

---

**Johdatus tilastotieteeseen**

# **Tilastollinen riippuvuus ja korrelaatio**

# Tilastollinen riippuvuus ja korrelaatio

---

**Tilastollinen riippuvuus, korrelaatio ja regressio**  
**Kahden muuttujan havaintoaineiston kuvaaminen**  
**Pearsonin korrelaatiokertoimen estimointi ja testaus**  
**Järjestyskorrelaatiokertoimet**

# Tilastollinen riippuvuus ja korrelaatio: Mitä opimme?

---

- Tarkastelemme tässä luvussa tilastollisia tutkimusasetelmia, joissa on mukana **kaksi tai useampia muuttujia**.
- Pyrimme vastaamaan seuraaviin kysymyksiin:
  - Miten **kahden** (tai useamman) **muuttujan** samanaikainen tarkastelu vaikuttaa tilastolliseen analyysiin?
  - Miten kahden (tai useamman) muuttujan tilastollista aineistoa **kuvataan**?
  - Mitä tarkoitetaan kahden muuttujan **tilastollisella riippuvuudella** ja miten se eroaa **eksaktista riippuvuudesta**?
  - Mitä tarkoitetaan **korrelaatiolla**?
  - Mikä on **korrelaation** ja **riippuvuuden** suhde?
  - Miten korrelaatio **estimoidaan**?
  - Miten korrelaatiota koskevia **hypoteeseja testataan**?

# Tilastollinen riippuvuus ja korrelaatio: Esitiedot

---

- Esitiedot: ks. seuraavia lukuja:
  - Tilastollisten aineistojen kerääminen ja mittaaminen**
  - Tilastollisten aineistojen kuvaaminen**
  - Estimointi**
  - Estimointimenetelmät**
  - Väliestimointi**
  - Tilastolliset testit**
  - Testit suhdeasteikollisille muuttujille**
- Tarvitset esitietoja myös seuraavista kalvokokoelman **Johdatus todennäköisyyslaskentaan** luvuista:
  - Satunnaismuuttujat ja todennäköisyysjakaumat**
  - Jakaumien tunnusluvut**
  - Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat**
  - Moniulotteisia todennäköisyysjakaumia**

# Tilastollinen riippuvuus ja korrelaatio: Lisätiedot

---

- Johdatus *regressioanalyysiin* esitetään luvussa  
**Johdatus regressioanalyysiin**
- Regressioanalyysia *yhden selittäjän lineaarisen regressiomallin* tapauksessa käsitellään luvussa  
**Yhden selittäjän lineaarinen regressiomalli**
- Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään luvuissa  
**Yleinen lineaarinen malli**  
**Regressiodiagnostiikka**  
**Regressiomallin valinta**  
**Regressioanalyysin erityiskysymyksiä**

# Tilastollinen riippuvuus ja korrelaatio

---

- >> Tilastollinen riippuvuus, korrelaatio ja regressio
  - Kahden muuttujan havaintoaineiston kuvaaminen
  - Pearsonin korrelaatiokertoimen estimointi ja testaus
  - Järjestyskorrelaatiokertoimet

# Tilastollinen riippuvuus, korrelaatio ja regressio

---

## *Avainsanat*

Eksakti riippuvuus

Korrelaatio

Korrelaatiokerroin

Regressioanalyysi

Regressiomalli

Testit korrelaatiokertoimille

Tilastollinen riippuvuus

Usean muuttujan havaintoaineiston  
kuvaaminen

## Tilastollinen riippuvuus, korrelaatio ja regressio

# Muuttujien väliset riippuvuudet tilastollisen tutkimuksen kohteena

---

- Tieteellisen tutkimuksen *tärkeimmät ja mielenkiintoisimmat kysymykset liittyvät* tavallisesti tutkimuksen kohteena olevaa ilmiötä kuvaavien **muuttujien välisiin riippuvuuksiin**.
- Jos tilastollisen tutkimuksen kohteena olevaan ilmiöön liittyy useampia kuin yksi muuttuja, *yhden muuttujan tilastolliset menetelmät* antavat tavallisesti vain *rajoittuneen kuvan* ilmiöstä.
- Sovellusten kannalta ehkä merkittävin osa tilastotiedettä käsittelee kahden tai useamman muuttujan välisten *riippuvuuksien kuvaamista ja mallintamista*.



# Tilastollinen riippuvuus, korrelaatio ja regressio

## Esimerkkejä riippuvuustarkasteluista

---

- Miten työttömyysaste Suomessa (% työvoimasta) *riippuu* BKT:n (bruttokansantuotteen) kasvuvauhdista Suomessa, Suomen viennin volyymista sekä BKT:n kasvuvauhdista muissa EU-maissa ja USA:ssa?
- Miten alkoholin kulutus (1 *per capita* vuodessa) *riippuu* alkoholijuomien hintatasosta, ihmisten käytettävissä olevista tuloista ja alkoholin saatavuudesta?
- Miten todennäköisyys sairastua keuhkosityöpään ( $p$ ) *riippuu* tupakoinnin määrästä ja kestosta?
- Miten vehnän hehtaarisato (t/ha) *riippuu* kesän keskilämpötilasta ja sademäärästä sekä maanmuokkauksesta, lannoituksesta ja tuholaisten torjunnasta?
- Miten betonin lujuus ( $\text{kg/cm}^2$ ) *riippuu* sen kuivumisajasta?
- Miten kemiallisen aineen saanto (%) *riippuu* valmistusprosessissa käytettävästä lämpötilasta?

# Tilastollinen riippuvuus, korrelaatio ja regressio

## Eksakti vs tilastollinen riippuvuus

---

- Tarkastelemme tässä esityksessä yksinkertaisuuden vuoksi pääasiassa *kahden* muuttujan välistä riippuvuutta:
  - (i) Muuttujien välinen *riippuvuus on eksaktia*, jos *toisen arvot voidaan ennustaa tarkasti toisen saamien arvojen perusteella*.
  - (ii) Muuttujien välinen *riippuvuus on tilastollista*, jos niiden välillä *ei ole eksaktia riippuvuutta*, mutta *toisen muuttujan arvoja voidaan käyttää apuna toisen muuttujan arvojen ennustamisessa*.

## Tilastollinen riippuvuus ja korrelaatio

---

- Kahden muuttujan välistä (lineaarista) *tilastollista riippuvuutta* kutsutaan tilastotieteessä tavallisesti **korrelaatioksi**.
- *Korrelaation* eli (lineaarisen) *tilastollisen riippuvuuden voimakkuutta* mittaavia tilastollisia tunnuslukuja kutsutaan **korrelaatiokertoimiksi**.
- Korrelaatiot muodostavat *perustan muuttujien välisten riippuvuuksien ymmärtämiselle*.

## Tilastollinen riippuvuus ja regressio

---

- Vaikka korrelaatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle, riippuvuuksia halutaan tavallisesti *analysoida* myös *tarkemmin*.
- **Regressioanalyysi** on tilastollinen menetelmä, jossa jonkin, ns. *selitettävän muuttujan* tilastollista riippuvuutta joistakin toisista, ns. *selittävistä muuttujista* pyritään mallintamaan **regressiomalliksi** kutsutulla tilastollisella mallilla; ks. lukua **Johdatus regressioanalyysiin**.
- Huomautus:

Tässä luvussa rajoitutaan tarkastelemaan *korrelaatioiden estimointia ja testaamista*.

## Kahden muuttujan havaintoaineiston kuvaaminen

---

- Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useamman muuttujan havaintoaineistojen kuvaamiselle muodostaa tutustuminen **havaintoarvojen jakaumaan**.
- Havaintoarvojen jakaumaa voidaan kuvailla ja esitellä *tiivistemällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:
  - Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valituilla **graafisilla esityksillä**.
  - Havaintoarvojen *jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **otostunnusluvuilla**.

## Kahden muuttujan havaintoaineiston kuvaaminen: Graafiset menetelmät

---

- Koska useampi- kuin kaksiulotteisten kuvioiden tekeminen ei ole käytännössä mahdollista, kolmen tai useamman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että *muuttujia tarkastellaan pareittain*.
- Kahden *järjestys-*, *välimatka-* tai *suhdeasteikoillisen* muuttujan havaittujen arvojen pareja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan **pistediagrammiksi**.
- Huomautus:  
*Monimuuttujamenetelmissä on kehitetty myös sellaisia tilastografiikan menetelmiä, joilla voidaan havainnollistaa useampi- kuin kaksiulotteisia aineistoja.*

## Kahden muuttujan havaintoaineiston kuvaaminen: Tunnusluvut

---

- Usean muuttujan havaintoaineistojen karakteristisia ominaisuuksia voidaan kuvata *muuttujakohtaisilla otostunnusluvuilla*.
- Muuttujakohtaiset otostunnusluvut *eivät* kuitenkaan *voi antaa informaatiota muuttujien välisistä riippuvuuksista*.
- Muuttujien *pareittaisia tilastollisia riippuvuuksia* voidaan *kuvata* sopivasti valitulla **korrelaation mitalla**.

## Kahden muuttujan havaintoaineiston kuvaaminen: Korrelaatio

---

- Tutkittavien muuttujien *mitta-asteikolliset ominaisuudet ohjaavat korrelaation mitan valintaa*:
  - **Välimatka- ja suhdeasteikollisille muuttujille** käytetään tavallisesti **Pearsonin korrelaatiokerrointa**.
  - **Järjestysasteikollisille muuttujille** käytetään tavallisesti **Spearmanin tai Kendallin järjestyskorrelaatiokerrointa**.



## Testit korrelaatiolle

---

- *Satunnaismuuttujien väliseen korrelaatioon* voidaan kohdistaa erilaisia *tilastollisia testejä*.
- Tässä esityksessä tarkastellaan seuraavia *Pearsonin korrelaatiokertoimelle* sopivia testejä:
  - **Yhden otoksen testi korrelaatiokertoimelle**
  - **Korrelaatiokertoimien vertailutesti**
  - **Testi korreloimattomuudelle**
- Tässä esityksessä tarkastellaan seuraavia *Spearmanin ja Kendallin järjestyskorrelaatiokertoimille* sopivia testejä:
  - **Testit korreloimattomuudelle**

# Tilastollinen riippuvuus ja korrelaatio

---

**Tilastollinen riippuvuus, korrelaatio ja regressio**

- >> Kahden muuttujan havaintoaineiston kuvaaminen**
- Pearsonin korrelaatiokertoimen estimointi ja testaus**
- Järjestyskorrelaatiokertoimet**

# Kahden muuttujan havaintoaineiston kuvaaminen

---

## *Avainsanat*

**Aikasarjadiagrammi**

**Aritmeettinen keskiarvo**

**Keskihajonta**

**Korrelaatio**

**Otoskovarianssi**

**Pearsonin otoskorrelaatiokerroin**

**Pistediagrammi**

**Varianssi**

## Kahden muuttujan havaintoaineiston kuvaaminen

# Pistediagrammi

---

- Tarkastellaan tilannetta, jossa tutkimuksen kohteina olevista *havaintoyksiköistä* on mitattu *kahden järjestys-, välimatka- tai suhdeasteikollisen* muuttujan  $x$  ja  $y$  arvot.
- Muuttujien  $x$  ja  $y$  arvojen samaan havaintoyksikköön liittyvien *parien* muodostamaa havaintoaineistoa voidaan kuvata graafisesti *pistediagrammilla*.
- Pistediagrammi sopii erityisesti kahden muuttujan välisen *riippuvuuden* havainnollistamiseen.
- Pistediagrammi on keskeinen työväline *korrelaatio- ja regressioanalyysissä*.

## Pistediagrammi:

### Määritelmä

---

- Olkoot  $x$  ja  $y$  *järjestys-, välimatka- tai suhdeasteikollisia* muuttujia, joiden havaitut arvot ovat

$$x_1, x_2, \dots, x_n$$

$$y_1, y_2, \dots, y_n$$

- Oletetaan lisäksi, että havaintoarvot  $x_i$  ja  $y_i$  liittyvät *samaan havaintoyksikköön* kaikille  $i = 1, 2, \dots, n$ .
- Havaintoarvojen  $x_1, x_2, \dots, x_n$  ja  $y_1, y_2, \dots, y_n$  parien **pistediagrammi** saadaan esittämällä *lukuparit*

$$(x_i, y_i), i = 1, 2, \dots, n$$

pisteinä avaruudessa  $\mathbb{R}^2$ .

