

FORELESNING 5

MET#31

EIVIND ERIKSEN

FEB 02 2012

STATISTIKK

PLAN:

- ① MER OM SENTRALMÅL. REPETISJON
- ② VARIASJONSMÅL. STANDARDAVVIK

[T] 3.2

[T] 3.3

① SENTRALMÅL:

Repetisjon: Viktige sentralmål = uttrykk for senteret i dataene

(a) Gjennomsnitt av variabelen x i datasettet:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

(b) Median av variabelen x i datasettet:

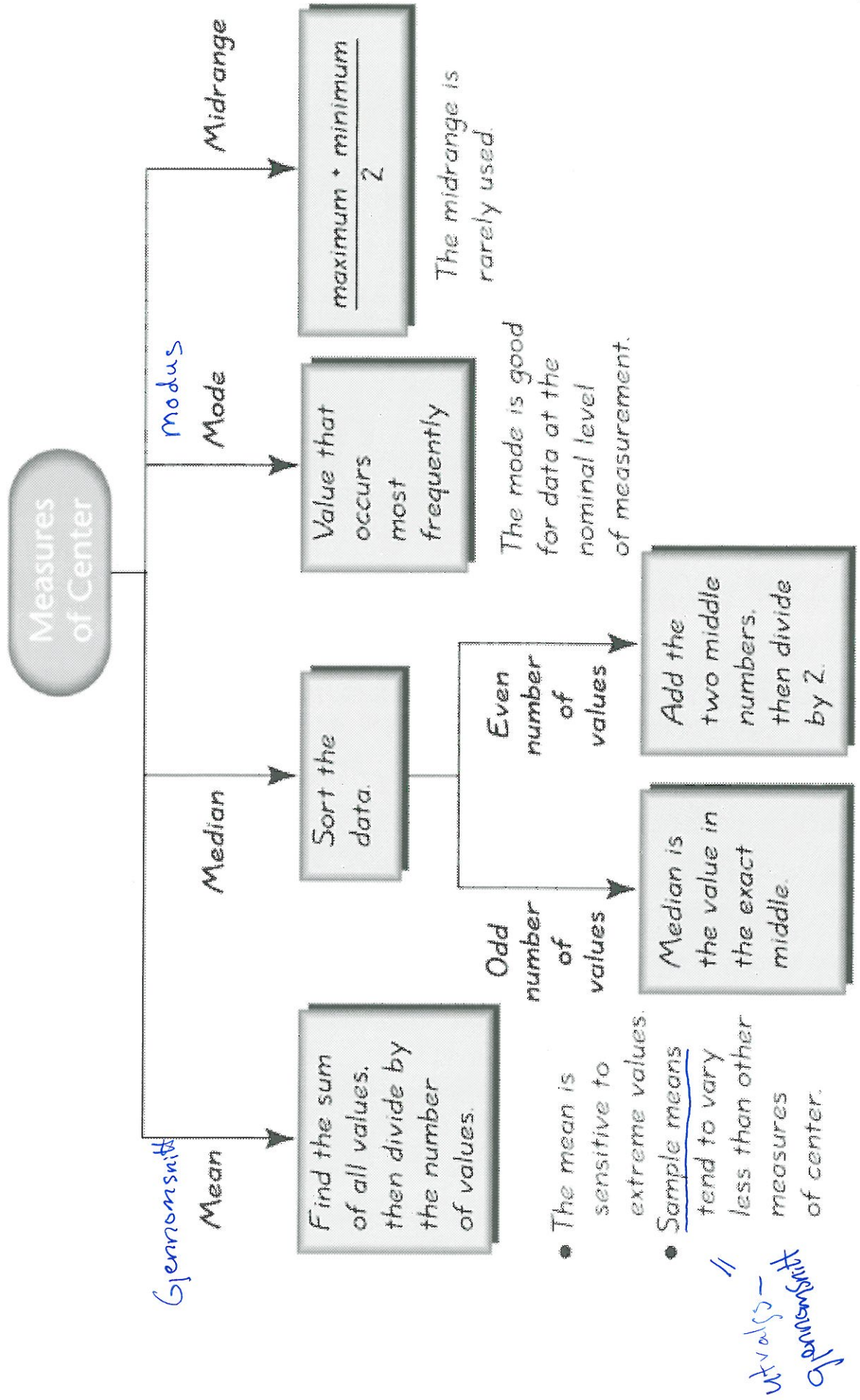
$\tilde{x}$ = den midterske verdien når observasjonene er ordnet i stigende rekkefølge

