

# Medisinsk statistikk del I, høst 2009 (KL MED 8004 / KLH3004)

Eirik Skogvoll

Overlege/ 1.amanuensis dr.med.

Det medisinske fakultet, NTNU

Klinikk for anestesi og akuttmedisin, St. Olavs Hospital  
Trondheim

## Innføring i medisinsk statistikk

### Del I (KL MED8004/ KLH3004)

#### en variabel

- Deskriptiv statistikk
- Grafisk fremstilling
- Sannsynlighet
- Tilfeldig variabel
- Sannsynlighetsfordeling
- Hypotesetesting
- Konfidensintervall
- Ett- og to-utvalgsproblem
- Ikke-parametriske metoder

### Del II (KL MED8005) "flervariabel"

- Tabellanalyse
- Korrelasjonsanalyse
- Regresjon
- ANOVA
- Epidemiologiske metoder
- Overlevelsesanalyse
- Logistisk regresjon

## Statistikk-kurs ved DMF

- KL MED8004 Medisinsk statistikk del I. Høst
  - Dr. grads studenter
- KLH3004 Medisinsk statistikk. Høst.
  - Master klinisk helsevitenskap
- KL MED8005 Medisinsk statistikk del II. Vår
  - Dr. grads studenter
- KL MED8006 Anvendt medisinsk statistikk. Vår Seminarserie.
  - Dr. grads studenter. Enkeltforelesninger kan følges med utbytte.

## Hva kursene *ikke* handler om...

- Registrering, håndtering og manipulering av "data" generelt
- Oppbygging av databaser, kobling mot diagnoseregistre
- Utforming og tolkning av spørreskjema
- Epidemiologi

## Dag 1

- Praktisk om kurset
- Hva er statistikk?
- Statistisk terminologi og notasjon
- Deskriptiv statistikk
- Grafisk fremstilling

## Praktisk...

- Kursets hjemmeside:  
[http://folk.ntnu.no/slyderse/medstat/medstatl\\_h09.html](http://folk.ntnu.no/slyderse/medstat/medstatl_h09.html)
- Forelesning onsdager kl. 1215-1400, + i blokkuke (KLH3004)  
Sted: Sjekk hjemmesiden!
- Tillegg:  
Intro til R (24.08), Stata (19.08), SPSS (25.08 og 31.08).  
Ikke veldig viktig for dette kurset.  
Se hjemmesiden for detaljer og litteraturforslag.
- Kursleder: Professor Stian Lydersen ([stian.lydersen@ntnu.no](mailto:stian.lydersen@ntnu.no)),  
AKF, tlf. 725 75428
- Kurssekretær: Mariann Hansen ([mariann.hansen@ntnu.no](mailto:mariann.hansen@ntnu.no)),  
IKM, tlf. 725 98660

## Praktisk...

- 3 obligatoriske øvinger, hver med 8 deloppgaver
  - anvendelse av teori (referanse for senere bruk).
  - sikrer jevnt arbeid underveis
  - gruppevis samarbeid og innlevering
- Grupper á max. 8 personer:
  - Gruppen arbeider selvstendig med oppgaven før veiledning (min. 2 samlinger anbefales)
- Veiledning ca. hver 3. uke (KL MED8004: uke 39, 44 og 47, KLH3004: uke 38, 41 og 46). Felles veiledning for to og to grupper:
  - 1,2: Mandag 1415 -1600
  - 3,4: Onsdag 1215 -1400 (In English)
  - 5,6: Onsdag 1415 -1600
  - 7,8: Onsdag 1515 -1700
  - 9+10: Onsdag 1315 -1500
  - 11,12,13,14: Torsdag 0815-1000
  - Obligatorisk oppmøte!
  - Skal avklare problemer som ennå ikke er løst...

## Praktisk...

- Veileder vil ...  
notere oppmøte, motta gruppens øvinger, rette, godkjenne og returnere til gruppekontakt.
- Hver deltaker skal sikre seg en (godkjent) kopi.

