KORRELATSIOON-ANALÜÜS

TEEMAD

- Korrelatsiooni mõiste
- Kovariatsioon
- Lineaarne korrelatsioonikordaja
- Korrelatsiooni statistilise olulisuse testimine
- Korrelatsioonimaatriks
- Lineaarse korrelatsioonikordaja puudused
- Astakkorrelatsioon ja Spearmani korrelatsioonikordaja

2 TUNNUSE VAHELINE SEOS

Kas keskmine töötuse kestvus sõltub piirkonnast?
Töötuse kestvus kvantitatiivne tunnus (intervallskaalas).
Piirkond kvalitatiivne tunnus (nimiskaalas).

2 piirkonda: keskväärtuste võrdlemise *t*-test

Töötuse kestvus kuudes	
Põhja-Eesti	Kirde-Eesti
25	58
13	40
16	37
…	…

Rohkem kui 2 piirkonda: dispersioonanalüüs

Töötuse kestvus kuudes				
Põhja-Eesti	Kirde-Eesti	Kesk-Eesti	Lääne-Eesti	Lõuna-Eesti
25	58	73	7	10
13	40	20	67	56
16	37	39	11	19
…	…	…	…	…

Kas toote meeldivus sõltub inimese soost?
Meeldivus kvalitatiivne tunnus (järjestusskaalas).
Sugu kvalitatiivne tunnus (nimiskaalas).

χ2-test

	mehed	naised
meeldib	17	14
neutraalne	29	45
ei meeldi	14	31

2 KVANTITATIIVSET TUNNUST

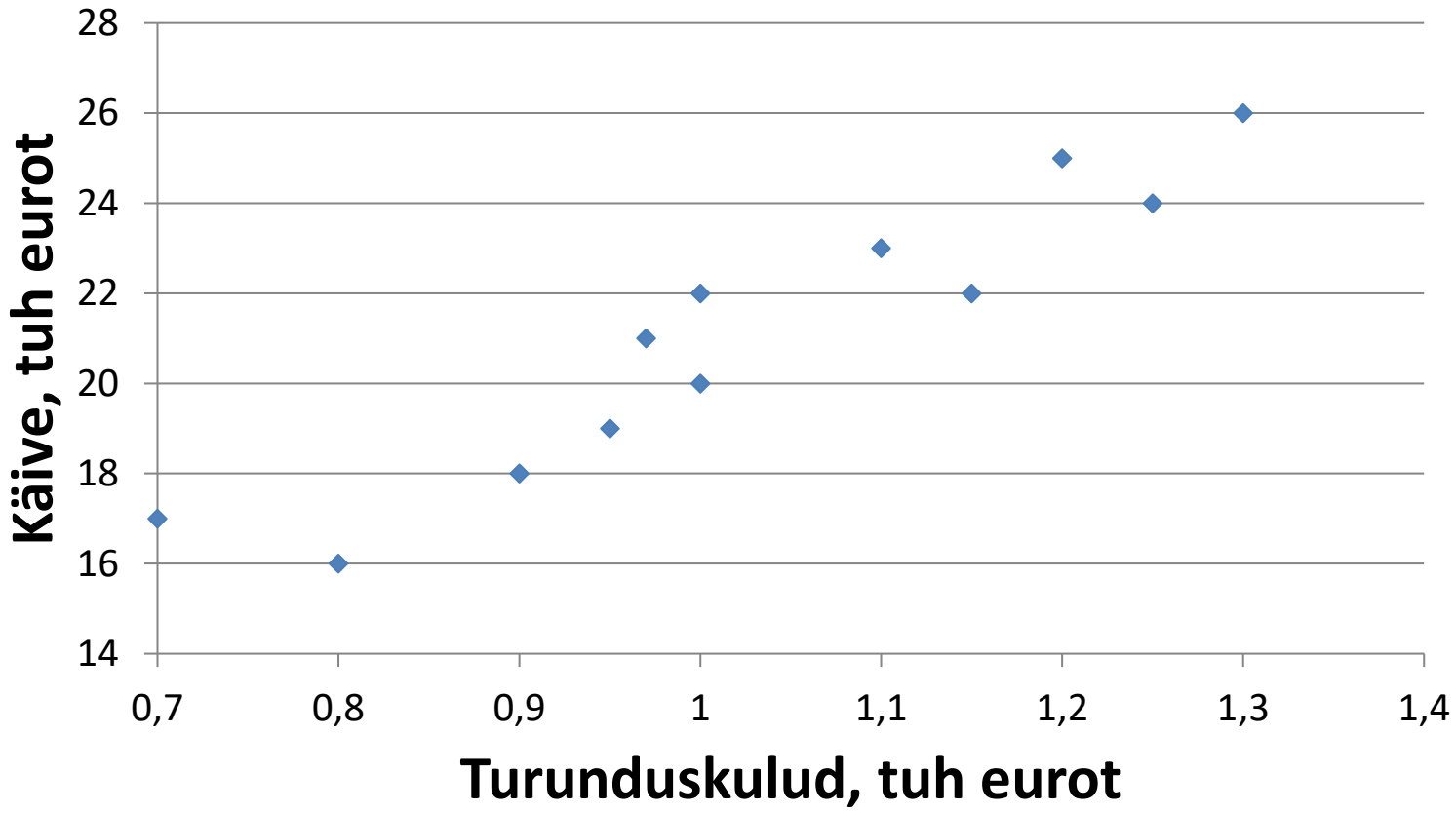
Kuidas hinnata seost kahe kvantitatiivse tunnuse (intervallskaalas) vahel?

Näiteks: kas on seos ettevõtte turunduskulude ja käibe vahel?
Mõlemad on kvantitatiivsed tunnused (intervallskaalas).

Kuu	Turunduskulud, tuh eurot	Käive, tuh eurot
jaan	1,0	20
veebr	1,1	23
märts	1,2	25
apr	0,9	18
...	...	...

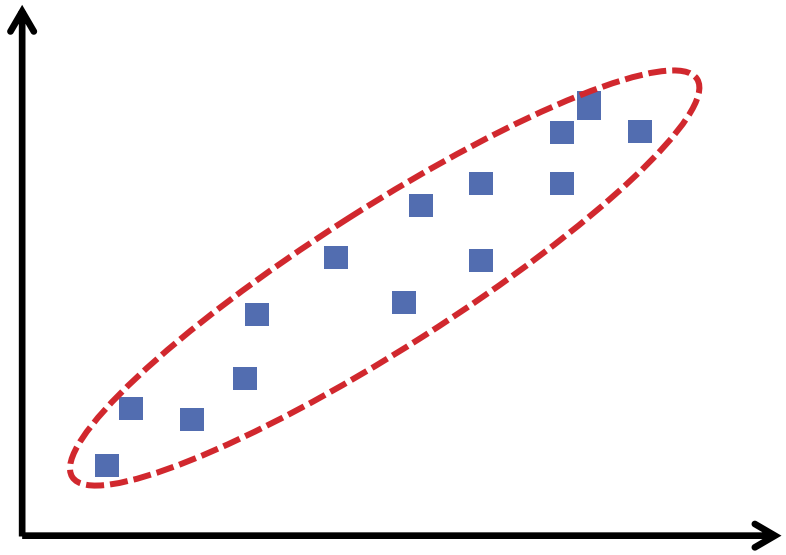
HAJUMISDIAGRAMM

scatter diagram

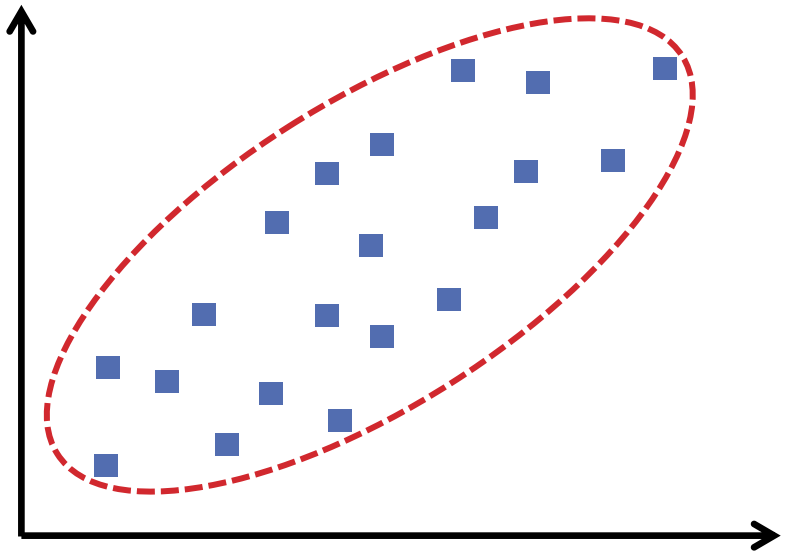


Turunduskulude kasvades keskmiselt kasvab ka käive.

POSITIIVNE KORRELATSIOON



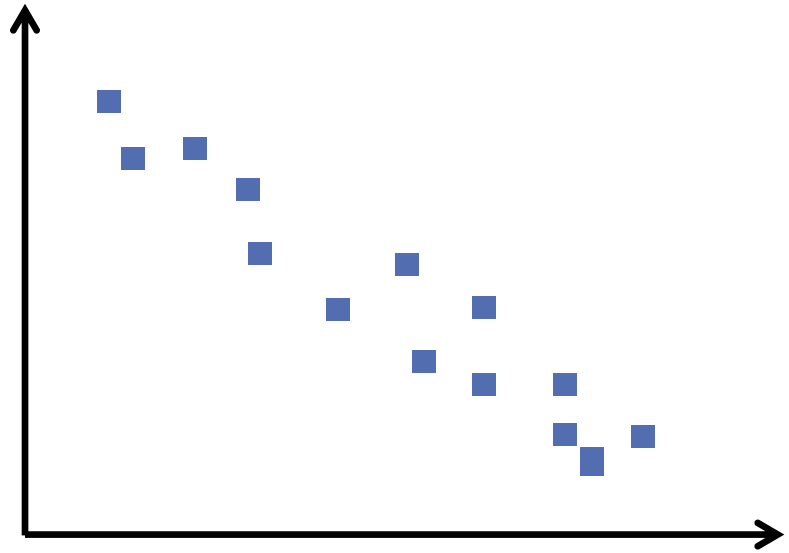
Tugev positiivne korrelatsioon



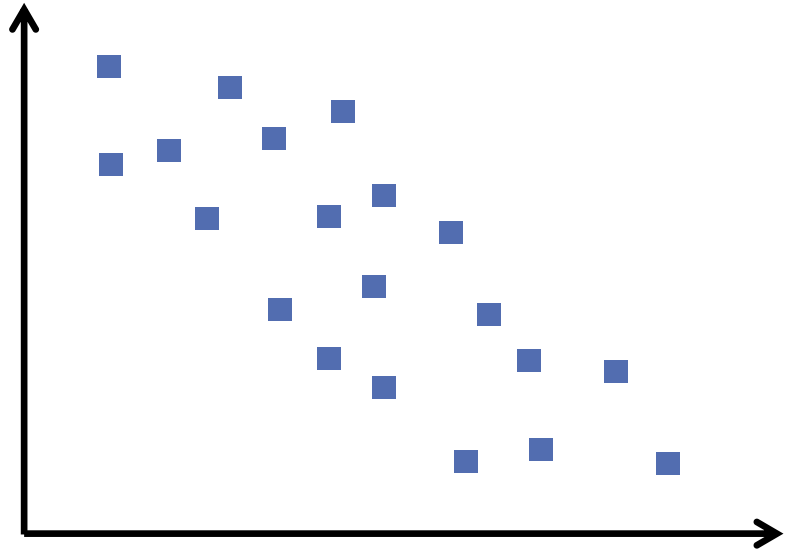
Nõrk positiivne korrelatsioon

Ühe suuruse kasvades teine suurus keskmiselt samuti kasvab.

NEGATIIVNE KORRELATSIOON



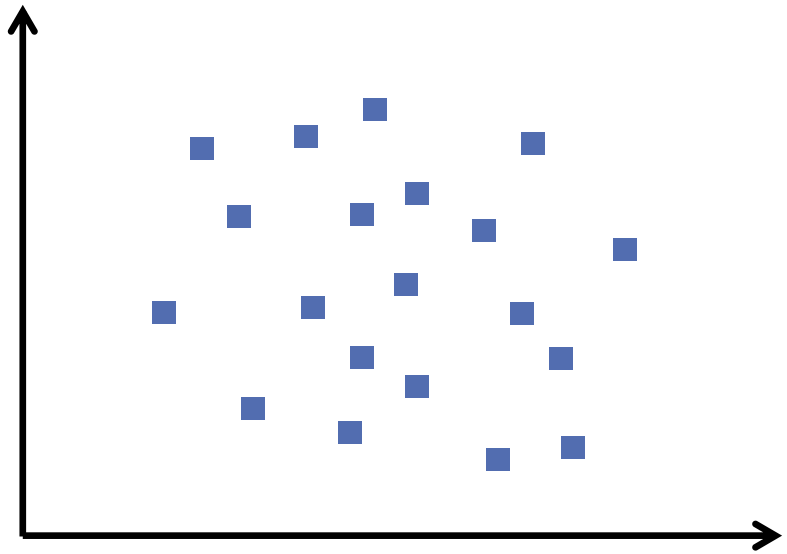
Tugev negatiivne korrelatsioon



Nõrk negatiivne korrelatsioon

Ühe suuruse kasvades teine suurus keskmiselt kahaneb.

