FORELESNING 19

MAR 29 2012

MET 3451

STATISTIKK

EIVIND ERIKSEN

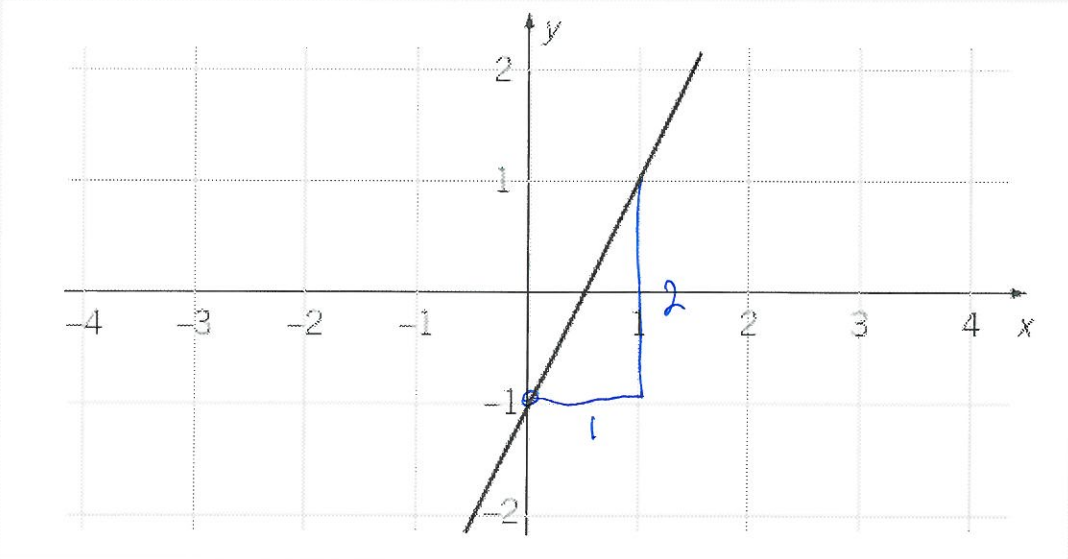
PLAN:

① Lineær korrelasjon [T] 10.1-10.2

1. 10-2: Korrelasjon

2. 10-3: Regresjon

Example



- Krysser y-aksen i -1: $b_0 = -1$
- Stiger med 2 hver gang x øker med 1: $b_1 = 2$
- Formelen til linja er derfor

$$y = -1 + 2x$$

Eksempel

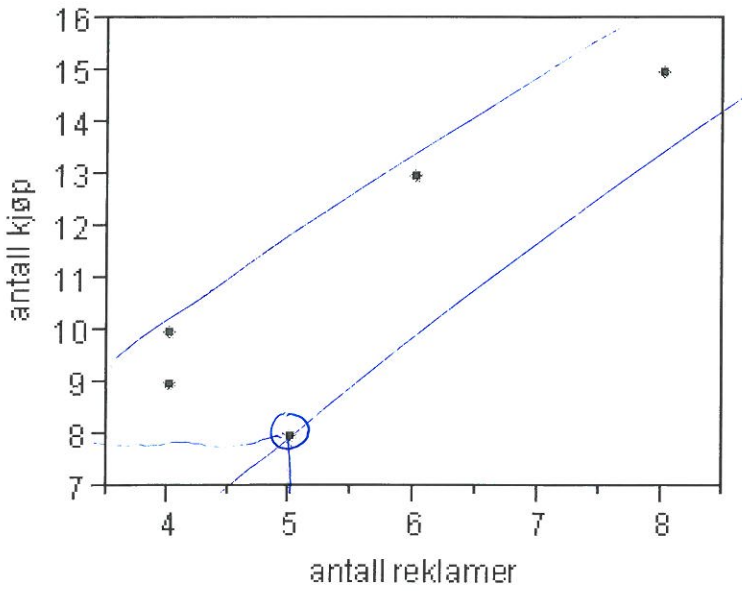
Example

Vi lar fem personer se en konfektreklame og noterer hvor mange konfektesker de kjøper i løpet av en måned:

Person	1	2	3	4	5	Gj.snitt	std.avvik
Antall reklamer	5	4	4	6	8	5.4	1.67
Antall kjøp	8	9	10	13	15	11.0	2.92

- Finnes det et tall som måler sammenhengen?
- Kan vi finne en hypotesetest som avgjør om det er sammenheng?

