

# Kapitel 10

## Simpel korrelation

Peter Tibert Stoltze  
stat@peterstoltze.dk

Elementær statistik  
F2011

1 / 14

## Indledning

- ▶ Korrelation mellem to variable betyder, at en ændring i den ene variabel giver en forudsigelig (mere eller mindre) ændring i den anden variabel
- ▶ En høj grad af korrelation kan ikke bruges til at postulere nogen årsagssammenhæng (kausalitet)
- ▶ Ved beregning af korrelation er det ikke nødvendigt at tage stilling til, hvilken variabel der er afhængig, og hvilken der er uafhængig — dette er heller ikke altid helt oplagt. . .
- ▶ Vi vil se på definition, egenskaber, beregning, fortolkning og signifikanstest for
  - ▶ Pearsons korrelationskoefficient  $r$
  - ▶ Spearmans rangkorrelationskoefficient  $\rho$

2 / 14

## Pearsons $r$

- ▶ Ved korrelationen mellem  $x$  og  $y$  er det tit underforstået, at der er tale om Pearsons lineære produktmoment korrelationskoefficient
- ▶ Beskriver den **lineære sammenhæng** mellem to variabler
- ▶ Pearsons  $r$  er et parametrisk mål, der kan anvendes når både  $x$  og  $y$  er målt på interval- eller ratioskala

3 / 14

## Pearsons $r$ (fortsat)

- ▶ Pearsons  $r$  varierer mellem  $-1$  og  $1$ 
  - ▶  $r = 1$  betyder perfekt positiv korrelation
  - ▶  $r = -1$  betyder perfekt negativ korrelation
  - ▶  $r = 0$  betyder fuldstændigt ukorreleret
- ▶ Pearsons  $r$  er et estimat for korrelationskoefficienten  $\rho$ :

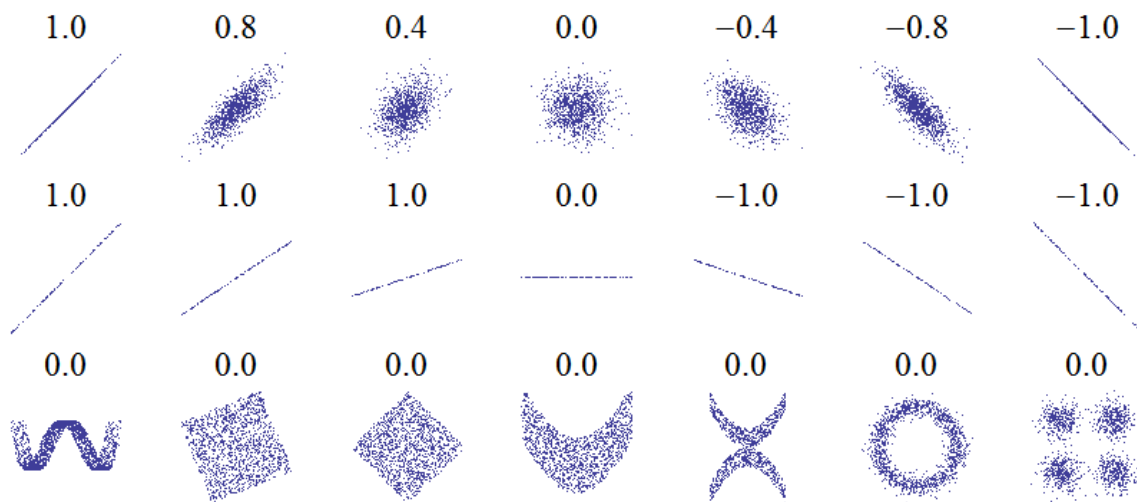
$$\hat{\rho} = r$$

- ▶ Korrelationskoefficienten  $\rho$  er defineret som

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

4 / 14

## Grafisk fortolkning af Pearsons $r$



[http://en.wikipedia.org/wiki/File:Correlation\\_examples.png](http://en.wikipedia.org/wiki/File:Correlation_examples.png)

5 / 14

## Beregning af Pearsons $r$

- ▶ Korrelationen mellem de  $n$  datapar  $(x_i, y_i)$  estimeres ved følgende formel

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

- ▶ Det er ikke så slemt, hvis du starter med at beregne summerne, kvadratsummerne og produktsummen...