# Pistediagrammi: Havainnollistus

---

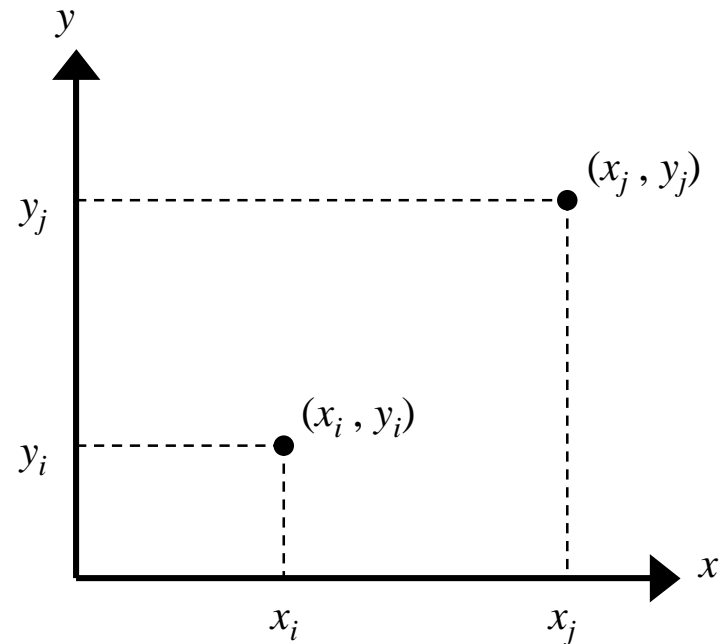
- Kuvio oikealla esittää lukuparien

$(x_i, y_i)$

ja

$(x_j, y_j)$

määrittelemien pisteiden  
esittämistä tasokoordinaatistossa.



## Pistediagrammi:

### 1. esimerkki – 1/2

---

- *Hooken lain* mukaan kierrejousen pituus riippuu *lineaarisesti* jouseen ripustetusta painosta.
- Oikealla on tulokset kokeesta, jossa Hooken lain pätevyyttä tutkittiin ripustamalla jouseen 6 erikokoista painoa.
- Merkitään:

$$(x_i, y_i), i = 1, 2, 3, 4, 5, 6$$

jossa

$$x_i = \text{paino } i$$

$$y_i = \text{jousen pituus, kun painona on } x_i$$

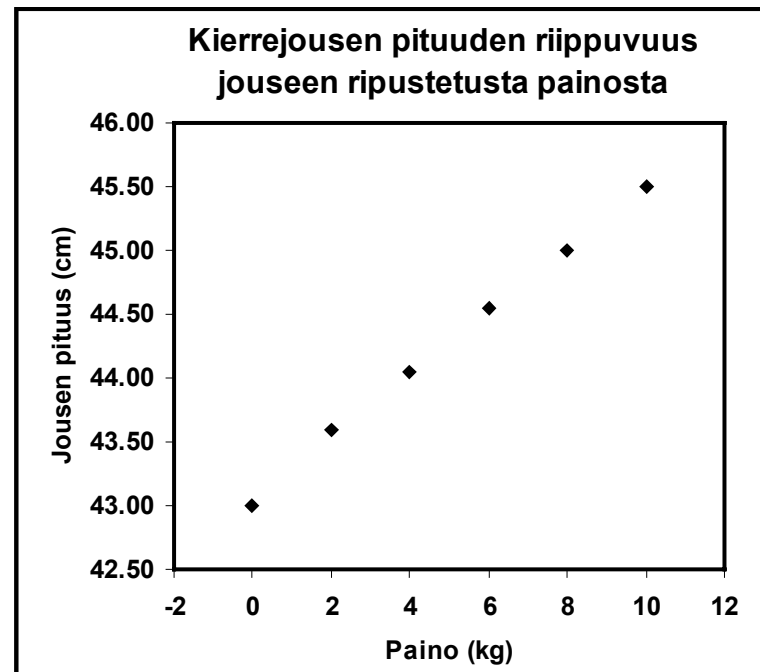
Paino (kg)	Pituus (cm)
0	43.00
2	43.60
4	44.05
6	44.55
8	45.00
10	45.50

## Kahden muuttujan havaintoaineiston kuvaaminen

# Pistediagrammi:

## 1. esimerkki – 2/2

- Pistediagrammi oikealla havainnollistaa koetuloksia graafisesti.
- Ovatko havainnot *sopusoinnussa* Hooken lain kanssa?
- Vastausta tarkastellaan luvuissa **Johdatus regressioanalyysiin ja Yhden selittäjän lineaarinen regressiomalli.**





## Pistediagrammi:

### 2. esimerkki – 1/2

- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:n isän ja heidän poikiensa pituuksien muodostamasta lukuparista

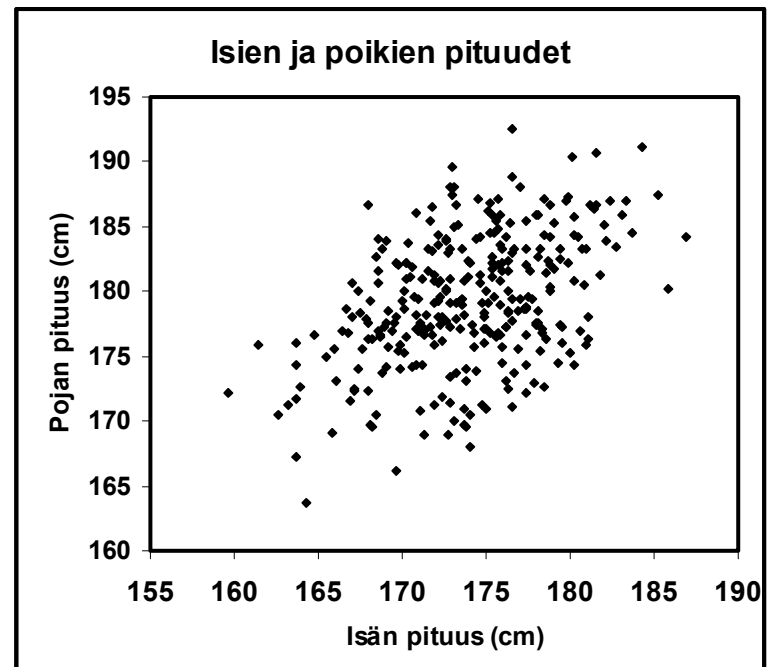
$$(x_i, y_i), i = 1, 2, \dots, 300$$

jossa

$x_i$  = isän  $i$  pituus

$y_i$  = isän  $i$  pojan pituus

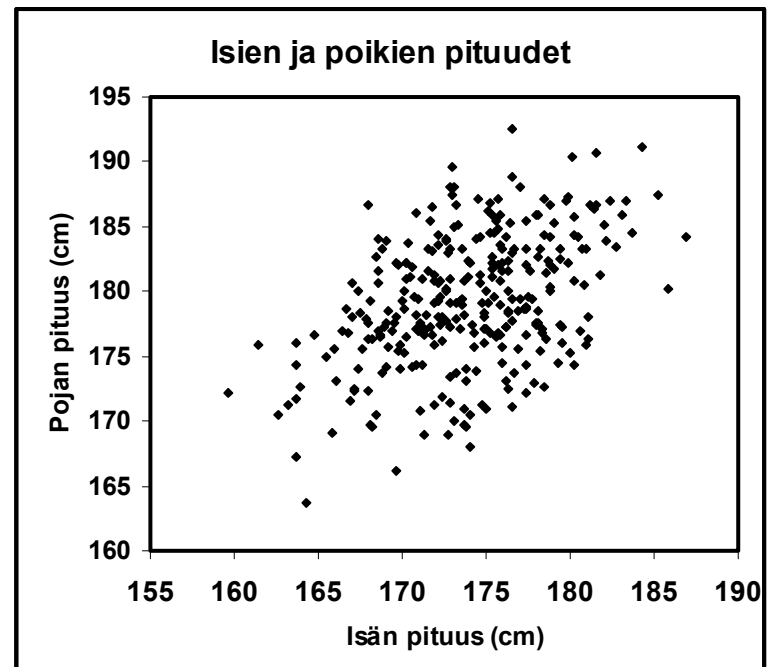
- Ks. pistediagrammia oikealla.



## Pistediagrammi:

### 2. esimerkki – 2/2

- Yhtä pitkällä isillä näyttää olevan monen mittaisia poikia.
- Mutta: Lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkällä isillä ja pitkällä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.
- Tällaisten *tilastollisten riippuvuuksien* analysoimista lineaaristen regressiomallien avulla tarkastellaan luvuissa **Johdatus regressioanalyysiin ja Yhden selittäjän lineaarinen regressiomalli.**



## Pistediagrammi:

### 3. esimerkki – 1/2

---

- Onko keuhkosityöpä yleisempää sellaisissa maissa, joissa tupakoidaan paljon?
- Oikealla on tiedot savukkeiden kulutuksesta ja keuhkosityövän yleisyydestä 10:ssä maassa.
- Havaintoaineisto koostuu 10:stä lukuparista

$$(x_i, y_i), i = 1, 2, \dots, 10$$

jossa

$x_i$  = savukkeiden kulutus maassa  $i$  1930

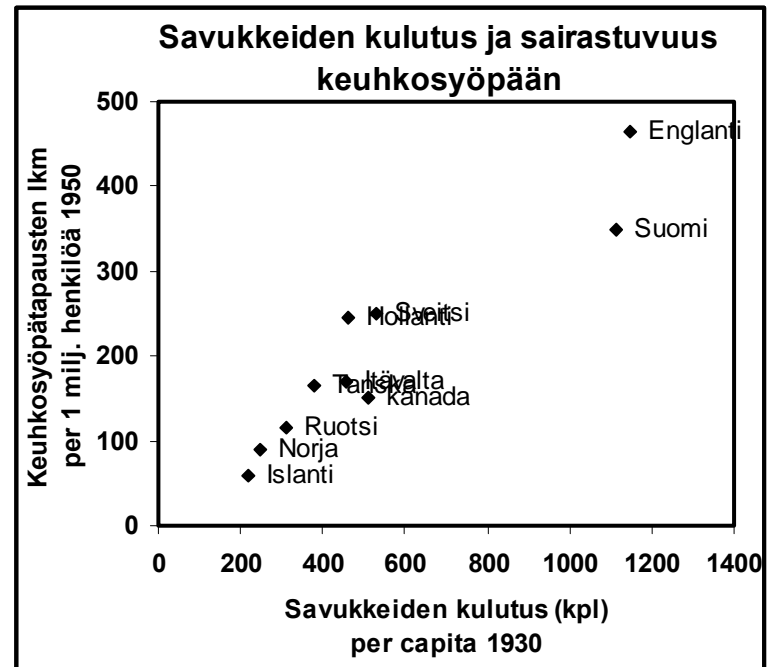
$y_i$  = sairastuvuus keuhkosityöpiin maassa  $i$  1950

Maa	Savukkeiden kulutus (kpl) per capita 1930	Keuhkosityöpätapausten lkm per 1 milj. henkilöä 1950
Islanti	220	58
Norja	250	90
Ruotsi	310	115
Kanada	510	150
Tanska	380	165
Itävalta	455	170
Hollanti	460	245
Sveitsi	530	250
Suomi	1115	350
Englanti	1145	465

## Pistediagrammi:

### 3. esimerkki – 2/2

- Pistediagrammi oikealla havainnollistaa savukkeiden kulutuksen ja keuhkosyövän yleisyyden välistä yhteyttä.
- Sairastuvuus keuhkosyöpään näyttää olevan *keskimäärin* korkeampaa sellaisissa maissa, joissa savukkeiden kulutus on ollut *keskimääräistä* suurempaa.
- Tällaisten *tilastollisten riippuvuuksien* analysoimista lineaaristen regressiomallien avulla tarkastellaan luvussa **Yhden selittäjän lineaarinen regressiomalli.**



## Pistediagrammi:

### 4. esimerkki – 1/2

- Kokeessa tutkittiin betonin vetolujuuden riippuvuutta betonin kuivumisajasta.
- Havaintoaineisto koostuu 21:stä lukuparista