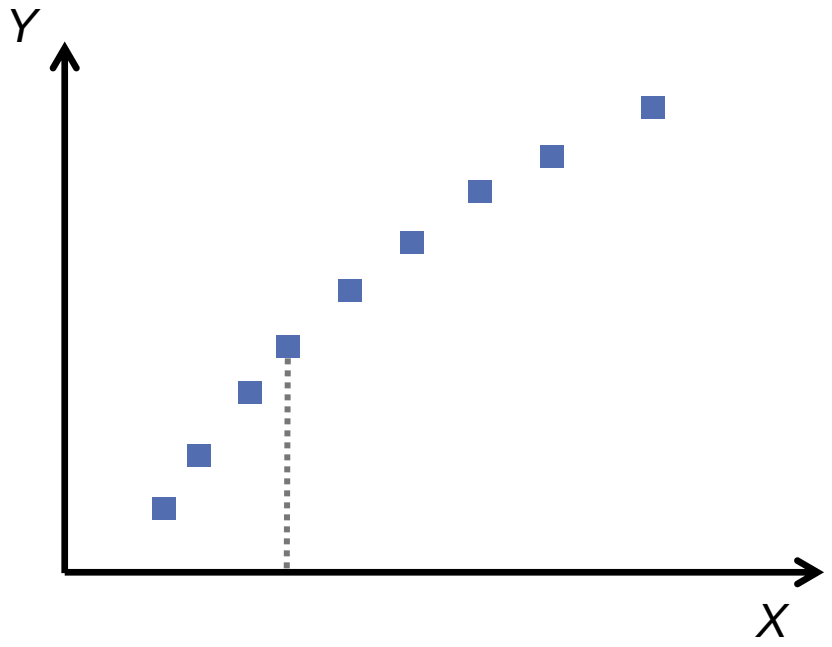
KORRELATSIOON PUUDUB



KORRELATSIOONI MÕISTE

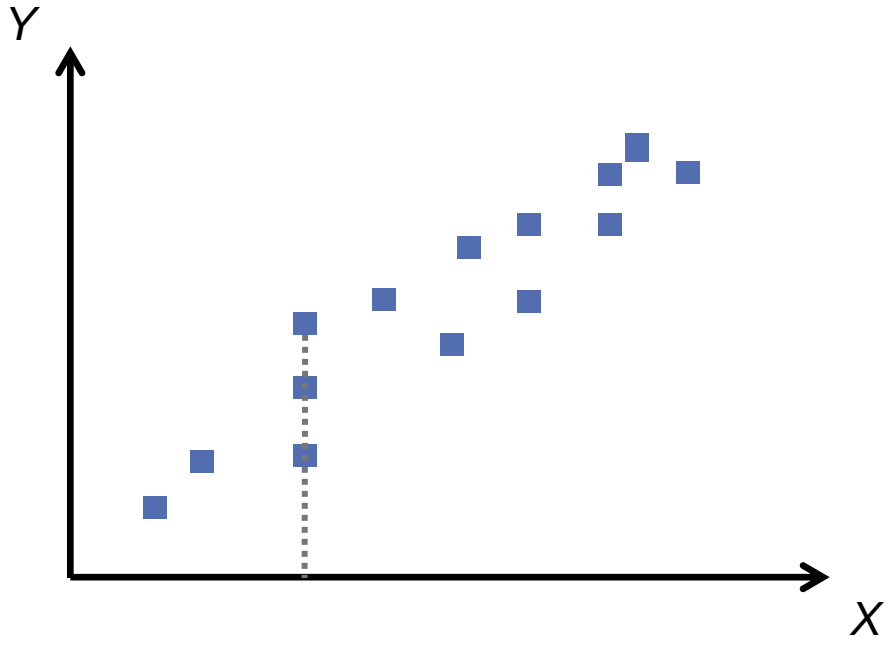
Korrelatsioon on juhuslike suuruste X ja Y vahel esinev statistiline seos.

FUNKTSIONAALNE SEOS



Ühele ja samale X väärtusele vastab üks ja ainult üks suuruse Y väärtus.

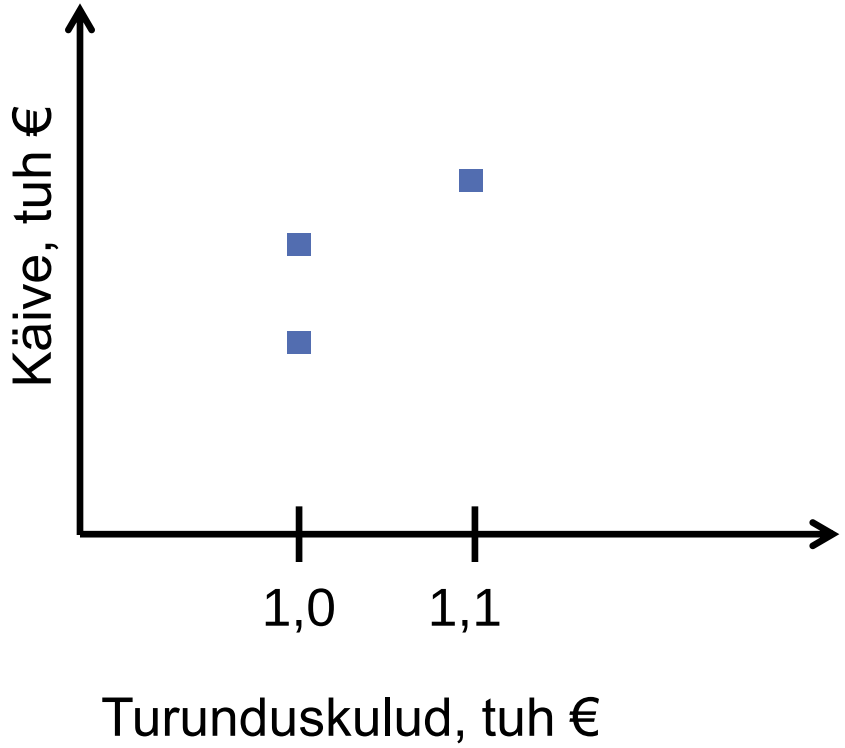
STATISTILINE SEOS



Ühele ja samale X väärtusele võib vastata mitu suuruse Y väärtust.

STATISTILINE SEOS

Kuu	Turundus-kulud, tuh €	Käive, tuh €
jaan	1,0	20
veebr	1,1	23
…	…	…
juuli	1,0	22



Samade turunduskulude korral võib käibe väärtus olla erinev.

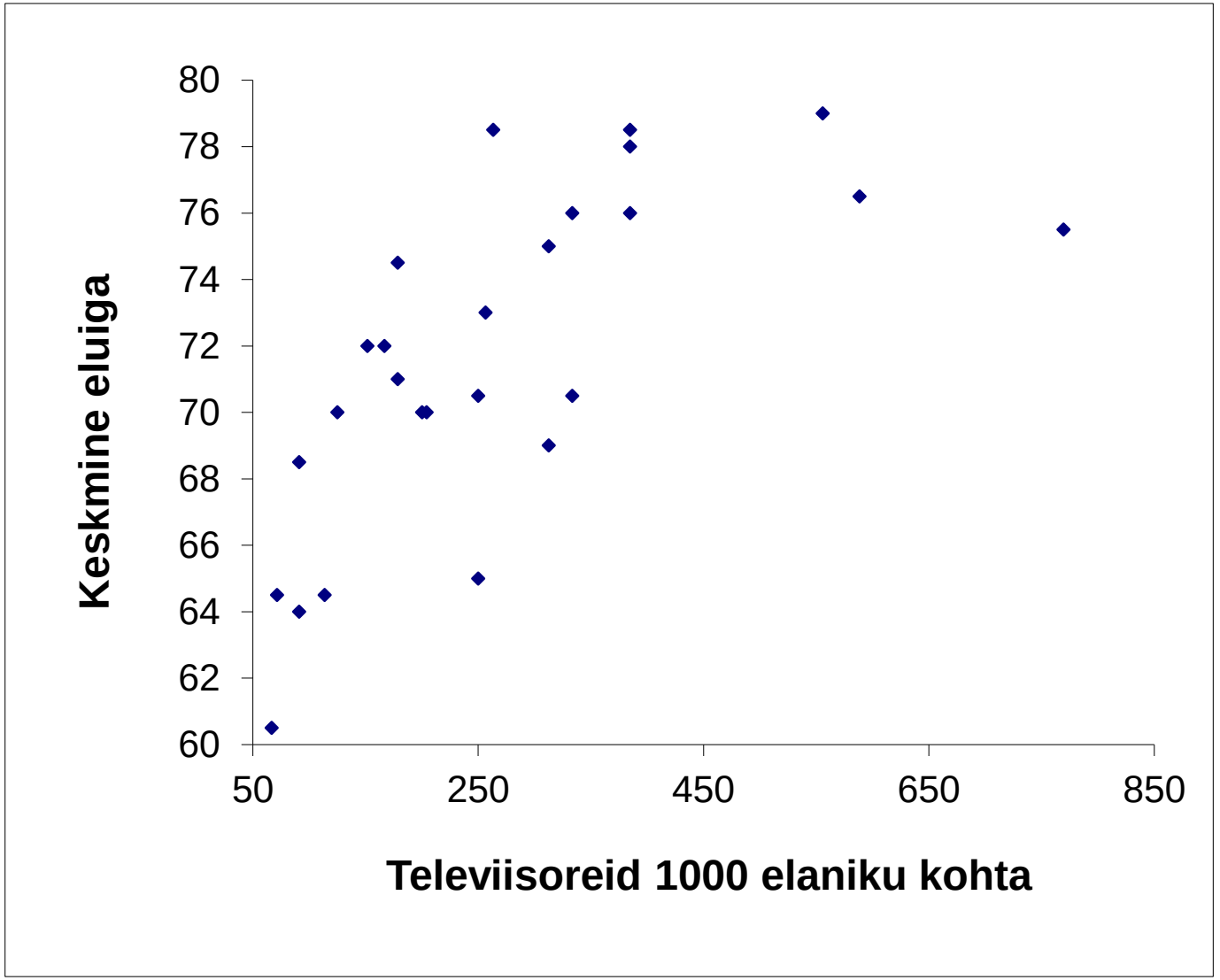
NÄIDE: KESKMINE ELUIGA JA TELERITE ARV

Valim 38 riiki.

- Keskmine eluiga;
- telerite arv 1000 elaniku kohta.

Kas keskmise eluea tõstmiseks peab televiisoreid olema rohkem?

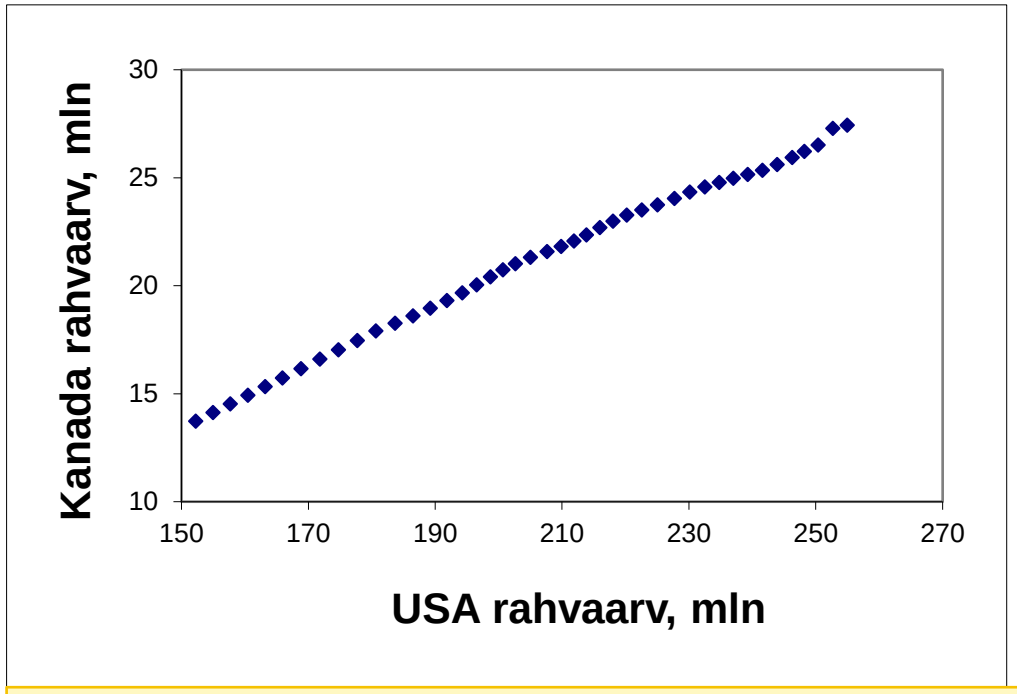
Näiv korrelatsioon



NÄIDE: USA JA KANADA RAHVAARV

USA ja Kanada rahvaarv aastatel 1950 – 1992.