- / verdien i midten når n er oddetall
- \ gjennomsnitt av de to i midten når n er partall

(c) Modus av variabelen x i datasettet:

modus = den verdien som forekommer flest ganger (=sterst frekvens)

Oppsummering



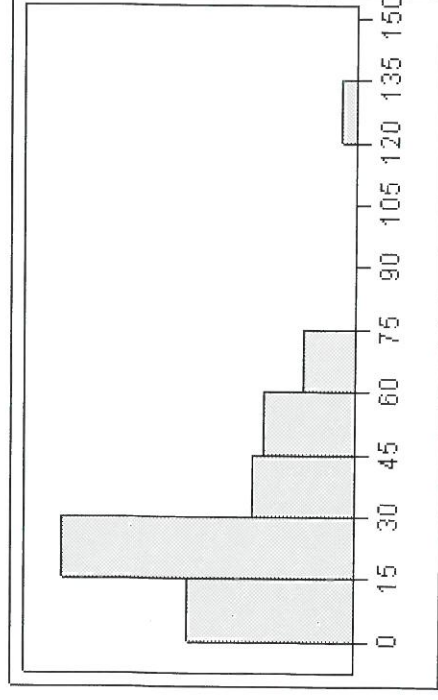
Skjev eller symmetrisk?

Symmetrisk fordeling av data

Dersom venstre halvdel av histogrammet er omtrent et speilbilde av høyre halvdel. Eksempel: normalfordelte data.

Skjev fordeling av data

Dersom ikke symmetrisk og en side av dataene er lenger enn den andre. Reisetid i minutter er skjevt fordelt mot høyre.

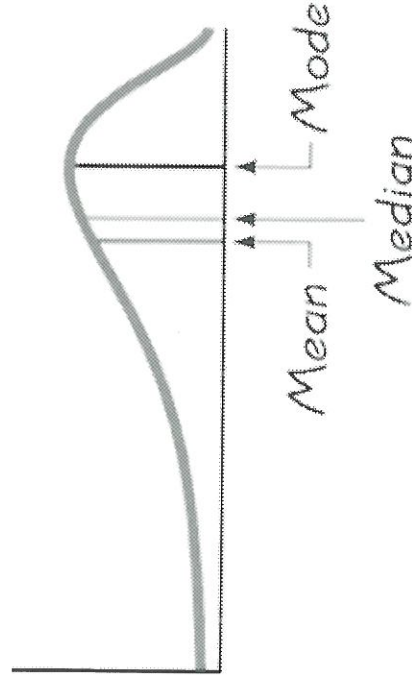


Symmetrisk eller skjev fordeling?

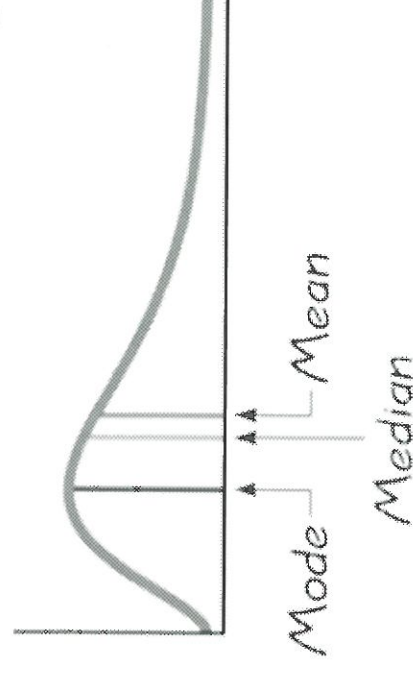
Symmetrisk



Venstreskjev



Høyreskjev



Gjennomsnitt fra en frekvenstabell

Beregne $\bar{x}$ fra en frekvenstabell

I hver klasse, anta at alle verdiene har midtverdien.

