smo poklice uredili od najmanj odgovornega do najbolj odgovornega in podobno od najmanj fizično napornega do najbolj napornega.

Poklicem smo torej priredili range po odgovornosti (R_0) in po napornosti (R_N) od 1 do 6.

Podatki so podani v tabeli:

poklic	R_0	R_N
A	1	6
D	2	4
C	3	5
D	4	2
E	5	3
F	6	1

Povezanost med spremenljivkama lahko merimo s koeficientom korelacije rangov r_s (Sperman), ki je definiran takole:

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

kjer je d_i razlika med **rangoma** v i-ti enoti.

Koeficient korelacije rangov lahko zavzame vrednosti v intervalu $[-1, 1]$. Če se z večanjem rangov po prvi spremenjivki večajo rangi tudi po drugi spremenljivki, gre za pozitivno povezanost. Tedaj je koeficient pozitiven in blizu 1. Če pa se z večanjem rangov po prvi spremenljivki rangi po drugi spremenljivki manjšajo, gre za negativno povezanost. Koeficient je tedaj negativen in blizu -1. V našem preprostem primeru gre negativno povezanost. Če ne gre za pozitivno in ne za negativno povezanost, rečemo, da spremenljivki nista povezani.

Izračunajmo koeficient korelacije rangov za primer šestih poklicev:

poklic	R_0	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			0	66

$$r_s = 1 - \frac{6 \cdot 66}{6 \cdot 35} = 1 - 1,88 = -0,88.$$

Res je koeficient blizu, kar kaže na močno negativno povezanost teh 6-ih poklicev.

Omenili smo, da obravnavamo 6 slučajno izbranih poklicev.

Zanima nas, ali lahko na osnovi tega vzorca posplošimo na vse poklice, da sta odgovornost in fizična napornost poklicev (negativno) povezana med seboj.

Upoštevajmo 5% stopnjo značilnosti.

Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho_s = 0$ (spremenljivki nista povezani)
H_1: $\rho_s \neq 0$ (spremenljivki sta povezani)

kjer populacijski koeficient označimo s ρ_s.

Pokaže se, da se statistika

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

porazdeljuje približno po porazdelitvi z $m = (n-2)$ prostostnimi stopnjami. Ker gre za dvostranski test, sta kritični vrednosti enaki

$$\pm t_{\alpha/2} = \pm t_{0,025}(4) = \pm 2,776.$$

Eksperimentalna vrednost statistike je za naš primer

$$t_e = \frac{-0,88 \times 2}{\sqrt{1-(-0,88)^2}} = \frac{-1,76}{0,475} = -3,71.$$

Eksperimentalna vrednost pade v kritično območje. Pri 5% stopnji značilnosti lahko rečemo, da sta odgovornost in fizična napornost (negativno) povezani med seboj.

Če je ena od obeh spremenljivk številska, moramo vrednosti pred izračunom d_i rangirati. Če so kakšne vrednosti enake, zanje lzračunamo povprečne pripadajoče range.

Preverjanje domneve o povezanosti dveh številskih spremenljivk

Vzemimo primer dveh številskih spremenljivk:

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

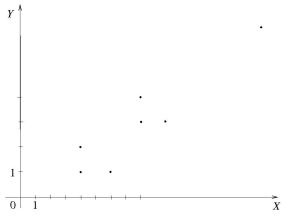
Podatki za 8 slučajno izbranih oseb so:

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

Grafično lahko ponazorimo povezanost med dvema številskima spremenljivkama z razsevnim grafikonom.

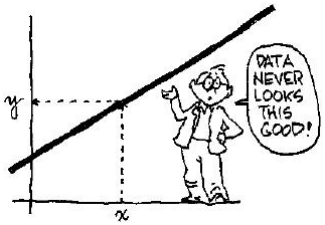
To je, da v koordinatni sistem, kjer sta koordinati obe spremenljivki, vrišemo enote s pari vrednosti.

V našem primeru je izgleda razsevni grafikon takole:

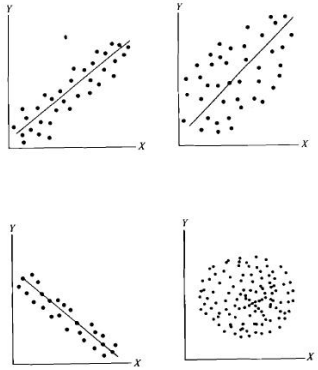


Tipi povezanosti:

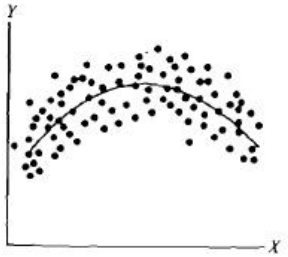
- **funkcijska** povezanost: vse točke ležijo na krivulji:
- **korelacijska** (stohastična] povezanost: točke so od krivulje bolj ali manj odklanjajo (manjša ali večja povezanost).



Tipični primeri linearne povezanosti spremenljivk:



Primer nelinearne povezanosti spremenljivk:



Kovarianca

$$\mathsf{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_X)\cdot(y_i-\mu_Y)$$

meri linearno povezanost med spremenljivkama.
$\mathsf{Cov}(X,Y) > 0$ pomeni pozitivno linearno povezanost,
$\mathsf{Cov}(X,Y) = 0$ pomeni da ni linearne povezanosti,
$\mathsf{Cov}(X,Y) > 0$ pomeni negativno linearno povezanost.
(Pearsonov) koeficient korelacije je

$$\rho_{XY} = \frac{\mathsf{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sum_{i=1}^{N}(x_i-\mu_X)\cdot(y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_X)^2\cdot\sum_{i=1}^{N}(y_i-\mu_Y)}}.$$

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

Če se z večanjem vrednosti prve spremenljivke
večajo tudi vrednosti druge spremenljivke,
gre za *pozitivno* povezanost.
Tedaj je koeficient povezanosti blizu 1.

