STATISTIKA

1 Kaj je statistika

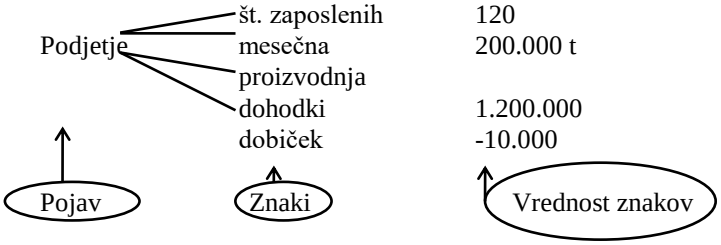
- številčni podatki ali statistika (ind., kmetijska statistika)
- delo pri zbiranju statističnih podatkov (ind. statistika, kmet. statistika)
- organi, ki zbirajo statistične podatke
- znanost, ki uči teorijo in metode statističnega preučevanja

2 Evidenca in statistika

Če beležimo pojave se ukvarjamo z evidenco, če pa jih analiziramo, se ukvarjamo s statistiko.

3 Enota pojava, znak, vrednost znaka

Če gremo v neko podjetje preučevat določene zadeve, je to podjetje *pojav*. Če pa je to le eno od podjetij, ki jih preučujemo, pa je to *enota pojava*. Vsaka stvar, ki jo preučujemo (št. zaposlenih, mesečna proizvodnja) je *znak*. Vsak znak ima *vrednost*.



Slika 1

Pojav je treba opredeliti z *opredeljujočimi pogoji:*

- časovno (na dan 1.1.1997)
- krajevno (v Sloveniji)
- stvarno (tekstilna ind.)

Vsi pojavi sodijo v *populacijo*.

4 Populacijski vzorec

Običajno je ukvarjanje s populacijo dolgo in drago. Zato vzamemo del populacije - *vzorec*. Vzroki za to, da vzamemo vzorec:- populacija je prevelika (zamudno, drago)
- brezmejne populacije (merjenje temp. v Ljubljani)

Vzorec se uporablja za ocenjevanje vrednosti populacije. Zato je pomembno kako ga izberemo. Oblikovan mora biti naključno, izražati mora značilnosti populacije.
Stratificirani vzorec - vzorec, pri katerem smo uporabili podatke o pojavu, ki jih že poznamo (vemo, da je v populaciji 43% fantov in 57% deklet - tako sestavimo tudi populacijo.

5 Izražanje vrednosti znakov

5.1 Krajevni, časovni, stvarni znaki

1. kakorkoli povezan z lokacijo
2. kakorkoli povezan s časom
3. ostali znaki

5.2 Numerični, atributivni znaki

1. izraženi s številko (teža, količina, cena...)
2. izraženi opisno (barva, kvalifikacijska struktura, spol...)

5.3 Izmerljivi, neizmerljivi znaki

1. tisti, ki se dajo izmeriti (vsi numerični)
2. tisti, ki jih lahko uredimo po velikosti: -kvalifikacijska struktura (nekval., kval., visoko kval.) - primerljivi
   - spol (moški, ženski) - neprimerljivi

5.4 Zvezni, nezvezni znaki

1. pokrivajo vrednosti na celotnem intervalu (višina, starost)
2. zavzemajo le določene vrednosti (spol, št. članov družine - ne more biti 2,5)

5.5 Momentni, intervalni znaki

1. v določenem trenutku (št. zaposlenih na dan...)
2. v določenem obdobju (dohodek v podjetju)

5.6 Absolutni, relativni znaki

1. sami zase povedo nekaj (višina, št. prebivalcev, starost)
2. nanašajo se še na nekaj (hitrost - km/h, kol. padavin - mm/m2, rodnost - št. živorojenih/1000 prebivalcev)

6 Statistična vrsta, rang

	stat. vr.	rang	stat. vr.	rang	stat. vr.	rang	sta. vr.
	višina	R_{vis}	izp. ocena	R_{io}	kvalif. strukt.	$R_{kv.str.}$	spol
A	174	190	9	9	visoko kv.	visoko kv.	M
B	181	190	1	9	nekval.	kvalif.	Ž
C	190	186	7	7	nekval.	nekval.	Ž
D	172	182	9	3	nekval.	nekval.	Ž
E	186	181	3	1	kvalif.	nekval.	M
F	190	175					
G	167	174					
H	171	172					
I	182	171					
J	175	167					

Tabela 1

Ko začnemo statistišno vrstico razvrščati po velikosti, dobimo rang (ranžirno vrsto). Ranžirno vrsto pri atributivnih znakih lahko naradimo le, če imajo lastnost primerljivosti (pri spolu je ne moremo).
Časovna vrsta: statistična vrstica je urejena po času.

leto	prid. pšenice
1990	12 t
1991	17 t
1992	11 t

Tabela 2

7 Srednja vrednost

7.1 Medijana (Me)

Medijana vseh vrednosti nekega znaka je tista vrednost, ki razdeli ranžirno vrsto na dva enaka dela.

7.2 Kvantili

Splošen pojem, ko se ranžirna vrsta razdeli na enake dele. Medijana je eden izmed kvantilov.

- Q - kvartil - razdeli na 4 enake dele
- D - decil - razdeli na 10 enakih delov
- C - centil - razdeli na 100 enakih delov

190 190 186 182 181 175 174 172 171 167

Q_1 Me Q_2 D_5 C_{50} Q_3

7.3 Modus (gostiščnica)

To je tista vrednost statistične vrste, ki je najbolj pogosta. Statistična vrsta ima lahko 0 modusov ali pa več.

7.4 Aritmetična sredina

M_x - za populacije, $\bar{x}$ - za vzorce

Če ima nek znak vrednosti $x_1,x_2,x_3,...,x_n$, je njegova aritmetična sredina enaka:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Aritmetična sredina visin dentov:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{174 + 181 + 191 + 172 + 186 + 190 + 17 + 171 + 182 + 175}{10} = 178{,}8$$

Aritmetična sredina ocen:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{9 + 1 + 7 + 9 + 3}{5} = \frac{29}{5} = 5{,}8$$

7.4.1 Lastnosti aritmetične sredine

1. Če aritmetično sredino pomnožimo s številom enot, je to enako seštevku vseh vrednosti $\bar{x} \cdot n = \sum_{i=1}^{n} x_i$
2. Če aritmetični sredini odštejemo posamične vrednosti in te vrednosti med seboj seštejemo, je njihov seštevek enak 0. $\sum_{i=1}^{n} (\bar{x} - x_i) = 0$ Primer za ocene: (5,8-9) + (5,8-1) + (5,8-7) + (5,8-9) + (5,8-3) = -3,2 + 4,8 - 1,2 - 3,2 + 2,8 = 0
3. Če neki konstanti a odštejemo posamezne vrednosti in to kvadriramo ter seštejemo, je vrednost seštevka minimalna, če je a enak aritmetični sredini. $\sum_{i=1}^{n} (a - x_i)^2 = \min; a = \bar{x}$

