1.3 Beskrivende statistikk

- sentralmål: gjennomsnitt og median
- spredningsmål: standardavvik og kvartilbredde

- hensiktsmessig presentasjon av data

Histogram

- eksempel: datasett med alder på studenter:

19, 20, 22, 25, 29 osv.

- deler datamaterialet inn i intervaller/ "klasser", feks.

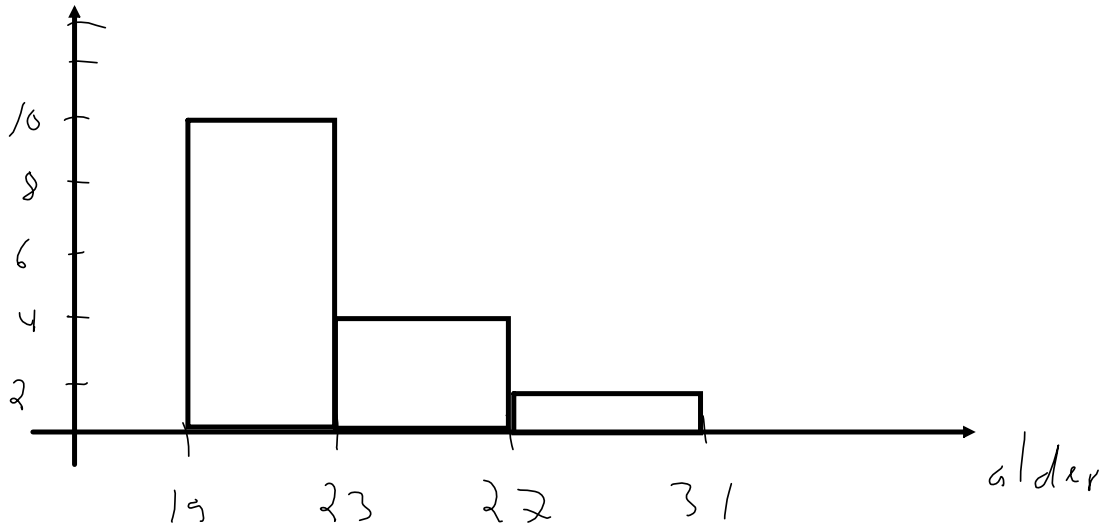
$[19, 23\rangle$, $[23, 27\rangle$, $[27, 31\rangle$

- klassebredden er 4
- utgangspunkt: frekvenstabell som viser antall i hver klasse:

klasse	antall
$[19, 23\rangle$	10
$[23, 27\rangle$	4
$[27, 31\rangle$	2

Histogram:

ant. obs.



Sentralmål (mål på hvor "sentrum" for et datasett er)

To vanligste sentralmål:

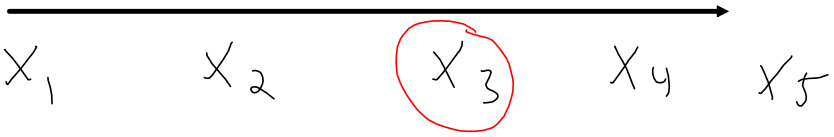
1) Gjennomsnittet av målingene $X_1, X_2, \ldots, X_n$

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- ulempe: gjennomsnitt er følsomt for utliggere/ekstremverdier

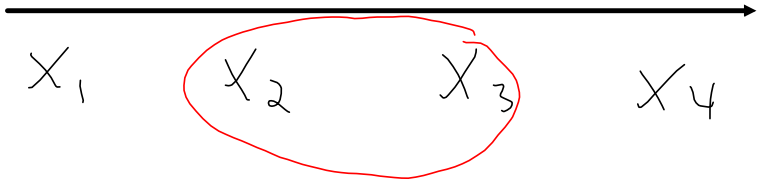
2) Median: observasjonen i midten av datasettet, når dette er sortert i stigende rekkefølge

Odde antall



Median

Partall antall



snittet av de to midterste

- fordel: median er mindre følsom for utliggere

Spredningsmål (mål på hvor "spredt" et datasett er)

1) Standardavvik / Varians

Varians for et datasett $X_1, X_2, \dots, X_n$ er definert som

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

WTF?