Če pa se z večanjem vrednosti prve spremenljivke
vrednosti druge spremenljivke manjšajo,
gre za *negativno* povezanost.
Koeficient je tedaj negativen in blizu -1.

Če ne gre za pozitivno in ne za negativno povezanost,
rečemo da spremenljivki nista povezani in koeficient je blizu 0.

Statistično sklepanje o korelacijski povezanosti:

Postavimo torej ničelno in osnovno domnevo:

H_0: $\rho = 0$ (spremenljivki nista linearno povezani)
H_1: $\rho \neq 0$ (spremenljivki sta linearno povezani)

Pokaže se, da se statistika

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

porazdeljuje po t porazdelitvi z $m = (n-2)$ prostostnimi stopnjami.

Z r označujemo koeficient korclacije na vzorcu in
z ρ koeficient korelacije na populaciji.

Primer: Preverimo domnevo, da sta izobrazba (število priznanih let šole) in število ur branja dnevnih časopisov na teden povezana med seboj pri 5% stopnji značilnosti. Najprej izračunajmo vzorčni koeficient korelacije:

x_i	y_i	$x_i-\mu_x$	$y_i-\mu_y$	$(x_i-\mu_x)^2$	$(y_i-\mu_y)^2$	$(x_i-\mu_x)\cdot(y_i-\mu_y)$
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

$$r = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2\sum_{i=1}^{n}(y_i-\mu_y)^2}}$$

$$= \frac{48}{\sqrt{104\cdot 26}} = 0,92.$$

Ker gre za dvostranski test, je kritično območje določeno s kritičnima vrednostima

$$\pm t_{\alpha/2}(n-2) = \pm t_{0,025}(6) = \pm 2,447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \frac{0,92\sqrt{8-2}}{\sqrt{1-0,92^2}} = 2,66.$$

Eksperimentalna vrednost pade v kritično območje.

Zaključek: ob 5% stopnji značilnosti lahko rečemo, da je izobrazba linearno povezana z branjem dnevnih časopisov.

Parcialna korelacija

Včasih je potrebno meriti zvezo med dvema spremenljivkama in odstraniti vpliv vseh ostalih spremenljivk.

To zvezo dobimo s pomočjo koeficienta parcialne korclacije. Pri tem seveda predpostavljamo, da so vse spremenljivke med seboj linearno povezane. Če hočemo iz zveze med spremenljivkama X in Y odstraniti vpliv tretje spremenljivke Z, je **koeficient parcialne korelacije**:

$$r_{XY,Z} = \frac{r_{XY} - r_{XZ}\, r_{YZ}}{\sqrt{1-r_{XZ}^2}\sqrt{1-r_{YZ}^2}}.$$

Tudi ta koeficient, ki zavzema vrednosti v intervalu $[-1,1]$, interpretiramo podobno kot običajni koeficient korelacije.

S pomočjo tega obrazca lahko razmišljamo naprej, kako bi izločili vpliv naslednjih spremenljivk.

Primer:
V neki ameriški raziskavi, v kateri so proučevali vzroke za kriminal v mestih, so upoštevali naslednje spremenljivke:

X : % nebelih prebivalcev,

Y : % kaznivih dejanj,

Z : % revnih prebivalcev,

U : velikost mesta.

Izračunali so naslednje koeficiente korelacije:

	X	Z	U	Y
X	1	$0,51$	$0,41$	$0,36$
Z		1	$0,29$	$0,60$
U			1	$0,49$
Y				1

Zveza med nebelim prebivalstvom in kriminalom je

$$r_{XY} = 0,36.$$

Zveza je kar močna in lahko bi mislili, da nebeli prebivalci povzročajo več kaznivih dejanj.

Vidimo pa še, da je zveza med revščino in kriminalom tudi precejšna

$$r_{YZ} = 0,60.$$

Lahko bi predpostavili, da revščina vpliva na zvezo med nebelci in kriminalom, saj je tudi zveza med revnimi in nebelimi precejšna $r_{XZ} = 0,51$.

Zato poskusim odstraniti vpliv revščine iz zveze:

"nebelo prebivalstvo : kazniva dejanja":

$$r_{XY,Z} = \frac{0,36 - 0,51 \cdot 0,60}{\sqrt{1-0,51^2}\sqrt{1-0,60^2}}.$$

Vidimo, da se je linearna zveza zelo zmanjšala.

Če pa odstranimo še vpliv velikosti mesta, dobimo parcialno korelacijo $-0,02$ oziroma zveze praktično ni več.

Regresijska analiza

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 regresijske 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 fukcija linearna:

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

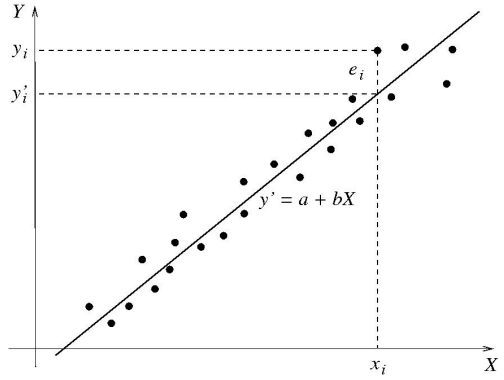
je regresijska odvisnost

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

oziroma za i to enoto

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

Regresijsko odvisnost si lahko zelo nazorno predstavimo v razsevnem grafikonu:



Regresijsko funkcijo lahko v splošnem zapišemo

$$Y' = f(X, a, b, \cdots),$$

kjer so $a, b, \ldots$ 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.

