Medisinsk statistikk Del II
Forelesning 25 februar 2009
Korrelasjon

av
Stian Lydersen og Eirik Skogvoll

1

Innhold

- Kovarians og korrelasjon (5.6.1)
- Pearson's r (11.7)
- T-test og z-test for korrelasjonskoeffisient (11.8)
- Konfidensintervall
- Rang-korrelasjon, Spearman's r_s (11.12)
- (Intra Class Correlation Coefficient ICC, 12.9)

2

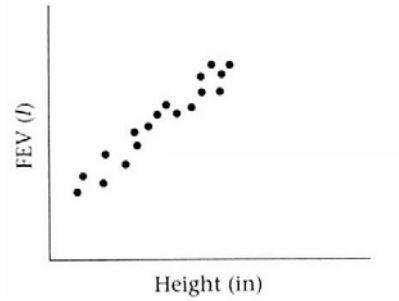
Korrelasjon

= sammenheng, samvariasjon
≠ kausalitet (årsak/virkning)

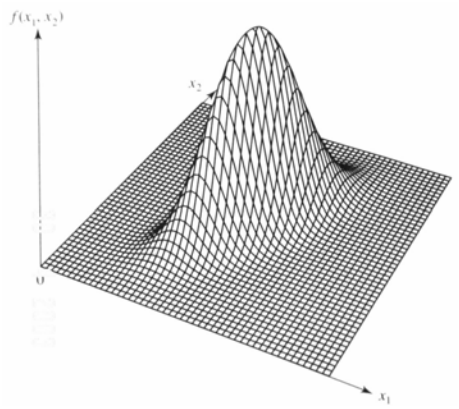
Hva skiller korrelasjon fra regresjon?

3

Eksempel: Fig. 5.16a



4



5

Kovarians

Def. 5.13 s 142 (5.12, s. 135 i 5th edition)

Cov $(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$

I klartekst ☺ :
"Forventningsverdien
til produktet
av avviket
for hhv. X og Y
fra sine respektive forventingsverdier"

6

Kovarians og uavhengihet

- Hvis X og Y er uavhengige, er cov $(X,Y) = 0$
- Omvendt gjelder ikke nødvendigvis!
- Men:
- Hvis (X, Y) er bivariat normalfordelt og cov $(X,Y) = 0$, så er X og Y uavhengige.

7

Kovarians og korrelasjon

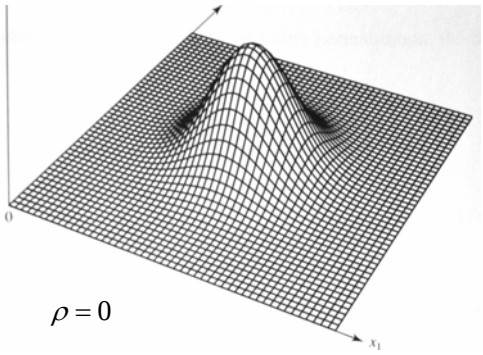
Problem med enhetene ...

- Hvis X og Y måler hhv. FEV (l) og høyde (m) får kovariansen enheten l*m
  (sml. med definisjon og enhet for varians)
- Løser dette ved å dividere med standardavvikene til X og Y

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

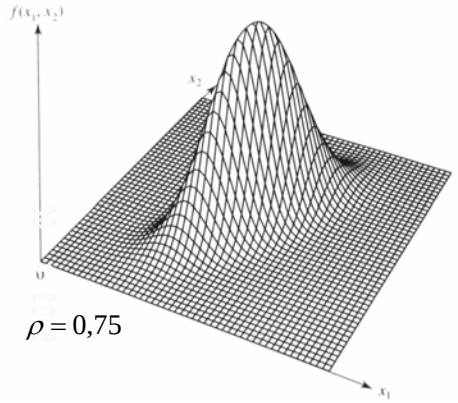
$$\rho \in (-1,1)$$

8



$\rho = 0$

9



$\rho = 0{,}75$

10

Kovarians beregnet fra et utvalg - empirisk kovarians:

Utvalgs-kovarians:

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

11

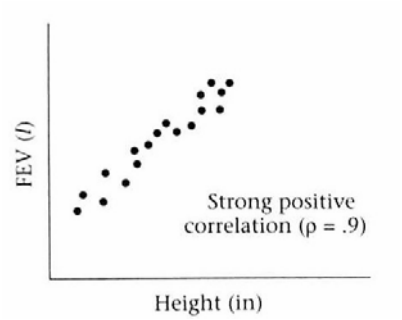
Korrelasjon beregnet fra et utvalg - empirisk korrelasjon: Pearson's *r*