$$\bar{x} = \frac{\sum f \cdot x_{\text{midten}}}{\sum f}$$

her er f frekvensen/antallet i hver klasse.

Gjennomsnitt med utgangspunkt i frekvenstabellen for datasettet: 22 Oppg. 18:

117	110-124	(2)
132	125-139	(7)
147	140-154	17
162	155-169	11
177	170-184	2
192	185-199	1

40

midtpunktet i hvert intervall

$$\frac{110 + 124}{2} = 117$$

$$\frac{125 + 139}{2} = 132 = 117 + 15$$

↑
intervall-
bredden

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_{40}}{40}$$

x_{midten}

$$\approx \frac{117 \cdot 2 + 132 \cdot 7 + 147 \cdot 17 + 162 \cdot 11 + 177 \cdot 2 + 192 \cdot 1}{40}$$

$$\approx \underline{149.6}$$

② Variationsmål:

[T] 3.3

- variationsbredd
- varians = variance
- standardavvik = standard deviation

Variationsmål mäter variationen i datasettet.

Hvor mye varierer dataene?

Variasjon

- Variasjon er grunntemaet i statistikk
- Seksjon 3-3 i boka er dermed en av de aller viktigste
- Legg stor vekt på tolkning av standardavviket

Et enkelt forslag...

Variasjonsbredde

Variasjonsbredde = Maksverdi - Minimumsverdi

Example

Stikkprøve med tre tall: **3, 7, -4**. Her er variasjonsbredden

$$x_{\max} - x_{\min} = 7 - (-4) = 11.$$

Primitivt

Variasjonsbredde er et veldig enkelt mål for variasjon i dataene. For enkelt.

Notasjon brukes i formler

Example

I en statistikk bok ser du formelen

$$s^2 = \text{var}(x) = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$s = \text{std. avvik}(x) = \sqrt{\text{var}(x)}$$

som skal regnes ut for en stikkprøve med tre tall: **3, 7, -4**. Du regner ut at gjennomsnittet er $\bar{x} = 2$ og setter $n = 3$ og får:

$$\underline{x_1 = 3} \quad \underline{x_2 = 7} \quad \underline{x_3 = -4} \quad n = 3$$

$$\bar{x} = \frac{3 + 7 + (-4)}{3} = \frac{6}{3} = 2$$

$$\frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(3-2)^2 + (7-2)^2 + (-4-2)^2}{3-1} = \frac{(3-2)^2 + (7-2)^2 + (-4-2)^2}{3-1}$$

$$= \frac{1^2 + 5^2 + (-6)^2}{2} = \frac{1 + 25 + 36}{2}$$

$$= 31$$

Notasjon brukes i formler

Example

I en statistikk bok ser du formelen

$$\frac{\sum (x_i - \bar{x})^2}{n - 1}$$

som skal regnes ut for en stikkprøve med tre tall: **3, 7, -4**. Du regner ut at gjennomsnittet er $\bar{x} = 2$ og setter $n = 3$ og får:

$$\begin{aligned} \frac{\sum (x_i - \bar{x})^2}{n - 1} &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{n - 1} \\ &= \frac{(3 - 2)^2 + (7 - 2)^2 + (-4 - 2)^2}{3 - 1} \\ &= \frac{1^2 + 5^2 + (-6)^2}{2} = \frac{1 + 25 + 36}{2} = 31 \end{aligned}$$