... Regresijska analiza

Pri dvorazsežno normalno porazdeljenem slučajnem vektorju $(X, Y) : N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ je, kot vemo

$$\mathsf{E}(Y|x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

Pogojna porazdelitev Y glede na X je tudi normalna:

$$N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y\sqrt{1-\rho^2}).$$

Regresija je linearna in regresijska krivulja premica, ki gre skozi točko (σ_x, σ_y).
Med Y in X ni linearne zveze, sta le 'v povprečju' linearno odvisni.
Če označimo z $\boldsymbol{\beta} = \rho\frac{\sigma_y}{\sigma_x}$ *regresijski koeficient*, $\boldsymbol{\alpha} = \mu_y - \beta\mu_x$ in $\boldsymbol{\sigma^2} = \sigma_y\sqrt{1-\rho^2}$, lahko zapišemo zvezo v obliki

$$y = \alpha + \beta x.$$

Preizkušanje regresijskih koeficientov

Po metodi momentov dobimo cenilki za α in β:

$$B = R\frac{C_y}{C_x} = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \overline{Y} - B\overline{X},$$

kjer so $C_x^2 = \sum_{i=1}^n (X_i - \overline{X})^2$, $C_y^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2$ in $C_{xy} = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$.

Kako sta cenilki B in A porazdeljeni?

$$B = \frac{C_{xy}}{C_x^2} = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}(Y_i - \overline{Y}) = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}Y_i.$$

Ker proučujemo pogojno porazdelitev Y glede na X (torej so vrednost X poznane), obravnavamo spremenljivke $X_1, \ldots, X_n$ kot konstante.
Ker je B linearna funkcija spremenljivk $Y_1, \ldots, Y_n$, ki so normalno porazdeljene $Y_i : N(\alpha + \beta X_i, \sigma)$, je tudi B normalno porazdeljena.

...Preizkušanje regresijskih koeficientov

Določimo parametra te porazdelitve:

$$\mathsf{E}B = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}\mathsf{E}Y_i = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}(\alpha + \beta X_i) = \sum_{i=1}^n \frac{X_i - \overline{X}}{C_x^2}\beta(X_i - \overline{X}) = \beta.$$

Pri tem upoštevamo, da je $\sum_{i=1}^n (X_i - \overline{X}) = 0$ in da sta α ter $\overline{X}$ konstanti.

$$\mathsf{D}B = \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{C_x^4}\mathsf{D}Y_i = \frac{\sigma^2}{C_x^2}.$$

Torej je $B : N\left(\beta, \frac{\sigma}{C_x}\right)$, oziroma $\frac{B - \beta}{\sigma}C_x : N(0, 1)$.

Podobno dobimo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\right).$$

. . . Preizkušanje regresijskih koeficientov

Težje se je dokopati do cenilke za parameter σ^2.
Označimo $Q^2 = \sum_{i=1}^{n}(Y_i - A - BX_i)^2$.

Po nekaj računanja se izkaže, da velja $\mathsf{E}\dfrac{Q^2}{\sigma^2} = n-2$.

Torej je $S^2 = \dfrac{Q^2}{n-2} = \dfrac{\sigma^2}{n-2}\,\dfrac{Q^2}{\sigma^2}$ nepristranska cenilka za σ^2.

S^2 je neodvisna od A in B. Testni statistiki za A in B sta tedaj

$$T_A = \frac{A - \mathsf{E}A}{\sqrt{\mathsf{D}A}} = \frac{A-\alpha}{S}\sqrt{\frac{nC_x^2}{C_x^2 + n\overline{X}^2}} = \frac{A-\alpha}{S}C_x\sqrt{\frac{n}{\sum_{i=1}^{n}X_i^2}},$$

$$T_B = \frac{B - \mathsf{E}B}{\sqrt{\mathsf{D}B}} = \frac{B-\beta}{S}C_x,$$

ki sta obe porazdeljeni po Studentu $S(n-2)$. Statistika za σ^2 pa je spremenljivka $\dfrac{Q^2}{\sigma^2} = (n-2)\dfrac{S^2}{\sigma^2}$, ki je porazdeljena po $\chi^2(n-2)$.

. . . Preizkušanje regresijskih koeficientov

Pokazati je mogoče tudi, da velja

$$Q^2 = C_y^2 - B^2C_x^2 = C_y^2(1-R^2).$$

To nam omogoča S v statistikah zapisati z C_y in R.

Te statistike uporabimo tudi za določitev intervalov zaupanja za parametre α, β in σ^2.

Linearni model

Pri proučevanju pojavov pogosto teorija postavi določeno funkcijsko zvezo med obravnavanimi spremenljivkami. Oglejmo si primer *linernega modela*, ko je med spremenljivkama x in y linearna zveza

$$y = \alpha + \beta x$$

Za dejanske meritve se pogosto izkaže, da zaradi različnih vplivov, ki jih ne poznamo, razlika $u = y - \alpha - \beta x$ v splošnem ni enaka 0, čeprav je model točen. Zato je ustreznejši *verjetnostni linearni model*

$$Y = \alpha + \beta X + U,$$

kjer so X, Y in U slučajne spremenljivke in $\mathsf{E}U = 0$ – model je vsaj v povprečju linearen.