Scatterplot



Visualisér dataene i et scatterplott

- Det ser ut som det er en korrelasjon
- Jo mer reklame, jo mere kjøp
- En positiv korrelasjon

Korrelasjon mellom to variable

Kapittel 10

- Dataene består av *par* (x, y)
- Hvordan samvarierer de to variablene x og y?
- Vi sier at en *korrelasjon* finnes mellom x og y dersom det er en sammenheng mellom dem på en eller annen måte
- Samvariasjonen (korrelasjonen) beskrives ved hjelp av den *lineære korrelasjonskoeffisienten* r

r

- Dataene er parvise: (x, y)
- Dataene brukes til å regne ut korrelasjonskoeffisienten r
- r måler hvor sterkt x og y er korrelerte

Lineært

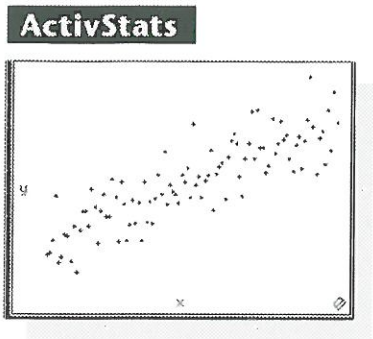
JMP

- *Analyze* > *Fit Y by X* og så velger du *Density Ellipse*
- $r = 0.87$

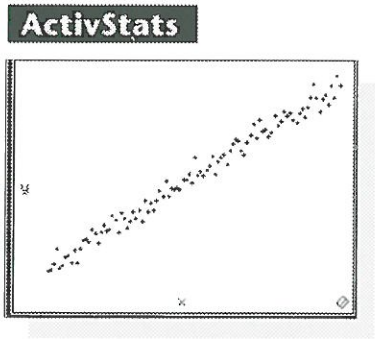
Correlation

Variable	Mean	Std Dev	Correlation	Signif. Prob	Number
antall reklamer	5.4	1.67332	0.871165	0.0544	5
antall kjøp	11	2.915476			

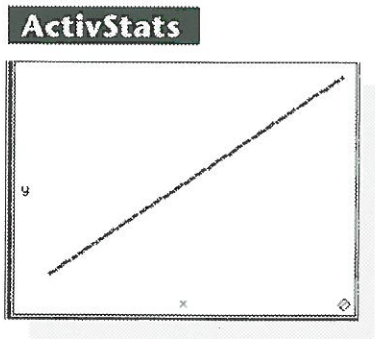
11
۲



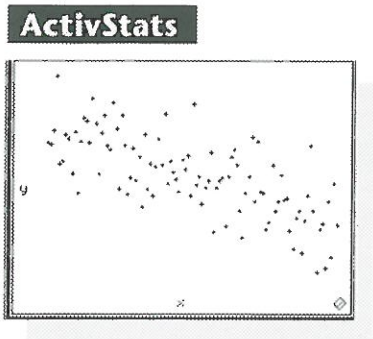
(a) Positive correlation:
$r = 0.851$



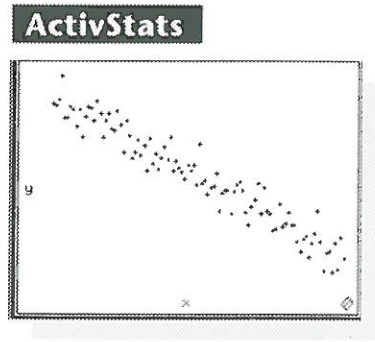
(b) Positive correlation:
$r = 0.991$



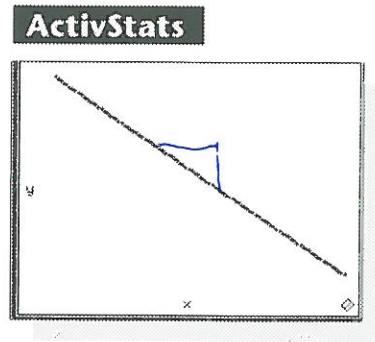
(c) Perfect positive correlation:
$r = 1$