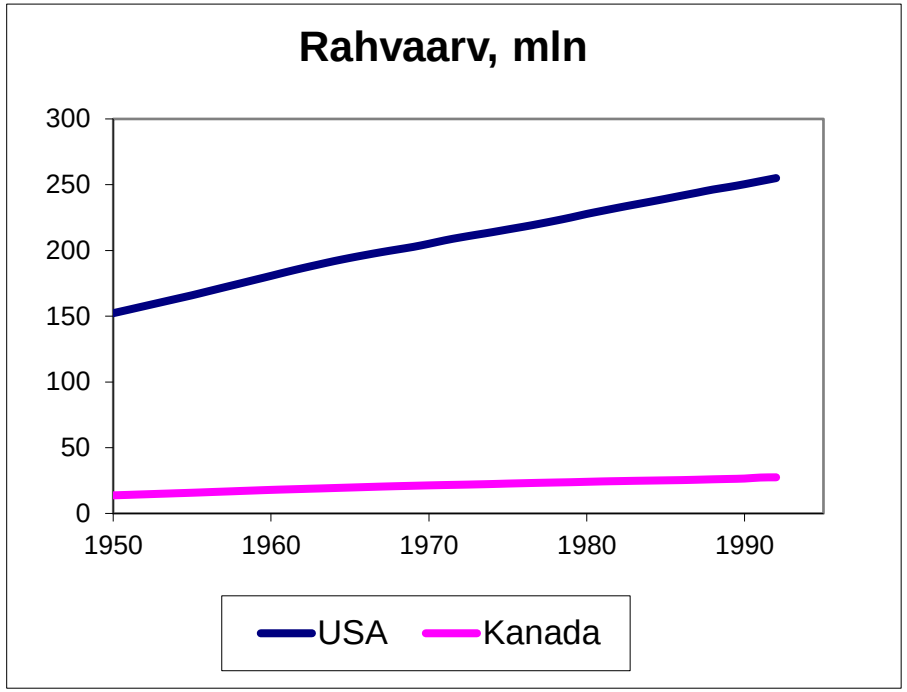
Hajumisdiagramm



Kas USA ja Kanada rahvaarv on omavahel seotud?

Näiv korrelatsioon

Aegread

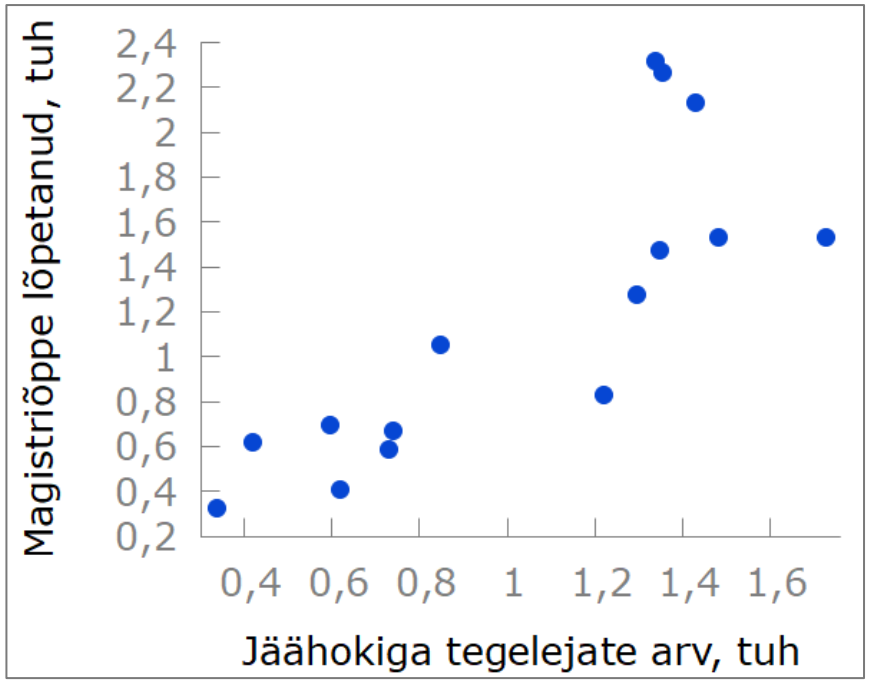


Nad muutuvad koos, neil on ligikaudu ühesugune trend.

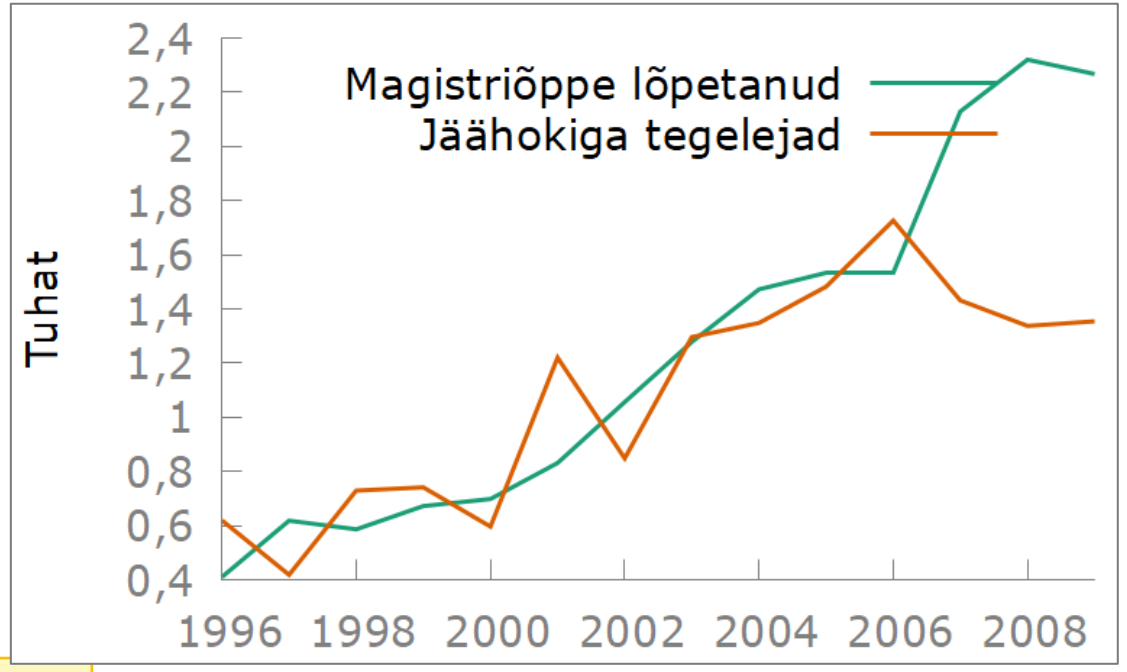
NÄIDE: MAGISTRIÕPPE LÕPETANUD JA JÄÄHOKIGA TEGELEJATE ARV

Magistriõppe lõpetanud ja jäähokiga tegelejate arv Eestis 1995 – 2009.

Hajumisdiagramm



Aegread



Kas jäähokiga tegelejate arv ja magistriõppe lõpetanute arv on omavahel seotud?

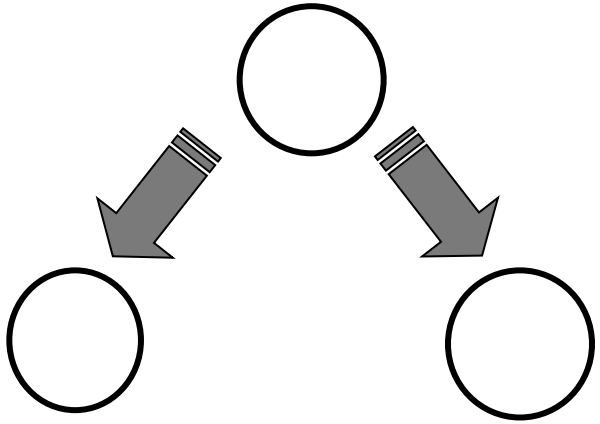
Näiv korrelatsioon

Nad muutuvad koos, neil on ligikaudu ühesugune trend.

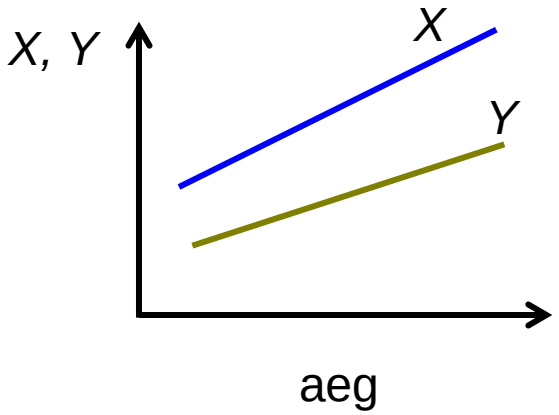
PÕHJUSLIK SEOS JA NÄIV KORRELATSIOON

Põhjuslik seos

Kui korrelatiivne seos on tugev:
- võib olla on põhjuslik seos;
- võib olla on näiv korrelatsioon.

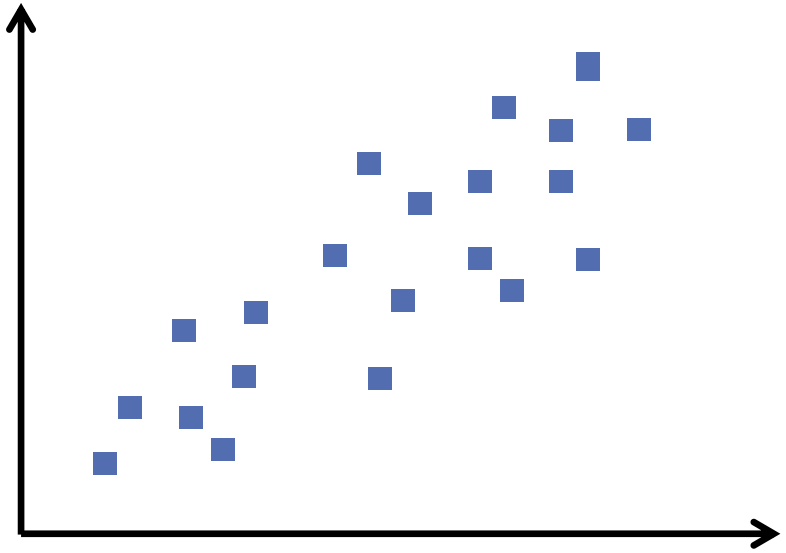


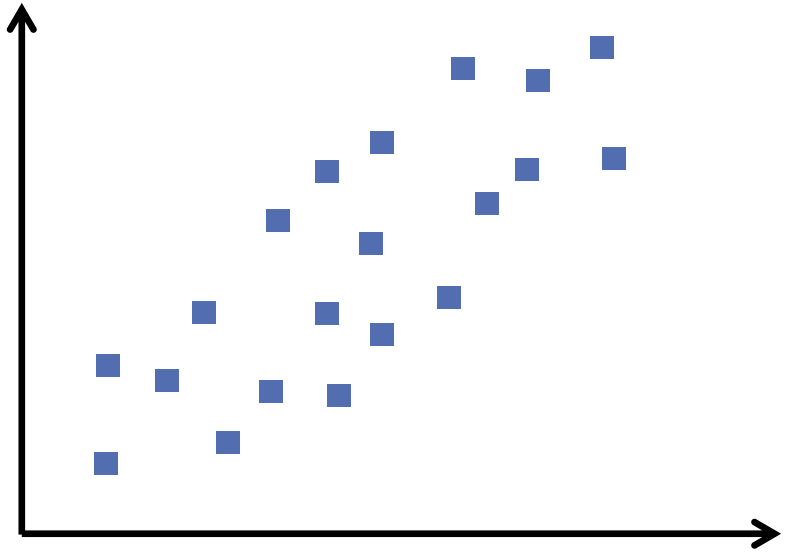
– mõlemale uuritavale tunnusele võib mõjuda hoopis kolmas, vaatluse alt välja jäänud tunnus



- mõlemal tunnusel on ühesugune trend

KORRELATSIOONI TUGEVUSE HINDAMINE





Kas vasakul ja paremal olev seos on ühesuguse tugevusega või mitte?

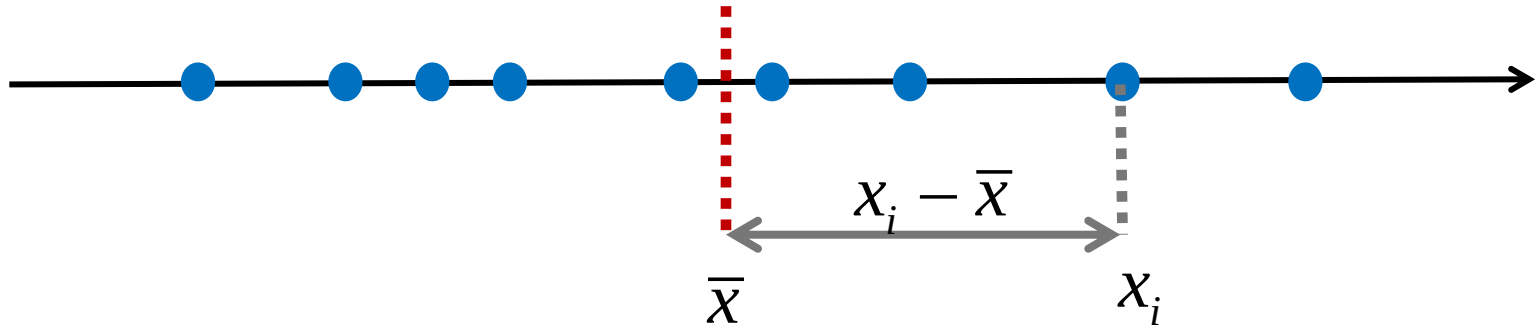
Seose tugevuse hindamine hajumisdiagrammi põhjal: subjektiivne.

Vaja objektiivset arvkarakteristikut.