. . . Linearni model

Slučajni vzorec (meritve) $(X_1, Y_1), \ldots, (X_n, Y_n)$ je realizacija slučajnega vektorja. Vpeljimo spremenljivke

$$U_i = Y_i - \alpha - \beta X_i$$

in predpostavimo, da so spremenljivke U_i med seboj neodvisne in enako porazdeljene z matematičnim upanjem 0 in disperzijo σ^2. Torej je:

$\mathsf{E}U_i = 0, \quad \mathsf{D}U_i = \sigma^2 \quad$ in $\quad \mathsf{E}(U_iU_j) = 0, \quad$ za $i \neq j$.

Običajno privzamemo še, da lahko vrednosti X_i točno določamo – X_i ima vedno isto vrednost. Poleg tega naj bosta vsaj dve vrednosti X različni.

Težava je, da (koeficientov) premice $y = \alpha + \beta x$ ne poznamo.
Recimo, da je približek zanjo premica $y = a + bx$.

. . . Linearni model

Določimo jo po *načelu najmanjših kvadratov* z minimizacijo funkcije

$$f(a,b) = \sum_{i=1}^{n}(y_i - (bx_i + a))^2.$$

Naloga zadošča pogojem izreka. Iz pogoja $\nabla P = 0$ dobimo enačbi

$$\begin{aligned}\frac{\partial f}{\partial a} &= \sum_{i=1}^{n} 2(y_i - (bx_i + a)) = 0,\\ \frac{\partial f}{\partial b} &= \sum_{i=1}^{n} 2(y_i - (bx_i + a))x_i = 0,\end{aligned}$$

z rešitvijo

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}, \qquad a = \frac{1}{n}\Big(\sum y - b\sum x\Big).$$

. . . Linearni model

oziroma, če vpeljemo oznako $\overline{z} = \frac{1}{n}\sum z$:

$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}, \qquad a = \overline{y} - b\overline{x}.$$

Poglejmo še Hessovo matriko

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial a^2} & \frac{\partial^2 f}{\partial a \partial b} \\ \frac{\partial^2 f}{\partial b \partial a} & \frac{\partial^2 f}{\partial b^2} \end{bmatrix} = 2\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}.$$

Ker je $\Delta_1 = 2\sum x^2 > 0$ in

$$\Delta_2 = 4(n\sum x^2 - (\sum x)^2) = 2\sum\sum(x_i - x_j)^2 > 0,$$

je matrika H pozitivno definitna in zato funkcija P strogo konveksna.
Torej je *regresijska premica* enolično določena.

. . . Linearni model

Seveda sta parametra a in b odvisna od slučajnega vzorca – torej slučajni spremenljivki. Iz dobljenih zvez za a in b dobimo že znani cenilki za koeficients α in β

$$B = \frac{C_{xy}}{C_x^2} \quad \text{in} \quad A = \overline{Y} - B\overline{X}.$$

Iz prej omenjenih predpostavk lahko (brez poznavanja porazdelitve Y in U) pokažemo

$$\mathsf{E}A = \alpha \quad \text{in} \quad \mathsf{D}A = \sigma^2\Big(\frac{1}{n} + \frac{\overline{X}^2}{C_x^2}\Big), \quad \mathsf{E}B = \beta \quad \text{in} \quad \mathsf{D}B = \frac{\sigma^2}{C_x^2},$$

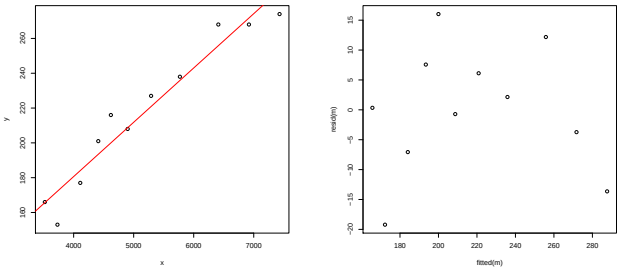
$$K(A,B) = -\sigma^2\frac{\overline{X}}{C_x^2}.$$

Cenilki za A in B sta najboljši linearni nepristranski cenilki za α in β.

. . . Linearni model

```
> x <- c(3520, 3730,4110,4410,4620,4900,5290,5770,6410,6920,7430)
> y <- c(166,153,177,201,216,208,227,238,268,268,274)
> l <- 1947:1957
> plot(y ~ x); abline(lm(y ~ x),col="red")
> m <- lm(y ~ x)
> summary(m)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2149  -5.4003   0.3364   6.8453  16.0204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.852675  14.491253   3.854  0.00388 **
x            0.031196   0.002715  11.492 1.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.18 on 9 degrees of freedom
Multiple R-Squared: 0.9362,     Adjusted R-squared: 0.9291
F-statistic: 132.1 on 1 and 9 DF,  p-value: 1.112e-06

> plot(fitted(m),resid(m))
```


. . . Linearni model




```
> coef(m)
(Intercept)           x
 55.8526752   0.0311963
```

To metodo ocenjevanja parametrov regresijske funkcije imenujemo **metoda najmanjših kvadratov**.