(d) Negative correlation:
$r = -0.702$

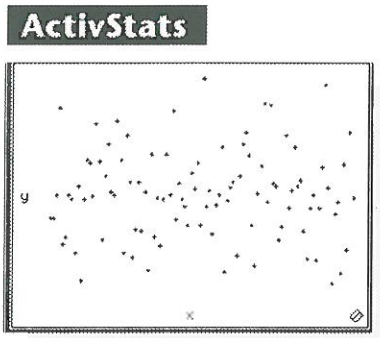


(e) Negative correlation:
$r = -0.965$

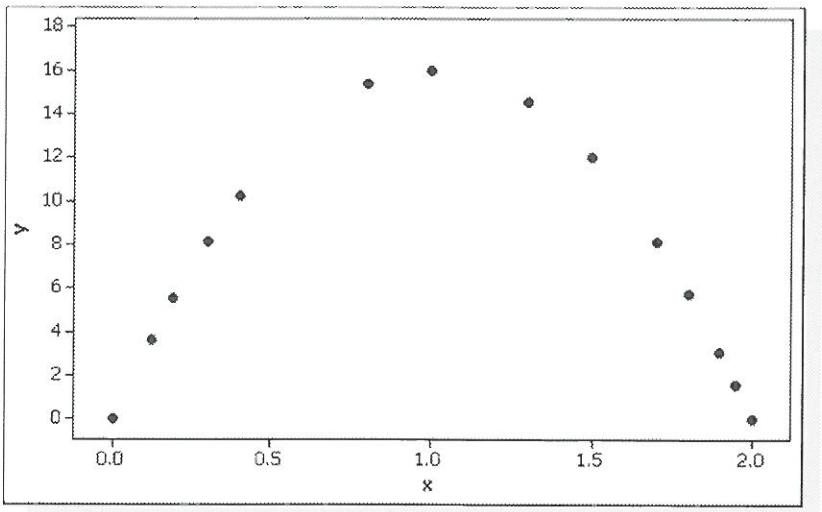


(f) Perfect negative correlation:
$r = -1$

r måler bare *lineær* korrelasjon



(g) No correlation: $r = 0$



(h) Nonlinear relationship: $r = -0.087$

Positiv, negativ eller ingen korrelasjon?

- Korrelasjonen er alltid et tall mellom -1 og $+1$:
  - Er r nær $+1$ så har vi positiv samvariasjon
  - Er r nær 0 så har vi ingen samvariasjon
  - Er r nær -1 så har vi negativ samvariasjon

Example

I konfekteksempelet var korrelasjonen $+0.87$. Det er en sterk *positiv* korrelasjon mellom antall reklamer man har sett og antall konfektesker man kjøper

Hva skal vi med r?

Den brukes til å avgjøre om det finnes samvariasjon mellom to ting. Vi skal snart se dette i t-testen for lineær samvariasjon...

$X \rightsquigarrow z_x = \frac{x - \bar{x}}{s_x}$

Formelen

$\sum \frac{z_x \cdot z_y}{n-1}$

//

Beregne r

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

Tolke r

- r regnes ut av kalkulator/JMP/Excel
- Vi fokuserer på *forståelsen og tolkningen* av r, det er viktigere enn hvordan den regnes ut

Egenskaper til r

- $-1 \leq r \leq 1$
- r måler styrken i den lineære korrelasjonen mellom x og y

Forklart varians

Tolkning av r^2

- r^2 er andelen av variasjon i y som blir forklart av variasjon i x

Example

Person	1	2	3	4	5	Gj.snitt	std.avvik
Antall reklamer	5	4	4	6	8	5.4	1.67
Antall kjøp	8	9	10	13	15	11.0	2.92

- Korrelasjonskoeffisienten er $r = +0.87$, og $r^2 = 0.757$
- Vi sier at 75.7% av variasjonen i kjøp skyldes variasjonen i sette reklamer
- Det innebærer at 24.3% av variasjonen i kjøp ikke har sammenheng med reklame

Typiske feil i korrelasjon

1. En typisk feil er å tro at korrelasjon betyr at det er en årsak-virkning effekt (kausalitet)
2. Det kan være en korrelasjon, selv om den ikke er lineær. r vil ikke fange opp dette. Se figur side 9.

Example

Det er ikke sikkert at det er reklamen som gjør at folk kjøper mer konfekt. Det kan tenkes at det er en *bakenforliggende* faktor, som gjør at folk både ser mer reklame og kjøper mer konfekt.