HAJUVUSKARAKTERISTIK ÜHE TUNNUSE KORRAL

Ühe tunnuse korral kirjeldab hajumist dispersioon

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$



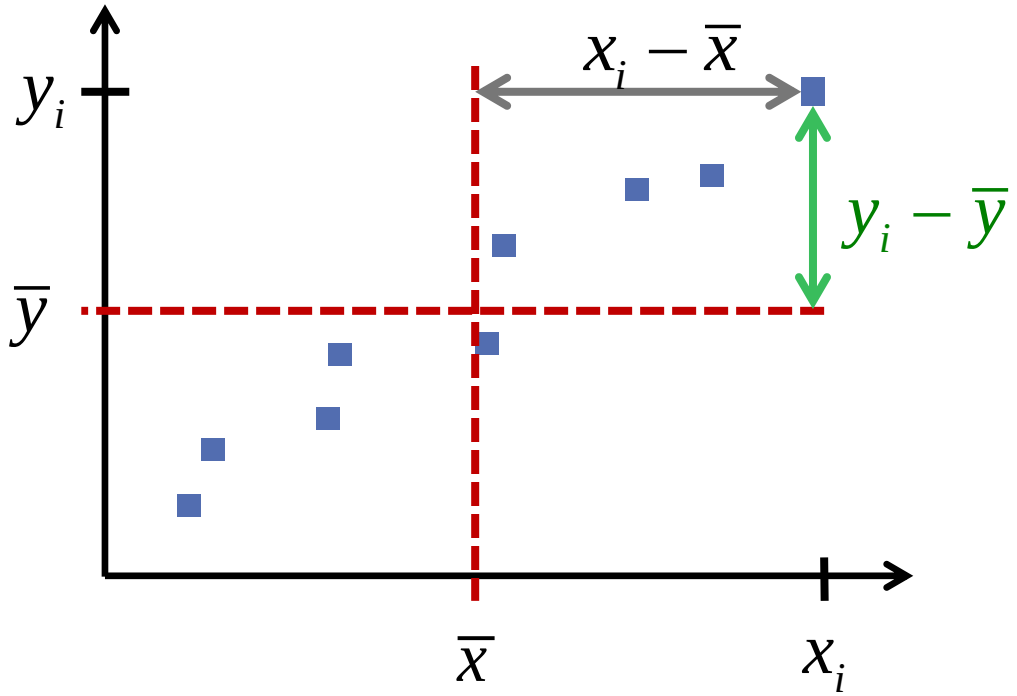
Ruudu esitame korrutisena $\sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})(x_i - \bar{x})$

KAHE SUURUSE KOOS MUUTUMINE

$$\sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})(x_i - \bar{x})$$

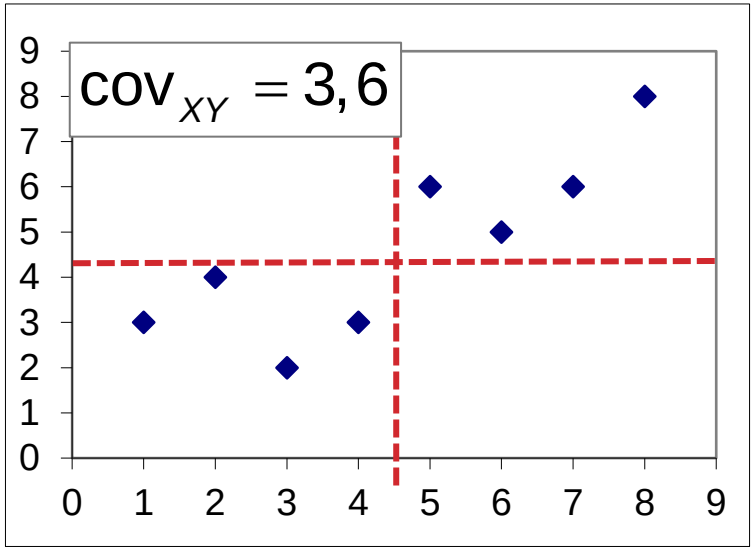
$$\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \mathrm{cov}_{XY}$$

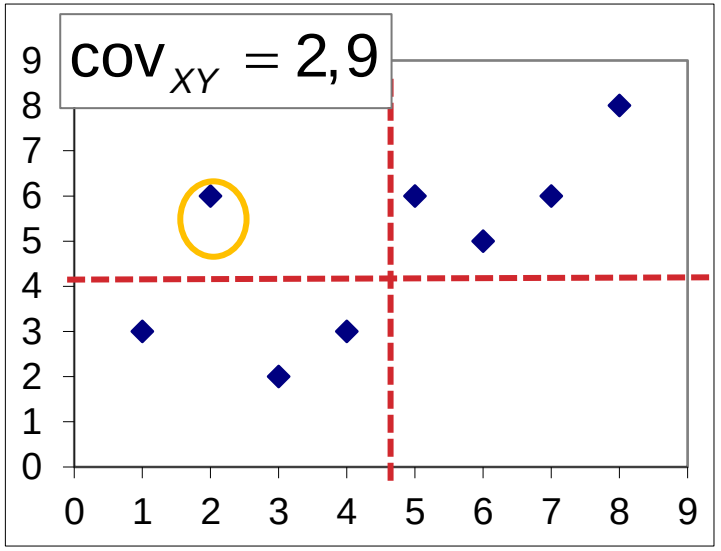
kovariatsioon



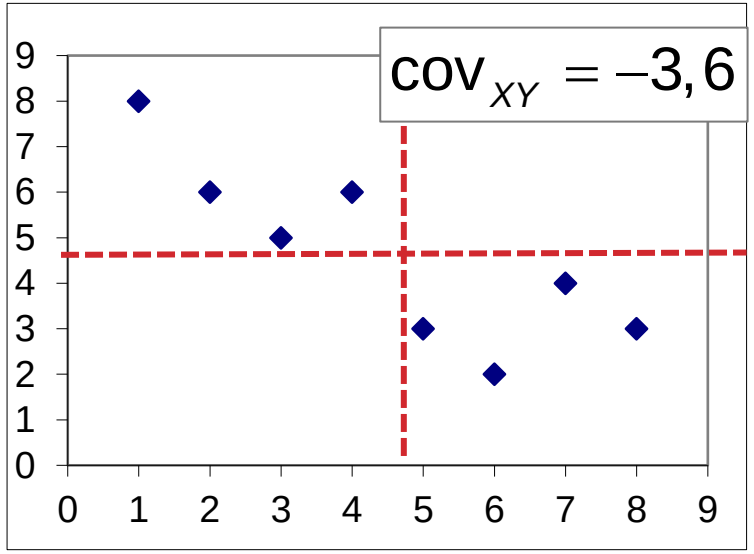
covariation – koos muutumine

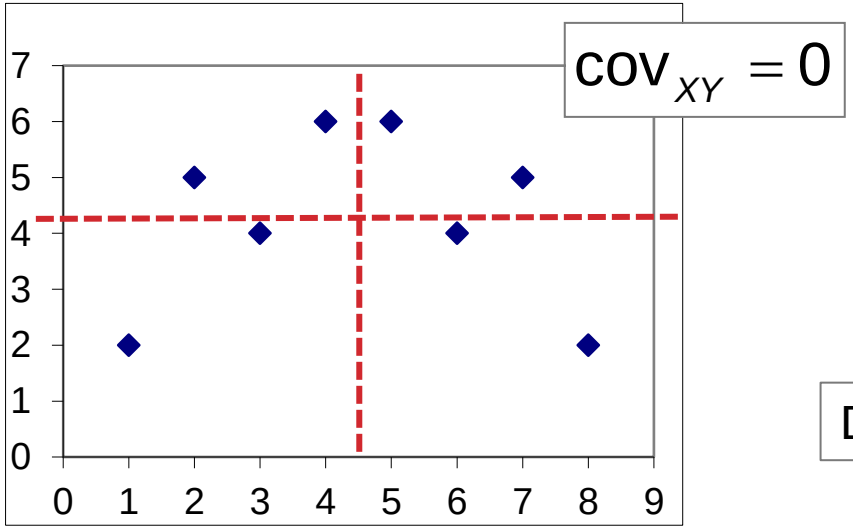
NÄIDE: KOVARIATSIOONI ERINEVAD VÄÄRTUSED





Kriipsjooned näitavad *x* ja *y* keskmisi väärtusi.





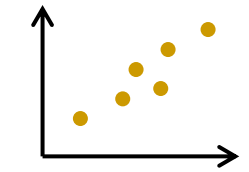
Demo: kovariatsioon

KOVARIATSIOONI OMADUSED, 1

$$\mathrm{cov}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

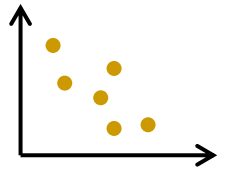
1. Kovariatsioon võib olla nii positiivne kui negatiivne $-\sigma_X \sigma_Y < \mathrm{cov}_{XY} < \sigma_X \sigma_Y$

Positiivne kovariatsioon: suurematele *X* väärtustele vastavad keskmiselt ka suuremad *Y* väärtused, väiksematele *X* väärtustele väiksemad *Y* väärtused.



$\mathrm{cov}_{XY} > 0$

Negatiivne kovariatsioon: suurematele *X* väärtustele vastavad keskmiselt väiksemad *Y* väärtused, väiksematele *X* väärtustele suuremad *Y* väärtused.



$\mathrm{cov}_{XY} < 0$

KOVARIATSIOONI OMADUSED, 2

$$\operatorname{cov}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

2. Sümmeetrilisus $\operatorname{cov}_{XY} = \operatorname{cov}_{YX}$

3. Kui X=Y, siis $\operatorname{cov}_{XX} = \sigma_X^2$

- Kovariatsioon on dispersiooni üldistus.
- Dispersioon on kovariatsiooni erijuht: kovariatsioon iseendaga.

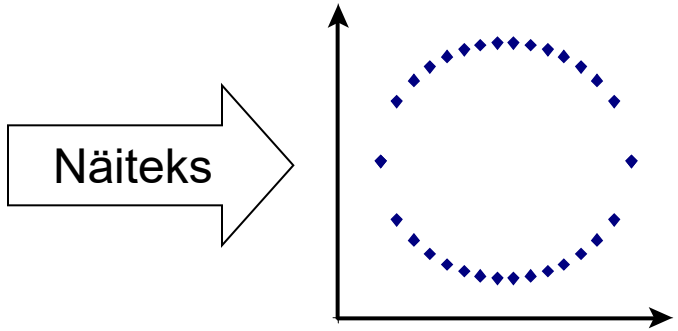
KOVARIATSIOONI OMADUSED, 3

$$\mathrm{cov}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

4. Sõltumatute juhuslike suuruste kovariatsioon on võrdne nulliga:

$$\mathrm{cov}_{XY} = 0$$

Vastupidine ei kehti, st kui kovariatsioon on null, ei pruugi suurused olla sõltumatud.



5. Kui $\mathrm{cov}_{XY} \neq 0$, siis nimetatakse suurusi *X* ja *Y* korreleeruvateks.

AKTSIAPORTFELLI RISK

Väärtpaberi riski mõõduks on tulumäära dispersioon.

Olgu meil 2 aktsiat:
A, tulumäär r_A, tulumäära dispersioon σ_A^2
B, tulumäär r_B, tulumäära dispersioon σ_B^2

Moodustame neist aktsiaportfelli.
Aktsia A osakaal w_A
Aktsia B osakaal $w_B = 1 - w_A$

Aktsiaportfelli tulumäär $r_P = w_A r_A + w_B r_B$

Kuidas leida aktsiaportfelli tulumäära dispersiooni, st aktsiaportfelli riski?

Tulumäärad r_A ja r_B on juhuslikud suurused.
Kuidas leida kahe juhusliku suuruse summa dispersiooni?

JUHUSLIKE SUURUSTE SUMMA DISPERSIOON

Kahe juhusliku suuruse X ja Y summaarne dispersioon on võrdne nende juhuslike suuruste dispersioonide summaga, millele on liidetud kahekordne nendevaheline kovariatsioon:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2cov_{XY}$$

Kahest väärtpaberist A ja B koosneva portfelli tulumäära dispersioon on

$$\sigma^2_P = w^2_A\sigma^2_A + w^2_B\sigma^2_B + 2w_A w_B cov_{AB}$$

NÄIDE: AKTSIAPORTFELLI DIVERSIFITSEERIMINE

Ajavahemikul 1.05.- 1.06.2012 olid kahe New Yorgi börsil kaubeldava aktsia tulumäär ja standardhälve ning nendevaheline kovariatsioon järgmised

	Keskmine tulumäär	Tulumäära standardhälve
JPMorgan Chase & Co (JPM)	-1,39%	2,64%
AT&T Inc. (T)	0,116%	0,66%
Kovariatsioon	$-2{,}42 \cdot 10^{-5}$	

Olgu aktsiaportfellis aktsia JPM osakaal 9% ja aktsia T osakaal 91%.

Aktsiaportfelli standardhälbe arvutus

$$\sigma_P^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_A w_B cov_{AB} =$$

$$= 0{,}09^2 \cdot 0{,}0264^2 + 0{,}91^2 \cdot 0{,}0066^2 + 2 \cdot 0{,}09 \cdot 0{,}91 \cdot (-2{,}42 \cdot 10^{-5}) = 3{,}78 \cdot 10^{-5}$$

$$\sigma_P = \sqrt{3{,}78 \cdot 10^{-5}} = 0{,}61\%$$

Väiksem kui kummagi aktsia tulumäära standardhälve.

Diversifitseerimine on investeerimisriski hajutamine.

KOVARIATSIOONI ÜHIKUD

$$\mathrm{cov}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Imiku vanus, kehakaal ja pikkus.

Vanus, kuud	Kaal, kg	Pikkus, cm
1	4,55	57
2	4,87	60
3	5,64	64
4	6,15	66
5	6,55	67
6	6,9	70

$$\mathrm{cov}_{\text{vanus, kaal}} = 1{,}44\ \text{kuud} \cdot \text{kg}$$

$$\mathrm{cov}_{\text{kaal, pikkus}} = 3{,}67\ \text{kg} \cdot \text{cm}$$

Kumb seos on tugevam?

