Fakulteta za organizacijske vede

Univerza v Mariboru

Statistika in verjetnost

Vaje za šolsko leto 2003/2004

Formule, ki smo jih spoznali na vajah

Kazalo

Frekvenčne porazdelitve

Diskretna frekvenčna porazdelitev

Med zbranimi podatki o diskretni spremenljivki X (zavzema samo cele vrednosti), preštejemo kolikokrat se pojavi posamezna vrednost spremenljivke X (npr. med ocenami 100 študentov, preštejemo koliko študentov je pisalo oceno 1, oceno 2, ..., oceno 10). Številu posamezneih vrednosti spremenljivke X pravimo frekvenca vrednosti spremenljivke (npr. če med 100 študenti preštejemo vse tiste, ki so pisali 6 in ugotovimo, da je takih študentov 40), potem je frekvenca šestice pri spremenljivki ocena enaka 40.

spremenljivka (X)	frekvenca (f)	relativna frekvenca (p)	kumulativna frekvenca (F)	relativna kumulativna frekvenca (H)
x_1	f_1	p_1	F_1	H_1
x_2	f_2	p_2	F_2	H_2
...	...	...	...	...
x_k	f_k	p_k	F_k	H_k
...	...	...	...	...
x_n	f_n	p_n	F_n	H_n
---	$\sum f = N$	$\sum p = 1$	---	---

Zvezna frekvenčna porazdelitev

Pri zvezni spremenljivki X, vrednosti spremenljivke lahko ležijo kjerkoli na določenem intervalu. Zato moramo zvezno spremenljivki najprej razdeliti na nekaj manjših podintervalov – razredov. Frekvence spremenljivke X potem dobimo tako, da preštejemo koliko enot pade v posamezen podinterval.

$$p = \frac{f}{N};$$

$$F_1 = f_1; F_i = F_{i-1} + f_i$$

$$H = \frac{F}{N}$$

možnost razdelitve zvezne spremenljivke v podintervale je s pomočjo **Sturgessovega pravila** →

→ število razredov $n = 1 + 3{,}3 \log N$, pri čemer je N število enot, ki jih proučujemo

Numerične mere

Ranžirne vrste

Rang

negrupirani podatki	grupirani podatki
$\frac{x-x_-}{R_x-R_-}=\frac{x_+-x_-}{R_+-R_-}$ → $R_x=R_-+\frac{x-x_-}{x_+-x_-}$	$R_x=\sum_{i=1}^{k-1} f_i+f_k\frac{x-r_{k-1}}{\Delta x_k}$

Kvantilni rang

$P_x=\frac{R_x-0{,}5}{N}$ → kvantilni rang se računa po istem postopku za negrupirane in grupirane podatke

Kvantili

Dan je kvantilni rang P (0<=P<=1)

negrupirani podatki	grupirani podatki
$x_P=x_{(n+1)*P}$	$x_P=r_{k-1}+\frac{NP-\sum_{i=1}^{k-1} f_i}{f_k}\Delta x_k$

Kvartili

So vrednosti, ki razdelijo proučevane podatke popualcije ali vzorca na 4 enake dela → Q1, Q2, Q3 … torej tem kvartilom pripadajo kvantilni rangi P_{Q1}=0.25, P_{Q2}=0.50, P_{Q3}=0.75,

Mere srednje vrednosti

Aritmetična sredina – točka ravnotežja

negrupirani podatki	grupirani podatki
$\bar{x}=\frac{\sum_i^n x_i}{n}$	$\bar{x}=\frac{\sum_i^K fx_{Si}}{n}$

Mediana

negrupirani podatki	grupirani podatki
$Me=x_{\frac{N+1}{2}}$	$Me=r_{k-1}+\frac{\frac{N}{2}-\sum_1^{k-1} f_i}{f_k}\Delta x_k$;

Modus

negrupirani podatki	grupirani podatki
→ najpogostejša vrednost, ki se pojavi v podatkih	$Mo=r_{k-1}+\frac{f_k-f_{k-1}}{(f_k-f_{k-1})+(f_k-f_{k+1})}\Delta x_k$

Harmonična sredina

negrupirani podatki	grupirani podatki
$x_H = \dfrac{N}{\sum_{i=1}^{N} \dfrac{1}{x_i}}$	$x_H = \dfrac{\sum_{i=1}^{K} f_i}{\sum_{i=1}^{K} \dfrac{f_i}{x_{Si}}}$

