

ISTT100y, Uke 1, 2024

DEL 1, Menti

Eks 1: emnekode, kategoriske data (4 kategorier)

litt logisk
sortering

Eks 2: Forboksstav, —"—— (29 kategorier)

veldig logisk sortering

Eks 3: Fødselsår, diskrete data (her 14 tall
1995 - 2008)

naturlig sortering

↳ heltall
langs tall-linja

For di vi nå jobber med tall kan vi også snakke
om oppsummeringstall slik som gjennomsnitt

- Hva er det vanligste fødselsåret?
- Gjennomsnittlig fødselsår? (gjennomsnittsalder)

Eks 4: *Tenkepause*

↳ MENTI

→ Gjennomsnitt?
Variasjon/spredning?

Høyde, diskret? kontinuerlig?

↳ her oppga
dere heltall

↳ ingen grunn til at vi ikke
kan måle med nøyaktig enn
det, naturlig å tenke på høyde som
kontinuerlig

DEL 2, regne for hånd

Eks: Sykle til jobb

a) Typetall

b) Median

c) Gjennomsnitt

Tall 1, 4, 4, 1, 3, 5, 2, 3, 1, 5
 x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}

= 1
1 det
som
opptre
ofte

Sortere 1, 1, 1, 2, 3, 3, 4, 4, 5, 5

$$m = \frac{3+3}{2} = 3$$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{29}{10} = 2,9$$

Sentral mid / mål på
lokasjon. Hvor på tallene
definerer observasjonene seg?

* FORMELARK *

d) Empirisk varians

e) Emp. Standardavvik

$$s^2 = \frac{1}{9} \sum_{i=1}^9 (x_i - \bar{x})^2 =$$
$$\frac{1}{9} [(1-2,9)^2 + (4-2,9)^2 + (4-2,9)^2 + \dots \text{osv}]$$

$$\approx \underline{\underline{2,54}}$$

$$s = \sqrt{s^2} \approx 1,59$$

hvorfor n-1?

Det lærer vi om noen
uker!

mål på spredning
NB: må ha minst
to observasjoner

DEL 3, Fruktbarhetstall

Gjennomsnitt i utvalg (dere) = _____

Diskuter hvorfor tallet vårt ble så feil.

LINEÆRE SAMMENHENGER

A) Korrelasjon ← sier noe om graden av linear sammenheng mellom to størrelser (x og y)

Ek: Hva tror vi om korrelasjonen mellom fruktbarhetstall (y) og år etter 2000 (x)?

- Negativ

* FORMELARK *

$$r = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^5 (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^5 (y_i - \bar{y})^2}}$$

$\bar{x} = 10$
 $\bar{y} = 1.76$
 ≈ -0.85

NB : Empirisk korrelasjon er et oppsummeringstall som sier noe om graden av linear sammenheng mellom to størrelser, og retningen.

Men korrelasjonen sier ingenting om hvor observasjonene ligger. Når to størrelser er korrelerte er naturlig å tenke seg en rett trendlinje slik som DENNE.

B) MINSTE KVADRATSUMS RETTE LINJE

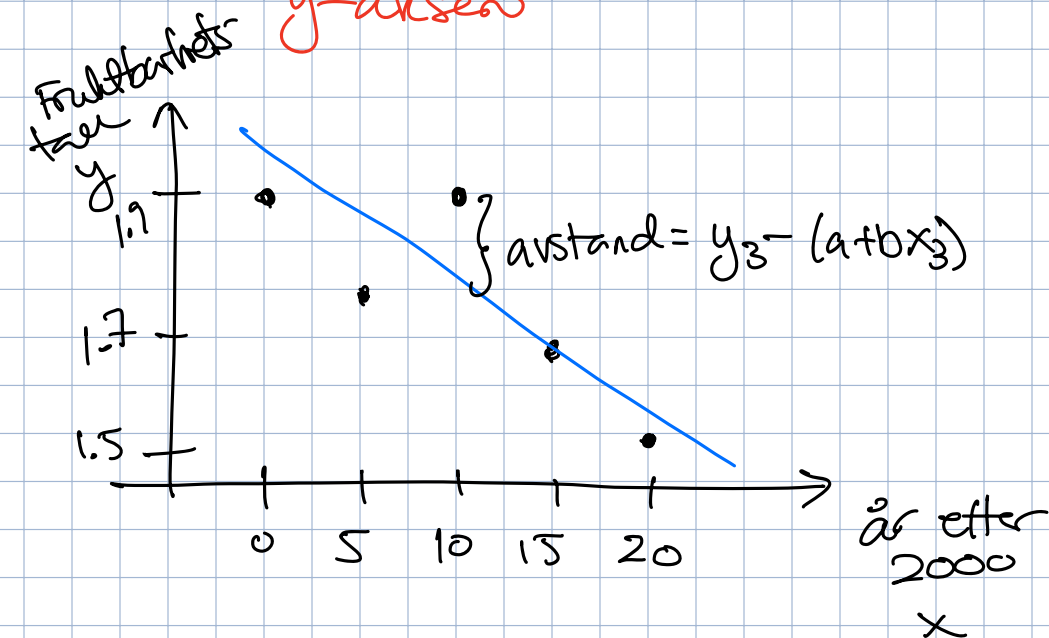
$$y = a + bx$$

↑
Skjæringspunkt
y-aksen

↑
Stigningstall

↑ linja som minimerer
summen av kvadratavvik
mellom linja og
observasjonene

↓
Den linja som beskriver den lineære
trenden i dataene "best"