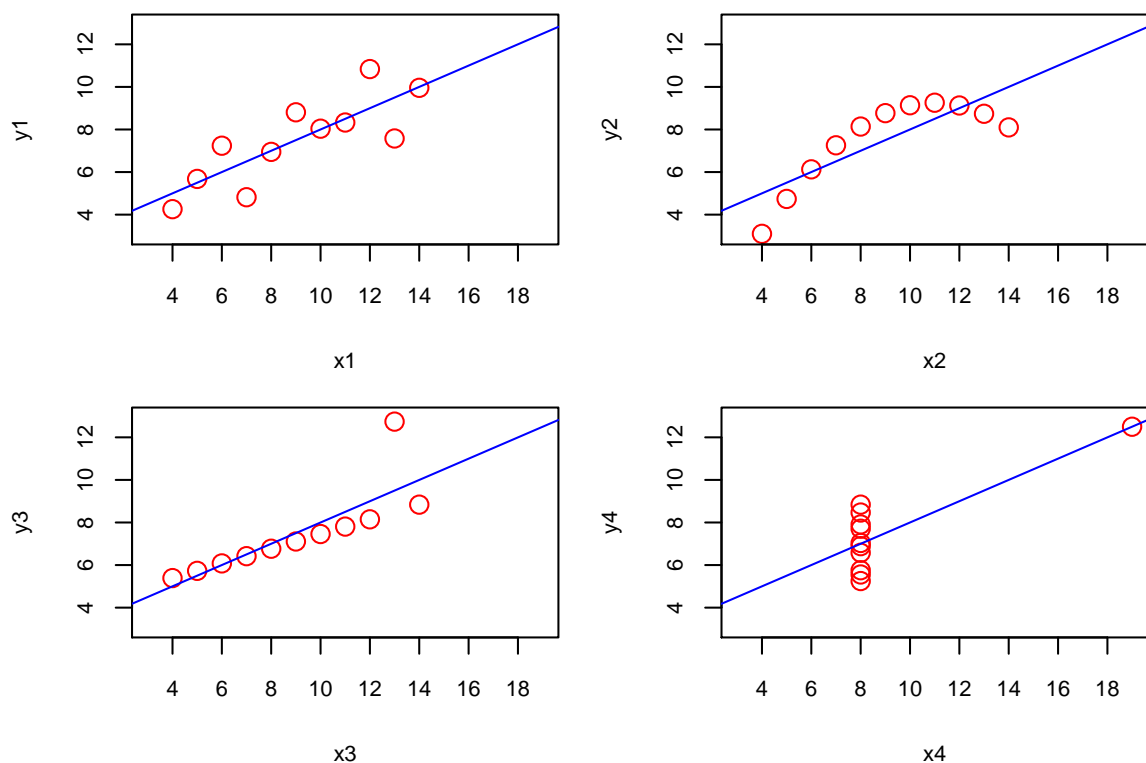
6 / 14

# Signifikanstest for Pearsons $r$

- ▶ Tabel G i Appendix 1 indeholder kritiske værdier for Pearsons  $r$  (numerisk værdi), idet  $df = n - 2$
- ▶ Hypoteser formuleres
  - ▶  $H_0 : r = 0; H_1 : r \neq 0$  (to-sidet alternativ)
  - ▶  $H_0 : r \leq 0; H_1 : r > 0$  (hvis vi har  $r > 0$ )
  - ▶  $H_0 : r \geq 0; H_1 : r < 0$  (hvis vi har  $r < 0$ )
- ▶ Husk **altid** at lave grafisk kontroll!

7 / 14

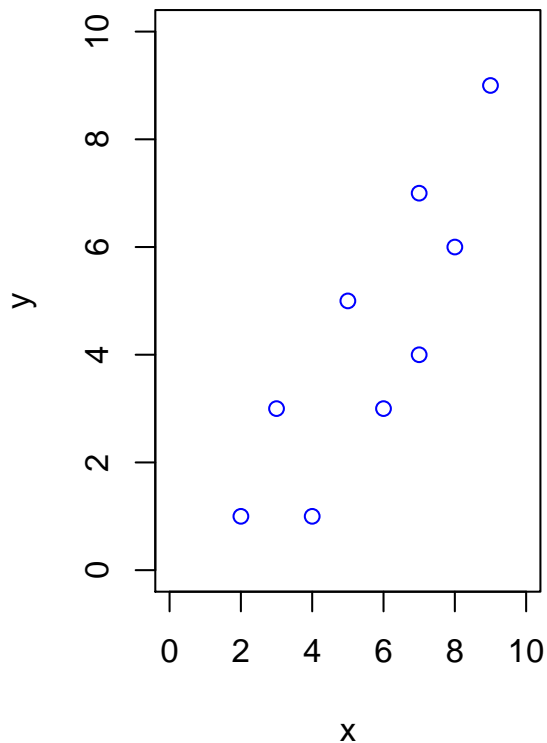
Pearsons  $r$  for Anscombes data er 0,87



Efter <http://en.wikipedia.org/wiki/Image:Anscombe.svg>

8 / 14

## Eksempel



► Data  $(x, y)$ :

(2, 1)	(3, 3)	(4, 1)
(5, 5)	(6, 3)	(7, 4)
(7, 7)	(8, 6)	(9, 9)

► Beregningshjælp:  $n = 9$ ,  
 $\sum x = 51$ ,  $\sum x^2 = 333$ ,  
 $\sum y = 39$ ,  $\sum y^2 = 227$