Geometrična sredina sredina

negrupirani podatki	grupirani podatki
$x_G = \sqrt[N]{x_1 x_2 \cdot \ldots \cdot x_N}$	$x_G = \sqrt[N]{x_{S1}^{f_1} x_{S2}^{f_2} \cdot \ldots \cdot x_{Sk}^{f_K}}$

Mere razpršenosti podatkov

Variacijski razmik

$R = x_{\max} - x_{\min}$

Varianca

Najpogostejša mera razpršenosti podatkov … računa povprečni odklon kvadriranih vrednosti od aritmetične sredine.

negrupirani podatki	grupirani podatki
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \dfrac{\sum_{i=1}^{N} x_i^2 - N\mu^2}{N}$ (populacija)	$\sigma^2 = \dfrac{\sum_{i=1}^{K} f(x_{Si} - \mu)^2}{N} = \dfrac{\sum_{i=1}^{K} fx_{Si}^2 - N\mu^2}{N}$ (populacija)
$s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \dfrac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$ (vzorec)	$s^2 = \dfrac{\sum_{i=1}^{K} f(x_i - \overline{x})^2}{n-1} = \dfrac{\sum_{i=1}^{K} fx_i^2 - n\overline{x}^2}{n-1}$ (vzorec)

Standardni odklon

$\sigma = \sqrt{\sigma^2}$ (populacija) $s = \sqrt{s^2}$ (vzorec)

Koeficient variacije

$k_v = \dfrac{s}{\overline{x}}$

N ali n=število podatkov, ki jih proučujemo (če delamo na podlagi vzorčnih podatkov, potem število podatkov označimo z n)

K=število razredov v frekvenčni porazdelitvi

k=razred na katerega se nanaša izračun

f_i=frekvenca i-tega razreda; f_k=frekvenca k-tega razreda

x_{Si}=sredina razreda

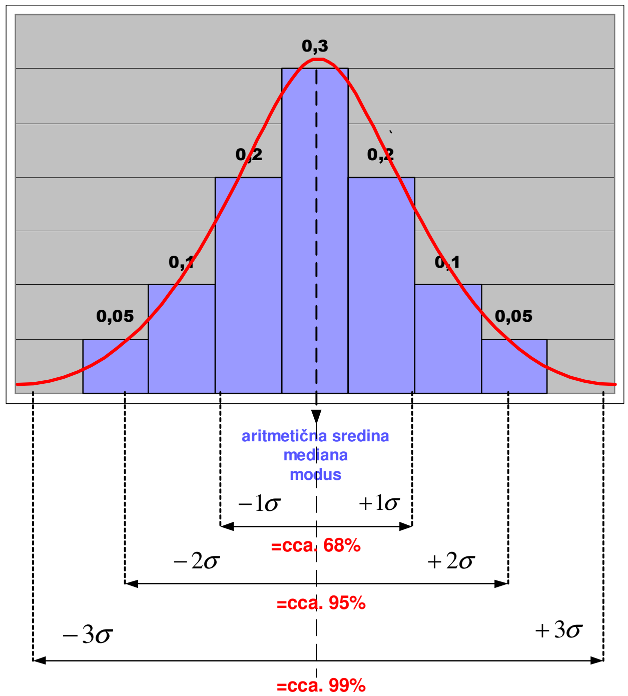
r_{k-1}=spodnja meja k-tega razreda

Δx_k=širina k-tega razreda

indeksi k-1 ali k+1 pomenijo vrednosti pred (k-1) oz. za (k+1) k-tim razredom

Opisovanje normalne frekvenčne porazdelitve

- ima en sam vrh, in sicer nad aritmetično sredino=mediana=modus
- glede na srednjo vrednost je popolnoma simetrična
- repi normalne porazdelitve grejo levo in desno od centra se asimptotično bližajo vodoravni osi



Mere oblik frekvenčnih porazdelitev

Asimetričnost

negrupirani podatki	grupirani podatki
$g_1 = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^3}}$	$g_1 = \dfrac{\frac{1}{N}\sum_{i=1}^{K}f_i(x_{Si}-\bar{x})^3}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{K}f_i(x_i-\bar{x})^2\right)^3}}$

Opomba: spodnji del formule predstavlja varianco!