1. Imamo statistično vrsto x_i

$x_i' = x_i + a \qquad \bar{x}_i' = \bar{x}_i + a$; naj bo a=1

x_i	9	1	7	9	3
x_i'	10	2	8	10	4

Tabela 3

$\bar{x} = 5{,}8$

$$\bar{x}' = \frac{34}{5} = 6{,}8 = \bar{x} + a = 5{,}8 + 1$$

$x_i'' = k \cdot x_i \Rightarrow \bar{x}'' = k \cdot \bar{x}$; naj bo k=2

x_i	9	1	7	9	3
x_i''	18	2	14	18	6

Tabela 4

$$\bar{x}'' = \frac{58}{5} = 11{,}6 = k \cdot \bar{x} = 2 \cdot 5{,}8$$

$y_i = k \cdot x_i + a \leftarrow$ linearna transformacija stat. vrste

$\bar{y} = k \cdot \bar{x} + a$

x_i	9	1	7	9	3
y_i	19	3	15	19	7

Tabela 5

$$\bar{y} = \frac{63}{5} = 12{,}6 = 1 + 2 \cdot 5{,}8 = k \cdot \bar{x} + a$$

7.4.2 Tehtana aritmetična sredina

f - teža znaka
x - znak

$$M_x = \bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Uporaba:

- Če imamo veliko števil, ki se ponavljajo. Dobimo enak rezultat kot pri navadni aritmetični sredini.

x_i	9	1	7	3
f_i	2	1	1	1

Tabela 6

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{2 \cdot 9 + 1 \cdot 3 + 1 \cdot 7 + 1 \cdot 1}{2 + 1 + 1 + 1} = 5{,}8$$

- Če želimo izračunati povprečje različnih znakov z različnimi deleži. Npr.: povprečno oceno, pri čemer ni bilo enako število posameznih ocen.

x_i	9	1	7	3	9
f_i	0,4	0,1	0,1	0,1	0,3

Tabela 7

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{0{,}4 \cdot 9 + 0{,}1 \cdot 3 + 0{,}1 \cdot 7 + 0{,}1 \cdot 1 + 0{,}3 \cdot 9}{0{,}4 + 0{,}1 + 0{,}1 + 0{,}1 + 0{,}3} = 7{,}4$$

- Kadar delamo z realnimi števili. Utež je vedno tisto, kar se pri realnih številih nahaja v imenovalcu. Npr.: proizvodnost delavca se izraža v št. proizvodov / zaposlenega (x_i). f_i - zaposleni

leto	1	2	3	4	5
x_i	5	3	4	6	2
f_i	10	20	15	5	8

Tabela 8

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{10 \cdot 5 + 20 \cdot 3 + 15 \cdot 4 + 5 \cdot 6 + 8 \cdot 2}{10 + 20 + 15 + 5 + 8} = 3{,}72$$

8 Indeks - vrednosti v času

	jan	*feb*	*mar*	*apr*	*maj*	*jun*	
$čas(t)$	0	1	2	3	4	5	
$vrednost(x_t)$	x_0	x_1	x_2	x_3	x_4	x_5	
mes. proiz	12	13	17	16	16	18	
$indeks.sta\ \ln o.bazo(I_B(0))$	$\frac{x_0}{x_0}$	$\frac{x_1}{x_0}$	$\frac{x_2}{x_0}$	$\frac{x_3}{x_0}$	$\frac{x_4}{x_0}$	$\frac{x_5}{x_0}$	$x100$
$I_B(2)$	$\frac{x_0}{x_2}$	$\frac{x_1}{x_2}$	$\frac{x_2}{x_2}$	$\frac{x_3}{x_2}$	$\frac{x_4}{x_2}$	$\frac{x_5}{x_2}$	$x100$
$verizni.indeks.(I_v)$	–	$\frac{x_1}{x_0}$	$\frac{x_2}{x_1}$	$\frac{x_3}{x_2}$	$\frac{x_4}{x_3}$	$\frac{x_5}{x_4}$	$x100$
$koeficient.rasti(KS)$	–	$\frac{x_1}{x_0}$	$\frac{x_2}{x_1}$	$\frac{x_3}{x_2}$	$\frac{x_4}{x_3}$	$\frac{x_5}{x_4}$	
$stopnja.rasti(SR)$	–	$\frac{x_1 - x_0}{x_0}$	$\frac{x_2 - x_1}{x_1}$	$\frac{x_3 - x_2}{x_2}$	$\frac{x_4 - x_3}{x_3}$	$\frac{x_5 - x_4}{x_4}$	

Tabela 9
Bazni indeks - izračunava se glede na bazno leto, ki je stalno.
Verižni indeks - primerjava dveh zaporednih vrednosti.

Če v zgornjo tabelo vstavimo števila, dobimo:

	x_0	x_1	x_2	x_3	x_4	x_5	*kritična.meja*
	12	13	17	16	16	18	
$I_B = 0$	100	108,3	141,2	133,3	133,3	150	100
$I_B = 2$	70,6	76,5	100	94,1	94,1	105,9	100
I_v	–	108,3	130,8	94,1	100	112,5	100
KS	–	1,083	1,308	0,941	1	1,125	1
SR	–	0,083	0,308	–0,059	0	0,125	0

$$I_B = \frac{x_t}{x_0} \cdot 100$$

$$I_v = \frac{x_t}{x_{t-1}} \cdot 100$$

$$KR = \frac{x_t}{x_{t-1}} = \frac{I_v}{100}$$

$$SR = \frac{x_t - x_{t-1}}{x_{t-1}}$$

Tabela 10
Če je vrednost indeksa nad kritično mejo, se je vrednost v tem obdobju povečala, če pa je pod to mejo, pa se je zmanjšala.

$$SR = KR - 1 = \frac{I_v}{100} - 1$$

I_B = 250 - indeks je v primerjavi z izhodiščnim letom 250
I_v = 250 - indeks je v primerjavi s predhodnim letom 250
KR = 2,5 - pojav se je v primerjavi s predhodnim letom povečal 2,5 krat
SR = 1,5 - pojav se je v primerjavi z izhodiščnim letom povečal 1,5 krat

9 Geometrijska sredina

Uporabimo jo v primeru, ko se podatki med seboj množijo.

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \quad ; x_1 \cdot x_2 \cdot \ldots \cdot x_n - koeficienti\ rasti$$

leto	90	91	92	93	94
x_i	2	5	4	3	6
KR	-	2,5	0,8	0,75	2,0

Tabela 11

$$G = \sqrt[4]{2{,}5 \cdot 0{,}8 \cdot 0{,}75 \cdot 2} = \sqrt[4]{3} = 1{,}316$$

Torej je na koncu (leta 94) količina pridelka enaka, kot če bi konstantno rastla vsako leto s koeficientom rasti 1,316. Dobili smo povprečen koeficient rasti (KR).