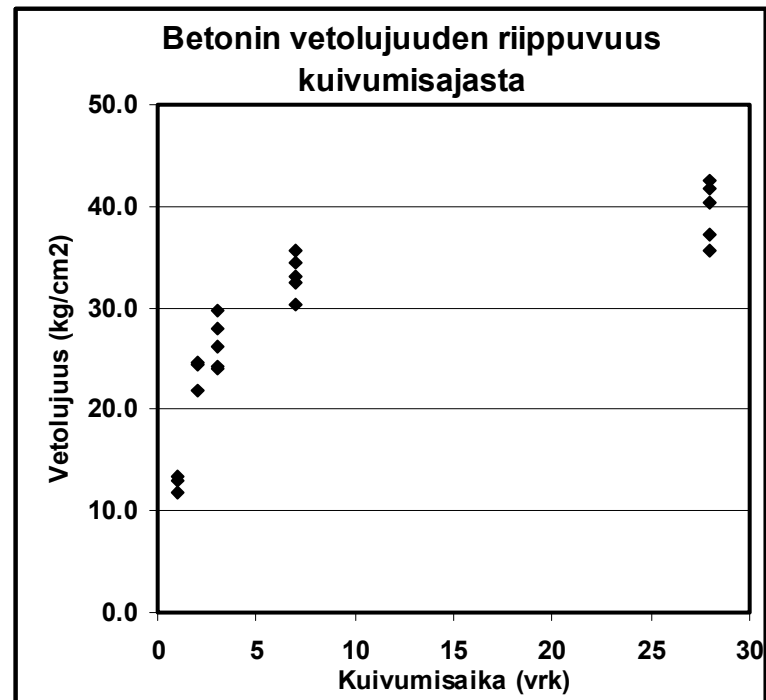
$$(x_i, y_i), i = 1, 2, \dots, 21$$

jossa

$x_i$  = betoniharkon  $i$   
kuivumisaika

$y_i$  = betoniharkon  $i$   
vetolujuus

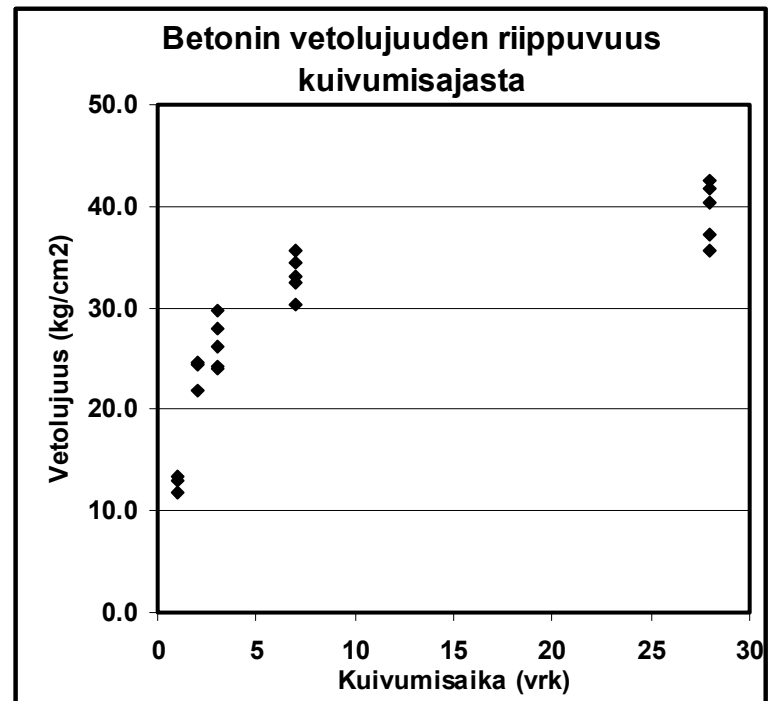
- Ks. pistediagrammia oikealla.



## Pistediagrammi:

### 4. esimerkki – 2/2

- Vetolujuus näyttää riippuvan kuivumisajasta *epälineaarisesti*.
- *Tässä tapauksessa* muuttujien välinen epälineaarinen riippuvuus voidaan kuitenkin *linearisoida*; ks. lukua **Johdatus regressio-analyysiin**.
- Linearisoinnin jälkeen riippuvuutta voidaan analysoida lineaaristen regressiomallien avulla.



## Aikasarjadiagrammi:

### Määritelmä 1/2

---

- Oletetaan, että *järjestys-*, *välimatka-* tai *suhdeasteikollisen* muuttujan  $x$  havaitut arvot

$$x_1, x_2, \dots, x_n$$

muodostavat *aikasarjan*.

- Tällä tarkoitetaan sitä, että havaintoarvot on indeksoitu niin, että indeksit viittaavat *peräkkäisiin* ajanhetkiin, jolloin havainnot ovat *aikajärjestyksessä*.

## Aikasarjadiagrammi:

### Määritelmä 2/2

---

- **Aikasarjadiagrammi** on pistediagrammi, jossa *lukuparit*  
 $(t, x_t)$ ,  $t = 1, 2, \dots, n$   
esitetään pisteinä avaruudessa  $\mathbb{R}^2$ .
- Tavallisesti *peräkkäisiin ajanhetkiin liittyvät pisteet*  
 $(t-1, x_{t-1})$ ,  $(t, x_t)$ ,  $t = 2, 3, \dots, n$   
*yhdistetään* aikasarjadiagrammissa toisiinsa *janoilla*.



# Aikasarjadigrammi: Havainnollistus

---

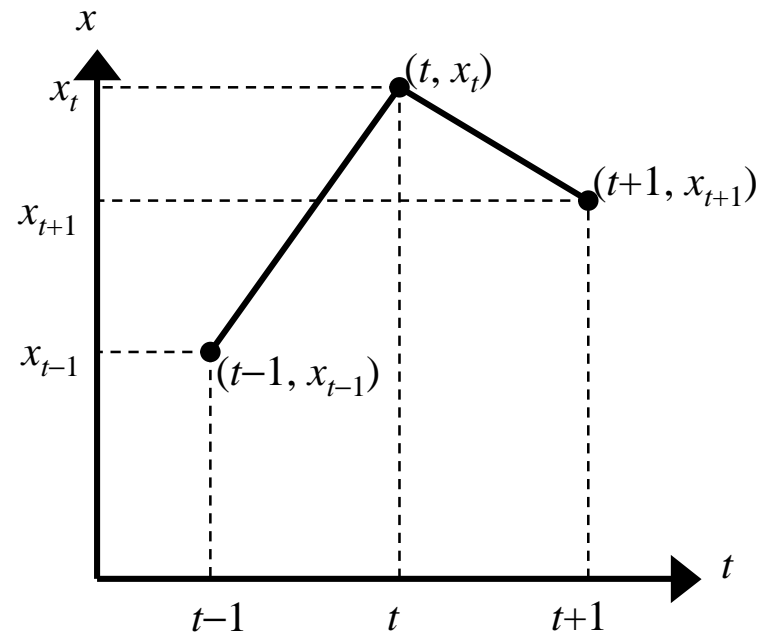
- Kuvio oikealla esittää aikasarjan

$$x_t, t = 1, 2, \dots, n$$

peräkkäisten havaintoarvojen

$$x_{t-1}, x_t, x_{t+1}$$

määrittelemien pisteiden  
esittämistä tasokoordinaatistossa.



# Aikasarjadiagrammi: Esimerkki

- Aikasarjadiagrammi oikealla esittää erään tukkukaupan kk-myyntin arvon vaihtelua.
- Havaintoaineisto koostuu 144:stä lukuparista

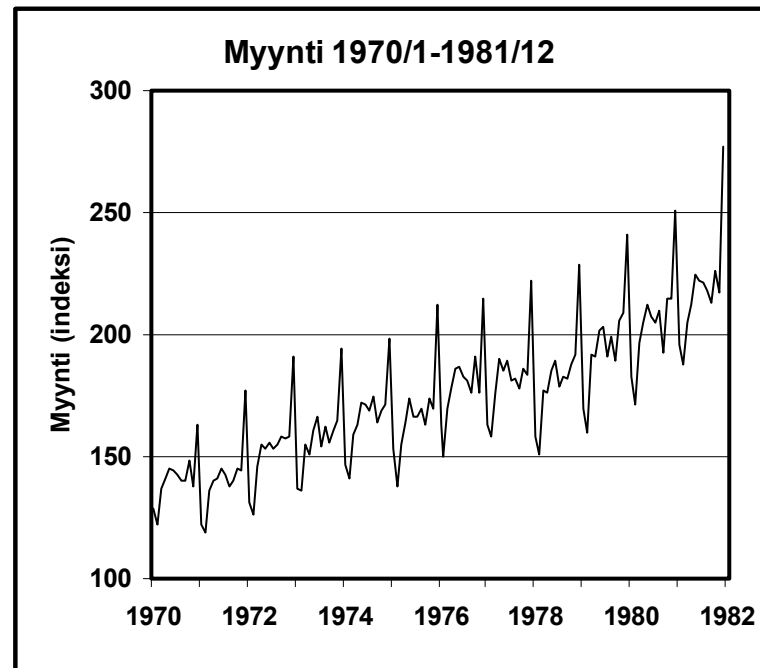
$$(t, x_t)$$

jossa

$t$  = aika (1970/1-1981/12)

$x_t$  = kk-myyntin arvoa  
kuvaava indeksi  
(1960/1 = 100)

- Huomaa, että kk-myyntissä on ollut *nouseva trendi* ja selvää *kausivaihtelua*.



## Kahden muuttujan havaintoaineiston kuvaaminen

# Tunnusluvut

---

- Kahden *välimatka-* tai *suhdeasteikollisen* muuttujan havaintoarvojen parien muodostamaa jakaumaa voidaan *karakterisoida* seuraavilla *tunnusluvuilla*:
  - Havaintoarvojen keskimääräistä *sijaintia* kuvataan **aritmeettisilla keskiarvoilla**.
  - Havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* kuvataan **keskihajonnoilla** tai **(otos-) variansseilla**.
  - Havaintoarvojen (lineaarista) riippuvuutta kuvataan **otoskovarianssilla** ja **otoskorrelaatiokertoimella**.

# Kahden muuttujan havaintoaineiston kuvaaminen

## Havainnot

---

- Olkoot

$$x_1, x_2, \dots, x_n$$

ja

$$y_1, y_2, \dots, y_n$$

*välimatka-* tai *suhdeasteikollisten* muuttujien  $x$  ja  $y$  havaittuja arvoja.

- Oletetaan lisäksi, että havaintoarvot  $x_i$  ja  $y_i$  liittyvät *samaan havaintoyksikköön  $i$  kaikille  $i = 1, 2, \dots, n$ .*

## Aritmeettiset keskiarvot: Määritelmät

---

- Havaintoarvojen  $x_1, x_2, \dots, x_n$  **aritmeettinen keskiarvo** on

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

- Havaintoarvojen  $y_1, y_2, \dots, y_n$  **aritmeettinen keskiarvo** on

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{y_1 + y_2 + \dots + y_n}{n}$$

## Aritmeettiset keskiarvot: Tulkinnat

---

- Havaintoarvojen pareista

$$(x_i, y_i), i = 1, 2, \dots, n$$

laskettujen aritmeettisten keskiarvojen  $\bar{x}$  ja  $\bar{y}$  muodostama lukupari

$$(\bar{x}, \bar{y})$$

on havaintoarvojen parien muodostamien pisteiden *painopiste*.

- Havaintoarvojen aritmeettinen keskiarvo kuvaa havaintoarvojen *keskimääräistä sijaintia*.

## Varianssit: Määritelmät

---

- Havaintoarvojen  $x_1, x_2, \dots, x_n$  (otos-) **varianssi** on

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

jossa  $\bar{x}$  on  $x$ -havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen  $y_1, y_2, \dots, y_n$  (otos-) **varianssi** on

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

jossa  $\bar{y}$  on  $y$ -havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen varianssi mittaa havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havaintoarvojen aritmeettisen keskiarvon suhteen.

## Keskihajonnat:

### Määritelmät

---

- Havaintoarvojen  $x_1, x_2, \dots, x_n$  **keskihajonta** on

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

jossa  $\bar{x}$  on  $x$ -havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen  $y_1, y_2, \dots, y_n$  **keskihajonta** on

$$s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

jossa  $\bar{y}$  on  $y$ -havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen keskihajonta mittaa havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havaintoarvojen aritmeettisen keskiarvon suhteen.