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

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

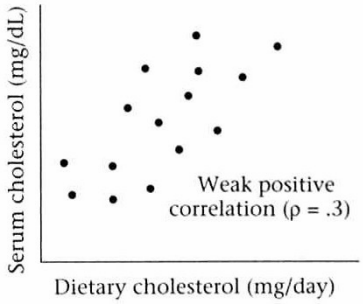
12

Eksempel: Fig. 5.16a



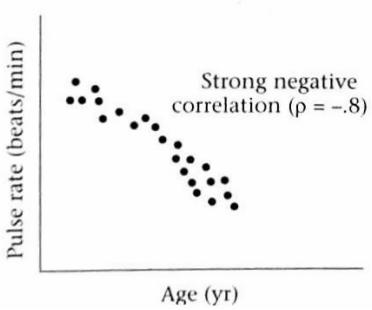
13

Eksempel: Fig. 5.16b



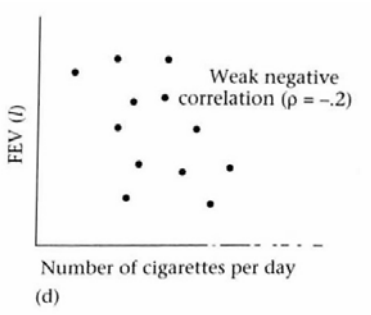
14

Eksempel: Fig. 5.16c



15

Eksempel: Fig. 5.16d



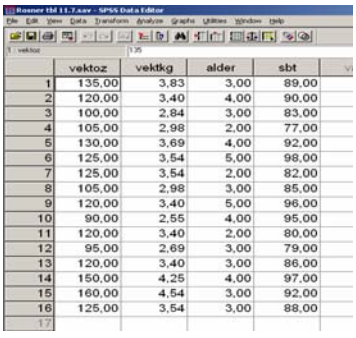
16

Noen egenskaper til korrelasjonen

- $-1 \leq r \leq 1$
- r=1 hvis og bare hvis alle observasjonene ligger langs en linje med postivt stigningstall
- r=-1 hvis og bare hvis alle observasjonene ligger langs en linje med negativt stigningstall

17

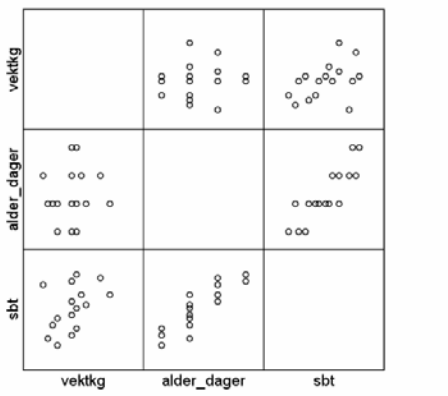
Korrelasjon - eksempel



	vektoz	vektkg	alder	sbt
1	135,00	3,83	3,00	89,00
2	120,00	3,40	4,00	90,00
3	100,00	2,84	3,00	83,00
4	105,00	2,98	2,00	77,00
5	130,00	3,69	4,00	92,00
6	125,00	3,54	5,00	98,00
7	125,00	3,54	2,00	82,00
8	105,00	2,98	3,00	85,00
9	120,00	3,40	5,00	96,00
10	90,00	2,55	4,00	95,00
11	120,00	3,40	2,00	80,00
12	95,00	2,69	3,00	79,00
13	120,00	3,40	3,00	86,00
14	150,00	4,25	4,00	97,00
15	160,00	4,54	3,00	92,00
16	125,00	3,54	3,00	88,00

Rosner Table 11.9 s 511 (5th ed table 11.7, s. 468)

18



19

Korrelasjonsanalyse - bruk

- Screene ”store mengder” data for sammenhenger - hva henger sammen med hva?
- Før inklusjon av prediktor-variable i evt. multippel regresjonsmodell - hvilke variable måler omtrent det samme?
- Alminnelig i psykologisk forskning - (spørreskjema f.eks.) hvilke egenskaper er korrelert?
  → *faktor-analyse, prinsipal komponent-analyse*

20

Korrelasjonsanalyse - misbruk

Bland, J. M , Altman, D. G.:
Statistical methods for assessing agreement between two methods of clinical measurements
The Lancet, 8 Febr 1986, p 307-310

(Sitert 13143 ganger ISI 23 februar 2009)

21

Ikke-paramterisk korrelasjonskoeffisient

- Kan være å foretrekke når:
  - Sammenhengen er ikke-lineær
  - Data er ordinale (kvalitative)
  - Ved avvik fra normalfordelingen(?)
- Alternativer:
  - Spearman’s rho
  - Kendall’s tau
  - Liten forskjell på dem, men kanskje en viss preferanse for Kendall's tau.