Standardavviket er da rota av variansen:

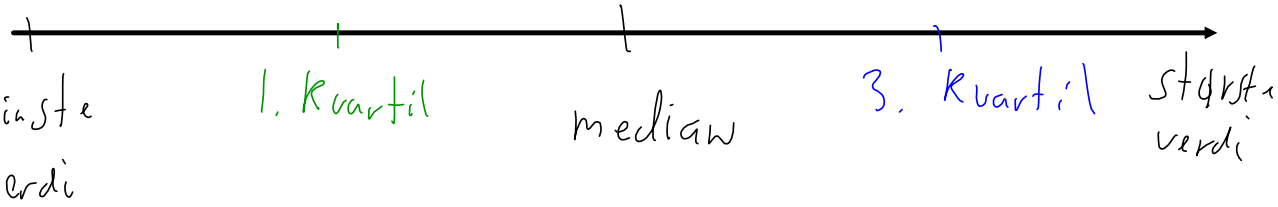
$$S = \sqrt{\text{Varians}} = \sqrt{S^2}$$

- ulempe: følsomt for utliggere

2) Kvartilbredde

En kvartil ("fjerdedel") for et datasett er en observasjon som deler datasettet (sortert i stigende rekkefølge) i fire like store deler:

$Q_1 \approx$ obs. nr. $\frac{n+1}{4}$ $\quad Q_3 =$ obs. nr. $3 \cdot \left(\frac{n+1}{4}\right)$