Linearni model

Če izračunana parametra vstavimo v regresijsko funkcijo, dobimo:

$$Y = \mu_Y + \frac{\mathsf{Cov}(X,Y)}{\sigma_X^2}(X - \mu_X).$$

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

Podobno bi lahko ocenili linearno regresijsko funkcijo

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

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

$$X = \mu_X + \frac{\mathsf{Cov}(X,Y)}{\sigma_Y^2}(Y - \mu_Y).$$

To funkcijo imenujemo **druga** regresijska funkcija,

Primer: Vzemimo primer 8 osob, ki smo ga obravnavali v poglavju o povezanosti dveh številskih spremenljivk.

Spremenjivki sta bili:
X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.
Spomnimo se podatkov za teh 8 slučajno izbranh osob:

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

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

Ko smo računali koeficient korelacije smo že izračunali aritmetični sredini

$$\mu_X = \frac{64}{8} = 8, \qquad \mu_Y = \frac{24}{8} = 3,$$

vsoti kvadratov odklonov od aritmetične sredine za obe spremenljivki

$$\sum_{i=1}^{n}(x_i - \mu_X)^2 = 104, \qquad \sum_{i=1}^{n}(y_i - \mu_Y)^2 = 26$$

in vsoto produktov odklonov od obeh aritmetičnih sredin

$$\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) = 48.$$

Potem sta regresijski premici

$$Y = \mu_Y + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\sum_{i=1}^{N}(x_i-\mu_X)^2}(X-\mu_X),$$

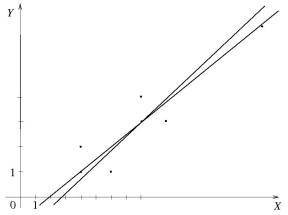
$$X = \mu_X + \frac{\sum_{i=1}^{N}(x_i-\mu_X)(y_i-\mu_Y)}{\sum_{i=1}^{N}(y_i-\mu_Y)^2}(Y-\mu_Y),$$

oziroma

$$Y = 3+\frac{48}{104}(X-8) = -0,68+0,46X,$$

$$X = 8+\frac{48}{26}(Y-3) = -2,46+1,85X.$$

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



Regresijski premici se sečeta v točki, določeni z aritmetičnima sredinama spremenljivk X in Y.
Dokažite, da se premici vedno sečeta v tej točki.

Statistično sklepanje o regresijskem koeficientu

Vpeljmo naslednje oznake:

$Y = \alpha + \beta X$ regresijska premica na populaciji,
$Y = a + bX$ regresijska premica na vzorcu.

Denimo, da želimo preveriti domnevo o regresijskem koeficientu β. Postavimo ničelno in osnovno domnevo takole:

H_0: $\beta = \beta_0$,
H_1: $\beta \neq \beta_0$.

Nepristranska cenilka za regresijski koeficient β je $b = \mathsf{Cov}(X,Y)/s_X^2$, ki ima matematično upanje in standardno napako:

$$\mathsf{E}b) = \beta; \quad \mathsf{SE}(b) = \frac{s_Y\sqrt{1-r^2}}{s_X\sqrt{n-2}}.$$

Testna statistika za zgornjo ničelno domnevo je:

$$t = \frac{s_Y\sqrt{n-2}}{s_X\sqrt{1-r^2}}(b-\beta_0),$$

ki se porazdeljuje po t-porazdelitvi z $m = (n-2)$ prostostnimi stopnjami.

Primer: Vzemimo primer, ki smo ga že obravnavali.

Spremenljivki sta

X - izobrazba (število priznanih let šole),
Y - št. ur branja dnevnih časopisov na teden.

Podatke za slučajno izbrane enote ($n = 8$) najdemo na prejšnjih prosojnicah.

Preverimo domnevo, da je regresijski koeficient različen od 0 pri $\alpha = 5\%$.

Postavimo najprej ničelno in osnovno domnevo:

H_0: $\beta = 0$,
H_1: $\beta \neq 0$.

Gre za dvostranski test. Zato je ob 5% stopnji značilnosti kritično območje določeno s kritičnimia vrednostima:

$$\pm t_{\alpha/2}(n-2) = \pm t_{0,025}(6) = \pm 2,447.$$

Eksperimentalna vrednost statistike pa je:

$$t_e = \sqrt{\frac{104\cdot(8-2)}{26\cdot(1-0,92^2)}}\cdot(0,46-0) = 5,8.$$

Regresijski koeficient je statistično značilno različen od 0.

Pojasnjena varianca (ang. ANOVA)

Vrednost odvisne spremenljivke Y_i lahko razstavimo na tri komponente:

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

kjer so pomeni posameznih komponent

μ_Y : rezultat splošnih vplivov,

$(y_i' - \mu_Y)$: rezultat vpliva spremenljivke X (regresija),

$(y_i - y_i')$: rezultat vpliva drugih dejavnikov (napake/motnje).

Če zgornjo enakost najprej na obeh straneh kvadriramo, nato seštejemo po vseh enotah in končno 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 spremenljvke Y,

σ_e^2 : nepojasnjena varianca spremnenljivke Y.

Delež pojasnjene variance spremenljivke Y s spremenljivko X je

$$R = \frac{\sigma_Y'^2}{\sigma_Y^2}.$$

Imenujemo ga **determinacijski koeficient** in je definiran na intervalu $[0, 1]$.

Pokazati se da, da je v primeru linearne regresijske odvisnosti determinacijski koeficient enak

$$R = \rho^2,$$

kjer je ρ koeficient korelacije.

Kvadratni koren iz nepojasnjene variance σ_e imenujemo **standardna napaka regresijske ocene**, ki meri razpršenost točk okoli regresijske krivulje.