22

"Our results suggest that Kendall's tau, has many advantages over Pearson's and Spearman's r; when applied to psychiatric data, tau, maintained adequate control of type I errors, was nearly as powerful as Pearson's r, provided much tighter confidence intervals and had a clear interpretation.“

Arndt S, Turvey C, Andreasen NC: Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation. JOURNAL OF PSYCHIATRIC RESEARCH Volume: 33 Issue: 2 Pages: 97-104, 1999

23

Denne, derimot, skriver at begge er ganske bra:

Kraemer HC: Correlation coefficients in medical research: from product moment correlation to the odds ratio. STATISTICAL METHODS IN MEDICAL RESEARCH Volume: 15 Issue: 6 Pages: 525-545, 2006

24

Rang-korrelasjon, Spearman's r_s

- Observasjonene ordnes i stigende rekkefølge
- Regn ut Pearson's korrelasjonskoeffisient for rangene (i)

Transformasjon: $x_i \xrightarrow[ordnes]{} x_{(i)}$

T-test for korrelasjonskoeffisient

(Rosner Eq. 11.20)

- Beregn Pearson's r
- Transformer denne til T:

$$T_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$ ...og i dette tilfelle

$$T_{n-2} = \frac{\frac{r-0}{\sqrt{1-r^2}}}{\sqrt{n-2}} = \frac{r}{\frac{\sqrt{1-r^2}}{\sqrt{n-2}}} = \frac{r(\sqrt{n-2})}{\sqrt{1-r^2}}$$

Z-test for korrelasjonskoeffisient

(Rosner Eq. 11.21 – 11.22)

- Beregn Pearson's r
- Fisher's z-transformasjon gir en tilnærmet normalfordelt variabel, Z:

$$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Under H_0 ($\rho=\rho_0$) har Z forventningsverdi og varians

$$E(Z_0) = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

$$\text{Var}(Z_0) = \frac{1}{n-3}$$

Konfidensintervall for ρ

- (Rosner eq. 11.23)
- Utregnet Pearson's r_0 fra et utvalg med n observasjoner, bruker Fisher's Z–transformasjon

$$z_0 = \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right)$$

95 % CI for $Z = (z_1, z_2) = z_0 \pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$

95 % CI for $\rho = (\rho_1, \rho_2) = \left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z2}+1}\right)$

Sammenlikning av to uavhengige korrelasjonskoeffisienter: Fisher's z-test (Eqn 11.27)

H_0: $\rho_1 = \rho_2$ versus $\rho_1 \neq \rho_2$

Under H_0 så gjelder tilnærmet

$$\frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1)$$

Sammenlikning av to avhengige korrelasjonskoeffisienter

H_0: $\rho_{XZ} = \rho_{YZ}$ versus $\rho_{XZ} \neq \rho_{YZ}$

Wolfe's test (Rosner Eqn 11.28) forutsetter bl.a. $\sigma_X = \sigma_Y$.

Bedre:
En metode basert på Fisher's z-transformasjon, se

Meng, X. L., Rosenthal, R., Rubin, D. B.:
Comparing correlated correlation coefficients.
Psychological bulletin, 1992, Vol 111, No 1, p 172-175.
(Referert 562 ganger ISI 16 februar 2009)

NTNU

To sum up, the z test in Equation 11.22 is used to test hypotheses about nonzero null correlations, whereas the t test in Equation 11.20 is used to test hypotheses about null correlations of zero. The z test can also be used to test correlations of zero under the null hypotheses, but the t test is slightly more powerful in this case and is preferred. However, if $\rho_0 \neq 0$, then the one sample z test is very sensitive to non-normality of either x or y. This is also true for the two sample correlation test presented later in this section.
Rosner (6th ed) s 501

31

NTNU

Noen sammenhenger i enkel lineær regresjon:

- Hypotesen
  $\beta = 0$ er ekvivalent med $\rho = 0$. En F-test (eller t-test) for $\beta = 0$ er ekvivalent med en t-test for $\rho = 0$
- $b = r\, s_y/s_x$
- r^2 dvs. "coefficient of determination" tolkes best i regresjon som andelen "forklart varians"

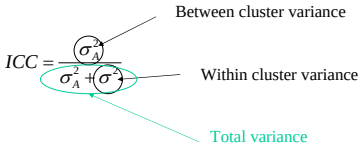
32

NTNU

Intraclass correlation coefficient (ICC)
(Rosner 6th ed: 12.9, page 613-617)
Correlation between two measurements in the same subjects (or clusters).

One-way random effects ANOVA:
Measurement number j at subject (or cluster) number i:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \alpha_i \sim N(0, \sigma_A^2), \quad \varepsilon_i \sim N(0, \sigma^2)$$



33