Erinevates ühikutes olevaid suurusi ei saa võrrelda!

KOVARIATSIOONI ÜHIKUD

$$\mathrm{cov}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Imiku vanus, kehakaal ja pikkus.

Vanus, kuud	Kaal, kg	Pikkus, m
1	4,55	0,57
2	4,87	0,60
3	5,64	0,64
4	6,15	0,66
5	6,55	0,67
6	6,9	0,70

$$\mathrm{cov}_{\text{vanus, kaal}} = 1{,}44\ \text{kuud} \cdot \text{kg}$$

$$\mathrm{cov}_{\text{kaal,pikkus}} = 0{,}0367\ \text{kg} \cdot \text{m}$$

Teisendame ühikuid, kovariatsiooni arvväärtus muutub.

KOVARIATSIOONI SUURUS

Keskmine brutokuupalk ja sündinud ettevõtete arv 2009. aastal Eesti maakondades (va Harju maakond).

Maakond	Keskmine brutokuupalk (eurot)	Sündinud ettevõtted
Hiiu maakond	607	26
Ida-Viru maakond	637	246
Jõgeva maakond	578	52
Järva maakond	625	64
Lääne maakond	619	81
Lääne-Viru maakond	624	168
Põlva maakond	641	60
Pärnu maakond	659	318
Rapla maakond	641	115
Saare maakond	652	100
Tartu maakond	749	676
Valga maakond	574	66
Viljandi maakond	643	145
Võru maakond	647	94

cov = 5563 eurot

Kas seos on tugev või nõrk?

KOVARIATSIOONI PUUDUSED

- Kovariatsiooni väärtus sõltub valitud ühikutest.
- Absoluutväärtus võib olla väga suur => raske hinnata seose tugevust.

Lahendus: normeeritakse nii, et absoluutväärtuse maksimaalne väärtus oleks 1.

KORRELATSIOONIKORDAJA

Normeerime kovariatsiooni, nii et muutuks vahemikus -1 kuni 1.

$$-\sigma_X \sigma_Y < \mathrm{cov}_{XY} < \sigma_X \sigma_Y \qquad |: \sigma_X \sigma_Y$$

$$-1 < \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} < 1$$

Korrelatsioonikordaja

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} \qquad -1 < r_{XY} < 1$$

$$r_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_X \sigma_y}$$

Lineaarne ehk Pearsoni korrelatsioonikordaja

KORRELATSIOONIKORDAJA OMADUSED

- Ühikuta suurus.
- Absoluutväärtus näitab lineaarse seose tugevust.
- Märk näitab seose suunda: positiivne või negatiivne.

$r = 0$	korrelatsioon puudub
$\lvert r \rvert = 1$	täielikult korreleeruvad suurused
$0 < r < 1$	positiivne korrelatsioon
$-1 < r < 0$	negatiivne korrelatsioon

Demo: korrelatsioon

NÄIDE $r_{AB} = 0{,}58$ $r_{AC} = -0{,}87$

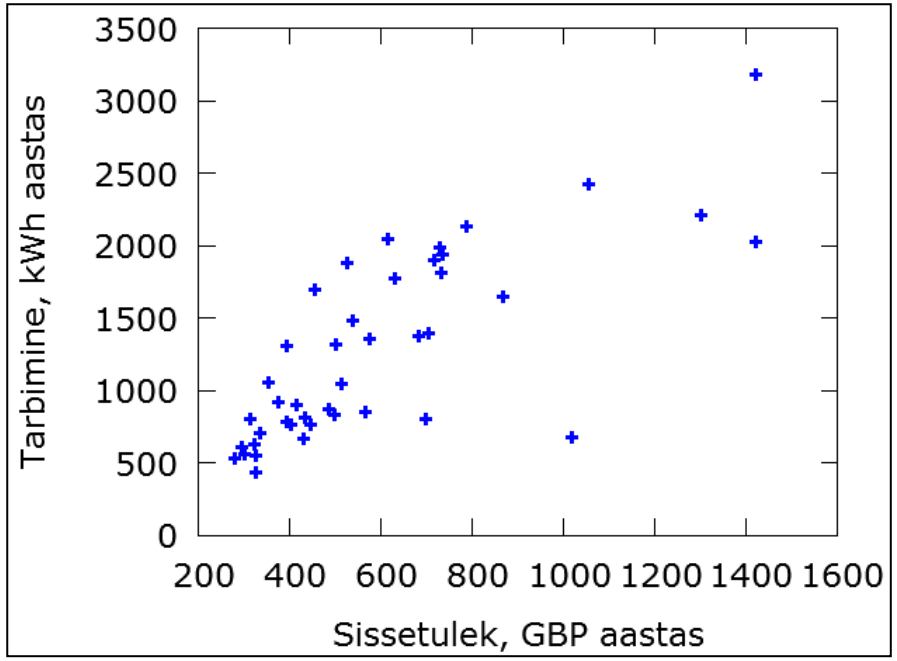
Kumb seos on tugevam?

A ja C vahel on tugevam seos kui A ja B vahel.

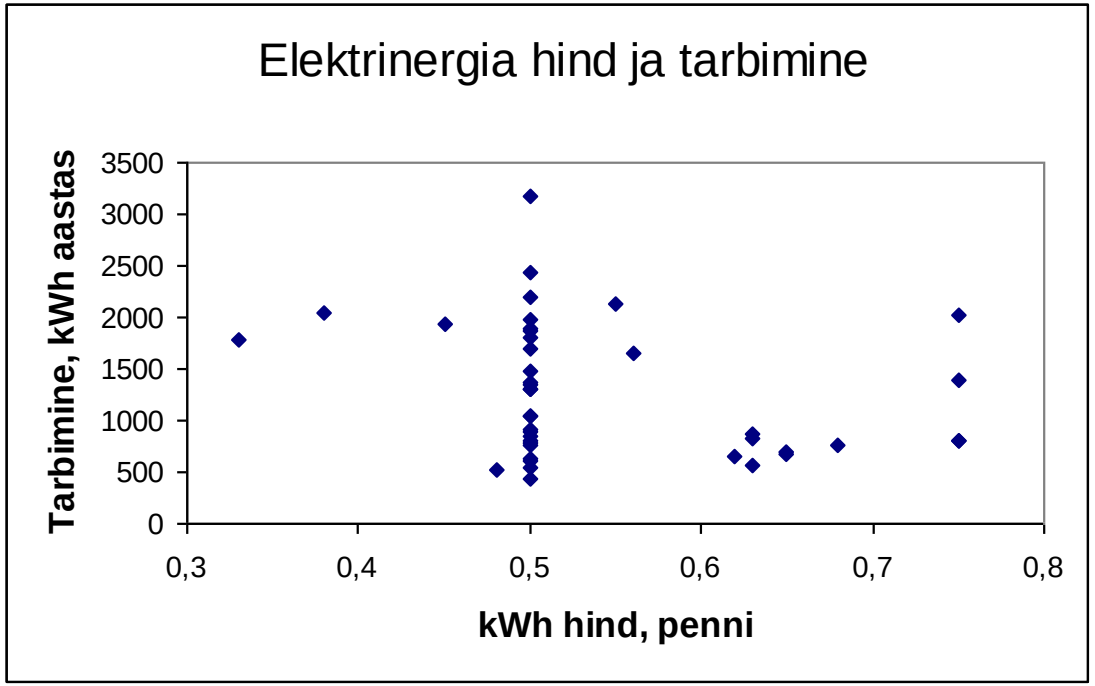
NÄIDE: POSITIIVNE JA NEGATIIVNE KORRELATSIOON

Elektrienergia tarbimine Suurbritannia erinevates linnades 1930. aastate lõpus. Andmed pärinesid 48 linnast.

Tarbija sissetuleku ja tarbimise vahel on positiivne korrelatsioon,
r = 0,767



Hinna ja tarbimise vahel on negatiivne korrelatsioon,
r = - 0,274



Allikas: Houthakker, H. S. 1951. Some Calculations on Electricity Consumption in Great Britain. Journal of the Royal Statistical Society. Series A (General) Vol. 114, 359-371.

KORRELATSIOONIMAATRIKS, 1

Kui objekte iseloomustavaid kvantitatiivseid tunnuseid on palju, siis nendevaheliste seoste tugevuse hindamiseks kasutatakse korrelatsioonimaatriksit.

Korrelatsioonimaatriks on juhuslike suuruste $X_1, X_2, \ldots X_K$ vahelist statistilist sõltuvust iseloomustav ruutmaatriks, mille elementideks on lineaarsed korrelatsioonikordajad:

$$r_{ij} = r_{X_i X_j}$$

kus *i* loendab ridu ja *j* veerge.

	X_1	X_2	X_3
X_1	r_{11}	r_{12}	r_{13}
X_2	r_{21}	r_{22}	r_{23}
X_3	r_{31}	r_{32}	r_{33}

Korrelatsioonikordaja iseendaga on 1

$$r_{11} = r_{22} = r_{22} = r_{ii} = 1$$

Korreratsioonikordaja on sümmeetriline

$$r_{12} = r_{21}, \quad r_{13} = r_{31}, \quad r_{23} = r_{32}$$

$$r_{ij} = r_{ji}$$

KORRELATSIOONIMAATRIKS, 2

Korrelatsioonimaatriks on sümmeetriline ruutmaatriks, mille peadiagonaalil on ühed.

Kuna sümmeetriline, siis tavaliselt ei esitata ülalpool peadiagonaali olevaid elemente, sest need on võrdsed vastavate allpool peadiagonaali olevate elementidega.

	X_1	X_2	X_3
X_1	1		
X_2	r_{21}	1	
X_3	r_{31}	r_{32}	1

NÄIDE: MÕNINGAD NÄITAJAD EESTI MAAKONDADES

Aastal 2009: töötuse määr, keskmine brutokuupalk, sündinud ettevõtete arv aastas, hõivatute osatähtsus sekundaarsektoris (töötlev tööstus), hõivatute osatähtsus tertsiaalsektoris (teenindus).

Mille vahel on kõige tugevam seos?

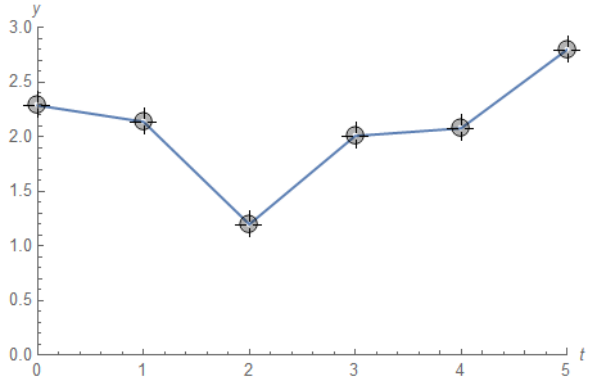
Mille vahel on kõige nõrgem seos?

	Töötuse määr	Keskmine palk	Sündinud ettevõtted	Sekundaar-sektor	Tertsiaal-sektor
Töötuse määr	1				
Keskmine palk	-0,521	1			
Sündinud ettevõtted	-0,219	0,852	1		
Sekundaar-sektor	0,175	-0,173	-0,260	1	
Tertsiaal-sektor	-0,043	0,310	-0,467	-0,891	1

AUTOKORRELATSIOON

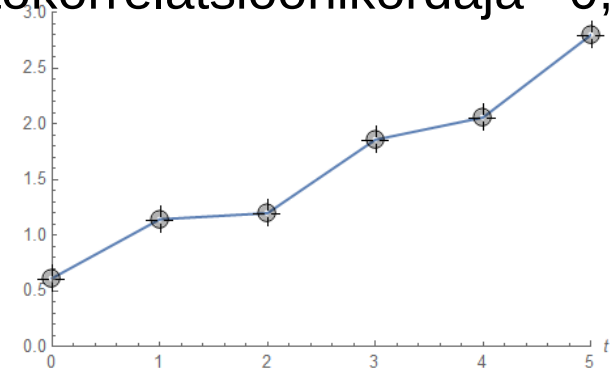
Aegridade analüüsimisel kasutatakse mingi suuruse muutumise juhuslikkuse või mittejuhuslikkuse hindamisel **autokorrelatsiooni.**

Juhuslikkus suur,
autokorrelatsioonikordaja ≈0



aeg *t*

Autokorrelatsioonikordaja =0,9



aeg *t*

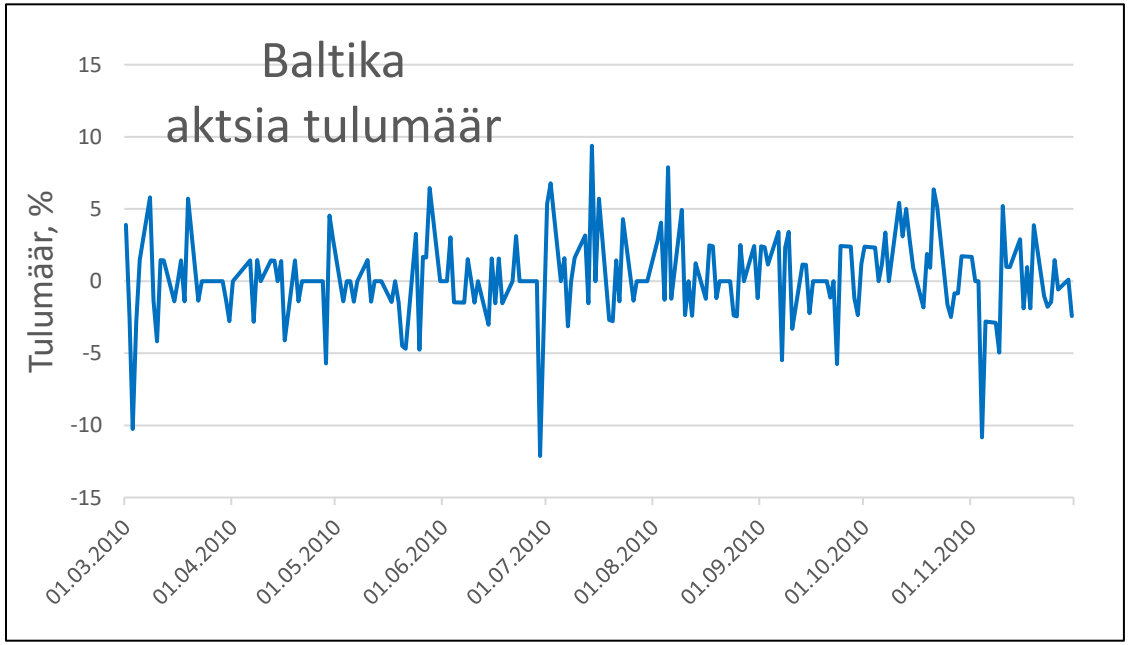
Demo: autokorrelatsioon

Aeg t	Suurus X	
1	x_1	
2	x_2	x_1
3	x_3	x_2
4	x_4	x_3
5	x_5	x_4

Leiame
korrelatsioonikordaja

NÄIDE: AUTOKORRELATSIOON

Harju Elektri ja Baltika aktsiate tulumäärad 1.03.-30.11.2010



Autokorrelatsioonikordaja $r = -0{,}149$ $r = -0{,}009$

Baltika aktsia tulumäära muutumises oli juhuslikkust rohkem.