Standarna napaka ocene meri kakovost ocenjevanja vrednosti odvisne spremenljivke z regresijsko funkcijo.

V primeru linearne regresijske odvisnosti je standardna napaka enaka:

$$\sigma_e = \sigma_Y\sqrt{1-\rho^2}.$$

Primer: Vzemimo spremenljivki

X - število ur gledanja televizije na teden

Y - število obiskov kino predstav na mesec

Podatki za 6 oseb so:

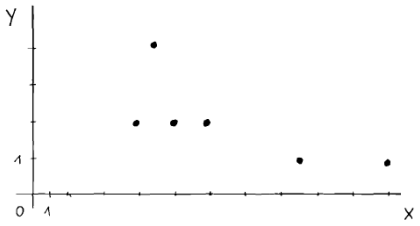
X	10	15	6	7	20	8
Y	2	1	2	4	1	2

Z linearno regresijsko funkcijo ocenimo, kolikokrat bo šla oseba v kino na mesec, če gleda 18 ur na teden televizijo.

Kolikšna je standardna napaka?

Kolikšen delež variance obiska kinopredstav lahko pojasnimo z gledanjem televizije?

Najprej si podatke predstavimo v razsevnem grafikonu:



Za odgovore potrebujemo naslednje izračune:

x_i	y_i	$x_i - \mu_X$	$y_i - \mu_Y$	$(x_i - \mu_X)^2$	$(y_i - \mu_Y)^2$	$(x_i - \mu_X)(y_i - \mu_Y)$
10	2	-1	0	1	0	0
15	1	4	-1	16	1	-4
6	2	-5	0	25	0	0
7	4	-4	2	16	4	-8
20	1	9	-1	81	1	-9
8	2	-3	0	9	0	0
66	12	0	0	148	6	21

$$Y' = 2 - \frac{21}{148}\,(X - 11) = 3,54 - 0,14X$$

$$y'(18) = 3,54 - 0,14 \cdot 18 = 1,02$$

$$\rho = \frac{21}{146 \cdot 6} = -0,70$$

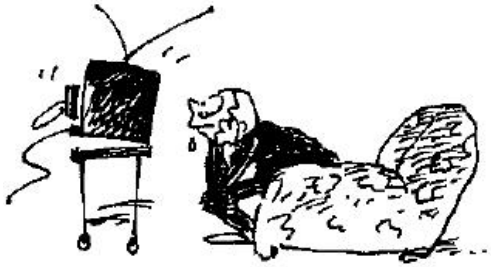
$$\sigma_e^2 = \frac{6}{6}\sqrt{1 - (-0,70)^2} = \sqrt{0,51} = 0,71$$

$$R = (0,70)^2 = 0,49$$

Če oseba gleda 18 ur na teden televizijo, lahko pričakujemo, da bo 1-krat na mesec šla v kino, pri čemer je standardna napaka $0,7$.

49% variance obiska kino predstav lahko pojasnimo z gledanjem televizije.

II.7 Časovne vrste



Družbeno-ekonomski pojavi so časovno spremenljivi. Spremembe so rezultat delovanja najrazličnejših dejavnikov, ki tako ali dugače vplivajo na pojave. Sliko dinamike pojavov dobimo s časovimi vrstami.

Časovna vrsta jo niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke.

Osnovni namen analize časovnih vrst je

- opazovati časovni razvoj pojavov,
- iskati njihove zakonitosti in
- predvidevati nadaljni razvoj.

Seveda to predvidevanje ne more biti popolnoma zanesljivo, ker je skoraj nemogoče vnaprej napovedati in upoštevati vse faktorje, ki vplivajo na proučevani pojav. Napoved bi veljala strogo le v primeru, če bi bile izpolnjene predpostavke, pod katerimi je napoved izdelana.

Časovne vrste prikazujejo individualne vrednosti neke spremenljivke v času. Čas lahko interpretiramo kot trenutek ali razdobje; skladno s tem so časovne vrste

- trenutne, npr. število zaposlenih v določenem trenutku:
- intervalne, npr. družbeni proizvod v letu 1993.

Časovne vrste analiziramo tako, da opazujemo spreminjanje vrednosti členov v časovih vrstah in iščemo zakonitosti tega spreminjanja. Naloga enostavne analize časovnih vrst je primerjava med členi v isti časovni vrsti.

Z metodami, ki so specifične za analizo časovnih vrst, analiziramo zakonitosti dinamike ene same vrste, s korelacijsko analizo pa zakonitosti odvisnosti v dinamiki več pojavov, ki so med seboj v zvezi.

Primer:

Vzemimo število nezaposlenih v Sloveniji v letih od 1981 do 1990.

V metodoloških pojasnilih v Statističnem letopisu Republike Slovenije 1991, so nezaposlni (spremenljivka X) opredeljeni takole:

Brezposelna oseba je oseba, ki je sposobna in voljna delati ter je pripravljena sprejeti zaposlitev, ki ustreza njeni strokovni izobrazbi oz. z delom pridobljeni delovni zmožnosti, vendar brez svoje krivde nima dela in možnosti, da si z delom zagotavlja sredstva za preživetje in se zaradi zaposlitve prijavi pri območni enoti Zavoda za zaposlovanje (do leta 1989 skupnosti za zaposlovanje).

leto	X_k
1981	12.315
1982	13.700
1983	15.781
1984	15.300
1985	11.657
1986	14.102
1987	15.184
1988	21.311
1989	28.218
1990	44.227

Primerljivost členov v časovni vrsti

Kljub temu, da so členi v isti časovni vrsti istovrstne količine, dostikrat niso med seboj neposredno primerljivi.