Standardavviket

Standardavvik

- Det desidert mest brukte målet for variasjon i dataene er *standardavviket*
- Engelsk: standard deviation, std.dev
- Symbolet er s

Formelen

Standardavviket måler hvor mye verdiene ligger ifra gjennomsnittet:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Utrekning av s for hånd

Example

Stikkprøve med tre tall: **3, 7, -4**.

$$\begin{aligned}s &= \sqrt{\frac{(3-2)^2 + (7-2)^2 + (-4-2)^2}{3-1}} \\&= \sqrt{\frac{1+25+36}{3-1}} \\&= \sqrt{31} \approx 5.6\end{aligned}$$

Selv om JMP eller Excel regner ut s, så skal dere kunne gjøre dette for hånd for enkle eksempler.

Varians:

Symbol: S^2

Defn: Variansen til et datasett er

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

Eks: Data: $x_1 = 3$ $x_2 = 7$ $x_3 = -4$

Varians: $S^2 = \frac{(3-2)^2 + (7-2)^2 + (-4-2)^2}{3-1} = 31$

Std.
avvik: $S = \sqrt{S^2} = \sqrt{31} \approx \underline{\underline{5.6}}$

Eks: $x_1 = 4$ $x_2 = 6$ $\bar{x} = \frac{4+6}{2} = 5$

$$(x_1 - \bar{x}) = 4 - 5 = -1 \quad (x_1 - \bar{x})^2 = (-1)^2 = 1$$

$$(x_2 - \bar{x}) = 6 - 5 = 1 \quad (x_2 - \bar{x})^2 = 1^2 = 1$$

Hvis vi hadde brukt $\rightarrow \frac{\sum (x_i - \bar{x})}{n-1} = \frac{(x_1 - \bar{x}) + (x_2 - \bar{x})}{1}$

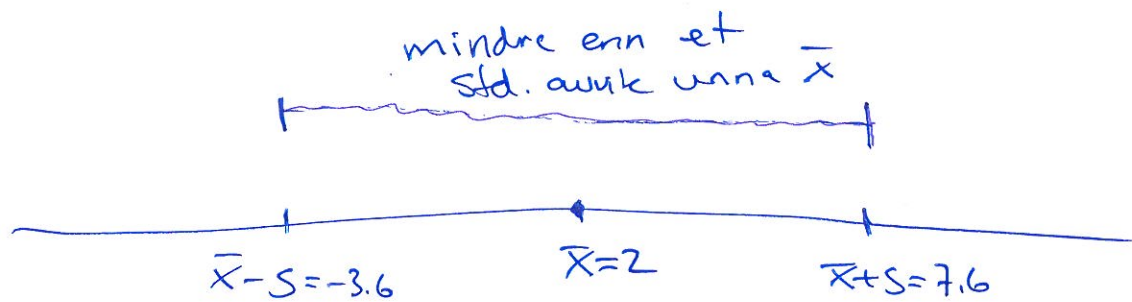
ville dette ikke faset $= \frac{-1 + 1}{1} = 0$

opp all variasjon i datasett

Vi bruker derfor $S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2}{1} = \frac{(-1)^2 + 1^2}{1} = 2$

og $S = \sqrt{S^2} = \sqrt{2} \approx 1.4$

Hva betyr det når $\bar{X} = 2$ og $s = 5.6$



Før normalfordeling:

Innenfor 1 std avvik :	68%	av observasjoner
— — 2 — — :	95%	— —
— — 3 — — :	99.7%	— —

Egenskaper til standardavviket

- s måler typisk avstand ifra gjennomsnittet
- s er som regel positiv $s \geq 0$ og $s = 0$ betyr at alle data er like
- s kan øke veldig mye dersom en uteligger tas med
- s er målt i samme enheter som verdiene i stikkprøven

observasjoner i meter $\Rightarrow$ std. avvik i meter
varians i $\text{lev. meter} = m^2$