KORRELATSIOONI STATISTILINE OLULISUS

Demo: korrelatsiooni juhuslikkus

Korrelatsioonikordaja on nullist erinev ka täiesti juhuslike arvupaaride korral.

Kui suur peab korrelatsioonikordaja olema, et võiksime öelda: seos on olemas?

Vaja kriteeriumi!

HÜPOTEESI KONTROLLIMINE KORRELATSIOONIKORDAJA JAOKS

Kasutatakse *t* testi, teststatistiku arvutusvalem

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

Nullhüpotees H_0 (korrelatsioon puudub):	$r = 0$
Sisukas hüpotees H_1 (korrelatsioon esineb):	$r \neq 0$

Olulisuse nivoole α vastavad kriitilised väärtused *t* -jaotusest.

Kriitiline piirkond (võtta vastu H_1): $|t| > t_{kr}$

Demo: korrelatsiooni juhuslikkus

NÄIDE: MEESTE KESKMINE ELUIGA JA RKT

91 riigi andmed

korrelatsioonikordaja	$r = 0{,}643$
valimi maht	$n = 91$

Nullhüpotees H_0 (korrelatsioon puudub): $r = 0$

Sisukas hüpotees H_1 (korrelatsioon esineb): $r \neq 0$

t- statistiku empiiriline väärtus $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 7{,}92$

Kriitilised väärtused -1,987 1,987

Nullhüpotees on tagasi lükatud.
Meeste keskmine eluiga on seotud rahvusliku kogutoodanguga ühe elaniku kohta.

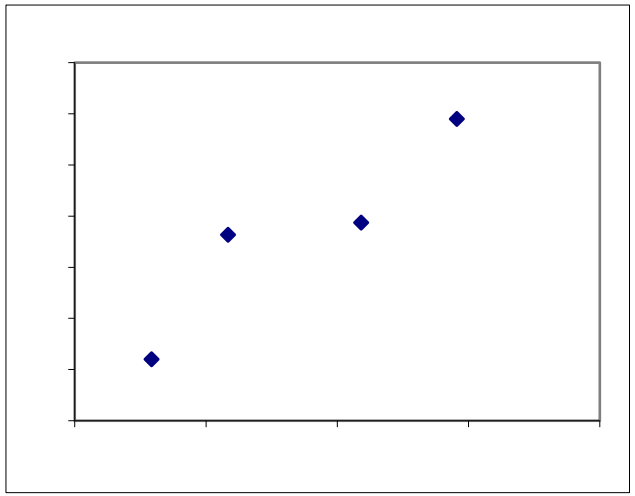
MILLEST SÕLTUB KORRELATSIOONI OLULISUS

Millest sõltub *t*-statistik? $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

korrelatsioonikordaja *r*
punktipaaride arv *n*

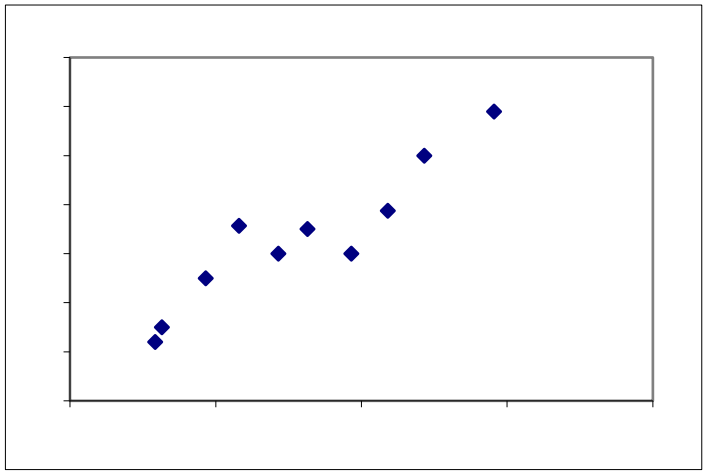
r = 0,931 *t* = 3,60
n = 4 t_{kr} = 4,30

H_0 korrelatsioon ei ole oluline



r = 0,931 *t* = 7,24
n = 10 t_{kr} = 2,31

H_1 korrelatsioon on oluline



NÄIDE: KESKMINE ÕHUTEMPERATUUR JA BÖRSIINDEKS

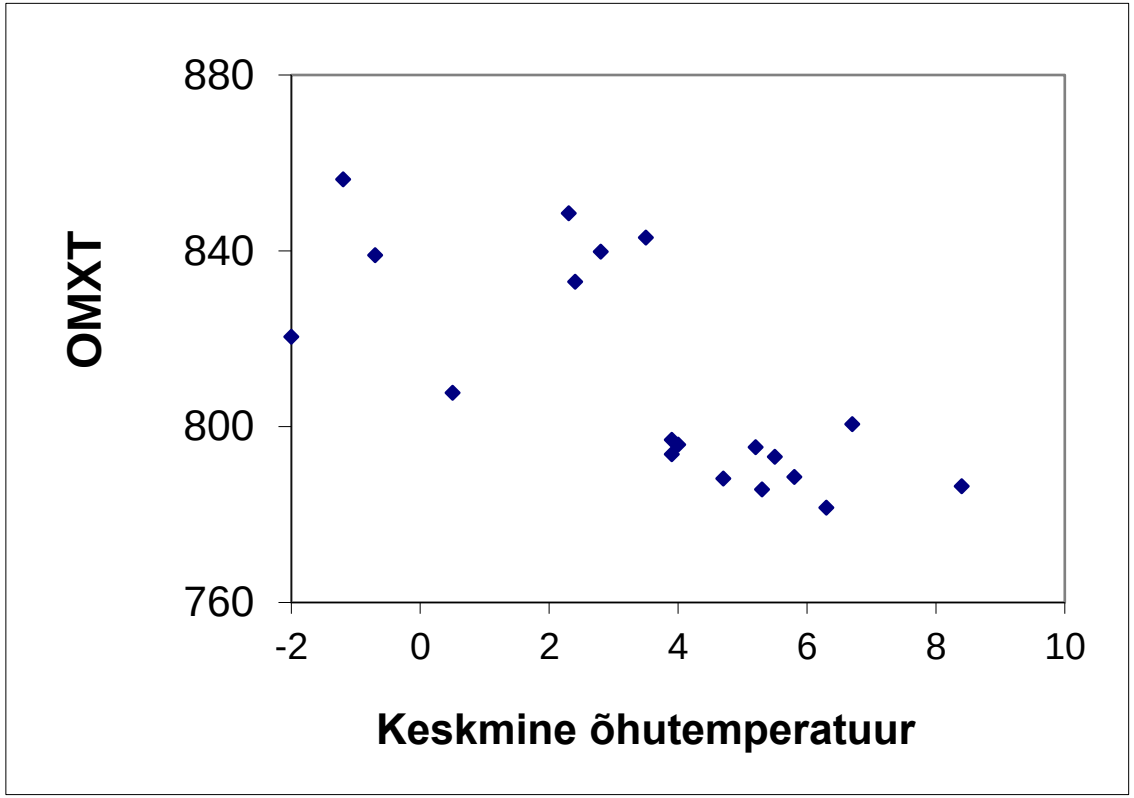
Detsember 2006: keskmine õhutemperatuur Tallinnas ja OMXT indeks.
19 päeva andmed.

Korrelatsioonikordaja - 0,718

Korrelatsiooni olulisuse hüpoteesi kontrollimine nivool 5%

t = - 4,25

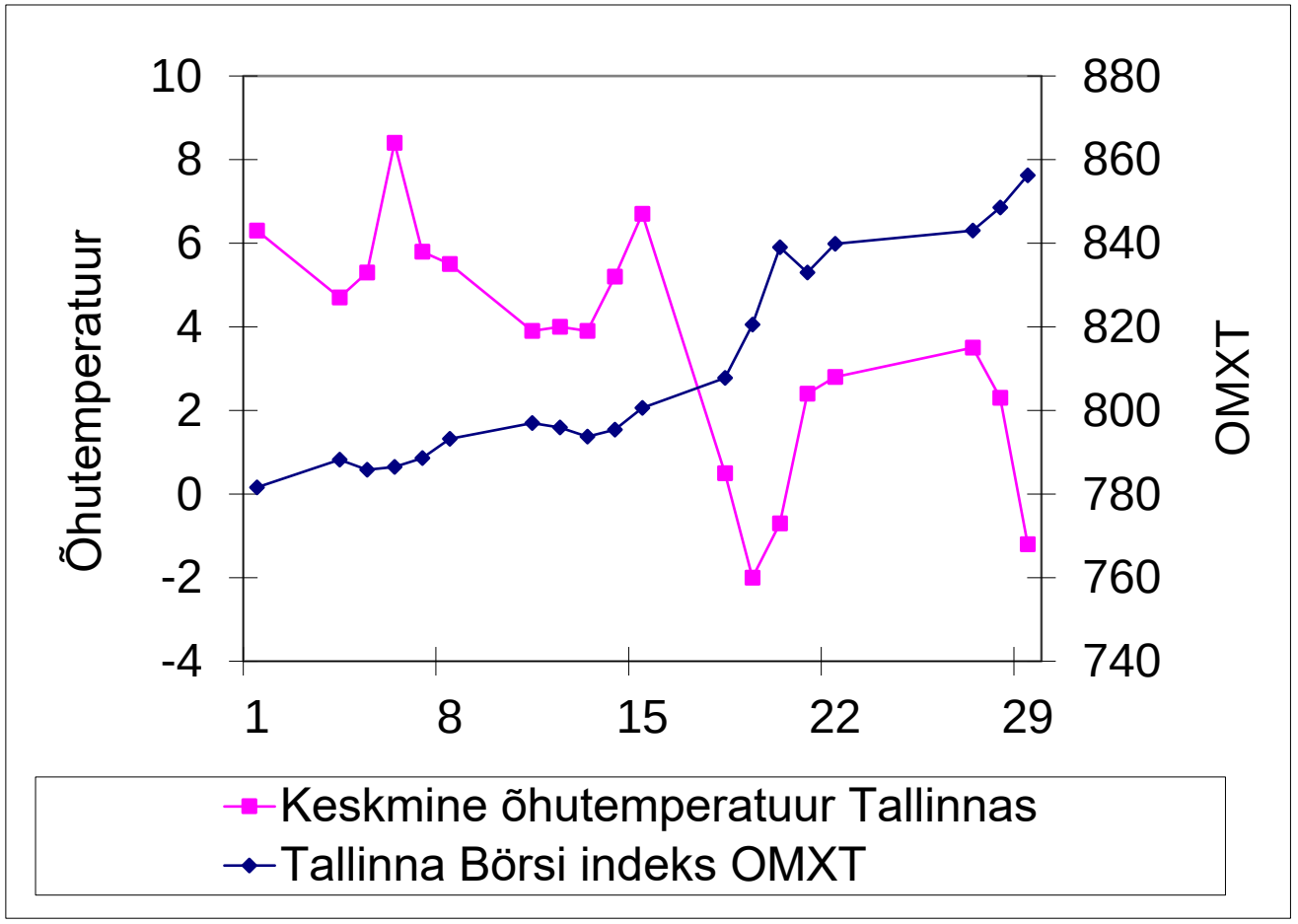
Kriitilised -2,11 2,11



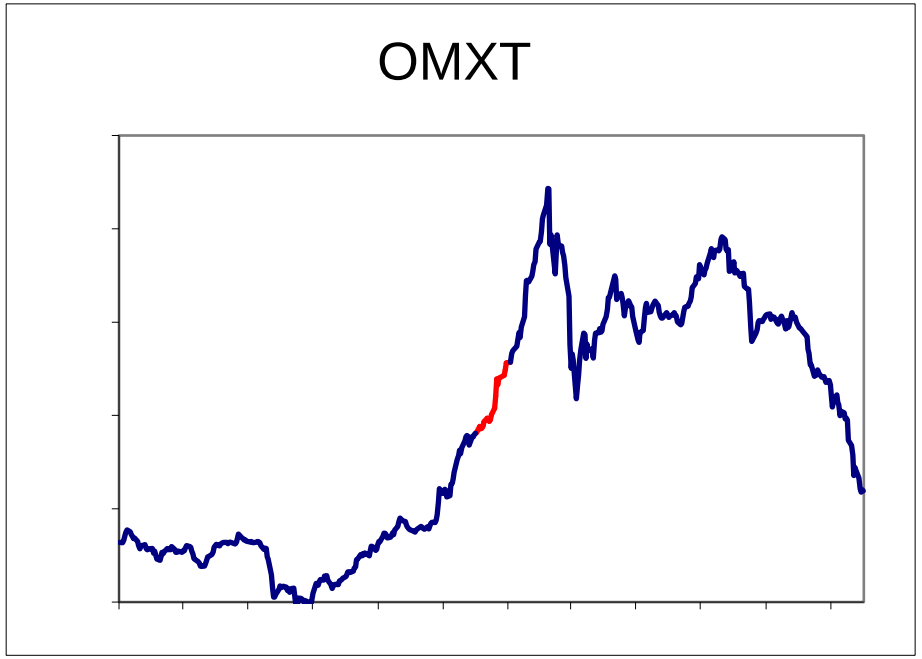
Vastu võtta sisukas hüpotees: **korrelatsioon on statistiliselt oluline.**

NÄIDE: KESKMINE ÕHUTEMPERATUUR JA BÖRSIINDEKS

Muutumine ajas



NÄIV KORRELATSIOON



KORRELATSIOONIMAATRIKS JA KORRELATSIOONI OLULISUS

Korrelatsioonimaatriksis on valimi maht *n* kõikide korrelatsioonikordajate korral ühesugune.