## Evaluering

- Øvingene danner grunnlag for godkjent kurs!
- Eksamen = framlegg av oppgaver (15-20 min, faglærer og sensor):
  - Komplette øvingssett skal legges fram
  - Én oppgave (blant 24) trekkes ut og presenteres ...
  - Supplerende teoretiske spørsmål knyttet til oppgaven

## Hva er statistikk?

“Statistics” (eng.) betyr:

- Statistikk i vanlig forstand (tabeller, tellinger, “statistisk årbok”)
- En gren av matematikken med egen terminologi og metode
- Tallstørrelser beregnet fra et utvalg (middelverdi, median, empirisk varians, rekkevidde etc. – *ikke noe helt godt norsk ord for dette: “observator”*)

## Hva er statistikk?

«Generelt er statistikk det vitenskapelige verktøy som brukes for å trekke slutninger (...) basert på data som er beheftet med usikkerhet.»

Årsberetning, Institutt for matematikk og statistikk,  
Universitetet i Trondheim, 1994

## Hva er statistikk?

«Data have no meaning in themselves; they are meaningful only in relation to a conceptual model of the phenomenon studied.»

Box, Hunter & Hunter: Statistics for experimenters,  
Wiley 1978

## Hva er statistikk?

«I am dismayed by how often my clients ask whether a particular approach would be "legal" or "against the rules" rather than "accurate" or "misleading."»

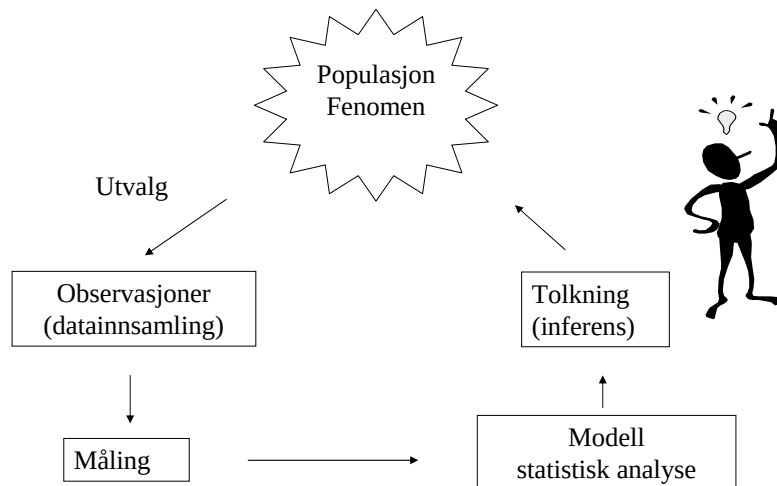
This misunderstanding of statistics as a body of *seemingly arbitrary dogma* leads many (...) to perceive violations even when the research has not actually been harmed.»

Professor Peter Bacchetti, Dept. of Epidemiology and Biostatistics, University of California (BMJ, 2002)

## Statistisk terminologi

- Populasjon
- Utvalg
- Variabel
- Sannsynlighetsfordeling
- Parameter

## Populasjon og utvalg (sample)



## Variabel

Definisjon:

En observasjon (en størrelse) som ikke er konstant, men kan ta forskjellige verdier

Numerisk variabel

- Kontinuerlig: kan anta alle verdier på tallinjen

Eksempler:

- Vekt (g)
- Høyde (cm)
- inntekt (kr)

- Diskret: en tellbar mengde (kan nummereres)

Eksempler:

- Celler pr. synsfelt
- Antall døde
- Resultat av terningkast



# Variabel

## Numerisk variabel

Andre aspekter.

- Naturlig nullpunkt?

Eksempel: 200 K er dobbelt av 100 K, mens 30 °C ikke er "dobbelt så varmt" som 15 °C

- Sensurert observasjon?

Vet bare om den er over eller under en viss grenseverdi.

Eksempler:

- CRP < 5 mg/l (deteksjonsgrense)
- Overlevelsestid > 120 dager (siste kontroll)

# Variabel

## Kategorisk variabel

- Ordinal: rangorden, men ikke definert avstand ("3 er mer enn 2, men hvor mye mer?")

Eksempler:

- APGAR score (0-10),
- NYHA-klassifisering (I-IV)

- Nominal: ingen ordning

Eksempler:

- livssyn (jøde, muslim, kristen, ateist, buddhist)
- kjønn (gutt, jente)
- bosted (by, tettsted, landsbygd)

# Variabel

Stokastisk (tilfeldig) variabel

- kan anta forskjellige verdier. Beskrives matematisk med sannsynlighetsfordeling (normal-,  $\chi^2$ , etc.)