> Osnovni pogoj za primerljivost členov v isti časovni vrst je pravilna in nedvoumna opredelitev pojava, ki ga časovna vrsta prikazuje. Ta opredelitev mora biti vso dobo opazovanja enaka in se ne sme spreminjati.

Ker so spremembe pojava, ki ga časovna vrsta prikazuje bistveno odvisne od časa, je zelo koristno, če so **časovni razmiki med posameznimi členi enaki**. Na velikost pojavov dostikrat vplivajo tudi **administrativni** ukrepi, ki z vsebino proučevanja nimajo neposredne zveze.

En izmed običajnih vzrokov so upravnoteritorialne spremembe, s katerimi se spremeni geografska opredelitev pojava, ki onemogoča primerljivost podatkov v časovni vrsti. V tem primeru je potrebno podatke časovne vrste za nazaj preračunati za novo območje.

Grafični prikaz časovne vrste

Kompleksen vpogled v dinamiko pojavov dobimo z grafičnim prikazom časovnih vrst v koordinatnem sistemu, kjer nanašamo na abscisno os čas in na ordinatno vrednosti dane spremenljivke. V isti koordinatni sistem smemo vnašati in primerjati le istovrstne časovne vrste.

Primer:
Grafično prikažimo število brezposelnih v Sloveniji v letih od 1981 do 1990.

Indeksi

Denimo, da je časovna vrsta dana z vrednostmi neke spremenljivke v časovnih točkah takole:

$$X_1, X_2, \ldots, X_n$$

o indeksih govorimo, kadar z relativnimi števili primerjamo istovrstne podatke.

Glede na to, kako določimo osnovo, s katero primerjamo člene v časovni vrsti, ločimo dve vrsti indeksov:

- **Indeksi s stalno osnovo**
  Člene časovnih vrst primerjamo z nekim stailnim členom v časovni vrsti, ki ga imenujemo osnova X_0

$$I_{k/0} = \frac{X_k}{X_0} \cdot 100.$$

- **Verižni indeksi**
  Za dano časovno vrsto računamo vrsto verižnih indeksov tako, da za vsak člen vzamemo za osnovo predhodni člen

$$I_k = \frac{X_k}{X_{k-1}} \cdot 100.$$

  člene časovne vrste lahko primerjamo tudi z absolutno in relativno razliko med členi:

- **Absolutna razlika**

$$D_k = X_k - X_{k-1}.$$

- **Stopnja rasti** (relativna razlika med členi)

$$T_k = \frac{X_k - X_{k-1}}{X_{k-1}} \cdot 100 = I_k - 100.$$

Interpretacija indeksov

	pojav		
indeks	raste	stagnira	pada
s stalno osnovo	$I_{k+1/0} > I_{k/0}$	$I_{k+1/0} = I_{k/0}$	$I_{k+1/0} < I_{k/0}$
verižni indeks	$I_k > 100$	$I_k = 100$	$I_k < 100$
stopnja rasti	$T_k > 0$	$T_k = 0$	$T_k < 0$

Primer: Izračunajmo omenjene indekse za primer brezposelnih v Sloveniji:

leto	X_k	$I_{k/0}$	I_k	T_k
1981	12.315	100	–	–
1982	13.700	111	111	11
1983	15.781	128	115	15
1984	15.300	124	97	−3
1985	11.657	119	96	−4
1986	14.102	115	97	−3
1987	15.184	124	107	7
1988	21.311	173	141	41
1989	28.218	229	132	32
1990	44.227	359	157	57

Rezultati kažejo, da je bila brezposelnost v letu 1990
kar 3,5 krat večja kot v letu 1981
(glej indeks s stalno osnovo).

Iz leta 1989 na leto 1990 je bil
prirast nezposlenih 57%
(glej stopnjo rasti).

Sestavine dinamike v časovnih vrstah

Posamezne vrednosti časovnih vrst so rezultat številnih dejavnikov,
ki na pojav vplivajo.
Iz časovne vrste je moč razbrati skupen učinek dejavnikov,
ki imajo širok vpliv na pojav, ki ga proučujemo.
Na časovni vrsti opazujemo naslednje vrste sprememb:

1. Dolgoročno gibanje ali trend - X_T
   podaja dolgoročno smer razvoja.
   Običajno ga je mogoče izraziti s
   preprostimi rahlo ukrivljenimi krivuljami.

2. Ciklična gibanja - X_C,
   so oscilarijo okoli trenda.
   Poriode so ponavdi daljše od enega leta
   in so lahko različno dolge.

3. Sezonske oscilacije - X_S
   so posledice vzrokov, ki se pojavljajo na stalno razdobje.
   Periode so krajše od enega leta, ponavadi sezonskega značaja.