	X	Y	Z
X	1		
Y	r_{XY}	1	
Z	r_{XZ}	r_{YZ}	1

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

Kriitiline *t* väärtus $t_{\alpha/2}(v)$ on kõikide korrelatsioonikordajate jaoks ühesugune, *v=n-2.*

$$\Downarrow$$

$$r_{kr} = \frac{1}{\sqrt{1+\frac{\upsilon}{t^2_{\alpha/2}(\upsilon)}}}$$

Saame leida korrelatsioonikordaja kriitilise väärtuse antud korrelatsioonimaatriksi jaoks.

Kui $|r| > r_{kr}$, siis korrelatsioon esineb (H_1).

NÄIDE: KORRELATSIOONIMAATRIKS JA KRIITILINE KORRELATSIOONIKORDAJA

Andmed 2009: töötuse määr; keskmine brutokuupalk; sündinud ettevõtete arv aastas, hõivatute osatähtsus sekundaarsektoris (töötlev tööstus), hõivatute osatähtsus tertsiaalsektoris (teenindus).

n=15, α=0,05
Kriitiline korrelatsioonikordaja
$r_{kr} = 0{,}514$

	Töötuse määr	Keskmine palk	Sündinud ettevõtted	Sekundaar-sektor	Tertsiaal-sektor
Töötuse määr	1				
Keskmine palk	-0,521**	1			
Sündinud ettevõtted	-0,219	0,852**	1		
Sekundaar-sektor	0,175	-0,173	-0,260	1	
Tertsiaal-sektor	-0,043	0,310	-0,467	-0,891**	1

Tärnidega statistiliselt oluline korrelatsioon.

MÕNED KORRELATSIOONIKORDAJA KRIITILISED VÄÄRTUSED

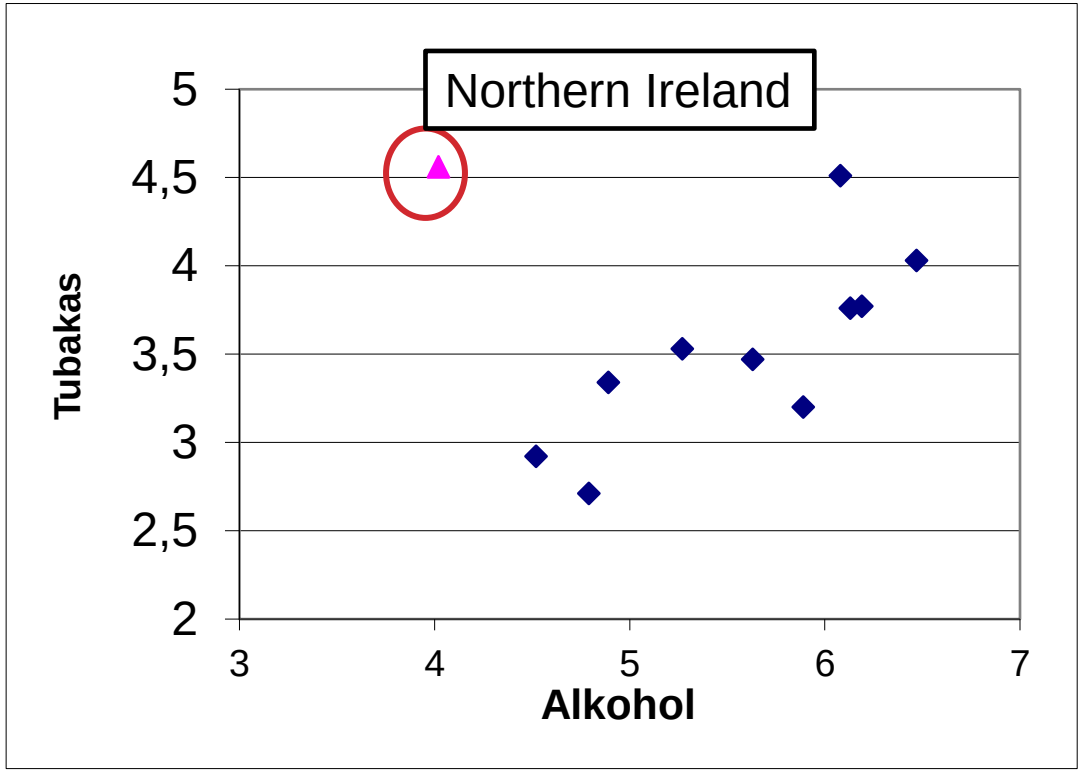
n	r_{kr} (α=0,05)
5	0,878
10	0,632
20	0,444
50	0,279
100	0,197
1000	0,062

Vt Statistika õpik tabel B.7 lk 728.

NÄIDE: KULUD ALKOHOLILE JA TUBAKALE, 1

Uuring Suurbritannias: kas kulutused tubakale ja alkoholile on seotud?

Piirkond	Kulud alkoholile	Kulud tubakale
North	6,47	4,03
Yorkshire	6,13	3,76
Northeast	6,19	3,77
East Midlands	4,89	3,34
West Midlands	5,63	3,47
East Anglia	4,52	2,92
Southeast	5,89	3,2
Southwest	4,79	2,71
Wales	5,27	3,53
Scotland	6,08	4,51
Northern Ireland	4,02	4,56



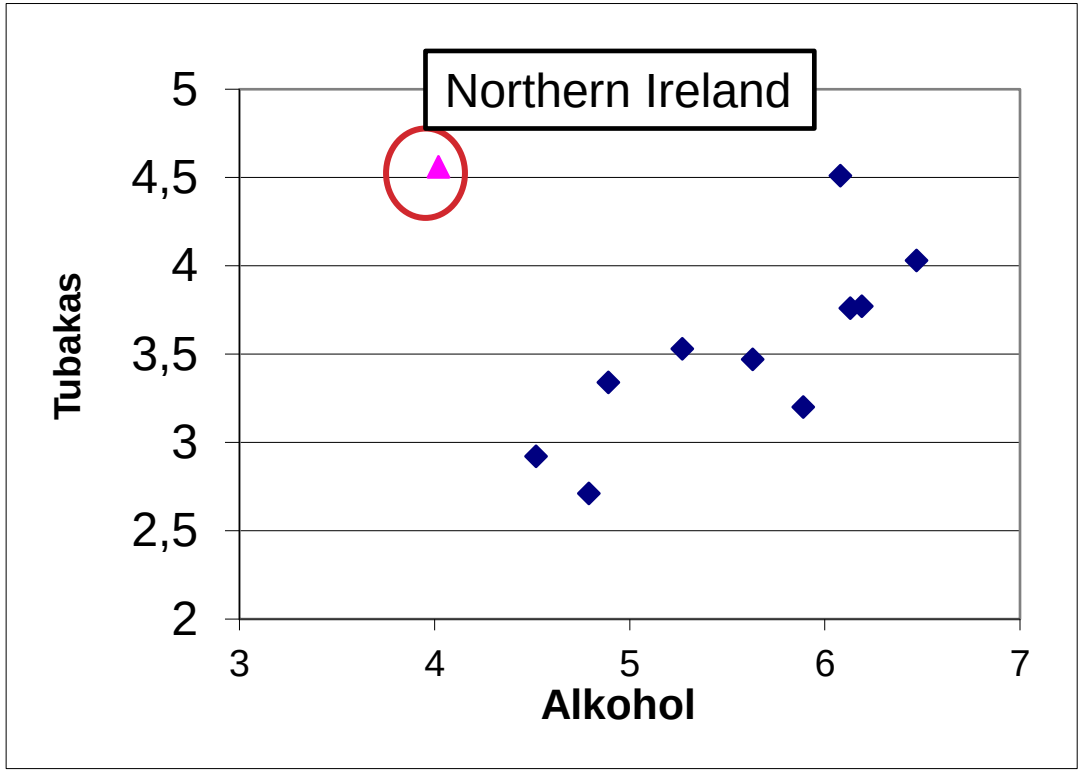
ERIND

r = 0,224, nõrk seos

NÄIDE: KULUD ALKOHOLILE JA TUBAKALE, 2

Uuring Suurbritannias: kas kulutused tubakale ja alkoholile on seotud?

Piirkond	Kulud alkoholile	Kulud tubakale
North	6,47	4,03
Yorkshire	6,13	3,76
Northeast	6,19	3,77
East Midlands	4,89	3,34
West Midlands	5,63	3,47
East Anglia	4,52	2,92
Southeast	5,89	3,2
Southwest	4,79	2,71
Wales	5,27	3,53
Scotland	6,08	4,51
~~Northern Ireland~~	~~4,02~~	~~4,56~~



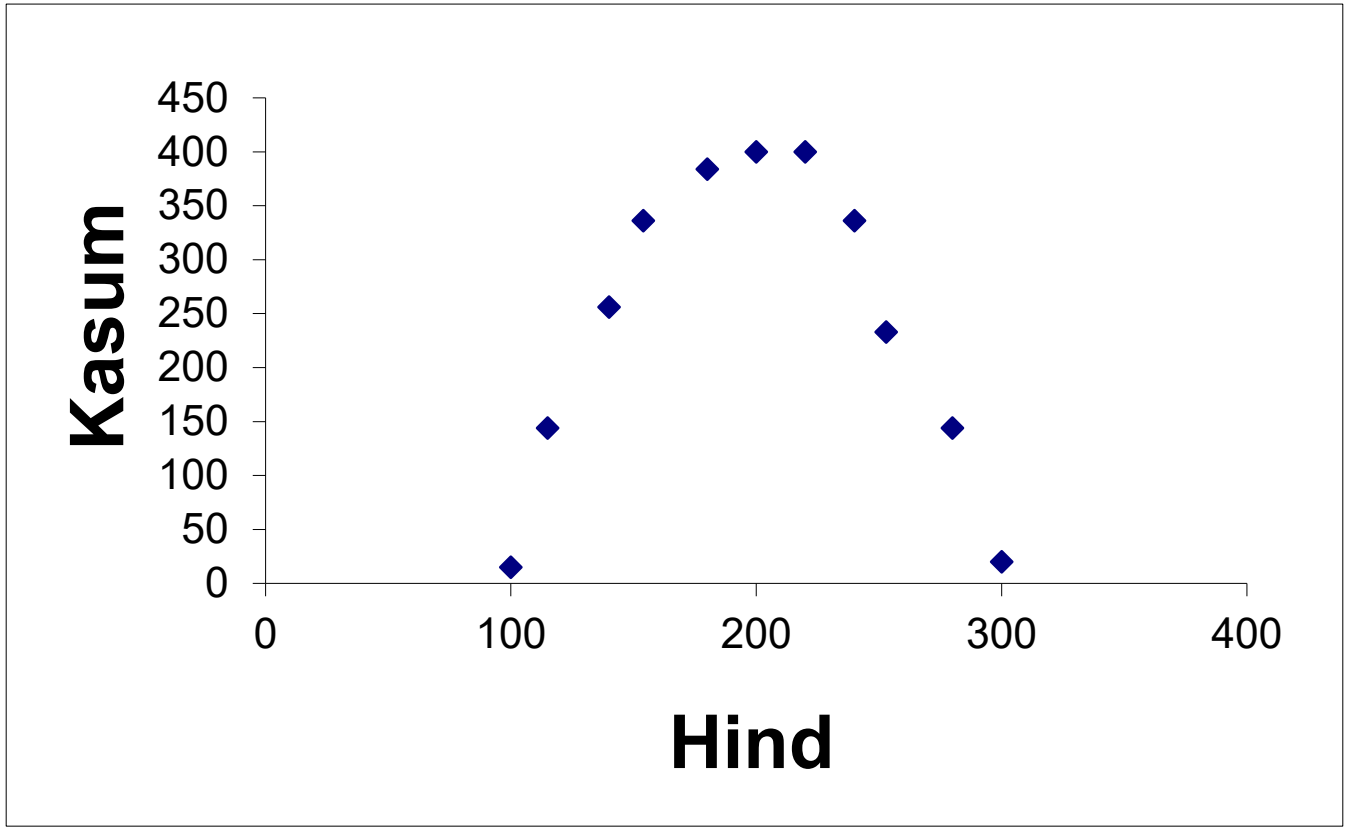
Ilma erindita

r = 0,784, tugev seos

ERIND

Demo: korrelatsioonikordaja ja erind

NÄIDE: MITTELINEAARNE SEOS



$$r = -0,0058$$

Lineaarne korrelatsioonikordaja näitab, et seos puudub.
Aga tegelikult seos on, see on mittelineaarne.