Avhengig variabel, utfallsvariabel

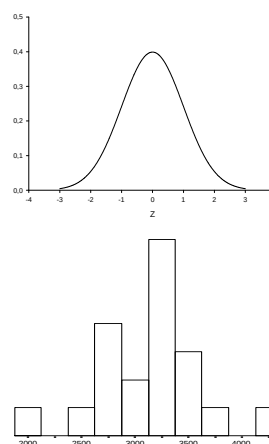
- stokastisk, "den vi måler og er interessert i"

Uavhengig variabel

- kjent, determinert, vanligvis ikke stokastisk, evt. grupperingsvariabel.

# Sannsynlighetsfordeling

- En beskrivelse av hvilke verdier den tilfeldige variabel kan ta, og sannsynligheten for hver enkelt verdi teoretisk →
- Et histogram beskriver empirisk (observert) hyppighet →



## Parameter

- En fast tallstørrelse, en konstant
- Definerer egenskapene til en sannsynlighetsfordeling  
Eksempel: variabelen "høyde" er normalfordelt med to parametre:  
forventningsverdi ( $\mu$ ) og varians ( $\sigma^2$ )
- Estimeres (beregnes) som regel fra data

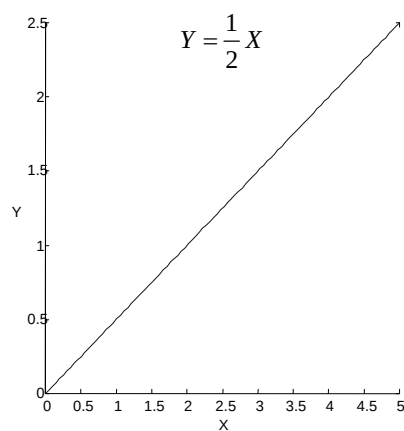
Parameter  $\neq$  variabel !

## Variabel vs. parameter

Likning for en rett linje:

$$Y = \alpha + \beta X$$

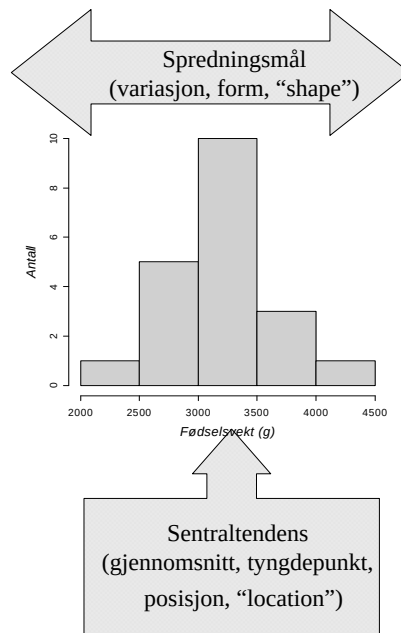
- Avhengig og uavhengig variabel?
- Parameter(e)?



# Deskriptiv statistikk

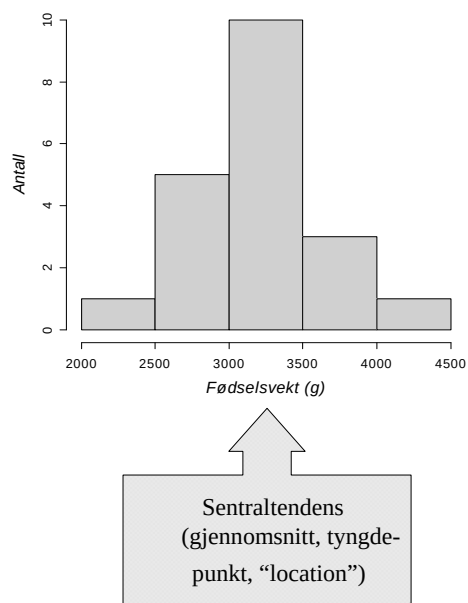
Oppsummere datamaterialet v.h.j.a. noen få, meningsfylte størrelser (tall):

- Sentraltendens (ofte middelerdi)
- Spredningsmål (ofte standard avvik)



## Mål på sentraltendens

- Aritmetisk\_middelerdi
- Justert (trimmet) middelerdi
- Median
- Mode
- (Midrange)
- (Geometrisk middelerdi)



# Aritmetisk middelværdi

## Synonymer

- Gjennomsnitt
- Middelværdi
- Mean value
- Sample mean
- Utvalgsmiddelværdi
- Empirisk middelværdi