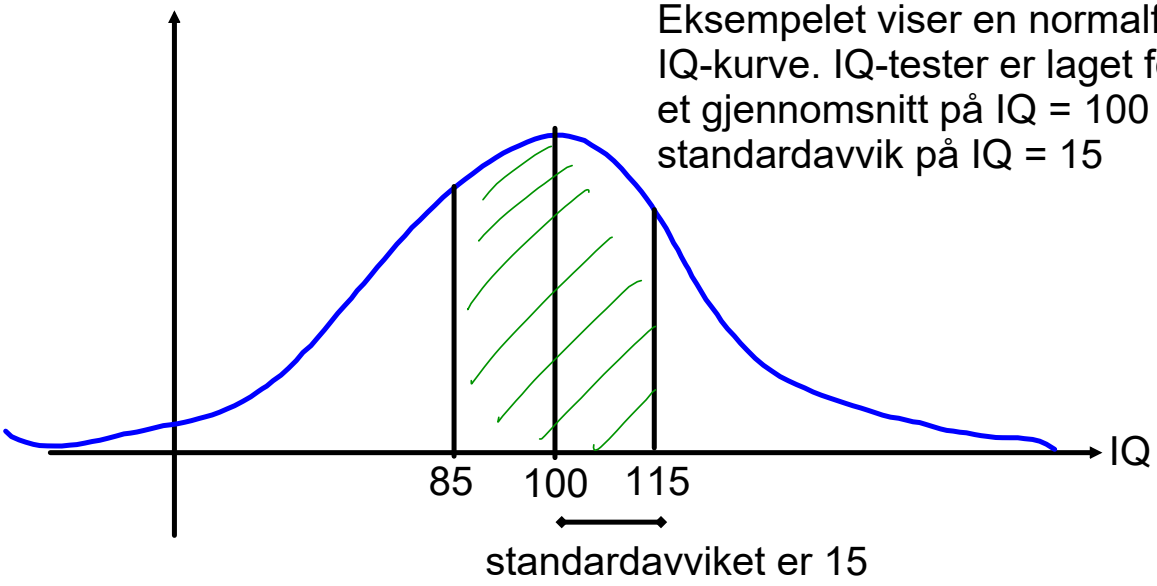


Spredningsmålet kvartilbredde er definert som

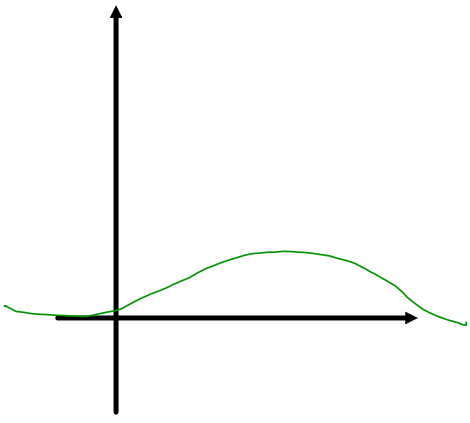
$$Q_3 - Q_1$$

- fordel: mindre følsomt for utliggere enn standardavvik

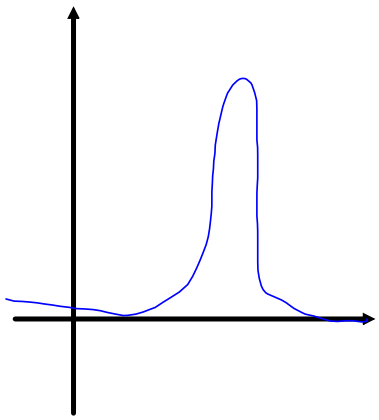
Figurer som illustrerer sentral- og spredningsmål



Standardavvikets betydning for utseendet til fordelingen



stort standardavvik
= stor "spredning"



lite standardavvik
= liten "spredning"

Oppsummering av nyttige Excel-formler for sentral- og spredningsmål og datavisualisering

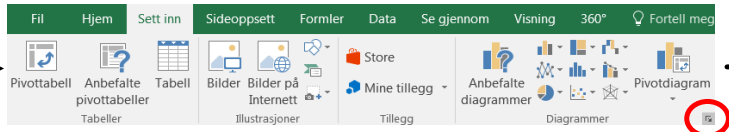
Anta at dataene ligger i én kolonne, i området A1-A100.

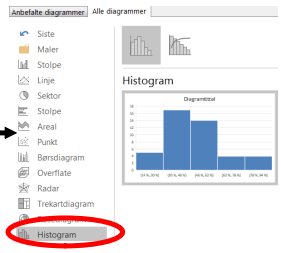
Gjennomsnitt	=GJENNOMSNITT(A1:A100)
Standardavvik	=STDAV.S(A1:A100)
Median	=MEDIAN(A1:A100)
Øvre kvartil Q3	=KVARTIL.EKS(A1:A100;3)
Nedre kvartil Q1	=KVARTIL.EKS(A1:A100;1)
Kvartilbredde	(ber Excel beregne Q3-Q1)

Forskjellen mellom "inklusiv" og "ekslusiv" kvartil er ikke noe vi skal bry oss om, men handler om hvordan man "vekter" dersom kvartilene havner mellom forskjellige observasjoner

Sette inn histogram:

Merk data-området

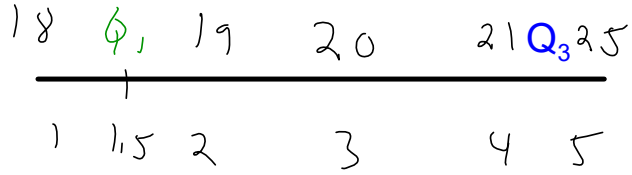
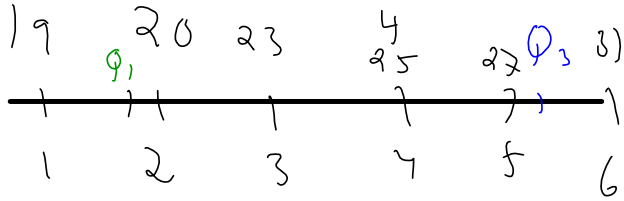




Velg Sett inn -meny og trykk den lille pila for å se alle diagrammer

Eksempel på beregning av median/kvartiler

Alderen til noen studenter er gjengitt i datasettene under:

Odde antall obs.	Partall antall obs.
18, 19, 20, 21, 25	19, 20, 23, 25, 27, 31
Median: $M = 20$	Median: $M = \frac{23+25}{2} = 24$
Q_1 = obs. nr. $\frac{(5+1)}{4}$ = nr. 1,5	Q_1 = obs. nr. $\frac{(6+1)}{4}$ = nr. 1,75
	
Her er $Q_1 = 18,5$	Her er $Q_1 = 19 + 0,75 \cdot (20-19) = 19,75$
Q_3 = obs. nr. $\frac{3(5+1)}{4} = 4,5$	Q_3 = obs. nr. $\frac{3(6+1)}{4} = 5,25$
Her er $Q_3 = 23$	Her er $Q_3 = 27 + 0,25 \cdot (31-27) = 28$