10 Agregatni indeksi

Poznamo 4 tipe agregatnih indeksov; 2 sta cenovna (gre za spremembo cen pri določeni fiksni strukturi dobrin), 2 pa količinska (gre za spremembo potrošene količine).

x_i^j ; i = čas, j = dobrina

		$j \rightarrow$	*kruh*	*čevlji*	*kozmetika*
jan(0)	*cena*	x_0^j	20	500	100
	količina	f_0^j	30	1	1
feb(*i*)	*cena*	x_i^j	30	600	100
	količina	f_i^j	20	2	1

Tabela 12

$$I_A = \frac{\sum_{i=1}^{n} x_i^j \cdot f^j}{\sum_{i=1}^{n} x_0^j \cdot f^j} \cdot 100 \quad \text{- Agregatni indeks}$$

Laspeyeres-jev indeks - ponderska struktura (struktura košarice) se nanaša na izhodiščno leto:

$$I_L = \frac{\sum_{i=1}^{n} x_i^j \cdot f_0^j}{\sum_{i=1}^{n} x_0^j \cdot f_0^j} \cdot 100$$

Paasche-jev indeks - ponderska struktura se nanaša na tekoče leto.

$$I_P = \frac{\sum_{i=1}^{n} x_i^j \cdot f_i^j}{\sum_{i=1}^{n} x_0^j \cdot f_i^j} \cdot 100$$

$$I_L = \frac{30 \cdot 30 + 600 \cdot 1 + 100 \cdot 1}{20 \cdot 30 + 500 \cdot 1 + 100 \cdot 1} \cdot 100 = 133{,}3$$

$$I_P = \frac{30 \cdot 20 + 600 \cdot 2 + 100 \cdot 1}{20 \cdot 20 + 500 \cdot 2 + 100 \cdot 1} \cdot 100 = 126{,}7$$

Fischerjev indeks

$$I_F = \sqrt{I_P \cdot I_l}$$

Cenovna indeksa:

$$I_{L,C} = \frac{\sum_{i=1}^{n} C_i \cdot \boxed{K_0}}{\sum_{i=1}^{n} C_0 \cdot \boxed{K_0}} \qquad I_{P,C} = \frac{\sum_{i=1}^{n} C_i \cdot \boxed{K_i}}{\sum_{i=1}^{n} C_0 \cdot \boxed{K_i}}$$

količine baznega obdobja — količine tekočega obdobja

Količinska indeksa:

$$I_{L,C} = \frac{\sum_{i=1}^{n} K_i \cdot C_0}{\sum_{i=1}^{n} K_0 \cdot C_0} \qquad I_{P,C} = \frac{\sum_{i=1}^{n} K_i \cdot C_i}{\sum_{i=1}^{n} K_0 \cdot C_i}$$

11 Mere variacije

11.1 Enostavne mere variabilnosti

a) variacijski razpon: VR = x_M - x_m ; x_M - največja vrednost, x_m - najmanjša vrednost
b) kvartilni razpon: KR = Q_3 - Q_1; tretji kvartil minus prvi kvartil
c) decilni razpon: DR = Q_9 - Q_1; deveti decil minus prvi decil

11.2 Povprečni absolutni odklon

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

11.3 Varianca in standardni odklon

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

11.3.1 Lastnosti variance in standardnega odklona

a) Če posamične vrednosti odštejemo od konstante in jih kvadriramo, je ta varianca minimalna, če je konstanta enaka aritmetični sredini vrednosti.

$$\frac{1}{n}\sum_{i=1}^{n}(A - x_i)^2 = \min, \check{c}e.\ A = \bar{x}$$

b) Če statistično vrsto povečamo za konstanto, se variabilnost ne spremeni.
$x_i' = x_i + a$; $s' = s$
a = 2

x_i	2	4	6	8
x_i'	4	6	8	10

Tabela 13

Če statistično vrsto pomnožimo s koeficientom k, se variabilnost spremeni za k-krat.
$x_i'' = k*x_i$ $s'' = k*s$
k = 2

x_i	2	4	6	8
x_i''	4	8	12	16

Tabela 14

Linearna transformacija:
$y_i = a + k*x_i$; $s_y = k*s_x$

11.4 Koeficient variacije

Če želimo primerjati 2 statistični vrsti med seboj ni dober noben od zgoraj navedenih načinov računanja variabilnosti! Vsak znak bi morali deliti s povprečno vrednostjo.

$KV = \frac{s}{\bar{x}}$ - edina relativna mera za primerjavo dveh statističnih vrst.

Naloga: Imamo 2 skupini varčevalcev. Prva (A) varčuje v DEM (v tisočih), druga (B) pa v SIT (v tisočih). Izračunaj variacijski razpon, absolutni odklon, varianco, standardni odklon in koeficient variacije!

A: x_i: 10 5 1 3 6; $\sum_{i=1}^{n} x_1 = 25 \quad \bar{x} = 5$

B: x_i: 40 20 10 20 10; $\sum_{i=1}^{n} x_1 = 100 \quad \bar{x} = 20$

A

$VR_A = 10 - 1 = 9$

x_i	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
10	5	25
5	0	0
1	4	16
3	2	4
6	1	1
	12	46

Tabela 15

$$AD_A = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{5}\cdot 12 = 2{,}4$$

$$s_A^2 = \frac{1}{5}\cdot 46 = 9{,}2$$

$$s_A = \sqrt{9{,}2} = 3{,}03$$

$$KV_A = \frac{3{,}03}{5} = 0{,}61$$

B

$VR_B = 40\text{-}10 = 30$

x_i	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
40	20	400
20	0	0
10	10	100
20	0	0
10	10	100
	40	60

Tabela 16

$$AD_B = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = \frac{1}{5}\cdot 40 = 8$$

$$s_A^2 = \frac{1}{5}\cdot 600 = 120$$

$$s_A = \sqrt{120} = 10{,}95$$

$$KV_A = \frac{10{,}95}{20} = 0{,}55$$

12 Porazdelitev - distribucija

To je porazdelitev znakov v razrede. Sturgesovo pravilo pravi, da je optimalno št. razredov za N znakov 1+3,3logN.

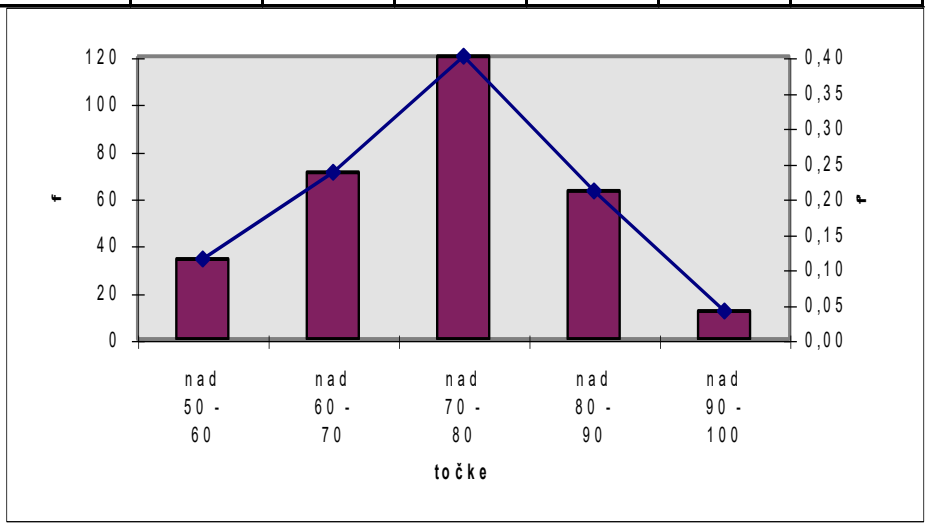
Večina pojavov se porazdeljuje tako, da so razredi enako veliki. Lahko pa so tudi različno široki razredi, ker bi sicer prišli vsi znaki v en razred ali pa bi bilo veliko razredov praznih. Ko tvorimo razrede, mora biti popolnoma jasno v kateri razred znak spada:

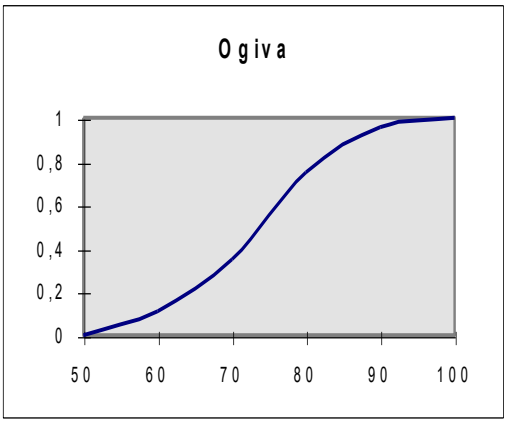
ni dobro	bolje		
150 - 160	151 - 160	nad 150 - 160	150 do 160
160 - 170	161 - 170	nad 160 - 170	160 do 170
170 - 180	171 - 180	nad 170 - 180	170 do 180

Tabela 17

12.1 Frekvenčna distribucija

x	f	f' = f/N	F	F' = F/N	50	0
nad 50 - 60	34	0,11	34	0,11	60	0,11
nad 60 - 70	71	0,24	105	0,35	70	0,35
nad 70 - 80	120	0,40	225	0,75	80	0,75
nad 80 - 90	63	0,21	288	0,96	90	0,96
nad 90 - 100	12	0,04	300	1	100	1
N	300	1				





Povprečno vrednost iz frekvenčne porazdelitve izračunamo s tehtano aritmetično sredino, kjer je x_i sredina razreda in f_i frekvenca v razredu.

$$M_x = \bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

13 Verjetnostni račun

13.1 Samostojna verjetnost P(A), P(B),...

točke	ocena	f_i	f_i'
51-60	6	34	0,11
61-70	7	7	0,24
71-80	8	120	0,40
81-90	9	63	0,21
91-100	10	12	0,04
		300	1

Tabela 18
A: nekdo dobi oceno 8
P(A) = 0,40 = 40%
B: nekdo dobi oceno 6
P(B) = 0,11 = 11%

13.2 Skupna verjetnost

P(A ali B) = P(A) + P(B)
A: dobi oceno 8
B: dobi oceno 10
P(A ali B) = 0,40 + 0,04 =0,44 = 44%

13.3 Družna verjetnost

Kolikšna je verjetnost, da bo od dveh študentov eden ocenjen z oceno 8, drugi pa z oceno 9?
P(A in B) = P(A) * P(B) = P(8) * P(9) = 0,40 * 0,21 = 0,084 = 8,4%
Za odvisne dogodke velja: P (A in B) = P(A) * P(B/A) = P(B) * P(A/B)
Za izključujoče dogodke velja: P(A in B) = 0;
A: nekdo je moški
B: nekdo je ženska
→ ne more biti hkrati moški in ženska

13.4 Pogojna verjetnost

Odvisni dogodki
Imamo 5 kroglic, 3 črne in 2 beli.

x_i	f_i	f_i'	
Č	3	0,6	P(Č) = 0,6
B	2	0,4	
Č	2	0,5	P(Č) = 0,5
B	2	0,5	
Č	1	1/3	P(Č) = 1/3
B	2	2/3	
Č	0	0	P(Č) = 0
B	2	1	

Pogojna verjetnost je verjetnost dogodka, pri čemer vemo, da se je prej zgodil nek drug dogodek.

točke	ocena	Ž	M	f_i - sk	$f_ž'$	f_m'	f_i'
51-60	6	30	4	34	0,09	0,02	0,11
61-70	7	46	25	7	0,16	0,08	0,24
71-80	8	105	15	120	0,35	0,05	0,40
81-90	9	53	10	63	0,18	0,03	0,21
91-100	10	6	6	12	0,02	0,02	0,04
		240	60	300	0,80	0,20	1

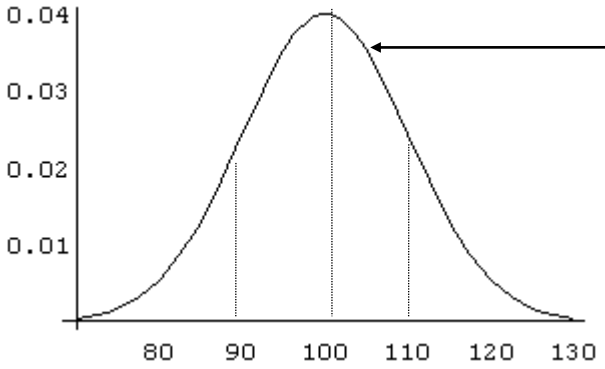
Tabela 19 - verjetnostna tabela
Verjetnost, da bo naključno izbrani ocenjen z oceno 10, pri pogoju, da smo izbrali žensko:

$$P(A/B) = \frac{P(AinB)}{P(B)} = \frac{0,02}{0,80} = 0,025$$

Neodvisni dogodki:

$$P(A/B)=\frac{P(AinB)}{P(B)}=\frac{P(A)\cdot P(B)}{P(B)}=P(A)$$

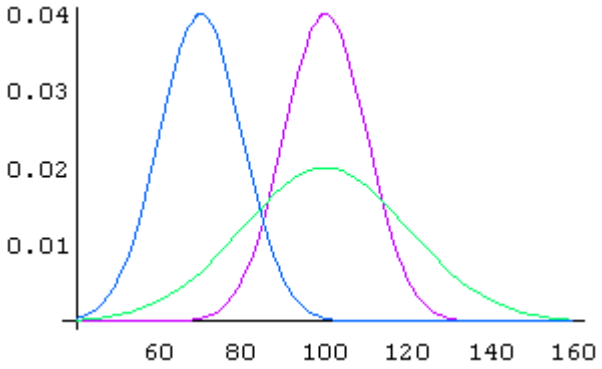
14 Normalna porazdelitev



P(x) - označuje gostoto frekvenčne porazdelitve
Mx = 100
$\sigma_x = 10$

$$P(x)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{1}{2}\cdot\frac{(x-Mx)^2}{\sigma_x^2}}$$

Za vsako normalno porazdelitev velja, da se v območju aritmetične sredine (Mx) ± en standardni odklon nahaja 68,27 % vseh enot, v območju Mx ± 2 standardna odklona se nahaja 95,45 vseh enot, v območju Mx ± 3 standardni odkloni pa se nahaja 99,75 vseh enot.



a) Mx = 100
$s_x = 10$
b) Mx = 100
$s_x = 20$
c) Mx = 70
$s_x = 10$

14.1 Lastnosti

- krivulja je *unimodalna* (ima en modus)
- je simetrična
- je zvonaste oblike

15 Standardizirana normalna porazdelitev

Standardizacija se opravi tako, da se za vsako enoto x_i izračuna $z_i=\frac{x_i-Mx}{\sigma_x}$ in $P(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$

Mz = 0
$\sigma_z = 1$

Vedno velja !!