Sploščenost

negrupirani podatki	grupirani podatki
$g_2 = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right)^2}}$	$g_2 = \dfrac{\frac{1}{N}\sum_{i=1}^{K}f_i(x_{Si}-\bar{x})^4}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{K}f_i(x_{Si}-\bar{x})^2\right)^2}}$

Opomba: spodnji del formule predstavlja varianco!

Fisherjev koeficient sploščenosti: $\gamma = g_2 - 3$

Opomba: za normalno porazdelitev je $g_2=3$!

Če se koeficinta asimetričnosti in sploščenosti bistveno razlikujeta od NIČ, se frekvenčna porazdelitev bistveno razlikuje od normalne!!!

Diskretne verjetnostne porazdelitve

Porazdelitveni zakon diskretne slučajne spremenljivke – verjetnostna porazdelitev

ξ	x_1	x_2	…	x_k	…	x_n	---
$P(\xi=x)$	p_1	p_2	…	p_k	…	p_n	$\sum_{i=1}^{n} p_i = 1$

Porazdelitvena funkcija diskretne slučajne spremenljivke	*Aritmetična sredina – matematično upanje diskretne slučajne spremenljivke*
$F(x) = \sum_{x_i \le x} p_i$	$\mu = M(\xi) = \sum_{i=1}^{n} x_i p_i$

Varianca diskretne slučajne spremenljivke
$\sigma^2 = V(\xi) = M\left(\xi - M(\xi)\right)^2 \Rightarrow \sigma^2 = V(\xi) = M(\xi^2) - M^2(\xi)$; pri čemer je $M(\xi^2) = \sum_{i=1}^{n} x_i^2 p_i$
Standardni odklon diskretne slučajne spremenljivke
$\sigma = \sqrt{V(\xi)}$

Posebne diskretne slučajne spremenljivke

Binomska slučajna spremenljivka

Porazdelitveni zakon	Matematično upanje	Varianca in standardni odklon
$P(\xi = k) = \binom{n}{k} p^k q^{n-k}$	$M(\xi)=np$	$V(\xi) = npq; \quad \sigma = \sqrt{V} = \sqrt{npq}$

Poissonova porazdelitev

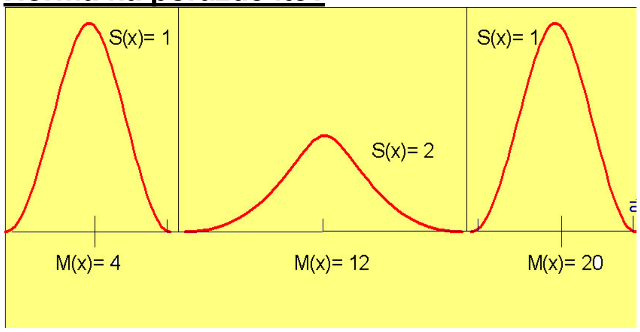
Porazdelitveni zakon	Matematično upanje	Varianca in standardni odklon
$P(\xi = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ rekurzivna formula: $P(\xi = k) = P(\xi = k-1)\frac{\lambda}{k}$	$M(\xi) = \lambda$	$V = \lambda; \quad \sigma = \sqrt{V} = \sqrt{\lambda}$

Negativna binomska porazdelitev

Porazdelitveni zakon	Matematično upanje	Varianca in standardni odklon
$P(\xi = k) = \binom{k-1}{k-r} p^r q^{k-r} = \binom{k-1}{r-1} p^r q^{k-r}$	$M(\xi) = \frac{r}{p}$	$V(\xi) = \frac{rq}{p^2}; \quad \sigma(\xi) = \sqrt{V} = \sqrt{\frac{rq}{p^2}}$