## Definisjon (Rosner def. 2.1, s. 9)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

# Aritmetisk middelværdi

## Egenskaper

- Alle observasjoner må være kjent
- Observasjonene trenger ikke ordnes etter størrelse
- Matematisk gunstig
- Følsom for "outliers": ekstreme, utypiske observasjoner

## Definisjon (Rosner def. 2.1, s. 9)

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

**TABLE 2.1** Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

| $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ |
|-----|-------|-----|-------|-----|-------|-----|-------|
| 1   | 3265  | 6   | 3323  | 11  | 2581  | 16  | 2759  |
| 2   | 3260  | 7   | 3649  | 12  | 2841  | 17  | 3248  |
| 3   | 3245  | 8   | 3200  | 13  | 3609  | 18  | 3314  |
| 4   | 3484  | 9   | 3031  | 14  | 2838  | 19  | 3101  |
| 5   | 4146  | 10  | 2069  | 15  | 3541  | 20  | 2834  |

Eksempel (Rosner expl. 2.4, s.11)

(fra Rosner table 2.1, s.10)

$$\bar{x} = \frac{3265 \text{ g} + 3260 \text{ g} + \dots + 2834 \text{ g}}{20} = 3166,9 \text{ g}$$

Hvis  $x_1 = 500 \text{ g}$  istedet for 3265 g ...

$$\bar{x} = \frac{500 \text{ g} + 3260 \text{ g} + \dots + 2834 \text{ g}}{20} = 3028,7 \text{ g}$$

...en rel. stor endring på grunn av en enkelt observasjon

## Aritmetisk middelværdi

### Egenskaper

- Ikke definert for sensurerte observasjoner
- Middelværdi fra flere utvalg kan kombineres:  $\rightarrow$
- "I det lange løp" vil middelværdien konvergere mot forventningsverdien til den tilfeldige variabel

$$\bar{\bar{x}} = \frac{\sum_{j=1}^k n_j \bar{x}_{n_j}}{\sum_{j=1}^k n_j} = \frac{n_1 \bar{x}_{n_1} + n_2 \bar{x}_{n_2} + \dots + n_k \bar{x}_{n_k}}{n_1 + n_2 + \dots + n_k}$$

$$E(\bar{X}) = \mu = \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{i=1}^n x_i \right)$$

Eksempel. Kombinasjon av flere middelveier :

$$\bar{x}_1 = 18,4 \quad n_1 = 22$$

$$\bar{x}_2 = 17,9 \quad n_2 = 11$$

$$\bar{x}_3 = 9,5 \quad n_3 = 16$$

$$\bar{x}_4 = 15,1 \quad n_4 = 5 \quad (\text{dvs. } k = 4)$$

$$\begin{aligned} \bar{\bar{x}} &= \frac{\sum_{j=1}^k \bar{x}_{n_j}}{\sum_{j=1}^k n_j} = \frac{n_1 \bar{x}_{n_1} + n_2 \bar{x}_{n_2} + \dots + n_k \bar{x}_{n_k}}{n_1 + n_2 + \dots + n_k} \\ &= \frac{22 \cdot 18,4 + 11 \cdot 17,9 + 16 \cdot 9,5 + 5 \cdot 15,1}{22 + 11 + 16 + 5} \\ &= \frac{404,8 + 196,9 + 152 + 75,5}{54} = \frac{829,2}{54} = 15,36 \end{aligned}$$

## Justert (trimmet) middelveier

- Aritmetisk middelveier basert på de sentrale 90 - 95 % av observasjonene; "halene" (5 - 10 %) er trimmet bort
- Mindre følsom for ekstreme observasjoner

# Median

## Synonym

- 50 percentil
- empirisk median

## Egenskaper

- Observasjonene ordnes i stigende rekkefølge
- “Median = den verdien som deler materialet i to”
- Ufølsom for ekstreme observasjoner

Transformasjon :

$$X_i \xrightarrow{\text{ordnes}} X_{(i)}$$

**TABLE 2.1** Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

| $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ | $i$ | $x_i$ |
|-----|-------|-----|-------|-----|-------|-----|-------|
| 1   | 3265  | 6   | 3323  | 11  | 2581  | 16  | 2759  |
| 2   | 3260  | 7   | 3649  | 12  | 2841  | 17  | 3248  |
| 3   | 3245  | 8   | 3200  | 13  | 3609  | 18  | 3314  |
| 4   | 3484  | 9   | 3031  | 14  | 2838  | 19  | 3101  |
| 5   | 4146  | 10  | 2069  | 15  | 3541  | 20  | 2834  |

Eksempel (Rosner expl. 2.5, s. 11)

(fra Rosner table 2.1, s. 10)

$$x_1 = 3265$$

$$x_2 = 3260$$

$$x_3 = 3245$$

$$x_4 = 3484$$

...