LINEAARSE KORRELATSIOONIKORDAJA PUUDUSED

1. Erindid mõjutavad tugevasti.
2. Iseloomustab vaid lineaarse seose tugevust.

Alternatiivne näitaja: astakkorrelatsiooni kordaja ehk Spearmani korrelatsioonikordaja.

SPEARMANI KORRELATSIOONIKORDAJA

Järjenumbrite ehk astakute korrelatsioonikordaja, mida nimetatakse Spearmani korrelatsioonikordajaks, leitakse valemist

$$r_S = 1 - \frac{6\sum d_i^{\,2}}{n(n^2 - 1)}$$

kus d_i on erinevates gruppides kõrvuti olevate järjekorranumbrite (astakute) vahe ja n väärtuspaaride arv.

NÄIDE: SPEARMANI KORRELATSIOONIKORDAJA

Tuli järjestada valikud A, B, C, D ja E.
Tabelites on võrreldud valikute järjestust erinevates valimites.
d_i on järjenumbrite vahe.

Kokkulangevad järjestused

Valik	Valim 1	Valim 2	d_i
A	1	1	0
B	2	2	0
C	3	3	0
D	4	4	0
E	5	5	0

$$r_S = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)}$$

$r_S = 1$

Vastupidised järjestused

Valim 1	Valim 3	d_i
1	5	-4
2	4	-2
3	3	0
4	2	2
5	1	4

$r_S = -1$

Osaliselt kokkulangevad järjestused

Valim 1	Valim 4	d_i
1	1	0
2	2	0
3	5	-2
4	4	0
5	3	2

$r_S = 0{,}6$

NÄIDE: SPEARMANI KORRELATSIOONIKORDAJA

Ettevõttes paluti järjestada keskastmejuhtide motivatsioonifaktorid (A kuni I) tippjuhtidel ja keskastmejuhtidel.
Tippjuhid pidid faktorid reastama nii, nagu arvasid keskastme juhte neid reastavat.
Tippjuhtide väljapakutud pingerida võrreldi sellega, kuidas keskastmejuhid ise samad motivatsioonifaktorid reastasid.

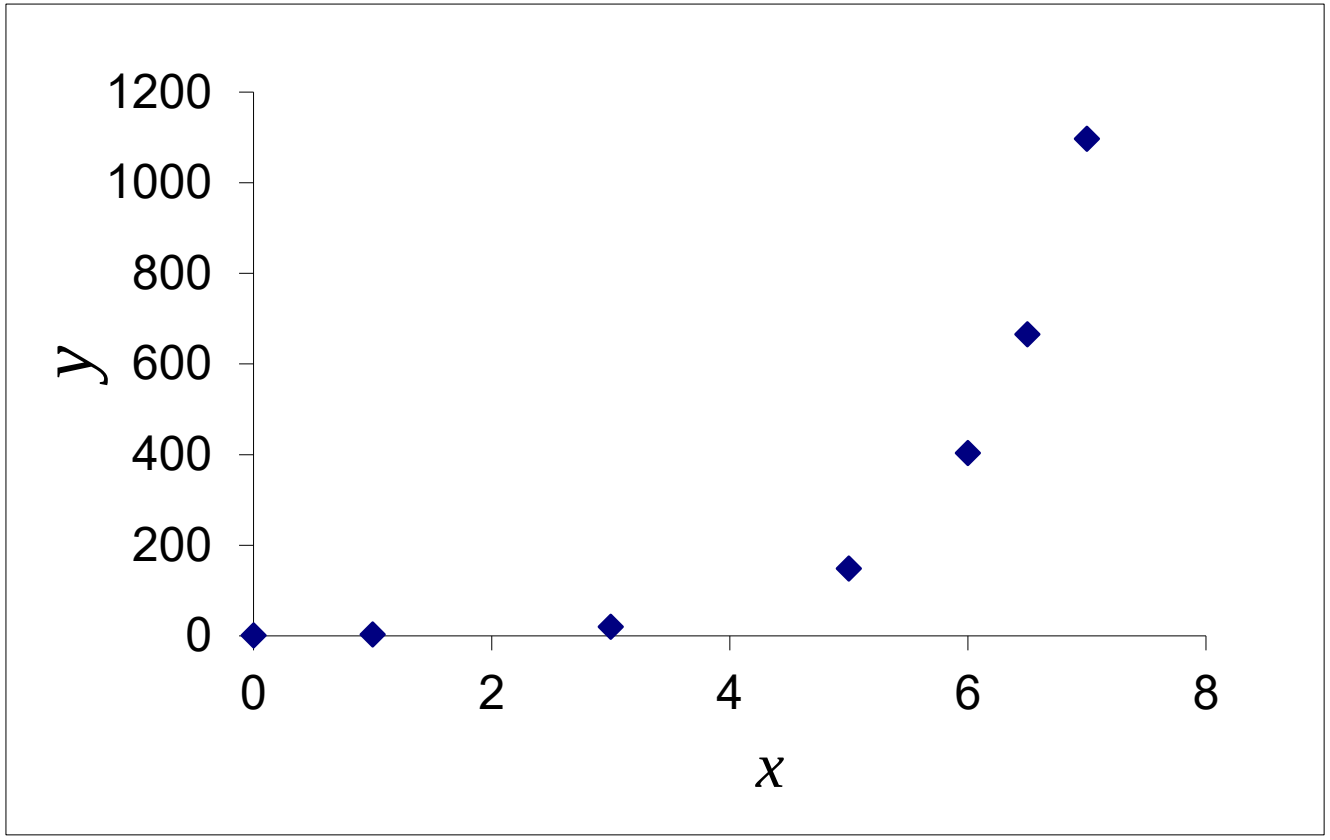
Faktor	Tippjuht	Kesk-astmejuht	Vahed d_i	Vahede ruudud
A	1	5	-4	16
E	4	8	-4	16
B	2	2	0	0
D	5	7	-2	4
H	3	6	-3	9
C	7	1	6	36
F	8	4	4	16
I	6	3	3	9

$$r_S = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)}$$

$$r_S = -0,26$$

Järeldus: tippjuhid ei tea, mismoodi motiveerida keskastmejuhte.

MONOTOONNE SEOS



Demo: astakkorrelatsioon

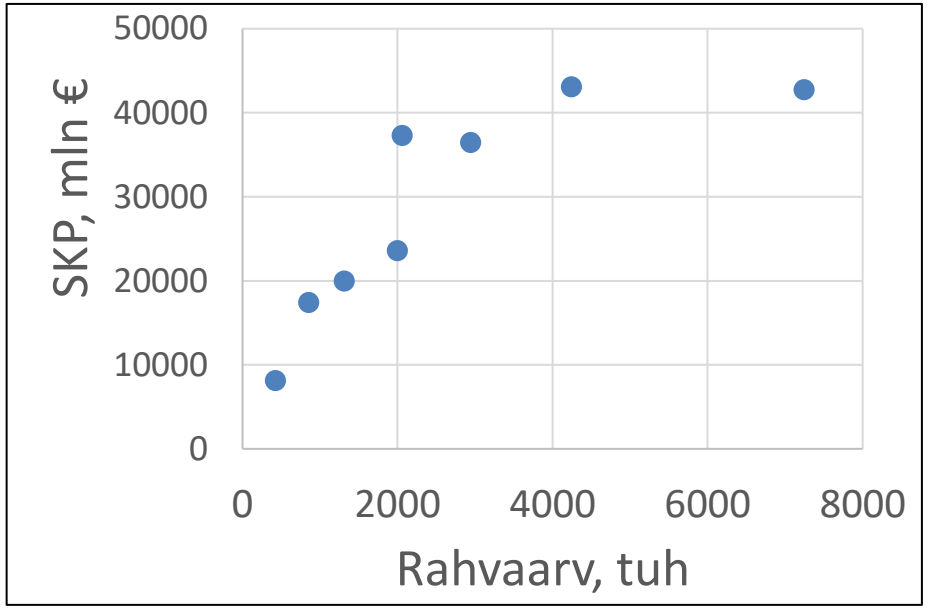
Lineaarne korrelatsioonikordaja on 0,82. Mõõdab lineaarse seose tugevust.

Spearmani korrelatsioonikordaja on 1, mõõdab monotoonse seose tugevust.

NÄIDE: RAHVAARV JA SKP

Suurema rahvaarvuga riikides peaks ka SKP olema suurem. Kas seos on lineaarne?
8 kõige väiksema SKP-ga riiki Euroopas aastal 2014.

Riik	Rahvaarv, tuh	SKP, mln eurot	Rahvaarvu astak	SKP astak	Astakute vahe
Horvaatia	4246,8	43084,8	2	1	1
Bulgaaria	7245,7	42750,9	1	2	-1
Sloveenia	2061,1	37303,2	4	3	1
Leedu	2943,5	36444,4	3	4	-1
Läti	2001,5	23580,9	5	5	0
Eesti	1315,8	19962,7	6	6	0
Küpros	858	17393,7	7	7	0
Malta	425,4	8106,1	8	8	0



Lineaarne korrelatsioonikordaja 0,818.
Spearmani korrelatsioonikordaja 0,952.

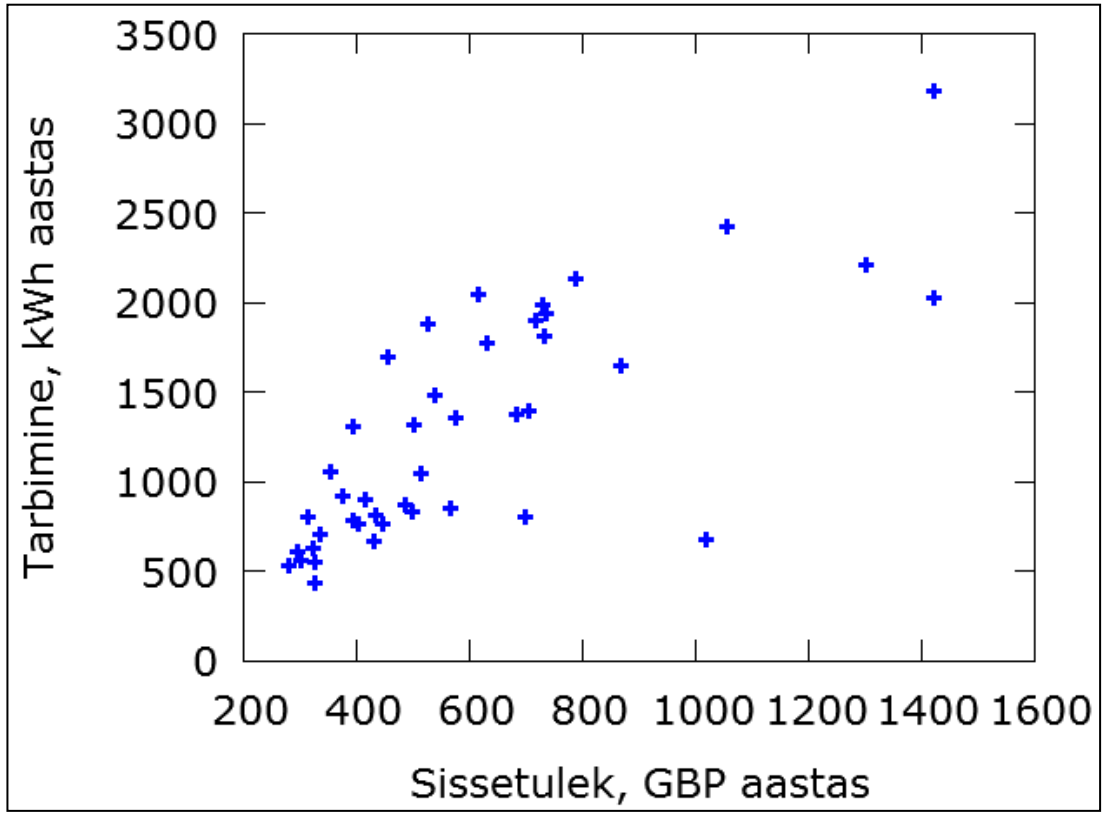
NÄIDE: SISSETULEK JA ELEKTRI TARBIMINE

Elektrienergia tarbimine Suurbritannia erinevates linnades 1930. aastate lõpus. Andmed pärinesid 48-st linnast.

Majapidamise sissetuleku ja tarbimise vahel on positiivne korrelatsioon, $r = 0{,}767$.

Kui majapidamise sissetulek on 1200 GBP aastas, milline oleks elektrienergia tarbimine?

Kuidas seost modelleerida? Järgmine teema: regressioonanalüüs.



Allikas: Houthakker, H. S. 1951. Some Calculations on Electricity Consumption in Great Britain. Journal of the Royal Statistical Society. Series A (General) Vol. 114, 359-371.