## Otoskovarianssi: Määritelmä

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettu **otoskovarianssi** on

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

jossa

$\bar{x}$  =  $x$ -havaintoarvojen aritmeettinen keskiarvo

$\bar{y}$  =  $y$ -havaintoarvojen aritmeettinen keskiarvo

- Huomaa, että  $x$ - ja  $y$ -havaintoarvojen otoskovarianssit niiden itsensä kanssa ovat niiden *variansseja*:

$$s_{xx} = s_x^2$$

$$s_{yy} = s_y^2$$

## Otoskovarianssi:

### Merkin määräytyminen 1/4

---

- Otoskovarianssin  $s_{xy}$  *merkin* määrää summalauseke

$$(1) \quad \sum (x_i - \bar{x})(y_i - \bar{y})$$

- Summalausekkeen (1)  $i$ . termin

$$(x_i - \bar{x})(y_i - \bar{y})$$

*itseisarvo*

$$|x_i - \bar{x}| |y_i - \bar{y}|$$

on sellaisen *suorakaiteen pinta-ala*, jonka sivujen pituudet ovat

$$|x_i - \bar{x}|$$

ja

$$|y_i - \bar{y}|$$

# Kahden muuttujan havaintoaineiston kuvaaminen

## Otoskovarianssi:

### Merkin määräytyminen 2/4

---

- Summalausekkeen (1)  $i$ . termin

$$(x_i - \bar{x})(y_i - \bar{y})$$

*merkki* määräytyy seuraavalla tavalla:

$$(x_i - \bar{x})(y_i - \bar{y}) \geq 0 \quad \begin{cases} \text{jos } x_i \geq \bar{x} \text{ ja } y_i \geq \bar{y} \\ \text{jos } x_i \leq \bar{x} \text{ ja } y_i \leq \bar{y} \end{cases}$$

$$(x_i - \bar{x})(y_i - \bar{y}) \leq 0 \quad \begin{cases} \text{jos } x_i \geq \bar{x} \text{ ja } y_i \leq \bar{y} \\ \text{jos } x_i \leq \bar{x} \text{ ja } y_i \geq \bar{y} \end{cases}$$

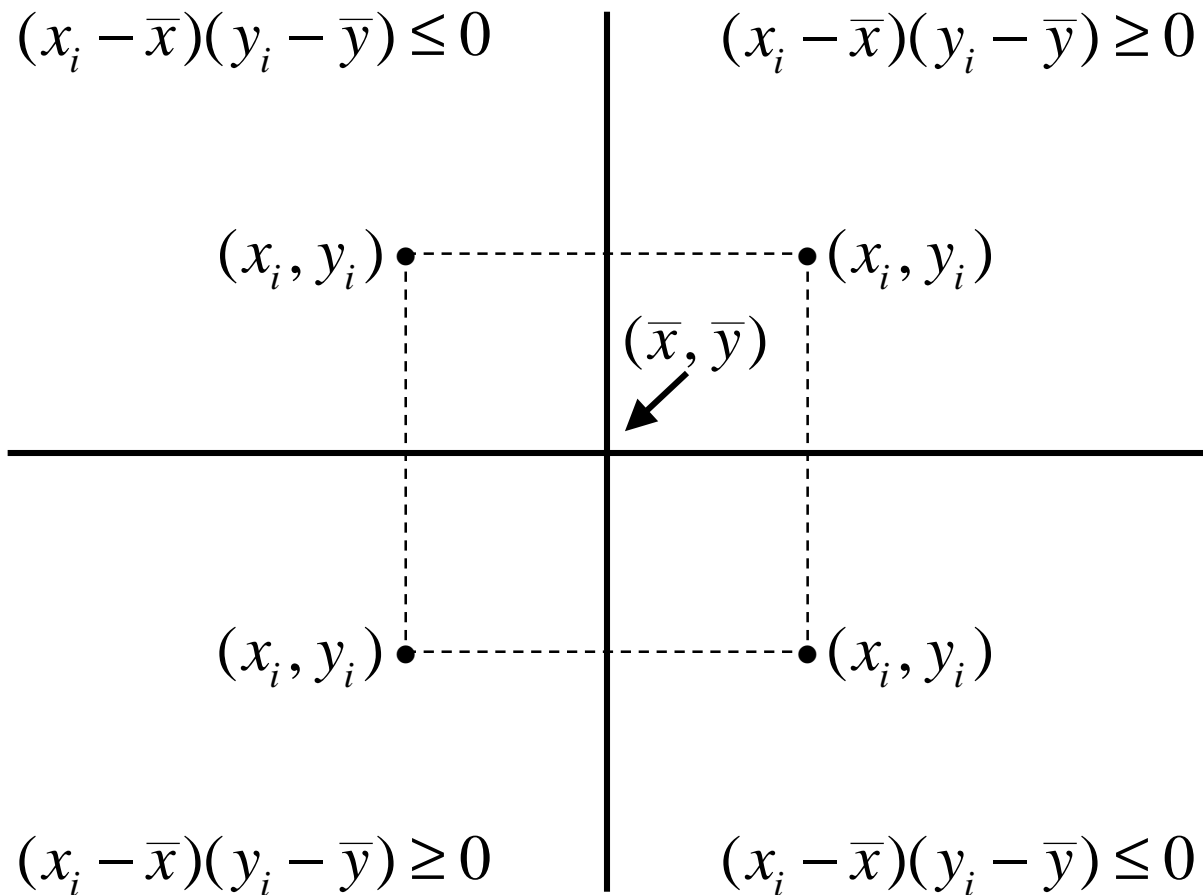
- Merkin määräytymistä voidaan *havainnollistaa geometrisesti* seuraavalla tavalla (ks. kuviota seuraavalla kalvolla):
  - (i) Jaetaan  $xy$ -taso neljään osaan eli *neljännekseen* pisteen  $(\bar{x}, \bar{y})$  kautta piirretyillä koordinaattiakseleiden suuntaisilla suorilla.
  - (ii) Termin  $(x_i - \bar{x})(y_i - \bar{y})$  *merkin* määrää se, *mihin neljännekseen havaintopiste*  $(x_i, y_i)$  *sijoittuu*.

# Kahden muuttujan havaintoaineiston kuvaaminen

## Otoskovarianssi:

### Merkin määräytyminen 3/4

---



# Kahden muuttujan havaintoaineiston kuvaaminen

## Otoskovarianssi:

### Merkin määräytyminen 4/4

---

- Jos positiiviset termit summalausekkeeseen

$$(1) \quad \sum (x_i - \bar{x})(y_i - \bar{y})$$

tuottavien suorakaiteiden yhteenlaskettu pinta-ala on *suurempi* (*pienempi*) kuin negatiiviset termit tuottavien suorakaiteiden yhteenlaskettu pinta-ala, otoskovarianssin  $s_{xy}$  merkki on *positiivinen* (*negatiivinen*).

- Siten otoskovarianssilla on taipumus saada *positiivisia* (*negatiivisia*) arvoja, jos havaintopisteiden muodostama pistepilvi tai -parvi *näyttää nousevalta* (*laskevalta*) *oikealle mentäessä*; ks. *pistediagrammin* ilmeen ja Pearsonin *otoskorrelaatiokertoimen* yhteyttä kuvaavia havainnollistuksia tässä kappaleessa.

## Otoskovarianssi: Tulkinta

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettu otoskovarianssi  $s_{xy}$  mittaa  $x$ - ja  $y$ -havaintoarvojen yhteisvaihtelua niiden aritmeettisten keskiarvojen ympärillä.
- Mitä suurempi on otoskovarianssin  $s_{xy}$  itseisarvo  $|s_{xy}|$  sitä voimakkaampaa on  $x$ - ja  $y$ -havaintoarvojen yhteisvaihtelu.

# Kahden muuttujan havaintoaineiston kuvaaminen

## Otoskovarianssi ja

## Pearsonin otoskorrelaatiokerroin

---

- Otoskovarianssin  $s_{xy}$  avulla voidaan määritellä  $x$ - ja  $y$ -havaintoarvojen *lineaarisen tilastollisen riippuvuuden voimakkuuden mittari*, jota kutsutaan **Pearsonin otoskorrelaatiokertoimeksi**.
- Pearsonin otoskorrelaatiokerroin  $r_{xy}$  saadaan otoskovarianssista  $s_{xy}$  *normeerausoperaatiolla*, jossa  $x$ - ja  $y$ -havaintoarvojen otoskovarianssi  $s_{xy}$  jaetaan  $x$ - ja  $y$ -havaintoarvojen keskihajonnoilla  $s_x$  ja  $s_y$ .

# Kahden muuttujan havaintoaineiston kuvaaminen

## Pearsonin otoskorrelaatiokerroin:

### Määritelmä 1/2

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettu **Pearsonin otoskorrelaatiokerroin** on

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

jossa

$s_{xy}$  =  $x$ - ja  $y$ -havaintoarvojen otoskovarianssi

$s_x$  =  $x$ -havaintoarvojen keskihajonta

$s_y$  =  $y$ -havaintoarvojen keskihajonta



## Kahden muuttujan havaintoaineiston kuvaaminen

# Pearsonin otoskorrelaatiokerroin:

### Määritelmä 2/2

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettu *Pearsonin otoskorrelaatiokerroin* voidaan kirjoittaa myös muotoon

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

jossa

$\bar{x}$  =  $x$ -havaintoarvojen aritmeettinen keskiarvo

$\bar{y}$  =  $y$ -havaintoarvojen aritmeettinen keskiarvo

## Kahden muuttujan havaintoaineiston kuvaaminen

# Pearsonin otoskorrelaatiokerroin:

## Ominaisuuksia

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  lasketulla *Pearsonin otoskorrelaatiokertoimella*  $r_{xy}$  on seuraavat ominaisuudet:
  - (i)  $-1 \leq r_{xy} \leq +1$
  - (ii)  $r_{xy} = \pm 1$ , jos ja vain jos
$$y_i = \alpha + \beta x_i$$
jossa  $\alpha$  ja  $\beta$  ovat reaalisia vakioita ja  $\beta \neq 0$ .
  - (iii) Korrelaatiokertoimella  $r_{xy}$  ja kovarianssilla  $s_{xy}$  on aina *sama merkki*.

## Kahden muuttujan havaintoaineiston kuvaaminen

# Pearsonin otoskorrelaatiokerroin: Tulkinta

---

- Havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  laskettu *Pearsonin otoskorrelaatiokerroin*  $r_{xy}$  mittaa  $x$ - ja  $y$ -havaintoarvojen *lineaarisen tilastollisen riippuvuuden voimakkuutta*.
- Jos  $r_{xy} = \pm 1$ , niin  $x$ - ja  $y$ -havaintoarvojen välillä *on eksakti eli funktionaalinen lineaarinen riippuvuus*, mikä merkitsee sitä, että kaikki havaintopisteet  $(x_i, y_i)$  asettuvat samalle suoralle.
- Jos  $r_{xy} = 0$ , niin  $x$ - ja  $y$ -havaintoarvojen välillä *ei voi olla eksaktia lineaarista riippuvuutta*.
- Vaikka  $r_{xy} = 0$ ,  $x$ - ja  $y$ -havaintoarvojen välillä *saattaa silti olla jopa eksakti epälineaarinen riippuvuus*.