Ordnet i stigende rekkefølge:

$$x_{(1)} = 2069$$

$$x_{(2)} = 2581$$

$$x_{(3)} = 2759$$

$$x_{(4)} = 2834$$

...



# Median

## Egenskaper

- Inntil halvparten av observasjonene kan være sensurerte
- Ugunstig matematisk!
- Median fra flere utvalg kan *ikke* kombineres

## Definisjon (Rosner def. 2.2, s. 11)

(1) Hvis ulike antall observasjoner ( $n = 3, 5, 7, \dots$ )

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

(2) Hvis like antall observasjoner ( $n = 2, 4, 6, \dots$ )

$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}$$

(andre definisjoner kan forekomme)

# Median

**TABLE 2.2**

Sample of admission white-blood counts ( $\times 1000$ ) for all patients entering a hospital in Allentown, PA, on a given day

| $i$ | $x_i$ | $i$ | $x_i$ |
|-----|-------|-----|-------|
| 1   | 7     | 6   | 3     |
| 2   | 35    | 7   | 10    |
| 3   | 5     | 8   | 12    |
| 4   | 9     | 9   | 8     |
| 5   | 8     |     |       |

## Eksempel (Rosner expl. 2.6, s.11)

(fra Rosner table 2.2, s. 12)

$n = 9$ , så definisjon (1) benyttes:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{9+1}{2}\right)} = x_{\left(\frac{10}{2}\right)} = x_{(5)} = 8$$

# Median

**TABLE 2.1** Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

| <i>i</i> | $x_i$ | <i>i</i> | $x_i$ | <i>i</i> | $x_i$ | <i>i</i> | $x_i$ |
|----------|-------|----------|-------|----------|-------|----------|-------|
| 1        | 3265  | 6        | 3323  | 11       | 2581  | 16       | 2759  |
| 2        | 3260  | 7        | 3649  | 12       | 2841  | 17       | 3248  |
| 3        | 3245  | 8        | 3200  | 13       | 3609  | 18       | 3314  |
| 4        | 3484  | 9        | 3031  | 14       | 2838  | 19       | 3101  |
| 5        | 4146  | 10       | 2069  | 15       | 3541  | 20       | 2834  |

Eksempel (Rosner expl. 2.5, s.11)

(fra Rosner table 2.1, s. 10)

$n = 20$ , så definisjon (2) benyttes:

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}}{2} = \frac{x_{(\frac{20}{2})} + x_{(\frac{20+2}{2})}}{2} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3245 \text{ g} + 3248 \text{ g}}{2} = 3246,5 \text{ g}$$

# Mode

Rosner def. 2.3

Den observasjon som forekommer hyppigst.

**TABLE 2.3** Sample of time intervals between successive menstrual periods (days) of college-age women

| Value | Frequency | Value | Frequency | Value | Frequency |
|-------|-----------|-------|-----------|-------|-----------|
| 24    | 5         | 29    | 96        | 34    | 7         |
| 25    | 10        | 30    | 63        | 35    | 3         |
| 26    | 28        | 31    | 24        | 36    | 2         |
| 27    | 64        | 32    | 9         | 37    | 1         |
| 28    | 185       | 33    | 2         | 38    | 1         |

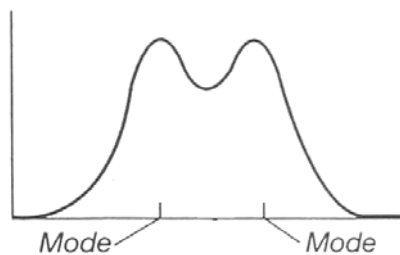
Rosner  
eks. 2.7,  
s. 13  
Mode =  
28

# Mode

- Lite i bruk. Ugunstig matematisk.
- Mer abstrakt: unimodal fordeling (en "topp"), bimodal (to topper) osv.