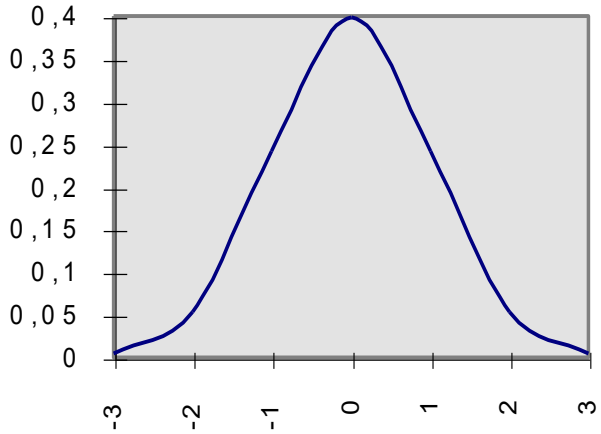
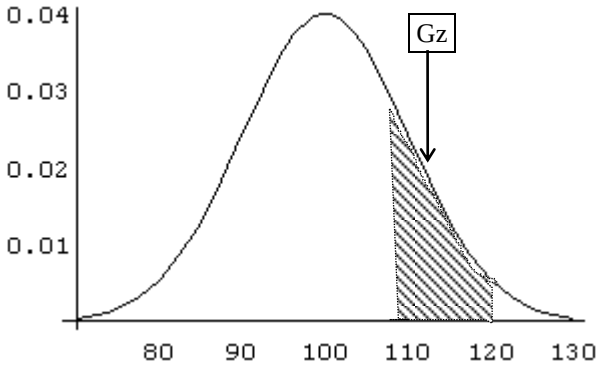


Tabela porazdelitve vsebuje število vrednosti od 0 do z. Odčitava se tako, da navpično raste celo število s prvo decimalko, vodoravno pa druga decimalka števila.

z	0	1	2	3	4	5	6	7	8	9
0.0										
0.1				0.13						
0.2										
...										
1.0										
1.1						1.15				
...										
1.5										
...										
2.0										
2.1										
...										
2.5								2.57		
2.6										
...										
3.5										

Tabela 20

Kolikšna je verjetnost, da je skok daljši od 110 m in krajši od 120 m, če je Mx=100m in s_x=10m in se skoki razporejajo normalno?



$$z_2 = \frac{x_2 - Mx}{\sigma_x} = \frac{120 - 100}{10} = 2$$

$$z_1 = \frac{x_1 - Mx}{\sigma_x} = \frac{110 - 100}{10} = 1$$

$$G(z) = H(z_2) - H(z_1) = 0{,}1359$$

Vrednosti $H(z_1)$ in $H(z_2)$ preberemo iz tabele.

16 Jakost povezanosti znakov

16.1 Atributivni znaki

16.1.1 Kontingenca

Ker se atributivnih znakov ne da seštevati in odštevati, jih je potrebno prevesti v numerične vrednosti. Oblikovati moramo kontingenčne tabele:
Ali je sodba odvisna od sodnika?

f_{gk}		k_1	k_2	k_3		
		Aco	Bine	Cene	Σ	
g_1	Oprostilna	130	162	90	382	f_{g1}
g_2	Pogojna	13	52	40	105	f_{g2}
g_3	Zaporna	7	16	20	43	f_{g3}
	Σ	150	230	150	530	
		f_{k1}	f_{k2}	f_{k3}	N	

Tabela 21 - absolutne frekvence

	Aco	Bine	Cene	Σ
Oprostilna	86,7	70,4	60,0	72,1
Pogojna	8,7	22,6	26,7	19,8
Zaporna	4,6	7,0	13,3	8,1
Σ	100,0	100,0	100,0	100,0

Tabela 22 - relativne frekvence

Če bi šlo za popolno odvisnost od sodnika, bi bila tabela lahko taka:

	Aco	Bine	Cene	Σ
Oprostilna	150	0	0	150
Pogojna	0	230	0	230
Zaporna	0	0	150	150
Σ	150	230	150	530

Tabela 23

$f_{gk}°$	Aco	Bine	Cene	Σ
Oprostilna	108	166	108	382
Pogojna	30	45	30	105
Zaporna	12	19	12	43
Σ	150	230	150	530

Tabela 24 - teoretične frekvence
Teoretične frekvence določijo, kako bi se vrednosti porazdeljevale, da ne bi bilo nobene odvisnosti od sodnika.
Primer izračuna (za Acota): 150 * 0,721 = 108
150 * 0,198 = 30
150 * 0,081 = 12

$$f_{gk}^{0} = \frac{f_k \cdot f_g}{N}$$

$$\frac{382 \cdot 150}{530} = 108$$

$$\frac{105 \cdot 150}{530} = 30$$

16.1.1.1 c^2 - hi kvadrat

$$\chi^2 = \sum \frac{\left(f_{gk} - f_{gk}^0\right)^2}{f_{gk}^0} \qquad 0 \le \chi^2 \le \infty$$

$$\chi^2 = \frac{(130-108)^2}{108} + \frac{(162-166)^2}{166} + \frac{(90-108)^2}{108} + \frac{(13-30)^2}{30} + \frac{(52-45)^2}{45} + \frac{(40-30)^2}{30} +$$

$$+ \frac{(7-12)^2}{12} + \frac{(16-19)^2}{19} + \frac{(20-12)^2}{12} = 4{,}48 + 0{,}10 + 3{,}00 + 9{,}63 + 1{,}09 + 3{,}33 + 2{,}08 + 0{,}47 + 3{,}33$$

$$\chi^2 = 29{,}51$$

16.1.1.2 C - Pearsonov koeficient kontingence

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{29{,}51}{29{,}51 + 530}} = \sqrt{0{,}0527} = 0{,}23 \qquad 0 \le C \le 1$$

$$C_M = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2}{3}} = 0{,}816 \quad \text{- maksimalna vrednost C za dani primer}$$

k je št. vrstic ali št. stolpcev v tabeli, tisto od teh dveh, ki je manjše

16.1.1.3 r_A - količnik korelacije atributov

$$r_A = \frac{C}{C_M} = \frac{0{,}23}{0{,}816} = 0{,}28 \qquad 0 \le r_A \le 1$$

Če je količnik bližje 1, je jakost povezanosti večja.