Minimer

$$\sum_{i=1}^S (y_i - (a + bx_i))^2$$

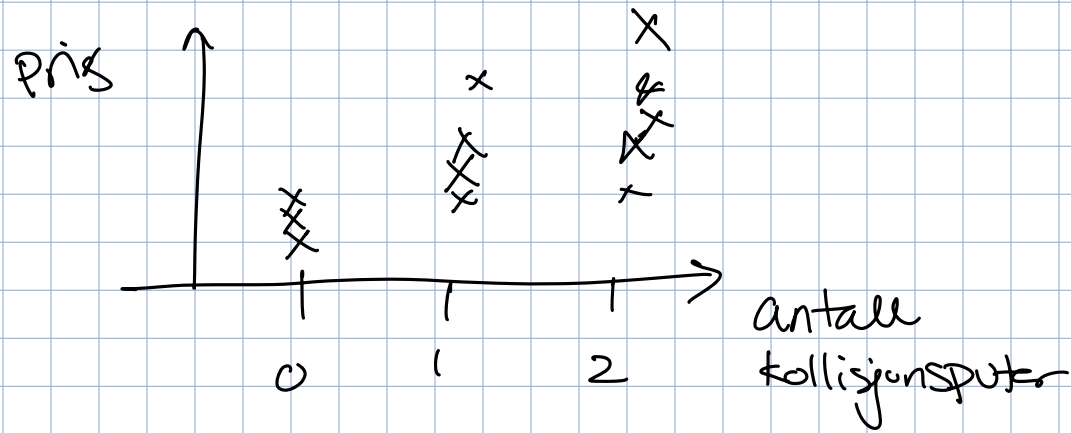
FORMELARK

$$b = -0.018, \quad a = 1.94 \quad (\text{slide})$$

↑
Negativt stigningstall
følger av negativ korrelasjon.

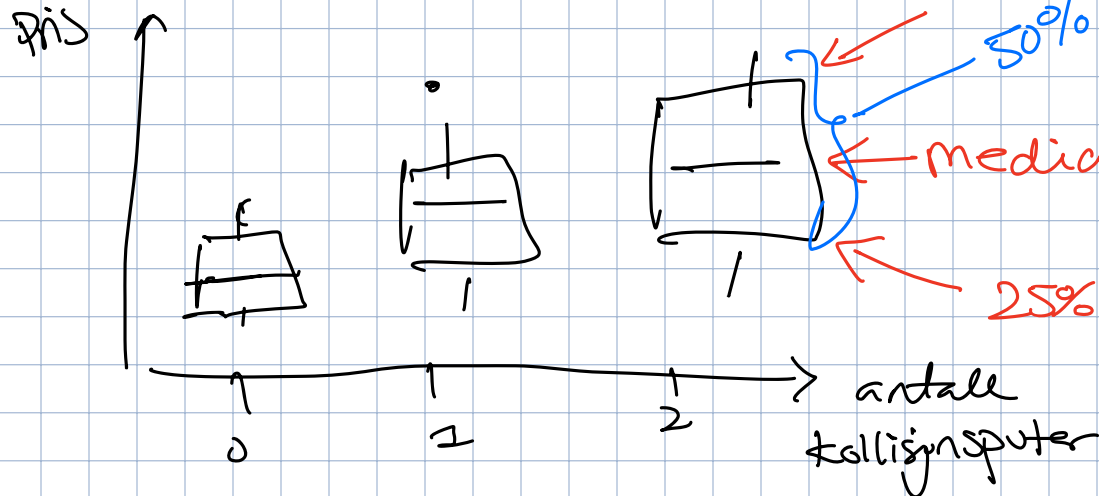
Eks 2: Pris og kollisjonspoter (og boksploTT!)

Figur 1



Hva ser vi?

Figur 2 — boksploTT



"uteligger"

75% kvantil, 75% av obs. under
50% av observasjoner "inni" boksen

median

halvparten av observasjonene
under og over

25% kvantil, 0.25-kvantil

25% av observasjoner under

Hvilken figur liker du best? Fordeler / Ulemper?

- * Figur I blir umulig med mange observasjoner
- * Vanskelig å se "midten" i Figur I

Boksplott viser medianen, hva tror du om gjennomsnitt?

ca like fordi dataene fordeler seg ganske Symmetrisk.

NB: også her har vi en linear sammenheng / trend.