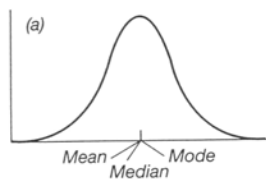
Eksempel på  
bimodal  
sannsynlighets-  
fordeling

(Zar 1999,  
fig. 3.2 b)

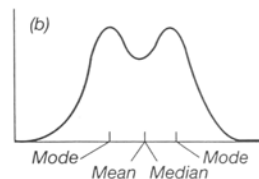


## Mål på sentral tendens - sammenstilling

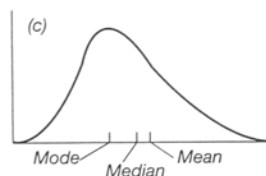
Unimodal,  
symmetrisk



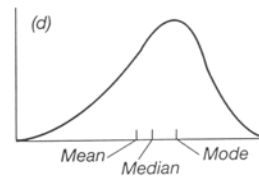
Bimodal,  
symmetrisk



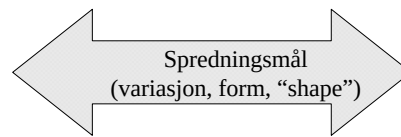
Unimodal,  
høyre (pos.)  
- skjev



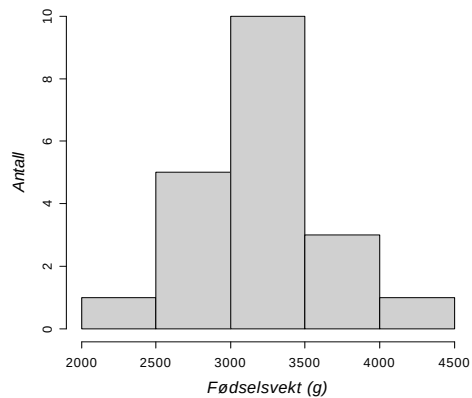
Unimodal,  
venstre (neg.)  
- skjev



# Spredningsmål



- Rekkevidde (range)
- Kvantiler, percentiler: interkvartil avstand
- Varians
- Standard avvik
- Variasjonskoeffisient



## Rekkevidde

### Synonym

- Range
- Variasjonsbredde

### Egenskaper

- Samme enhet som observasjonene
- Benytter bare to av dem
- Følsom for ekstreme observasjoner
- Øker med utvalgsstørrelsen  $n$

### Definisjon (BR def. 2.5, s. 18)

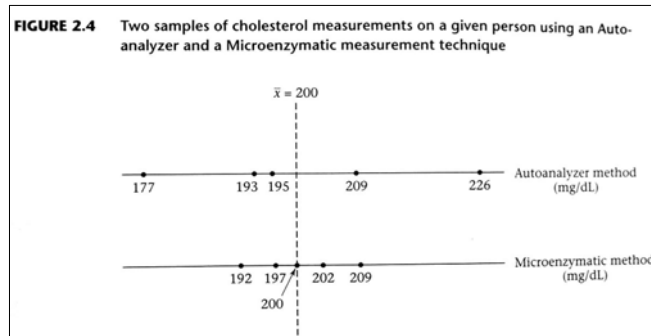
$$Range = X_{(n)} - X_{(1)}$$

### Eksempel (BR expl. 2.14, s. 18)

(fra BR table 2.1, s. 10)

$$\begin{aligned} Range &= x_{(20)} - x_{(1)} = 4146 \text{ g} - 2069 \text{ g} \\ &= 2077 \text{ g} \end{aligned}$$

## Rekkevidde (range)



Eksempel (BR expl. 2.15, s. 18)

(fra BR Fig 2.4, s. 18)

Range (auto-) =  $x_{(5)} - x_{(1)} = 226 \text{ mg/dl} - 177 \text{ mg/dl} = 49 \text{ mg/dl}$

Range (micro-) =  $x_{(5)} - x_{(1)} = 209 \text{ mg/dl} - 192 \text{ mg/dl} = 17 \text{ mg/dl}$

## Kvantiler og percentiler

Definisjon (Rosner def 2.6)

“Den tallverdi  $V_p$  som har en viss andel ( $p$ ) av (de ordnede) observasjonene under seg”,  $0 < p < 1$

(En observasjon teller omtrent  $1/n$ -del. Hvis  $n = 100$  teller hver observasjon 1 %. Observasjon nr. 20 svarer da til 20-percentilen.)