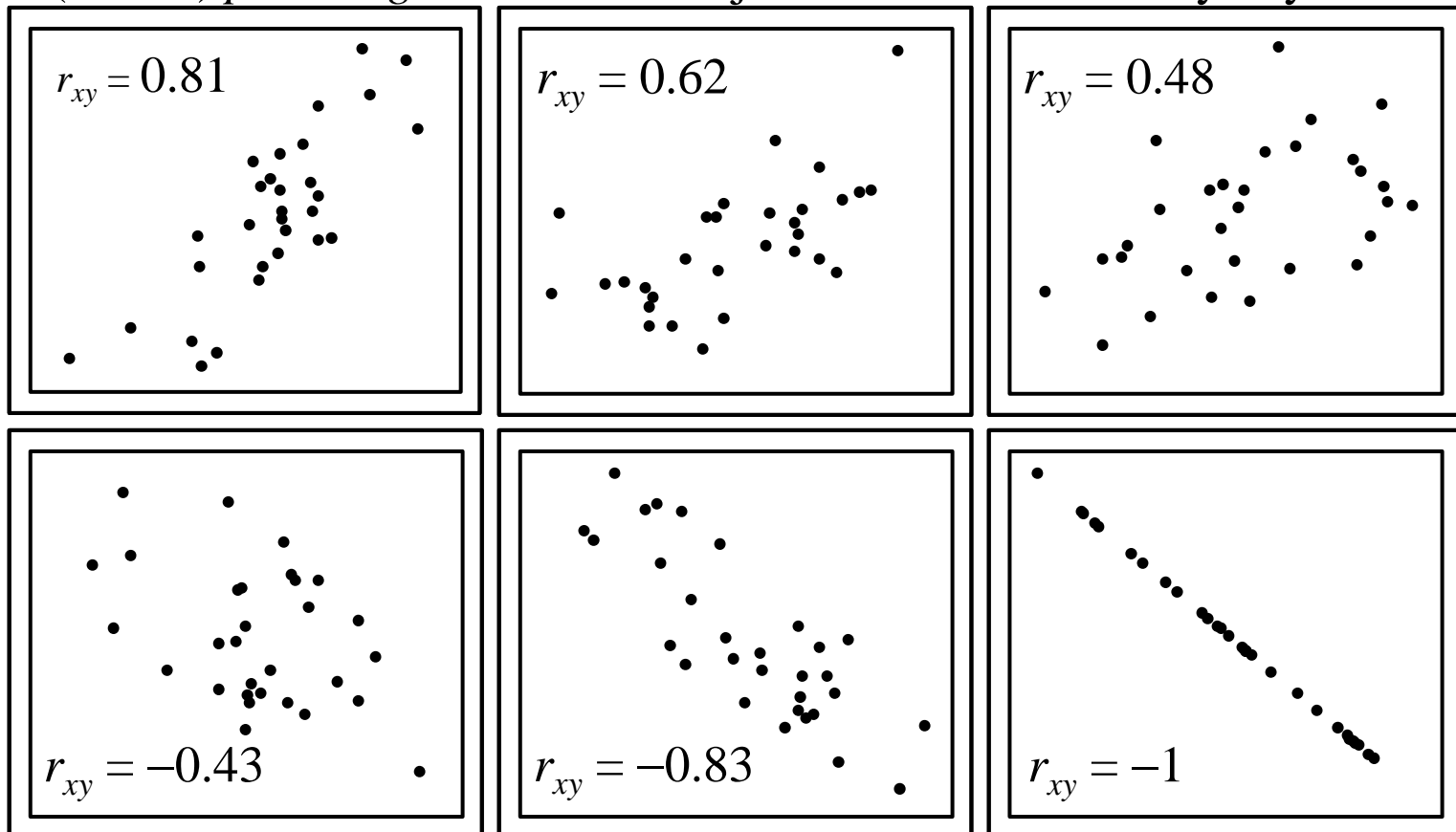
# Kahden muuttujan havaintoaineiston kuvaaminen

## Pearsonin otoskorrelaatiokerroin:

### Havainnollistus

---

- Kuviot alla havainnollistavat kahden muuttujan havaittujen arvojen ( $n = 30$ ) pistediagrammin ilmeen ja korrelaation välistä yhteyttä.



## Kahden muuttujan havaintoaineiston kuvaaminen

# Tunnuslukujen laskeminen 1/4

---

- Oletetaan, että haluamme laskea havaintoarvojen pareista  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  seuraavat otostunnusluvut *käsin* tai käyttämällä *laskinta*:
  - (i) *Aritmeettiset keskiarvot*:  $\bar{x}$ ,  $\bar{y}$
  - (ii) *Varianssit*:  $s_x^2$ ,  $s_y^2$
  - (iii) *Keskihajonnat*:  $s_x$ ,  $s_y$
  - (iv) *Kovarianssi*:  $s_{xy}$
  - (v) *Korrelaatio*:  $r_{xy}$
- Tällöin tarvittavat laskutoimitukset on mukavinta järjestää seuraavalla kalvolla esitettävän kaavion muotoon.

# Kahden muuttujan havaintoaineiston kuvaaminen

## Tunnuslukujen laskeminen 2/4

---

- Määrätään ensin havaintoarvojen *summat*, *neliösummat* ja *tulosumma*:

$i$	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	$x_1$	$y_1$	$x_1^2$	$y_1^2$	$x_1 y_1$
2	$x_2$	$y_2$	$x_2^2$	$y_2^2$	$x_2 y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$x_n$	$y_n$	$x_n^2$	$y_n^2$	$x_n y_n$
Summa	$\sum_{i=1}^n x_i$	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n x_i^2$	$\sum_{i=1}^n y_i^2$	$\sum_{i=1}^n x_i y_i$

## Kahden muuttujan havaintoaineiston kuvaaminen

### Tunnuslukujen laskeminen 3/4

---

- Havaintoarvojen *aritmeettiset keskiarvot*, *variانسsit* ja *kovarianssi* saadaan havaintoarvojen *summista*, *neliösummista* ja *tulosummasta* alla esitetyillä kaavoilla:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad s_x^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad s_y^2 = \frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2 \right)$$

$$s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) \right)$$

## Kahden muuttujan havaintoaineiston kuvaaminen

# Tunnuslukujen laskeminen 4/4

---

- Havaintoarvojen *keskihajonnat* ja *Pearsonin otoskorrelaatiokerroin* saadaan havaintoarvojen *variansseista* ja *kovarianssista* alla esitetyillä kaavoilla:

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

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

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$



# Kahden muuttujan havaintoaineiston kuvaaminen

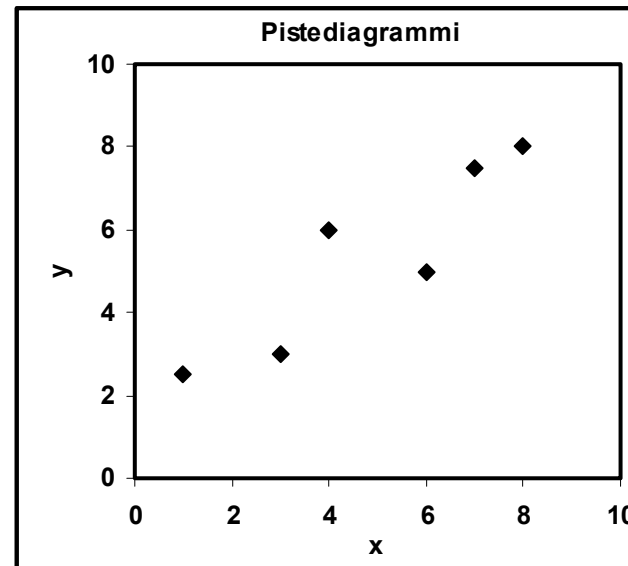
## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 1/5

---

- Taulukossa oikealla on keinotekoisien kahden muuttujan aineiston havaintoarvot ( $n = 6$ ).
- Aineistoa kuvaava *pistediagrammi* on oikealla alhaalla.

$i$	$x$	$y$
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8



# Kahden muuttujan havaintoaineiston kuvaaminen

## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 2/5

---

- Alla olevassa taulukossa on laskettu muuttujien  $x$  ja  $y$  havaittujen arvojen *summat*, *neliösummat* ja *tulosumma*.

$i$	$x$	$y$	$x^2$	$y^2$	$xy$
1	1	2.5	1	6.25	2.5
2	3	3	9	9	9
3	4	6	16	36	24
4	6	5	36	25	30
5	7	7.5	49	56.25	52.5
6	8	8	64	64	64
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>175</b>	<b>196.5</b>	<b>182</b>

- Muuttujien  $x$  ja  $y$  havaittujen arvojen *aritmeettiset keskiarvot*, *otosvarianssit*, *keskihajonnat*, *otoskovarianssi* ja *otoskorrelaatio* voidaan laskea näistä viidestä summasta; ks. seuraavaa kalvoa.

# Kahden muuttujan havaintoaineiston kuvaaminen

## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 3/5

---

- Keskiarvot, otosvarianssit ja otoskovarianssi:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} \times 29 = 4.833$$

$$s_x^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right) = \frac{1}{6-1} \left( 175 - \frac{1}{6} \times 29^2 \right) = 6.967$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{6} \times 32 = 5.333$$

$$s_y^2 = \frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2 \right) = \frac{1}{6-1} \left( 196.5 - \frac{1}{6} \times 32^2 \right) = 5.167$$

$$s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) \right) = \frac{1}{6-1} \left( 182 - \frac{1}{6} \times 29 \times 32 \right) \\ = 5.467$$

# Kahden muuttujan havaintoaineiston kuvaaminen

## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 4/5

---

- Otoskeskihajonnat ja otoskorrelaatio:

$$s_x = \sqrt{s_x^2} = \sqrt{6.967} = 2.639$$

$$s_y = \sqrt{s_y^2} = \sqrt{5.167} = 2.273$$

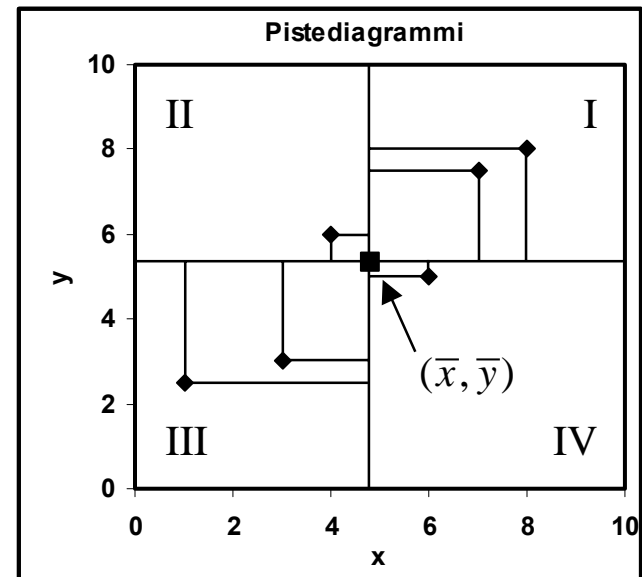
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{5.467}{2.639 \times 2.273} = 0.9112$$

# Kahden muuttujan havaintoaineiston kuvaaminen

## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 5/5

- Kuvioon oikealla on lisätty havaintopisteiden *painopiste*  
 $(\bar{x}, \bar{y}) = (4.833, 5.333)$
- Lisäksi kuvioon on piirretty painopisteen kautta kulkevat koordinaattiakselien suuntaiset suorat sekä *kovarianssin* ja *korrelaation merkin määräytymistä* havainnollistavat suorakaiteet.
- Kovarianssi (ja siten myös korrelaatio) on *positiivinen*, koska I ja III neljänneksen suorakaiteiden yhteenlaskettu pinta-ala on *suurempi* kuin II ja IV neljänneksen suorakaiteiden yhteenlaskettu pinta-ala; ks. tässä kappaleessa esitettyä selitystä *kovarianssin merkin määräytymisestä*.



# Tilastollinen riippuvuus ja korrelaatio

---

**Tilastollinen riippuvuus, korrelaatio ja regressio**

**Kahden muuttujan havaintoaineiston kuvaaminen**

**>> Pearsonin korrelaatiokertoimen estimointi ja testaus**

**Järjestyskorrelaatiokertoimet**

# Pearsonin korrelaatiokertoimen estimointi ja testaus

---

## *Avainsanat*

Fisherin z-muunnos

Korrelaatio

Korrelaatiokertoimen testaaminen

Korrelaatiokertoimien vertailutesti

Korreloimattomuuden testaaminen

Pearsonin korrelaatiokerroin

Pearsonin korrelaatiokertoimen  
estimointi

Pearsonin korrelaatiokertoimen  
luottamusväli

Pearsonin otoskorrelaatiokerroin

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Korrelaation estimointi ja testaus

---

- Tarkastellaan *välimatka-* tai *suhdeasteikollisten* satunnaismuuttujien  $X$  ja  $Y$  Pearsonin (tulomomentti-) korrelaatiokertoimen  $\rho_{XY}$  **estimointia** sekä seuraavia testejä korrelaatiokertoimelle  $\rho_{XY}$  :
  - **Yhden otoksen testi korrelaatiokertoimelle**
  - **Korrelaatiokertoimien vertailutesti**
  - **Korreloimattomuuden testaaminen**
- Lisätietoja moniulotteisista satunnaismuuttujista ja jakaumista: Ks. kalvokokoelman **Johdatus todennäköisyyslaskentaan lukuja Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat ja Moniulotteisia jakaumia.**



## Satunnaismuuttujien kovarianssi ja korrelaatio 1/2

---

- Olkoon

$$(X, Y)$$

*satunnaismuuttujien X ja Y muodostama järjestetty pari.*

- Olkoot

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

*satunnaismuuttujien X ja Y odotusarvot ja*

$$\sigma_X^2 = \text{Var}(X) = D^2(X) = E[(X - \mu_X)^2]$$

$$\sigma_Y^2 = \text{Var}(Y) = D^2(Y) = E[(Y - \mu_Y)^2]$$

*satunnaismuuttujien X ja Y varianssit.*

## Satunnaismuuttujien kovarianssi ja korrelaatio 2/2

---

- Määritellään satunnaismuuttujien  $X$  ja  $Y$  **kovarianssi**  $\sigma_{XY}$  kaavalla

$$\sigma_{XY} = \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- Määritellään satunnaismuuttujien  $X$  ja  $Y$  **korrelaatio**  $\rho_{XY}$  kaavalla

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

jossa

$$\sigma_X = D(X) = \sqrt{\sigma_X^2}$$

$$\sigma_Y = D(Y) = \sqrt{\sigma_Y^2}$$

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Satunnaismuuttujien korrelaatio

---

- Satunnaismuuttujien  $X$  ja  $Y$  korrelaatiota

$$\rho_{XY} = \text{Cor}(X, Y)$$

kutsutaan tavallisesti **Pearsonin (tulomomentti-) korrelaatiokertoimeksi**.