4. Naključne spremembe - X_E
   so spremembe, ki jih ne moremo razložiti s
   sistematičnimi gibanji (1, 2 in 3).

Časovna vrsta ne vsebuje nujno vseh sestavin. Zvezo med sestavinami je mogoče prikazati z nekaj osnovnim modeli. Npr.:

$$X = X_T + X_C + X_S + X_E$$

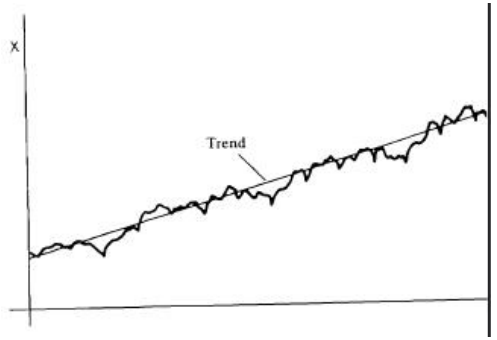
ali

$$X = X_T \cdot X_C \cdot X_S \cdot X_F;$$

ali

$$X = X_T \cdot X_C \cdot X_S + X_E.$$

Primer časovne vrste z vsemi štirimi sestavinami:



Ali je v časovni vrsti trend?

Obstaja statistični test, s katerim preverjamo ali trend obstaja v časovni vrsti. Med časom in spremenljivko izračunamo koeficient korelacije rangov

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

kjer je d_i, razlika med rangoma i tega časa in pripadajoče vrednosti spremenljivke. Ničelna in osnovna domneva sta:

H_0: $\rho_e = 0$ trend ne obstaja
H_1: $\rho_e \neq 0$ trend obstaja

Ustrezna statistika je

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

ki se porazdeluje približno po t porazdelitvi z $(n-2)$ prostostnimi stopnjami.

Metode določanja trenda

- Prostoročno
- Metoda drsečih sredin
- Metoda najmanjših kvadratov
- Druge analitične metode

Drseče sredine

Metoda drsečih sredin lahko pomaga pri določitvi ustreznega tipa krivulje trenda. V tem primeru namesto člena časovne vrste zapišemo povprečje določenega števila sosednjih članov. Če se odločimo za povprečje treh členov, govorimo o tričlenski vrsti drsečih sredin. Tedaj namesto članov v osnovni časovni vrsti X_k: tvorimo tričlenske drseče sredine X :

$$X'_k = \frac{X_{k-1} + X_k + X_{k+1}}{3}.$$

V tem primeru prvega in zadnjega člena časovne vrste moramo izračunati.

- Včasih se uporablja obtežena aritmetična sredina, včasih celo geometrijska za izračun drsečih sredin.
- Če so v časovni vrsti le naključni vplivi, dobimo po uporabi drsečih sredin ciklična gibanja (učinek Slutskega).
- Če so v časovni vrsti stalne periode, lahko drseče sredine zabrišejo oscilacije v celoti.
- V splošnem so drseče sredine lahko dober približek pravemu trendu.

Primer: Kot primer drsečih sredin vzemimo zopet brezposelne v Sloveniji. Izračunajmo tričlensko drsečo sredino:

T	X_k	tričl. drs. sred.
1981	12.315	–
1982	13.700	13.032
1983	15.781	14.030
1984	15.240	15.249
1985	15.300	14.710
1986	14.657	14.678
1987	14.102	15.184
1988	21.341	21.581
1989	28.218	31.262
1990	44.227	–

Analitično določanje trenda

Trend lahko obravnavamo kot posebni primer regresijske funkcije, kjer je neodvisna spremenljivka čas (T). Če je trend

$$X_T = f(T),$$

lahko parametre trenda določimo z metoda najmanjših kvadratov

$$\sum_{i=1}^{n}(X_i - X_{iT})^2 = \min.$$

V primeru linearnega trenda

$$X_T = a + bT,$$

$$\sum_{i=1}^{n}(X_i - a - bT_i)^2 = \min.$$

dobimo naslednjo oceno trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X})(T_i - \overline{T})}{\sum_{i=1}^{n}(T_i - \overline{T})^2}\,(T - \overline{T}).$$

Ponavadi je čas T transformiran tako, da je $t = 0$. Tedaj je ocena trenda

$$X_T = \overline{X} + \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \cdot t_i}{\sum_{i=1}^{n} t_i^2}\,t.$$

Standardna napaka ocene, ki meri razpršenost točk okoli trenda, je

$$\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{iT})^2}.$$

Primer:

Kot primer ocenimo število doktoratov znanosti v Sloveniji v razdobju od leta 1986 do 1990. Z linearnim trendom ocenimo koliko doktorjev znanosti je v letu 1991. Izračunajmo tudi standardno napako ocene.

Izračunajmo najprej trend:

T	Y_i	t_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})t_i$	t_i^2
1986	89	-2	$-19,8$	$39,6$	4
1987	100	-1	$-8,8$	$8,8$	1
1988	118	0	$9,2$	0	0
1989	116	1	$7,2$	$7,2$	1
1990	121	2	$12,2$	$24,4$	4
	544	0		80	10

$$\overline{Y} = \frac{544}{4} = 1088,$$

$$Y_T = 108,8 + \frac{80}{10}\,t = 108,8 + 8t,$$

$$Y_T(1991) = 108,8 + 8 \cdot 3 = 132,8.$$

Ocena za leto 1991 je približno 133 doktorjev znanosti.

Zdaj pa izračunajmo standardno napako ocene.
Za vsako leto je potrebno najprej izračunat

T	Y_i	Y_{iT}	$Y_i - Y_{iT}$	$(Y_i - Y_{iT})^2$
1986	89	$92,8$	$-3,8$	$14,14$
1987	100	$100,8$	$-0,8$	$0,64$
1988	118	$108,8$	$9,2$	$84,64$
1989	116	$116,8$	$-0,8$	$0,64$
1990	121	$124,8$	$-3,8$	$14,44$
	544	544	0	$114,8$

$$\sigma_e = \sqrt{\frac{114,8}{5}} = 4,8.$$