Formelt

$n$  ordnede observasjoner :  $X_{(1)}, X_{(2)}, X_{(3)}, \dots, X_{(k)}, X_{(k+1)}, \dots, X_{(n)}$

(1)  $V_p = X_{(k+1)}$  når  $np$  ikke er et heltall ("integer"),  $k < np < k+1$

(2)  $V_p = \frac{X_{(np)} + X_{(np+1)}}{2}$  når  $np$  er et heltall

## Kvantiler og percentiler

Eksempel (Rosner, expl. 2.16)

Finn  $V_{0,1}$  og  $V_{0,9}$  (10- og 90 percentilene) basert på fødselsvekt-data i table 2.1.

$p = 0,1$  og  $np = 2$  (et heltall) så vi bruker (2) og får

$$V_{0,1} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{2581 + 2759}{2} = 2670g$$

$p = 0,9$  og  $np = 18$  (et heltall fortsatt), vi bruker (2) og får

$$V_{0,9} = \frac{x_{(18)} + x_{(19)}}{2} = \frac{3609 + 3649}{2} = 3629g$$

De sentrale 80 % av barna veide altså mellom 2670 g og 3629 g, dvs. en spredning på ca 1000 g.

## Kvantiler og percentiler

$V_{0,5}$  = median = 50-percentil. Definisjonene må stemme overens.

Eksempel (fra Rosner table 2.1, s.10)

$np = 10$ , så definisjon 2.6 (2) benyttes :

$$V_{0,5} = \tilde{x} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{3245 \text{ g} + 3248 \text{ g}}{2} = 3246,5 \text{ g}$$

# Interkvartil-avstand

## Synonym

- Interquartile range - IQR

## Egenskaper

- Samme enhet som observasjonene
- Omfatter de sentrale 50 %
- Mindre følsomt for ekstreme observasjoner
- Definerer "boksen" i et box plott

## Definisjon

$$IQR = V_{0,75} - V_{0,25}$$

$V_{0,75}$  - øvre (3.) kvartil

$V_{0,25}$  - nedre (1.) kvartil

# Interkvartil-avstand

Eksempel (fra Rosner table 2.1, s. 9)

Finn IQR for fødselsvekt.

$$IQR = V_{0,75} - V_{0,25} \quad (np = 15 \text{ og } 5, \text{ respektive})$$

$$V_{0,75} = \frac{x_{(15)} + x_{(16)}}{2} = \frac{3323 \text{ g} + 3484 \text{ g}}{2} = 3403,5 \text{ g}$$

$$V_{0,25} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{2838 \text{ g} + 2841 \text{ g}}{2} = 2839,5 \text{ g}$$

$$IQR = 3403,5 - 2839,5 = 564 \text{ g}$$

## Andre mulige spredningsmål...

Observasjoner:

$x_1, x_2, \dots, x_n$

-Gjennomsnittlig avvik?

$$\frac{1}{n} \sum_{i=1}^n x_i - \bar{x}$$

Umulig! Summerer seg til 0 (null).

-Gjennomsnittlig, absolutt avvik?

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

OK, men matematisk ugunstig.

-Gjennomsnittlig, *kvadrert* avvik?

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

Best!

## Empirisk (utvalgs-) varians

Definsjon (Rosner Def.2.7)

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

Kommentar

Bruk av  $(n-1)$  i stedet for  $n$  i nevneren skyldes at vi gjør bruk av utregnet middelerdi i beregningen.

Dermed har vi «brukt opp» noe av informasjonen (nemlig 1 "frihetsgrad") som lå i utvalget.

Resultatet er en noe større verdi for  $S^2$ . For stor  $n$  er forskjellen ubetydelig.



## Empirisk (utvalgs-) varians

Eksempel (Rosner Expl. 2.19, s 21- 22)

Auto - analyzer,  $\bar{x} = 200$  mg/dl

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{4} \cdot [(177 - 200)^2 + (193 - 200)^2 + (195 - 200)^2 + (209 - 200)^2 + (226 - 200)^2] \\ &= \frac{1}{4} \cdot (529 + 49 + 25 + 81 + 676) = \frac{1360}{4} = 340 \end{aligned}$$

Micro – enzymatic (samme  $\bar{x}$ )

$$\begin{aligned} S^2 &= \frac{1}{4} \cdot [(192 - 200)^2 + (197 - 200)^2 + (200 - 200)^2 + (202 - 200)^2 + (209 - 200)^2] \\ &= \frac{1}{4} \cdot (64 + 9 + 0 + 4 + 81) = \frac{158}{4} = 39,5 \end{aligned}$$

## Empirisk (utvalgs-) varians

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

Egenskaper

- Alltid positiv
- Sterkt avvikende observasjoner bidrar mest (kvadrering)
- Gunstige matematiske egenskaper
- Resultatet kan ikke tolkes på original skala

(kg  $\rightarrow$  kg<sup>2</sup>, mg/dl  $\rightarrow$  mg<sup>2</sup>/dl<sup>2</sup>, cm  $\rightarrow$  cm<sup>2</sup> osv.)