- Pearsonin korrelaatiokerroin  $\rho_{XY}$  mittaa satunnaismuuttujien  $X$  ja  $Y$  *lineaarisen riippuvuuden voimakkuutta*.

## Pearsonin korrelaatiokertoimen estimointi 1/3

---

- Oletetaan, että satunnaismuuttujien  $X$  ja  $Y$  muodostama järjestetty pari  $(X, Y)$  noudattaa *2-ulotteista normaali-jakaumaa*  $N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY})$ , jossa

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

$$\sigma_X^2 = \text{Var}(X)$$

$$\sigma_Y^2 = \text{Var}(Y)$$

$$\rho_{XY} = \text{Cor}(X, Y)$$

- Olkoon

$$(X_i, Y_i), i = 1, 2, \dots, n$$

*riippumaton* satunnaisotos satunnaismuuttujien  $X$  ja  $Y$  muodostaman parin  $(X, Y)$  jakaumasta.

## Pearsonin korrelaatiokertoimen estimointi 2/3

---

- Olkoot

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

tavanomaiset havaintoarvojen pareista  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n$  lasketut *otostunnusluvut*.

## Pearsonin korrelaatiokertoimen estimointi 3/3

---

- Satunnaismuuttujien  $X$  ja  $Y$  *Pearsonin (tulomomentti-) korrelaatiokerroin*

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

voidaan **estimoida** vastaavalla *Pearsonin otoskorrelaatiokertoimella*

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

- **Huomautus:**

Estimaattori  $r_{XY}$  voidaan johtaa sekä *momenttimenetelmällä* että *suurimman uskottavuuden menetelmällä*.

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Fisherin z-muunnos

---

- Määritellään **Fisherin z-muunnos** kaavalla

$$z = f(u) = \frac{1}{2} \log \left( \frac{1+u}{1-u} \right)$$

- Fisherin z-muunnosta soveltamalla **luottamusvälit** ja **testit Pearsonin tulomomenttikorrelaatiokertoimelle**  $\rho_{XY}$  voidaan konstruoida *samanlaisella tekniikalla* kuin luottamusvälit ja testit konstruoidaan *normaalijakauman odotusarvolle*; ks. lukuja **Väliestimointi** ja **Testit suhdeasteikollisille muuttujille**.

## Luottamusväli Pearsonin korrelaatiokertoimelle: Oletukset

---

- Oletetaan, että satunnaismuuttujien  $X$  ja  $Y$  muodostama järjestetty pari  $(X, Y)$  noudattaa *2-ulotteista normaali-jakaumaa*  $N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY})$ , jossa

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

$$\sigma_X^2 = \text{Var}(X)$$

$$\sigma_Y^2 = \text{Var}(Y)$$

$$\rho_{XY} = \text{Cor}(X, Y)$$

- Olkoon

$$(X_i, Y_i), i = 1, 2, \dots, n$$

*riippumaton satunnaisotos* satunnaismuuttujien  $X$  ja  $Y$  muodostaman parin  $(X, Y)$  jakaumasta.



## Luottamusväli Pearsonin korrelaatiokertoimelle: Parametrien estimointi

---

- Estimoidaan 2-ulotteisen normaalijakauman parametrit *tavanomaisilla estimaattoreillaan*:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \qquad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

## Luottamusväli Pearsonin korrelaatiokertoimelle: Fisherin $z$ -muunnos 1/2

---

- Sovelletaan *Fisherin  $z$ -muunnosta*  $z = f(u)$  otoskorrelaatiokertoimeen  $r_{XY}$  :

$$z = f(r_{XY}) = \frac{1}{2} \log \left( \frac{1 + r_{XY}}{1 - r_{XY}} \right)$$

- Voidaan osoittaa, että satunnaismuuttuja  $z$  noudattaa suurissa otoksissa *approksimatiivisesti normaalijakaumaa*:

$$z \sim_a N(\mu_z, \sigma_z^2)$$

jossa

$$\mu_z = \frac{1}{2} \log \left( \frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right) \quad \text{ja} \quad \sigma_z^2 = \frac{1}{n-3}$$

- Approksimaatio on käytännössä *riittävän hyvä*, kun  $n > 25$ .

## Luottamusväli Pearsonin korrelaatiokertoimelle: Fisherin z-muunnos 2/2

---

- Pearsonin korrelaatiokertoimelle  $\rho_{XY}$  voidaan konstruoida *approksimatiivinen luottamusväli Fisherin z-muunnoksen avulla.*

- Olkoon

$$\mu_z = \frac{1}{2} \log \left( \frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right) \quad \text{ja} \quad \sigma_z^2 = \frac{1}{n-3}$$

- Tällöin *standardoitu satunnaismuuttuja*

$$v = \frac{z - \mu_z}{\sigma_z}$$

noudattaa suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa*  $N(0,1)$ :

$$v \sim_a N(0,1)$$

## Luottamusväli Pearsonin korrelaatiokertoimelle: Luottamustaso

---

- Määrätään *approksimatiivinen* **luottamusväli Pearsonin korrelaatiokertoimelle**  $\rho_{XY}$ .
- Valitaan **luottamustasoksi**  
 $1 - \alpha$
- Luottamustason valinta kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää korrelaatiokertoimen  $\rho_{XY}$  oikean arvon.

## Luottamusväli Pearsonin korrelaatiokertoimelle: Luottamuskerroimet 1/2

---

- Olkoon luottamustasona  $(1 - \alpha)$ .
- Valitaan **luottamuskerroin** eli piste

$$+z_{\alpha/2}$$

siten, että se *erottaa standardoidun normaalijakauman*  $N(0, 1)$  *oikealle hänälle todennäköisyysmassan*  $\alpha/2$ .

- Koska normaalijakauma on *symmetrinen*, **luottamuskerroin** eli piste

$$-z_{\alpha/2}$$

*erottaa standardoidun normaalijakauman vasemmalle hänälle todennäköisyysmassan*  $\alpha/2$ .

## Luottamusväli Pearsonin korrelaatiokertoimelle: Luottamuskertoimet 2/2

---

- Siten luottamuskertoimet  $+z_{\alpha/2}$  ja  $-z_{\alpha/2}$  valitaan siten, että

$$\Pr(z \geq +z_{\alpha/2}) = \frac{\alpha}{2}$$

$$\Pr(z \leq -z_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja  $z$  noudattaa standardoitua normaalijakaumaa:

$$z \sim N(0,1)$$

- Huomaa, että

$$\Pr(-z_{\alpha/2} \leq z \leq +z_{\alpha/2}) = 1 - \alpha$$

# Luottamusväli Pearsonin korrelaatiokertoimelle: Parametrin $\mu_z$ luottamusväli 1/2

---

- **Parametrin**

$$\mu_z = \frac{1}{2} \log \left( \frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right)$$

*approksimatiivinen* **luottamusväli luottamustasolla**

$(1 - \alpha)$  on edellä esitetyn nojalla muotoa

$$\left( z - z_{\alpha/2} \frac{1}{\sqrt{n-3}}, z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \right)$$

## Luottamusväli Pearsonin korrelaatiokertoimelle: Parametrin $\mu_z$ luottamusväli 2/2

---

- Parametrin  $\mu_z$  approksimatiivisen luottamusvälin

$$\left( z - z_{\alpha/2} \frac{1}{\sqrt{n-3}}, z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \right)$$

kaavassa

$$z = f(r_{XY}) = \frac{1}{2} \log \left( \frac{1 + r_{XY}}{1 - r_{XY}} \right)$$

$n$  = havaintojen *lukumäärä*

$-z_{\alpha/2}$ ,  $+z_{\alpha/2}$  = luottamustasoon  $(1 - \alpha)$  liittyvät  
*luottamuskertoimet standardoidusta  
normaalijakaumasta  $N(0, 1)$*



## Luottamusväli Pearsonin korrelaatiokertoimelle: Parametrin $\mu_z$ luottamusvälin tulkinta

---

- Parametrin  $\mu_z$  approksimatiivisen luottamusvälin

$$\left( z - z_{\alpha/2} \frac{1}{\sqrt{n-3}}, z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \right)$$

konstruktiosta seuraa, että

$$\Pr \left( z - z_{\alpha/2} \frac{1}{\sqrt{n-3}} \leq \mu_z \leq z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \right) =_a 1 - \alpha$$

- Siten konstruoitu luottamusväli *peittää* parametrin  $\mu_z$  oikean arvon approksimatiivisesti todennäköisyydellä  $(1 - \alpha)$  ja se *ei peitä* parametrin  $\mu_z$  oikeata arvoa approksimatiivisesti todennäköisyydellä  $\alpha$ .

## Luottamusväli Pearsonin korrelaatiokertoimelle: Korrelaatiokertoimen $\rho_{XY}$ luottamusväli 1/2

---

- *Pearsonin korrelaatiokertoimen  $\rho_{XY}$  approksimatiivinen luottamusväli* saadaan parametrin  $\mu_z$  luottamusvälistä ratkaisemalla  $\rho_{XY}$  epäyhtälöketjusta

$$\begin{aligned} z - z_{\alpha/2} \frac{1}{\sqrt{n-3}} &= \frac{1}{2} \log \frac{1+r_{XY}}{1-r_{XY}} - z_{\alpha/2} \frac{1}{\sqrt{n-3}} \\ &\leq \mu_z = \frac{1}{2} \log \frac{1+\rho_{XY}}{1-\rho_{XY}} \\ &\leq z + z_{\alpha/2} \frac{1}{\sqrt{n-3}} = \frac{1}{2} \log \frac{1+r_{XY}}{1-r_{XY}} + z_{\alpha/2} \frac{1}{\sqrt{n-3}} \end{aligned}$$

## Luottamusväli Pearsonin korrelaatiokertoimelle: Korrelaatiokertoimen $\rho_{XY}$ luottamusväli 2/2

---

- **Pearsonin korrelaatiokertoimen  $\rho_{XY}$**   
*approksimatiiviseksi luottamusväliksi* saadaan

$$(lb, ub)$$

jossa

$$lb = \frac{(1 + r_{XY}) - (1 - r_{XY}) \exp\left(+2z_{\alpha/2} / \sqrt{n-3}\right)}{(1 + r_{XY}) + (1 - r_{XY}) \exp\left(+2z_{\alpha/2} / \sqrt{n-3}\right)}$$

on luottamusvälin *alaraja* ja

$$ub = \frac{(1 + r_{XY}) - (1 - r_{XY}) \exp\left(-2z_{\alpha/2} / \sqrt{n-3}\right)}{(1 + r_{XY}) + (1 - r_{XY}) \exp\left(-2z_{\alpha/2} / \sqrt{n-3}\right)}$$

on luottamusvälin *yläraja*.

## Luottamusväli Pearsonin korrelaatiokertoimelle: Korrelaatiokertoimen $\rho_{XY}$ luottamusvälin tulkinta

---

- Pearsonin korrelaatiokertoimen  $\rho_{XY}$  approksimatiivisen luottamusvälin

$$(lb, ub)$$

konstruktiosta seuraa, että

$$\Pr(lb \leq \rho_{XY} \leq ub) =_a 1 - \alpha$$

- Siten konstruoitu luottamusväli *peittää* korrelaatiokertoimen  $\rho_{XY}$  oikean arvon approksimatiivisesti todennäköisyydellä  $(1 - \alpha)$  ja se *ei peitä* korrelaatiokertoimen  $\rho_{XY}$  oikeata arvoa approksimatiivisesti todennäköisyydellä  $\alpha$ .

## Yhden otoksen testi korrelaatiokertoimelle: Testausasetelma

---

- Tarkastellaan *yhden otoksen testiä Pearsonin korrelaatiokertoimelle*.
- Fisherin  $z$ -muunnoksen avulla testi voidaan pukea tavanomaisen *t-testin* muotoon.