- Vil du påvise årsak-virkning, så er det best å gjøre et randomisert eksperiment.
- Observasjonelle studier gir mindre mulighet til å påvise kausalitet, pga sammenblanding (*eng: confounding*)

Hypotesetest for lineær korrelasjon

Notasjon

- r er korrelasjonen i **stikkprøven**, (en observator)
- ρ er korrelasjonen i **populasjonen**, (en parameter)

Nullhypotesen: *ingen korrelasjon*

En tosidig hypotesetest skrives da:

$$H_0 : \rho = 0 \quad \textit{vs.} \quad H_1 : \rho \neq 0$$

Ensidig:

$H_0: \rho = 0$ $\quad$ $H_1: \rho > 0$ $\quad$ (positiv lin. korrelasjon)

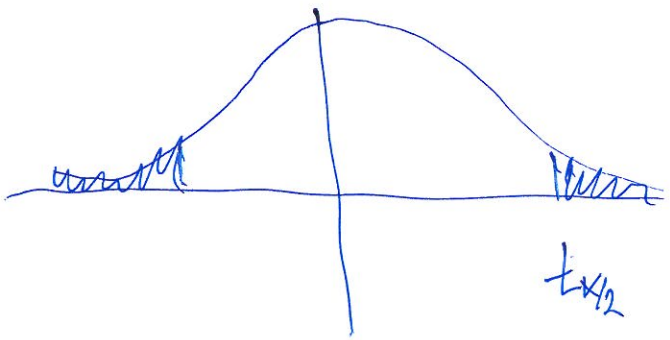
$H_0: \rho = 0$ $\quad$ $H_1: \rho < 0$ $\quad$ (negativ lin. korrelasjon)

Test for lineær korrelasjon

1. $H_0 : \rho = 0 \quad vs. \quad H_1 : \rho \neq 0$
2. Bestem signifikansnivået α
3. Regn ut r
4. Regn ut testobservatoren

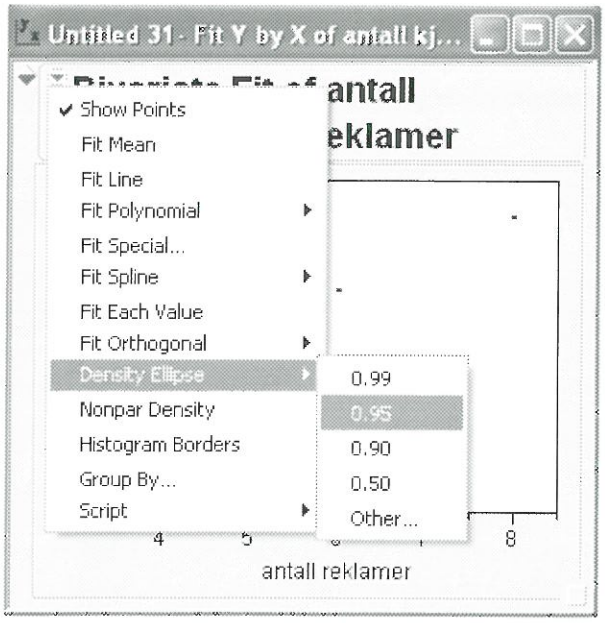
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

   har $n - 2$ frihetsgrader
5. Hvis t overstiger den kritiske verdien (tabell A3), forkast $H_0 : \rho = 0$. Hvis ikke, så kan vi ikke forkaste H_0
6.
   - Hvis H_0 forkastes, så er det en lineær korrelasjon mellom variablene
   - Hvis H_0 ikke forkastes, så kan vi ikke konkludere med at det er en lineær korrelasjon



Korrelasjon og regresjon i JMP

- *Analyze > Fit Y by X*
- Velg antall reklamer på x-aksen og antall kjøp på y-aksen
- For å få korrelasjonskoeffisienten r: Rød diamant og *Density ellipse..*



Reklame-Kjøp sammenhengen er signifikant

Example

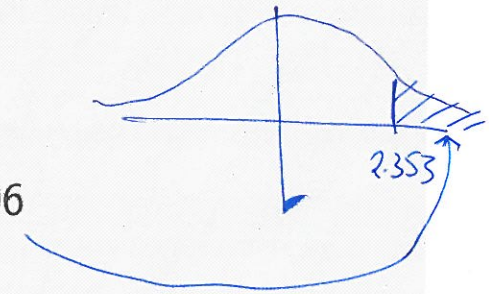
1. Enten virker ikke reklame, eller så økes kjøpelysten, dvs. ensidig test:

$$H_0 : \rho = 0 \quad vs \quad H_1 : \rho > 0$$

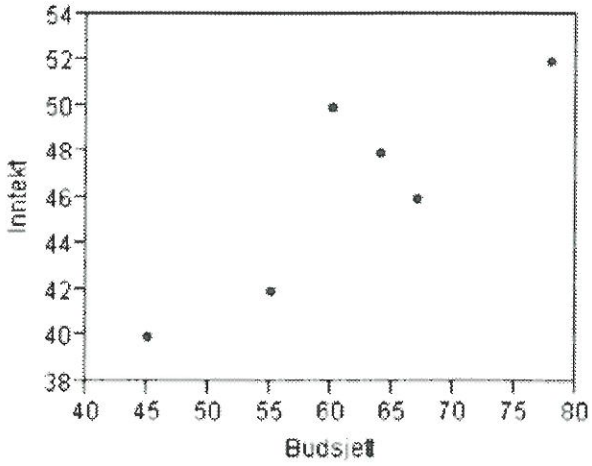
2. $\alpha = 0.05$ og vi har sett at $r = 0.87$

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

$$t = \frac{0.87}{\sqrt{\frac{1-0.87^2}{5-2}}} = 3.06$$



4. Kritisk verdi for ensidig test med 3 df er da (A3): $t_{0.05} = 2.353$
5. Siden testobservatoren 3.06 er større enn kritisk verdi: forkast H_0 positiv
6. Det er en lineær korrelasjon mellom antall sette reklamer og antall kjøp



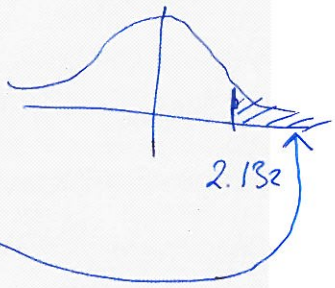
Example

Markedsføringsbudsjett	67	64	45	78	60	55
Innspillt første helg	46	48	40	52	50	42

- 6 Hollywood filmer.
- Er det en positiv korrelasjon mellom budsjett og billettinntekter?
- $H_0 : \rho = 0 \quad vs \quad H_1 : \rho > 0$

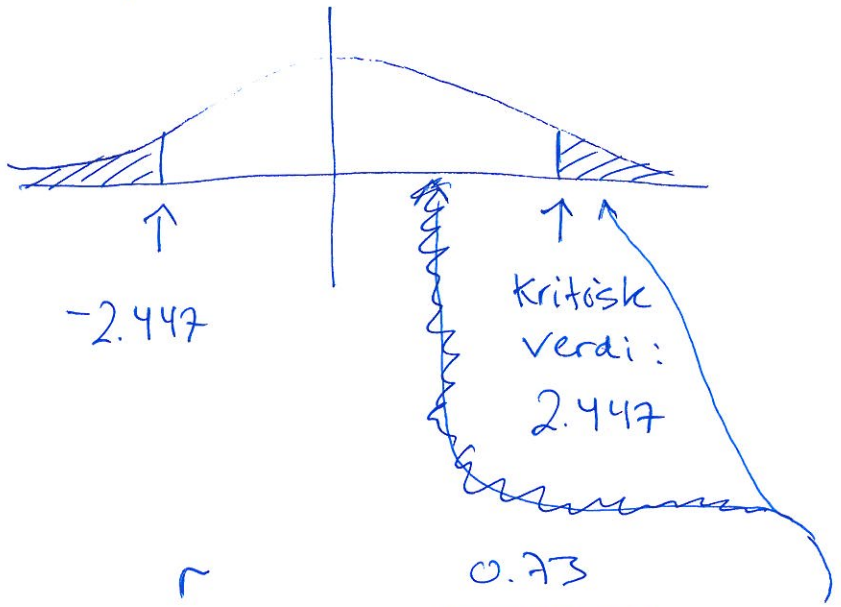
Example

1. $H_0 : \rho = 0 \quad vs \quad H_1 : \rho > 0$ (ensidig)
2. Signifikansnivået settes til $\alpha = 0.05$
3. Kalkulator/JMP/Excel regner ut $r = 0.861$
4. $$t = \frac{0.861}{\sqrt{\frac{1-0.861^2}{6-2}}} = 3.38$$
5. Kritisk verdi med 4 df er $t_{0.05} = 2.132$
6. Testobservator er høyere enn kritisk verdi: Forkast H_0
7. Vi har grunnlag til å hevde at det er en positiv korrelasjon mellom budsjett og inntekter



Eksamen 06/2011

6 b) $H_0: \rho = 0$ $H_1: \rho \neq 0$



$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.73}{\sqrt{\frac{1-0.73^2}{6}}} \simeq 2.616$$

8.2

Konklusjon: Vi forkaster H_0

Grunnlag for å si at det er en lin. korrelasjon mellom variablene.