16.1.2 Asociacija

16.1.2.1 Q - Yule-ov koeficient asociacije

a	b	a+b
c	d	c+d
a+c	b+d	N

	A	B	
oprostilna	130	162	292
kaznovalna	20	68	88
	150	230	380

Tabela 25

Uporabljamo le v primeru, ko imamo tabelo 2 x 2. Je manj natančen način.

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} = \frac{130 \cdot 68 - 20 \cdot 162}{130 \cdot 68 + 20 \cdot 162} = 0{,}46 \qquad -1 \le Q \le 1$$

Če je absolutna vrednost Q blizu 0, je stopnja povezanosti majhna, če pa je bližje 1 je velika.

16.2 Numerični znaki

16.2.1 Kovarianca

To je merilo jakosti povezanosti numeričnih znakov

$$C_{xy} = \frac{1}{N} \sum (x - \bar{x}) \cdot (y - \bar{y})$$

$$C_{xy} = \frac{10}{10} = 1$$

$$-\infty \le C_{xy} \le \infty$$

	x - statistika	y - matematika	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	8	7	0,5	0	0	0,25	0
B	6	6	-1,5	-1	1,5	2,25	1
C	10	8	2,5	1	2,5	6,25	1
D	6	6	-1,5	-1	1,5	2,25	1
E	7	6	-0,5	-1	0,5	0,25	1
F	9	9	1,5	2	3	2,25	4
G	8	6	0,5	-1	-0,5	0,25	1
H	8	9	0,5	2	1	0,25	4
I	6	7	-1,5	0	0	2,25	0
J	7	6	-0,5	-1	0,5	0,25	1
Σ	75	70	0	0	10	16,5	14
	7,5	7					

Tabela 26

	x - delovna doba	y - % slabih izdelkov	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	2	11	-6	5	-30	36	25
B	7	7	-1	1	-1	1	1
C	12	5	4	-1	-4	16	1
D	3	10	-5	4	-20	25	16
E	10	6	2	0	0	4	0
F	20	2	12	-4	-48	144	16
G	10	2	2	-4	-8	4	16
H	4	6	-4	0	0	16	0
I	1	7	-7	1	-7	49	1
J	11	4	3	-2	-6	9	4
Σ	80	60	0	0	-124	304	80
	8	6					

Tabela 27

$$C_{xy} = \frac{1}{N}\sum(x-\bar{x})\cdot(y-\bar{y}) = -\frac{124}{10} = -12{,}4$$

16.2.2 Korelacijski količnik

$$r_{xy} = \frac{C_{xy}}{s_x \cdot s_y} \qquad C_{xy} = \frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{N} \qquad s_x^2 = \frac{\sum (x-\bar{x})^2}{N} \qquad s_y^2 = \frac{\sum (y-\bar{y})^2}{N}$$

$$-1 \le r_{xy} \le 1 \qquad s_x = \frac{\sqrt{\sum (x-\bar{x})^2}}{\sqrt{N}} \qquad s_y = \frac{\sqrt{\sum (y-\bar{y})^2}}{\sqrt{N}}$$

$$r_{xy} = \frac{\frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{N}}{\frac{\sqrt{\sum (x-\bar{x})^2}}{\sqrt{N}} \cdot \frac{\sqrt{\sum (y-\bar{y})^2}}{\sqrt{N}}} = \frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2} \cdot \sqrt{\sum (y-\bar{y})^2}}$$

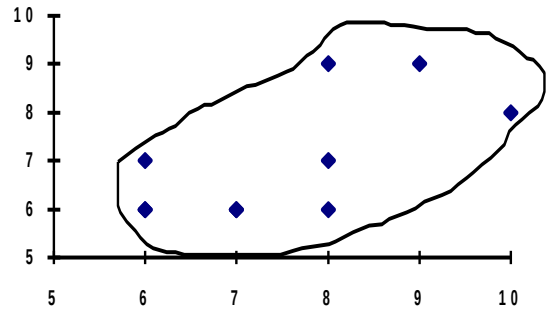
Za 1. tabelo:

$$r_{xy} = \frac{10}{\sqrt{16{,}5} \cdot \sqrt{14}} = \frac{10}{4{,}06 \cdot 3{,}74} = 0{,}66$$

Za 2. tabelo:

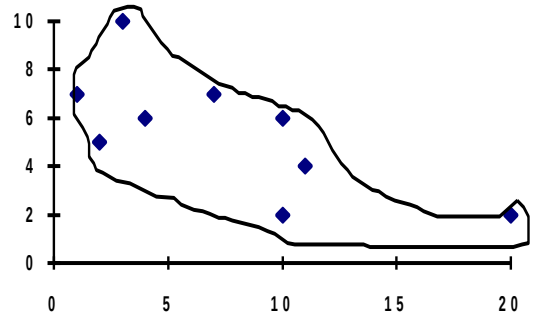
$$r_{xy} = \frac{-124}{\sqrt{304} \cdot \sqrt{80}} = \frac{-124}{17{,}44 \cdot 8{,}94} = -0{,}80$$

Jakost povezanosti je v obeh primerih relativno visoka, le da je v 1. primeru pozitivna (oba znaka se odklanjata v isto smer), v drugem primeru pa negativna (znaka se odklanjata v različni smeri).



Graf 1

Če je r > 0, ležijo vse točke v okviru, ki leži v enako smer kot premica s pozitivnim smernim koeficientom. Če pa je r = 1, ležijo vse točke na premici.



Graf 2

Če je r < 0, ležijo vse točke v okviru, ki leži v enako smer kot premica s negativnim smernim koeficientom. Če pa je r = -1, ležijo vse točke na premici.

16.2.3 Korelacija ranga

Uporablja se:
1. Kadar želimo hitreje priti do rezultata in imamo numerične znake.
2. Kadar znaki nimajo izmerljivih ampak le primerljive vrednosti.

$$r_R = 1 - \frac{6 \cdot \sum d^2}{N(N^2 - 1)} \qquad -1 \le r_R \le 1$$

$d = R_1 - R_2$; R_1, R_2 – ranga 1 in dva, od obeh vrst znakov

d - Spearmanov količnik korelacije ranga

	x - delovna doba	y - % slabih izdelkov	$R_1=R_{let}$	$R_2 = R_{slabi}$	$d=R_1-R_2$	d^2
A	2	11	2	10	-8	64
B	7	7	5	7,5	-2,5	6,25
C	12	5	9	4	5	25
D	3	10	3	9	-6	36
E	10	6	6,5	5,5	1	1
F	20	2	10	1,5	8,5	72,25
G	10	2	6,5	1,5	5	25
H	4	6	4	5,5	-1,5	2,25
I	1	7	1	7,5	-6,5	42,25
J	11	4	8	3	5	25
Σ	80	60			0	299

Tabela 28

$$r_R = 1 - \frac{6 \cdot 299}{10\left(10^2 - 1\right)} = 1 - \frac{1794}{990} = -0{,}81$$

Lahko pa imamo podan pridelek za leta 1991-95 in sicer:

91	10
92	12
93	16
94	15
95	10

Tabela 29

Vemo pa tudi, da je bila količina padavin porazdeljena: 94>93>95>91>92

Zato moramo obe vrsti rangirati:

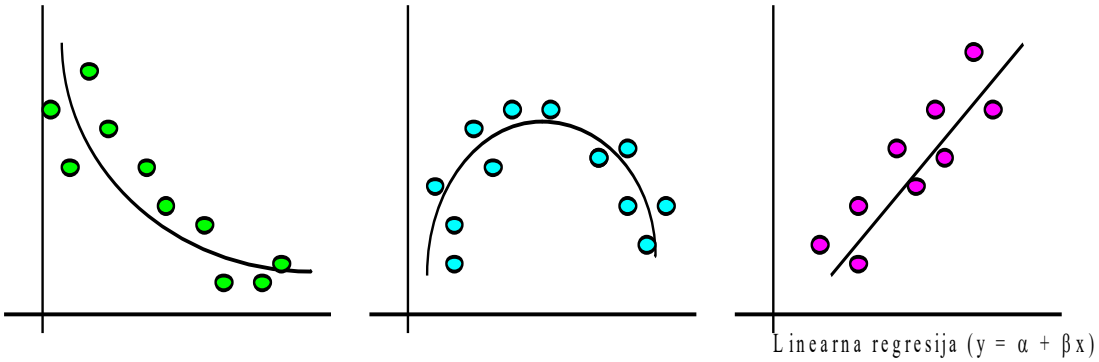
R_{pr}	R_{pad}	d	d^2
4,5	4	0,5	0,25
3	5	-2	4
1	2	-1	1
2	1	1	1
4,5	3	1,5	2,25
N=5			8,5

Tabela 30

$$r_R = 1 - \frac{6 \cdot \sum d^2}{N(N^2 - 1)} = 1 - \frac{6 \cdot 8{,}5}{5\left(5^2 - 1\right)} = 1 - \frac{51}{120} = 0{,}575$$

17 Regresija

Iščemo funkcijsko obliko povezanosti med odvisno spremenljivko y in neodvisno spremenljivko x.



Slika 2

17.1 Linearna regresija

$y = \alpha + \beta x + e$ ← enačba premice
e = slučajnostni odkloni - odkloni od najboljše premice, ki jo lahko prilagodimo danim podatkom
Pri regresijski premici velja, da so točke okoli premice porazdeljene normalno.

Predpostavke za odvisno spremanljivko x:

- Vsi y imajo za vse vrednosti x enako varianco (σ^2)
- Pričakovana vrednost E(y) leži na pravi regresijski premici
- Vrednosti y so med seboj neodvisne (pridelek v času t+1 ni odvisen od pridelka v času t)

Predpostavki za slučajnostni odklon:

- E(e) = 0
- Varianca od e je enaka σ^2

Naključni člen e je posledica:

- napak v merjenju
- nepravilne funkcijske povezave
- nevključitev vseh neodvisnih spremenljivk (x)
  (pridelek ni odvisen le od količine padavin, ampak še od drugih dejavnikov: gnojilo, struktura zemlje...)

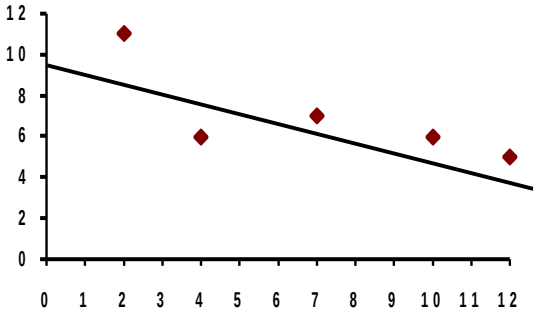
←Multipla regresija

	x - delovna doba	y - % slabih izdelkov	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$\hat{y}$
A	2	11	-5	4	-20	25	16	
B	7	7	0	0	0	0	0	
C	12	5	5	-2	-10	25	4	
E	10	6	3	-1	-3	9	1	
H	4	6	-3	-1	3	9	1	
Σ	35	35	0	0	-30	68	22	
	7	7						

Tabela 31

Premico lahko narišemo prostoročno ali pa analitično po metodi MNK (metoda najmanjših kvadratov).

Prostoročno:



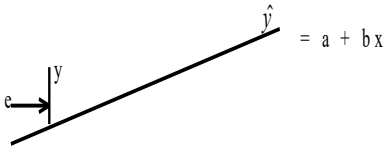
Graf 3

Pri **MNK** določimo enačbo premice:

y - dejanske vrednosti

$\hat{y}$ - ocenjene vrednosti

y - $\hat{y}$ = e



Slika 3

Premico moramo določiti tako, da bo **$\Sigma(y-\hat{y})^2$ = min → $\Sigma(y - a - bx)^2$ = min**

Zato odvajamo parcialno po a in po b in dobimo:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

Sedaj lahko izračunamo enačbo premice:

$$b = \frac{-30}{68} = -0{,}44$$

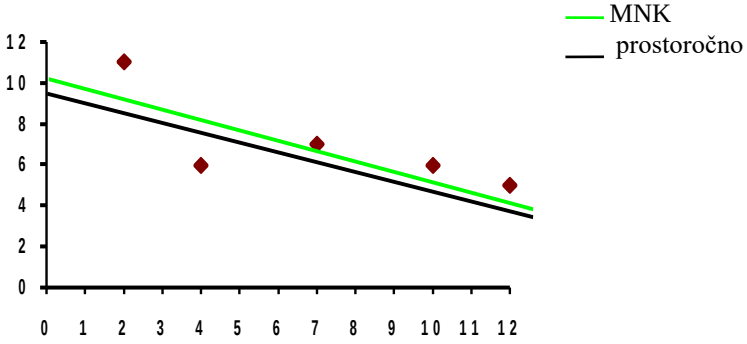
$$a = 7 - (-0{,}44 \cdot 7) = 7 + 3{,}08 = 10{,}08$$

$$\hat{y} = 10{,}08 - 0{,}44x$$

$$x_1 = 0 \qquad x_2 = 12$$

$$y_1 = 10{,}8 \qquad y_2 = 4{,}8$$

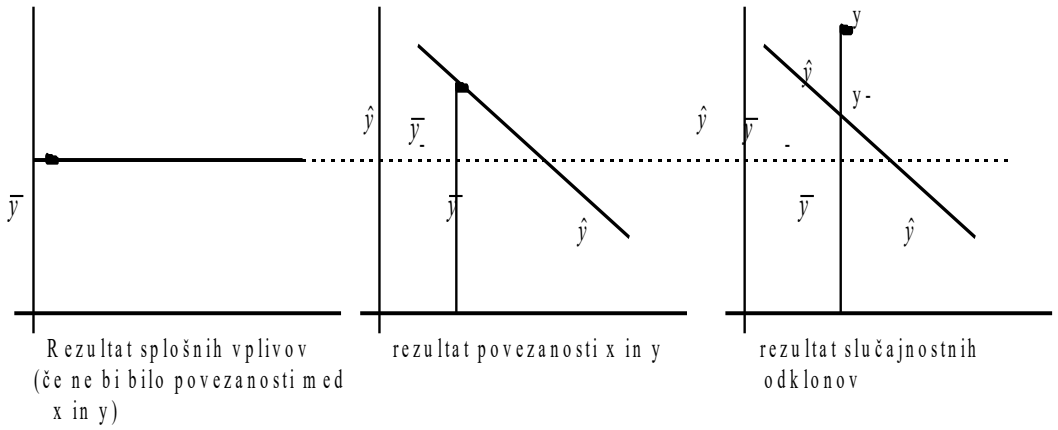
Sedaj, ko imamo enačbo premice, lahko narišemo graf:



Graf 4

17.2 Analiza variance

celotna varianca: $s_y^2 = \frac{1}{N}\sum(y-\bar{y})^2$



Slika 4

$$y = \bar{y} + (\hat{y} - \bar{y}) + (y - \hat{y})$$

$$s_{\hat{y}}^2 = \frac{1}{N}\sum(\hat{y}-\bar{y})^2$$ - pojasnjena varianca

$$s_e^2 = \frac{1}{N}\sum(y-\hat{y})^2$$ - nepojasnjena varianca

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

17.3 Determinacijski količnik (R^2)

Pove nam delež pojasnjenosti variabilnosti.

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = r^2 \qquad r = \sqrt{R^2}$$

$$0 \le R^2 \le 1 \qquad -1 \le r \le 1 \qquad -\infty \le b \le \infty$$

Za na imer:

$$s_y^2 = \frac{1}{5}\cdot 22 = 4{,}4$$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{-30}{\sqrt{68}\sqrt{22}} = -0{,}78$$

$$R^2 = r^2 = (-0{,}78)^2 = 0{,}60$$

$$s_{\hat{y}}^2 = R^2 \cdot s_y^2 = 0{,}60 \cdot 4{,}4 = 2{,}64$$

$$s_e^2 = s_y^2 - s_{\hat{y}}^2 = 4{,}4 - 2{,}64 = 1{,}76$$

17.4 Standardna napaka ocene (SNO, s_e)

$$s_e = \sqrt{s_e^2}$$

$$s_e = \sqrt{1{,}76} = 1{,}33$$

Pove nam, da če katerikoli točki odštejemo in prištejemo s_e, dobimo pas, v katerem leži 68,27% vseh enot okrog premice. Če odštejemo in prištejemo 2 s_e, je v tem pasu 95,45% vseh enot

3 s_e, je v tem pasu 99,71% vseh enot.

17.5 Linearni časovni trend

Gre za odvisno spremenljivko, ki se spreminja v času. Po metodi MNK je:

$$a = \bar{y} - b\bar{x} \Rightarrow a = \bar{y} - b\bar{t}$$

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \Rightarrow b = \frac{\sum (t - \bar{t})(y - \bar{y})}{\sum (t - \bar{t})^2}$$

leto	št zapornikov v 1000 [y]	t	$(y-\bar{y})$	$(t-\bar{t})$	$(y-\bar{y})(t-\bar{t})$	$(t-\bar{t})^2$
1987	5	0	-3	-2	6	4
1988	6	1	-2	-1	2	1
1989	9	2	1	0	0	0
1990	9	3	1	1	1	1
1991	11	4	3	2	6	4
Σ	40	10	0	0	15	10

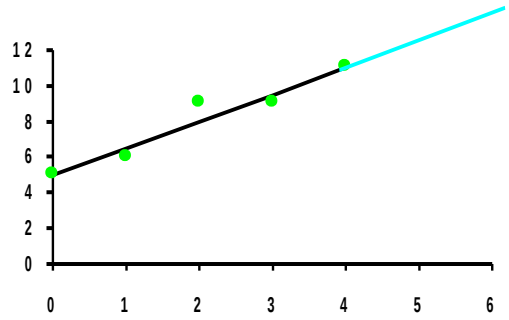
Tabela 32

$$\bar{y} = \frac{\sum y}{N} = 8 \qquad \bar{t} = \frac{\sum t}{N} = 2$$

$$\hat{y} = a + bt$$

$$b = \frac{15}{10} = 1{,}5 \qquad a = 8 - 1{,}5 \cdot 2 = 8 - 3 = 5$$

$$\hat{y} = 5 + 1{,}5t$$



Graf 5

Ekstrapolacija trenda - ocenimo bodoče gibanje pojavov, ko podaljšamo premico.

Leto 1992:
t = 5
y = 5 + 1,5 * 5 = 12,5

17.6 Značaj zvez med znaki

	x - statistika	y - matematika	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	8	7	0	-1	0	0	1
B	10	9	2	1	2	4	1
C	6	7	-2	-1	2	4	1
D	8	8	0	0	0	0	0
E	8	9	0	1	0	0	1
S	40	40	0	0	4	8	4

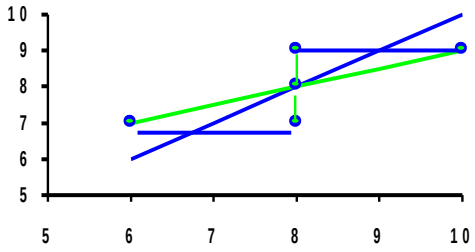
Tabela 33

$$\hat{y} = a_1 + b_1 x \qquad b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \qquad a_1 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{4}{8} = 0{,}5 \qquad a_1 = 8 - 0{,}5 \cdot 8 = 4 \qquad \hat{y} = 4 + 0{,}5x$$

$$\hat{x} = a_2 + b_2 y \qquad b_2 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} \qquad a_2 = \bar{x} - b_2 \bar{y}$$

$$b_2 = \frac{4}{4} = 1 \qquad a_2 = 8 - 1 \cdot 8 = 0 \qquad \hat{x} = y$$



Graf 6

17.7 Kako izbrati regresijsko zvezo

Mat → Stat ali Stat → Mat : v tem primeru je vseeno
Ni pa vseeno:
pridelek ← količine padavin
potrošnja ← plače
število nesreč ← gostota prometa

17.7.1 Kavzalna (vzročna) zveza

Plače → Cene
Cene → Plače

17.7.2 Nepristne in posredne zveze

A - obiskuje gledališče
B - obiskuje gledališče
Lahko bi preučevali katere predstave hodita gledat. Tu bi se lahko zmotili, saj bi bil lahko vzrok tudi drugje (kdaj imata prosti čas, vreme...)

18 Zanesljivost ocen

Populacija - vse enote, ki jih preučujemo
Vzorec - del populacije, ki ga preučujemo, da dobimo oceno vrednosti za populacijo
Veliki vzorci - za različne parametre so različni kriteriji (Mx - 30 enot, σ - 100 enot)
Majhni vzorci - verjetnost ocene je slabša. Pri majhnih vzorcih je treba za vsak parameter uporabit različno vrsto testa.
Pri velikih vzorcih je dober enak test za vse parametre.Pri velikih vzorcih izbiramo ven člene (vzorce) in računamo pri vsakem vzorcu aritmetično sredino ŕ $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \ldots, \bar{x}_n$ ß vzorčna populacija (enote so ocene iz vzorca, znaki so povprečja, proporci, mere variabilnosti)