## Yhden otoksen testi korrelaatiokertoimelle: Yleinen hypoteesi

---

- *Yleinen hypoteesi*  $H$  :

(i) Oletetaan, että satunnaismuuttujien  $X$  ja  $Y$  muodostama järjestetty pari  $(X, Y)$  noudattaa 2-ulotteista normaalijakaumaa, jonka parametrit ovat

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

$$\sigma_X^2 = \text{Var}(X)$$

$$\sigma_Y^2 = \text{Var}(Y)$$

$$\rho_{XY} = \text{Cor}(X, Y)$$

(ii) Olkoon

$$(X_i, Y_i), i = 1, 2, \dots, n$$

*riippumaton satunnaisotos* satunnaismuuttujien  $X$  ja  $Y$  muodostaman parin  $(X, Y)$  jakaumasta.

# Yhden otoksen testi korrelaatiokertoimelle: Nollahypoteesi ja vaihtoehtoinen hypoteesi

---

- *Nollahypoteesi*  $H_0$  :

$$H_0 : \rho_{XY} = \rho_0$$

- *Vaihtoehtoinen hypoteesi*  $H_1$  :

$$\left. \begin{array}{l} H_1 : \rho_{XY} > \rho_0 \\ H_1 : \rho_{XY} < \rho_0 \end{array} \right\} \text{1-suuntaiset vaihtoehtoiset hypoteesit}$$

$$H_1 : \rho_{XY} \neq \rho_0 \quad \text{2-suuntainen vaihtoehtoinen hypoteesi}$$

# Yhden otoksen testi korrelaatiokertoimelle: Parametrien estimointi

---

- Estimoidaan 2-ulotteisen normaalijakauman parametrit *tavanomaisilla estimaattoreilla*:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \qquad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \qquad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$



## Yhden otoksen testi korrelaatiokertoimelle: Fisherin z-muunnos 1/2

---

- Sovelletaan *Fisherin z-muunnosta*  $z = f(u)$  otoskorrelaatiokertoimeen  $r_{XY}$  :

$$z = f(r_{XY}) = \frac{1}{2} \log \left( \frac{1 + r_{XY}}{1 - r_{XY}} \right)$$

- Satunnaismuuttuja  $z$  noudattaa suurissa otoksissa *approksimatiivisesti normaalijakaumaa*:

$$z \sim_a N(\mu_z, \sigma_z^2)$$

jossa

$$\mu_z = \frac{1}{2} \log \left( \frac{1 + \rho_{XY}}{1 - \rho_{XY}} \right) \quad \text{ja} \quad \sigma_z^2 = \frac{1}{n-3}$$

- Approksimaatio on *riittävän hyvä*, kun  $n > 25$ .

## Yhden otoksen testi korrelaatiokertoimelle: Fisherin z-muunnos 2/2

---

- Testi nollahypoteesille

$$H_0 : \rho_{XY} = \rho_0$$

voidaan perustaa *Fisherin z-muunnoksen* käyttöön.

- *Jos nollahypoteesi  $H_0$  pätee,*

$$E(z) = \frac{1}{2} \log \left( \frac{1 + \rho_0}{1 - \rho_0} \right) = \mu_z^0$$

- Siten satunnaismuuttuja

$$v = \frac{z - \mu_z^0}{\sigma_z}$$

noudattaa *nollahypoteesin  $H_0$  pätiessä* suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa.*

## Yhden otoksen testi korrelaatiokertoimelle: Testisuure ja sen jakauma

---

- Määritellään **testisuure**

$$v = \frac{\frac{1}{2} \log \left( \frac{1 + r_{XY}}{1 - r_{XY}} \right) - \frac{1}{2} \log \left( \frac{1 + \rho_0}{1 - \rho_0} \right)}{\sqrt{\frac{1}{n-3}}}$$

- *Jos nollahypoteesi*

$$H_0 : \rho_{XY} = \rho_0$$

*pätee, testisuure  $v$  noudattaa suurissa otoksissa  
approksimatiivisesti standardoitua normaalijakaumaa:*

$$v \underset{a}{\sim} N(0,1)$$

## Yhden otoksen testi korrelaatiokertoimelle: Testi

---

- Testisuureen  $v$  normaaliarvo  $= 0$ , koska *nollahypoteesin*  $H_0 : \rho_{XY} = \rho_0$  *pätiessä*  
$$E(v) = 0$$
- Siten itseisarvoltaan *suuret* testisuureen  $v$  arvot viittaavat siihen, että *nollahypoteesi*  $H_0$  *ei päde*.
- Nollahypoteesi  $H_0$  *hylätään*, jos testin  $p$ -arvo on *kyllin pieni*.

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Korrelaatiokertoimien vertailutesti:

## Testausasetelma

---

- Tarkastellaan *vertailutestiä Pearsonin korrelaatiokertoimille*.
- Fisherin  $z$ -muunnoksen avulla testi voidaan pukeatavanomaisen *riippumattomien otosten  $t$ -testin* muotoon.

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaatiokertoimien vertailutesti:

### Hypoteesit

---

- *Yleinen hypoteesi*  $H$  :

Oletetaan, että käytössä on kaksi toisistaan riippumatonta yksinkertaista satunnaisotosta perusjoukoista, jotka noudattavat *2-ulotteisia normaalijakaumia*, joiden korrelaatiokertoimet ovat  $\rho_1$  ja  $\rho_2$  .

- *Nollahypoteesi*  $H_0$  :

$$H_0 : \rho_1 = \rho_2 = \rho_0$$

- *Vaihtoehtoinen hypoteesi*  $H_1$  :

$$\left. \begin{array}{l} H_1 : \rho_1 > \rho_2 \\ H_1 : \rho_1 < \rho_2 \end{array} \right\} \text{1-suuntaiset vaihtoehtoiset hypoteesit}$$

$$H_1 : \rho_1 \neq \rho_2 \quad \text{2-suuntainen vaihtoehtoinen hypoteesi}$$

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaatiokertoimien vertailutesti: Parametrien estimointi

---

- Olkoot

$$n_1 \text{ ja } n_2$$

*otoskoot* otoksista 1 ja 2.

- Olkoot

$$r_1 \text{ ja } r_2$$

otoksista 1 ja 2 lasketut *Pearsonin otoskorrelaatiokertoimet*.

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaatiokertoimien vertailutesti: Fisherin z-muunnokset

---

- Olkoon

$$z_k = f(r_k) = \frac{1}{2} \log \left( \frac{1+r_k}{1-r_k} \right)$$

*Fisherin z-muunnos* otoksesta  $k$  lasketulle otoskorrelaatiokertoimelle  $r_k$ ,  $k = 1, 2$ .

- *Jos nollahypoteesi*  $H_0 : \rho_1 = \rho_2 = \rho_0$  pätee, satunnaismuuttuja  $z_k$ ,  $k = 1, 2$  noudattaa suurissa otoksissa *approksimatiivisesti normaalijakaumaa*  $N(\mu_z^0, \sigma_k^2)$ , jossa

$$\mu_z^0 = \frac{1}{2} \log \left( \frac{1+\rho_0}{1-\rho_0} \right) \quad \text{ja} \quad \sigma_k^2 = \frac{1}{n_k - 3}$$

- Approksimaatio on *riittävän hyvä*, kun  $n_1 > 25$  ja  $n_2 > 25$ .



# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaatiokertoimien vertailutesti: Testisuure ja sen jakauma 1/2

---

- Koska satunnaismuuttujat  $z_1$  ja  $z_2$  ovat *riippumattomia*, satunnaismuuttuja

$$v = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

noudattaa *nollahypoteesin*  $H_0 : \rho_1 = \rho_2 = \rho_0$  *pätiessä* suurissa otoksissa *approksimatiivisesti standardoitua normaalijakaumaa*:

$$v \sim_a N(0,1)$$

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaatiokertoimien vertailutesti: Testisuure ja sen jakauma 2/2

---

- Määritellään **testisuure**

$$v = \frac{\frac{1}{2} \log \left( \frac{1+r_1}{1-r_1} \right) - \frac{1}{2} \log \left( \frac{1+r_2}{1-r_2} \right)}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

- *Jos nollahypoteesi*

$$H_0 : \rho_1 = \rho_2 = \rho_0$$

*pätee, testisuure  $v$  noudattaa suurissa otoksissa  
approksimatiivisesti standardoitua normaalijakaumaa:*

$$v \sim_a N(0,1)$$

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Korrelaatiokertoimien vertailutesti:

## Testi

---

- Testisuureen  $v$  *normaaliarvo*  $= 0$ , koska *nollahypoteesin*  $H_0 : \rho_1 = \rho_2 = \rho_0$  *pätiessä*  
$$E(v) = 0$$
- Siten itseisarvoltaan *suuret* testisuureen  $v$  arvot viittaavat siihen, että *nollahypoteesi*  $H_0$  *ei päde*.
- Nollahypoteesi  $H_0$  *hylätään*, jos testin  $p$ -arvo on *kyllin pieni*.

## Pearsonin korrelaatiokertoimen estimointi ja testaus

# Korreloimattomuuden testaaminen:

## Testausasetelma

---

- Monissa tutkimustilanteissa ollaan kiinnostuneita siitä ovatko satunnaismuuttujat  $X$  ja  $Y$  *korreloimattomia* vai ei.
- Huomautuksia:
  - **Satunnaismuuttujien  $X$  ja  $Y$  korreloimattomuudesta ei välttämättä seuraa niiden riippumattomuus, vaikka satunnaismuuttujien  $X$  ja  $Y$  riippumattomuudesta seuraa aina niiden korreloimattomuus.**
  - Jos satunnaismuuttujat  $X$  ja  $Y$  noudattavat *2-ulotteista normaali-jakaumaa*, satunnaismuuttujien  $X$  ja  $Y$  *korreloimattomuudesta* seuraa niiden *riippumattomuus*.
  - Monissa tutkimusasetelmissa toivotaan, että korreloimattomuus-oletus *tulee* testissä *hylätyksi*.

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korreloimattomuuden testaaminen:

### Yleinen hypoteesi

---

- *Yleinen hypoteesi*  $H$  :

(i) Oletetaan, että satunnaismuuttujien  $X$  ja  $Y$  järjestetty pari  $(X, Y)$  noudattaa *2-ulotteista normaalijakaumaa*, jonka parametrit ovat

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

$$\sigma_X^2 = \text{Var}(X)$$

$$\sigma_Y^2 = \text{Var}(Y)$$

$$\rho_{XY} = \text{Cor}(X, Y)$$

(ii) Olkoon

$$(X_i, Y_i), i = 1, 2, \dots, n$$

*riippumaton satunnaisotos* satunnaismuuttujien  $X$  ja  $Y$  muodostaman parin  $(X, Y)$  jakaumasta.

## Korreloimattomuuden testaaminen:

### Nollahypoteesi ja vaihtoehtoinen hypoteesi

---

- *Nollahypoteesi*  $H_0$  :

$$H_0 : \rho_{XY} = 0$$

- *Vaihtoehtoinen hypoteesi*  $H_1$  :

$$H_1 : \rho_{XY} > 0$$

$$H_1 : \rho_{XY} < 0$$

} 1-suuntaiset vaihtoehtoiset hypoteesit

$$H_1 : \rho_{XY} \neq 0$$

2-suuntainen vaihtoehtoinen hypoteesi

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korreloimattomuuden testaaminen: Parametrien estimointi

---

- Estimoidaan 2-ulotteisen normaalijakauman parametrit *tavanomaisilla estimaattoreilla*:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korreloimattomuuden testaaminen: Testisuure ja sen jakauma

---

- Määritellään ***t***-testisuure

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

- *Jos nollahypoteesi*

$$H_0 : \rho_{XY} = 0$$

*pätee*, testisuure *t* noudattaa Studentin *t*-jakaumaa, jonka vapausasteluku on  $n - 2$ :

$$t \sim t(n-2)$$



# Pearsonin korrelaatiokertoimen estimointi ja testaus

## Korrelaamattomuuden testaaminen: Testi

---

- Testisuureen  $t$  normaaliarvo  $= 0$ , koska *nollahypoteesin*  $H_0 : \rho_{XY} = 0$  *pätiessä*  
$$E(t) = 0$$
- Siten itseisarvoltaan *suuret* testisuureen  $t$  arvot viittaavat siihen, että *nollahypoteesi*  $H_0$  *ei päde*.
- Nollahypoteesi  $H_0$  *hylätään*, jos testin  $p$ -arvo on *kyllin pieni*.