Zvezne verjetnostne porazdelitve

Normalna porazdelitev



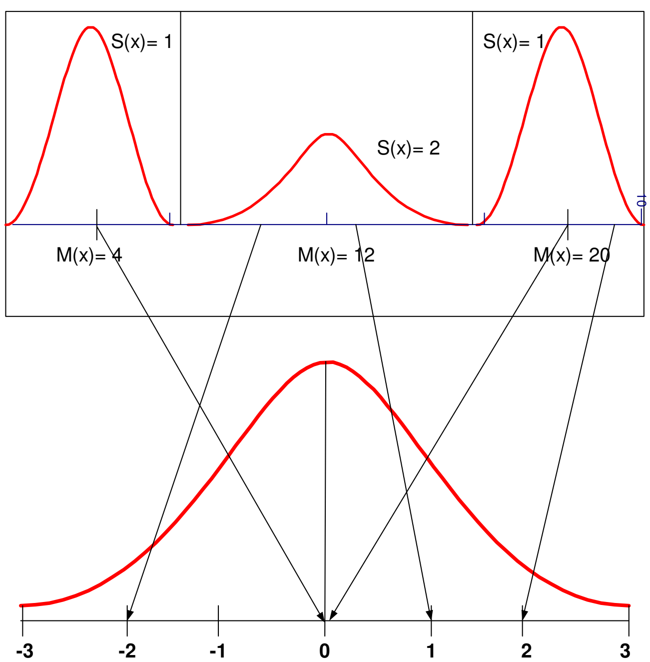
Značilnosti normalne porazdelitve:

- en vrh (M=Me=Mo)
- popolnoma simetrična
- repi krivulje se ne dotikajo osi x

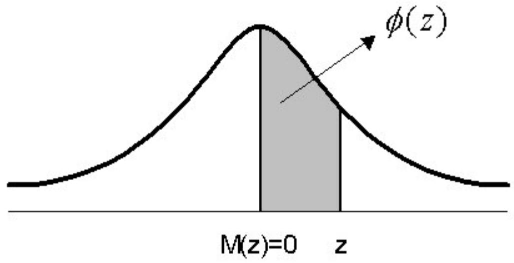
gostota verjetnosti → $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

Standardizirana normalna porazdelitev

$$z = \frac{x-\mu}{\sigma}$$



Ploščina pod krivuljo standardizirane normalne porazdelitve



Ploščina od M(z)=0 do z je $\phi(z)$.

Na primer: če je z=0,72 → $\phi(z)$ =0,2642

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Aproksimacija binomske v normalno

$z = \dfrac{x-\mu}{\sigma}$	pri čemer je:	$\mu = np$ $\sigma = \sqrt{npq}$

Ocenjevanje parametrov populacije

Točkasta ocena

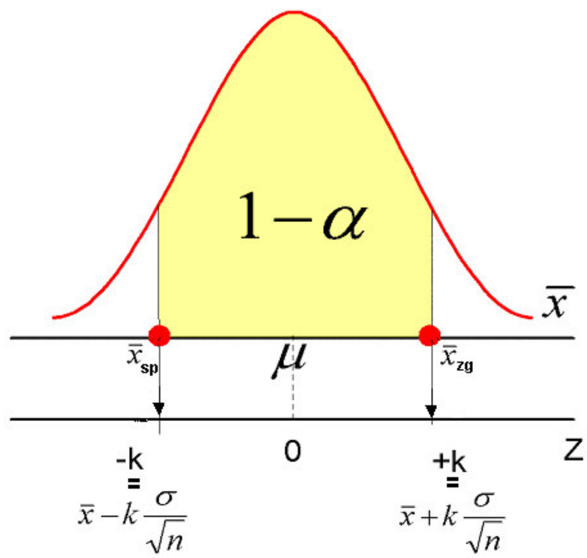
Izračunamo jo po formuli za parameter ... npr. za aritmetično sredino ...

Intervalska ocena

$$P(-k < z < k) = P(-k < \frac{\bar{x}-\mu}{\sigma}\sqrt{n} < k) = 1-\alpha \,.$$

Na podlagi tega pa lahko določimo **interval zaupanja za povprečje populacije**:

$$\bar{x} - k\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + k\frac{\sigma}{\sqrt{n}}$$



Velikost vzorca

Zahteva	→	$\lvert\bar{x}-\mu\rvert \le d$
Potrebna velikost vzorca	→	$n \ge \left(\dfrac{k\sigma}{d}\right)^2$

Testiranje hipotez o aritmetični sredini

Kaj je to … preverjanje predpostavk o parametrih populacije

1. Najprej postavimo ničelno in alternativno hipotezo

H_0: M=A

H_1: M $\neq$ A

2. Na podlagi stopnje zaupanja $1-\alpha$, oziroma stopnje tveganja (tudi napake 1. vrste) α določimo interval sprejemanja hipoteze → **-k** $\leq \frac{\bar{x}-A}{\sigma}\sqrt{n} \leq$ **k**