► Som noget nyt skal vi også bruge summen af produkterne

$$\sum_{i=1}^n x_i \cdot y_i = 264$$

9 / 14

## Eksempel (fortsat)

► Nu kan vi beregne Pearsons  $r$ :

$$r = \frac{9 \cdot 264 - 51 \cdot 39}{\sqrt{(9 \cdot 333 - 51^2)(9 \cdot 227 - 39^2)}} = 0,8512$$

► Sættet af hypoteser er  $H_0 : r \leq 0$ ;  $H_1 : r > 0$

► Med  $df = 9 - 2 = 7$  finder vi  $0,0005 < p < 0,005$  (enhalet), hvilket betyder klar afvisning af  $H_0$

10 / 14

## Spearmans $\rho$

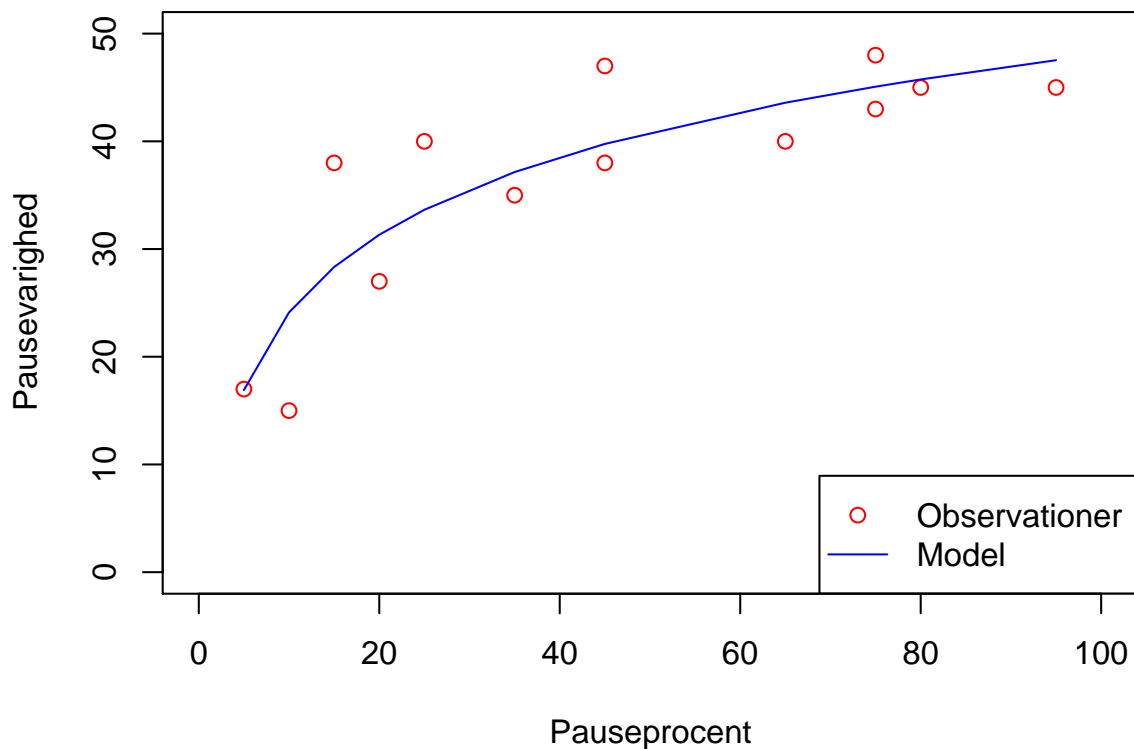
- ▶ Hvis en af variablene er målt på ordinalskala, eller hvis sammenhængen er ikke-lineær, så kan man ikke anvende Pearsons  $r$
- ▶ I stedet benyttes Spearmans  $\rho$  som beregnes af forskellen mellem rangværdierne for  $x$  og  $y$  som

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

idet  $d_i = \text{rang}(x_i) - \text{rang}(y_i)$  beregnes for alle  $n$  datapunkter

11 / 14

## Spearmans $\rho$ — eksempel på ikke-lineær sammenhæng



12 / 14

## Spearmans $\rho$ — eksempel (fortsat)

$n$	13
$\sum d^2$	63
Spearmans $\rho$	0,8269

- ▶ Hvis der er mange *ties* kan man lave en korrektion — eller beregne Pearsons  $r$  for rangværdierne i stedet

Pearsons $r$ (for rangværdier)	0,8257
Pearsons $r$ (for rådata)	0,7749

13 / 14

## Spearmans $\rho$ — signifikanstest

- ▶ Tabel G (kritiske værdier for Pearsons  $r$ ) kan benyttes hvis  $n > 10$  (husk at  $df = n - 2$ )
- ▶ For  $n < 10$  benyttes Tabel H hvor laves opslag efter  $n$
- ▶ I eksemplet er  $n = 13$  og  $\rho = 0,827$  så vi opstiller en nulhypotese og et alternativ:

$$H_0 : \rho = 0 \quad \text{og} \quad H_1 : \rho > 0$$

og vi finder  $p < 0,0005$  (énhalet,  $df = 11$ ) fra Tabel G

14 / 14