## Empirisk (utvalgs-) standardavvik

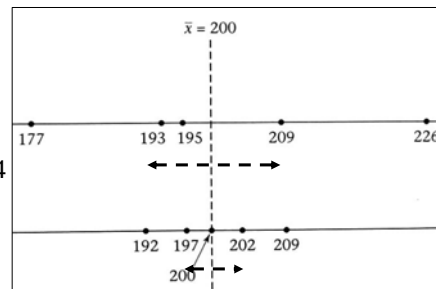
Definisjon (Rosner Def.2.8)

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

Eksempel (Rosner Expl. 2.19, s. 21)

Auto-analyzer:  $S = \sqrt{S^2} = \sqrt{340} = 18,4$

Micro – enz:  $S = \sqrt{S^2} = \sqrt{39,5} = 6,3$



## Empirisk (utvalgs-) standardavvik

Egenskaper

- Synonym: Standard Deviation (SD)
- Alltid positiv
- Sterkt avvikende observasjoner bidrar mest (kvadrering)
- Gunstige matematiske egenskaper
- Resultatet kan tolkes på opprinnelig skala!

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

## Empirisk (utvalgs-) standardavvik

- Omtrent 95 % av populasjonen befinner seg innenfor:

$$\bar{x} \pm 2SD$$

... som altså omtrent tilsvarer 2,5 og 97,5 percentilene

- Følger av normalfordelingens egenskaper
- Forutsetter noenlunde symmetri og unimodalitet

- “quick & dirty”:

$$SD \approx \frac{Range}{4}$$

- Nyttig f.eks. Hvis man må gjette på SD ved beregning av utvalgsstørrelse

## Grafisk framstilling

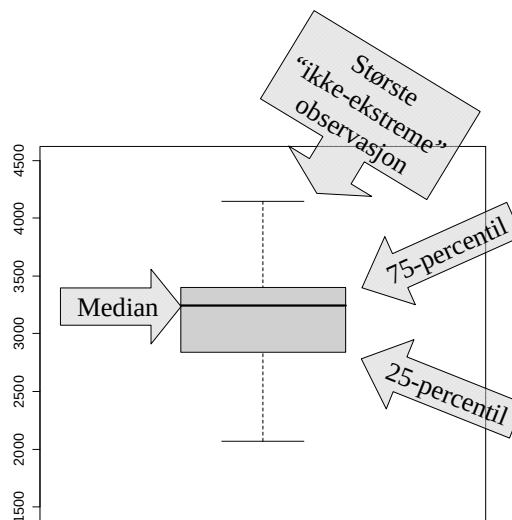
- Søylediagram
- Histogram
- Stem-and-leaf
- Box-plot

“En god figur sier mer  
enn 1000 T-tester!”

# Box-plott

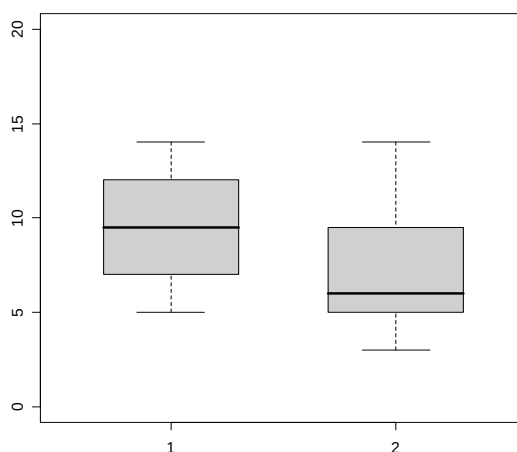
Egenskaper

- Raskt overblikk over materialets tyngdepunkt og spredning
- Ekstreme observasjoner (outliers) kan identifiseres



# Box-plott

Lett å sammenlikne grupper!



# Histogram

## Egenskaper

- Empirisk "sannsynlighetsfordeling"
- Viser generell form og tyngdepunkt
- Vanskelig å sammenlikne grupper
- Enkeltobservasjoner er usynlige