$$1-\alpha = 0{,}90 \Rightarrow \alpha = 0{,}10 \Rightarrow k = 1{,}64$$

$$1-\alpha = 0{,}95 \Rightarrow \alpha = 0{,}05 \Rightarrow k = 1{,}96$$

$$1-\alpha = 0{,}99 \Rightarrow \alpha = 0{,}01 \Rightarrow k = 2{,}58$$

3. S testom hipoteze **z** za preverjanje aritmetične sredine na podlagi vzorčnih podatkov izračunamo vrednost testa hipoteze:

$$z = \frac{\bar{x}-A}{\sigma}\sqrt{n}$$

4. Pogledamo ali izračunan test hipoteze **z** pade znotraj intervala sprejemanja hipoteze (-k,+k) → s stopnjo tveganja α **sprejememo sklep o sprejetju alternativne hipoteze** ("ničelno hipotezo sprejememo"), ali pade izven intervala sprejemanja hipoteze (-k,+k) → **ničelno hipotezo zavrnemo** pri stopnji tveganja α .

Regresija in korelacija

Enostavna linearna regresija in korelacija

X	Y	X^2	Y^2	XY
x_1	y_1	x_1^2	y_1^2	x_1y_1
x_2	y_2	x_2^2	y_2^2	x_2y_2
...	...	...	...	...
x_n	y_n	x_n^2	y_n^2	x_ny_n
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$

Najprej izračunamo delne vsote:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}$$

Na podlagi tako izračunanih delnih vsot najprej izračunamo oceno parametra β :

$b = \dfrac{S_{xy}}{S_{xx}}$, in nato še oceno parametra α : $\overline{y} = a + b\overline{x} \rightarrow a = \overline{y} - b\overline{x}$

Skupna ali začetna varianca: $s_Y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 = \dfrac{S_{yy}}{n-1}$

Nepojasnjena varianca: $s_e^2 = \dfrac{1}{n-2}\sum_{i=1}^{n} (y_i - (a + bx_i))^2 = \dfrac{1}{n-2}\left(S_{yy} - bS_{xy}\right)$

Pojasnjena varianca: $s_{xy}^2 = s_y^2 - s_e^2$

Determinacijski koeficient: $D = \dfrac{s_y^2 - s_e^2}{s_y^2} = 1 - \dfrac{s_e^2}{s_y^2}$

Korelacijski koeficient: $r = \dfrac{S_{xy}}{\sqrt{S_{yy}S_{xx}}}$ ... velja tudi: $D \approx r^2$

Koeficient korelacije ranžirnih vrst

... ali Spearmanov koeficient korelacije

$$r_S = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (w_i - v_i)^2$$

Časovne vrste

Metoda drsečih sredin

perioda drsenja: **m** → $1 \leq m \leq n$ (n je število podatkov, ki jih proučujemo)

Začetno napoved delamo za obdobje m+1:

$$y_{m+1} = \frac{y_1 + y_2 + \ldots + y_m}{m}$$

Nadaljnje napovedi pa lahko računamo po formuli:

$y'_{t+1} = y'_t + \frac{y_t - y_{t-m}}{m}$; če je m velik potem predpostavljamo veliko slučajnih vplivov

Metoda eksponentnega glajenja

faktor glajenja: **g** → $0 \leq g \leq 1$

Za začetno napoved vzamemo kar:

$$y'_1 = y_1$$

Nadaljnje napovedi pa računamo po formuli:

$y'_{t+1} = y'_t + g(y_t - y'_t)$; če je g majhen potem predpostavljamo veliko slučajnih vplivov

Model časovne vrste

Najprej obdobje opazovanja spremenimo v tehnični čas → pri katerem je njegova vsota enaka 0 …

T	Y	X	Y^2	YX	X2
T_1	y_1	x_1	y_1^2	y_1x_1	x_1^2
T_2	y_2	x_2	y_2^2	y_2x_2	x_2^2
…	…	…	…	…	…
T_n	y_n	x_n	y_n^2	y_nx_n	x_n^2
	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} x_i^2$

$a = \frac{\sum_{i=1}^{n} y_i}{n}$; $b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ → T = a + bx

Koeficienti dinamike časovnih vrst

Tempo rasti	Koeficient dinamike
$T_k = \frac{y_k - y_{k-1}}{y_{k-1}} \cdot 100$	$K_k = \frac{y_k}{y_{k-1}}$
Verižni indeks	**Bazni indeks**
$V_k = \frac{y_k}{y_{k-1}} \cdot 100$	$I_{k,0} = \frac{y_k}{y_0} \cdot 100$