# Tilastollinen riippuvuus ja korrelaatio

---

**Tilastollinen riippuvuus, korrelaatio ja regressio**

**Kahden muuttujan havaintoaineiston kuvaaminen**

**Pearsonin korrelaatiokertoimen estimointi ja testaus**

**>> Järjestyskorrelaatiokertoimet**

# Järjestyskorrelaatiokertoimet

---

## *Avainsanat*

Järjestyskorrelaatiokertoimet

Kendallin järjestyskorrelaatiokerroin

Korrelaatio

Korreloimattomuuden testaaminen

Spearmanin järjestyskorrelaatiokerroin

## Korreloimattomuuden testaaminen järjestysasteikollisilla muuttujilla

---

- Tarkastellaan *korrelaatiokertoimen määrittelemistä ja korreloimattomuuden testaamista järjestysasteikollisille muuttujille*.
- Tarkastelun kohteena ovat seuraavat *järjestyskorrelaatiokertoimet*:
  - **Spearmanin järjestyskorrelaatiokerroin**
  - **Kendallin järjestyskorrelaatiokerroin**
- Tarkasteltavat järjestyskorrelaatiokertoimet ja testit sopivat myös *välimatka- ja suhdeasteikollisille muuttujille*.

## Spearmanin järjestyskorrelaatiokerroin: Kertoimen idea

---

- *Spearmanin järjestyskorrelaatiokerroin  $\rho_S$  mittaa kahden muuttujan havaintoarvojen suuruusjärjestyksien yhteensopivuutta.*
- *Spearmanin järjestyskorrelaatiokerroin sopii järjestys-, välimatka- ja suhdeasteikollisille muuttujille.*
- *Spearmanin järjestyskorrelaatiokertoimella on samantapaiset ominaisuudet kuin Pearsonin otoskorrelaatiokertoimella.*

## Spearmanin järjestyskorrelaatiokerroin: Määritelmä 1/3

---

- Olkoot  $X_1, X_2, \dots, X_n$  ja  $Y_1, Y_2, \dots, Y_n$  järjestys-, välimatka- tai suhdeasteikollisten satunnaismuuttujien  $X$  ja  $Y$  havaittuja arvoja.
- Oletetaan lisäksi, että havainnot  $X_i$  ja  $Y_i$  liittyvät *samaan havaintoyksikköön kaikille  $i = 1, 2, \dots, n$ .*
- *Järjestetään* sekä  $X$ - että  $Y$ -muuttujan havaitut arvot *suuruusjärjestykseen* pienimmästä suurimpaan.

## Spearmanin järjestyskorrelaatiokerroin: Määritelmä 2/3

---

- Liitetään sekä  $X$ - että  $Y$ -muuttujan havaittuihin arvoihin niiden *suuruusjärjestyksien* mukaiset järjestysnumerot:

$R(X_i)$  = havainnon  $X_i$  järjestysnumero parissa  $i$

$R(Y_i)$  = havainnon  $Y_i$  järjestysnumero parissa  $i$

sekä määrittellään erotukset

$$D_i = R(X_i) - R(Y_i)$$

- Muuttujien  $X$  ja  $Y$  havaituille arvoille voidaan määrittellä *järjestyskorrelaatiokerroin* erotuksien  $D_i$  avulla.

## Spearmanin järjestyskorrelaatiokerroin: Määritelmä 3/3

---

- Määritellään **Spearmanin järjestyskorrelaatiokerroin**  $\rho_S$  eli *Spearmanin rho* kaavalla

$$\rho_S = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n^3 - n}$$

- Spearmanin järjestyskorrelaatiokerroin  $\rho_S$  voidaan laskea myös soveltamalla *Pearsonin otoskorrelaatiokertoimen* kaavaa muuttujien  $X$  ja  $Y$  havaittujen arvojen pareja  $(X_i, Y_i)$  vastaaviin järjestyslukujen eli *rankien* pareihin

$$(R(X_i), R(Y_i))$$



## Spearmanin järjestyskorrelaatiokerroin: Ominaisuudet 1/2

---

- Spearmanin järjestyskorrelaatiokertoimella  $\rho_S$  on kaikki *hyvältä korrelaation mitalta vaadittavat ominaisuudet*:

(i)  $-1 \leq \rho_S \leq +1$

- (ii) Jos muuttujien  $X$  ja  $Y$  havaittujen arvojen järjestysnumerot ovat jokaisessa havaintoparissa *samat*,

$$\rho_S = +1$$

- (iii) Jos muuttujien  $X$  ja  $Y$  havaittujen arvojen järjestysnumerot liittyvät toisiinsa *täysin satunnaisesti*,

$$\rho_S \approx 0$$

Jos  $\rho_S = 0$ , sanotaan, että muuttujat  $X$  ja  $Y$  ovat *korreloimattomia*.

## Spearmanin järjestyskorrelaatiokerroin: Ominaisuudet 2/2

---

- (iv) Jos sekä *suuret* muuttujien  $X$  ja  $Y$  järjestysnumerot että *pienet* muuttujien  $X$  ja  $Y$  järjestysnumerot liittyvät havaintopareissa  $(X_i, Y_i)$  toisiinsa, kertoimella  $\rho_S$  on taipumus saada *positiivisia* arvoja.
- (v) Jos *suuret* ja *pienet* muuttujien  $X$  ja  $Y$  järjestysnumerot liittyvät havaintopareissa  $(X_i, Y_i)$  toisiinsa, kertoimella  $\rho_S$  on taipumus saada *negatiivisia* arvoja.

## Spearmanin järjestyskorrelaatiokerroin: Korreloimattomuuden testaaminen 1/2

---

- Määritellään *t*-testisuure

$$z = \sqrt{n-2} \frac{\rho_s}{\sqrt{1-\rho_s^2}}$$

- *Jos nollahypoteesi*

$$H_0 : \text{Cor}(X, Y) = 0$$

*pätee, testisuure  $z$  noudattaa suurissa otoksissa  
approksimatiivisesti standardoitua normaalijakaumaa:*

$$z \sim_a N(0,1)$$

## Spearmanin järjestyskorrelaatiokerroin: Korreloimattomuuden testaaminen 2/2

---

- Testisuure  $z$  noudattaa *suurissa otoksissa approksimatiivisesti* standardoitua normaalijakaumaa  $N(0, 1)$ , jos *nollahypoteesi*  $H_0$  pätee.
- Approksimaatio on melko hyvä jo, kun  $n > 10$  ja riittävä lähes kaikkiin tarkoituksiin, kun  $n > 30$ .
- Testisuureen  $z$  normaaliarvo = 0, koska *nollahypoteesin*  $H_0$  *pätiessä*  
$$E(z) = 0$$
- Siten itseisarvoltaan *suuret* testisuureen  $z$  arvot viittaavat siihen, että *nollahypoteesi*  $H_0$  *ei päde*.
- Nollahypoteesi  $H_0$  *hylätään*, jos testin  $p$ -arvo on *kyllin pieni*.

## Kendallin järjestyskorrelaatiokerroin: Kertoimen idea

---

- *Kendallin järjestyskorrelaatiokerroin  $\tau$  mittaa kahden muuttujan havaintoarvojen suuruusjärjestyksien yhteensopivuutta.*
- *Kendallin järjestyskorrelaatiokerroin sopii järjestys-, välimatka- ja suhdeasteikollisille muuttujille.*
- *Kendallin järjestyskorrelaatiokertoimella on samantapaiset ominaisuudet kuin Pearsonin otoskorrelaatiokertoimella.*

## Kendallin järjestyskorrelaatiokerroin:

### Määritelmä 1/3

---

- Olkoot  $X_1, X_2, \dots, X_n$  ja  $Y_1, Y_2, \dots, Y_n$  järjestys-, välimatka- tai suhdeasteikollisten satunnaismuuttujien  $X$  ja  $Y$  havaittuja arvoja.
- Oletetaan lisäksi, että havainnot  $X_i$  ja  $Y_i$  liittyvät *samaan havaintoyksikköön kaikille*  $i = 1, 2, \dots, n$ .
- Järjestetään parit  $(X_i, Y_i)$  muuttujan  $X$  havaittujen arvojen mukaan *suuruusjärjestykseen* pienimmästä suurimpaan.
- Kendallin järjestyskorrelaatiokerroin perustuu tunnuslukuun, joka mittaa muuttujan  $Y$  arvojen *epäjärjestystä* muuttujan  $X$  arvoihin nähden.

## Kendallin järjestyskorrelaatiokerroin:

### Määritelmä 2/3

---

- *Järjestetään* parit  $(X_i, Y_i)$ ,  $i = 1, 2, \dots, n$  muuttujan  $X$  havaittujen arvojen mukaan siten, että ensimmäiseksi tulee pari, jossa muuttujan  $X$  arvo on pienin ja viimeiseksi pari, jossa muuttujan  $X$  arvo on suurin.
- Olkoon  $(X_k, Y_k)$  järjestetykseen asetetuista pareista numero  $k$ .
- Määritellään havaintoarvoon  $Y_k$  liittyvät *epäjärjestyspisteet*

$$S_{kl}, l = k + 1, k + 2, \dots, n, k = 1, 2, \dots, n - 1$$

seuraavalla tavalla:

$$S_{kl} = +1, \text{ jos } Y_l > Y_k$$

$$S_{kl} = -1, \text{ jos } Y_l < Y_k$$

## Kendallin järjestyskorrelaatiokerroin: Määritelmä 3/3

---

- Muuttujan  $Y$  arvojen *epäjärjestysmitta*  $S$  muuttujan  $X$  arvojen suhteen määritellään kaavalla

$$S = \sum_{k=1}^{n-1} \sum_{l=k+1}^n S_{kl}$$

- Määritellään **Kendallin järjestyskorrelaatiokerroin**  $\tau$  eli *Kendallin tau* kaavalla

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$



## Kendallin järjestyskorrelaatiokerroin: Ominaisuudet 1/2

---

- Kendallin järjestyskorrelaatiokertoimella  $\tau$  on kaikki *hyvältä korrelaation mitalta vaadittavat ominaisuudet*:
  - (i)  $-1 \leq \tau \leq +1$
  - (ii) Jos muuttujien  $X$  ja  $Y$  havaittujen arvojen järjestysnumerot ovat jokaisessa havaintoparissa *samat*,
$$\tau = +1$$
  - (iii) Jos muuttujien  $X$  ja  $Y$  havaittujen arvojen järjestysnumerot liittyvät toisiinsa *täysin satunnaisesti*,
$$\tau \approx 0$$

Jos  $\tau = 0$ , sanotaan, että muuttujat  $X$  ja  $Y$  ovat *korreloimattomia*.

## Kendallin järjestyskorrelaatiokerroin: Ominaisuudet 2/2

---

- (iv) Jos sekä *suuret* muuttujien  $X$  ja  $Y$  järjestysnumerot että *pienet* muuttujien  $X$  ja  $Y$  järjestysnumerot liittyvät havaintopareissa  $(X_i, Y_i)$  toisiinsa, kertoimella  $\tau$  on taipumus saada *positiivisia* arvoja.
- (v) Jos *suuret* ja *pienet* muuttujien  $X$  ja  $Y$  järjestysnumerot liittyvät havaintopareissa  $(X_i, Y_i)$  toisiinsa, kertoimella  $\tau$  on taipumus saada *negatiivisia* arvoja.

# Kendallin järjestyskorrelaatiokerroin: Korreloimattomuuden testaaminen 1/2

---

- Määritellään **testisuure**

$$z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n+1)}}}$$

- *Jos nollahypoteesi*

$$H_0 : \text{Cor}(X, Y) = 0$$

*pätee, testisuure  $z$  noudattaa suurissa otoksissa  
approksimatiivisesti standardoitua normaalijakaumaa:*

$$z \sim_a N(0,1)$$

## Kendallin järjestyskorrelaatiokerroin: Korreloimattomuuden testaaminen 2/2

---

- Testisuure  $z$  noudattaa *suurissa otoksissa approksimatiivisesti* standardoitua normaalijakaumaa  $N(0, 1)$ , jos *nollahypoteesi*  $H_0$  pätee.
- Approksimaatio on melko hyvä jo, kun  $n > 10$  ja riittävä lähes kaikkiin tarkoituksiin, kun  $n > 30$ .
- Testisuureen  $z$  normaaliarvo  $= 0$ , koska *nollahypoteesin*  $H_0$  *pätiessä*  
$$E(z) = 0$$
- Siten itseisarvoltaan *suuret* testisuureen  $z$  arvot viittaavat siihen, että *nollahypoteesi*  $H_0$  *ei päde*.
- Nollahypoteesi  $H_0$  *hylätään*, jos testin  $p$ -arvo on *kyllin pieni*.