Medical statistics part I, autumn 2009 (KLMED 8004 / KLH3004)

Eirik Skogvoll

Consultant/ assoc. professor
Faculty of Medicine, NTNU
Dept. of Anaesthesiology and Emergency Medicine,
St. Olav University Hospital
Trondheim

Overview of the introductory courses

Part I (KLMED8004/ KLH3004) one variable

- Descriptive statistics
- Graphics
- Probability
- Random variable
- Probability distribution
- Hypothesis testing
- Confidence intervals
- One- and two sample procedures
- Non-parametric methods

Part II (KLMED8005) "multi" variable

- Table analysis
- Correlation
- Linear regression
- Analysis of variance (ANOVA)
- Survival analysis
- Logistic regression

What we do not cover …

- Data management in general
- Registry and database design
- Questionnaire design and analysis
- Epidemiology

Day 1

- Practical and organizational details
- What is statistics?
- Statistical terminology and some basic mathematical notation
- Descriptive statistics
- Graphical display

Practical...

- Home page of the course:
  http://folk.ntnu.no/slyderse/medstat/medstatI_h09.html
- Lecture on Wednesdays, 1215-1400, + block weeks (KLH3004)
  Where? Check the home page!
- Additional lectures:
  Introduction to R (24.08), Stata (19.08), SPSS (25.08 og 31.08).
  These are not *that* important for the course. Check the home page for details and supplementary literature.
- Course manager: Professor Stian Lydersen
  (stian.lydersen@ntnu.no), phone 725 75428
- Course administrator: Mariann Hansen
  (mariann.hansen@ntnu.no), phone 725 98660

Practical...

- 3 mandatory exercises, each with 8 parts
  - Practical application of theory (for future reference)
  - Ensures continuous work throughout the course
  - Group-wise collaboration and hand-in
- Groups of (maximum) 8 persons:
  - Each group works independently (minimum two meetings), and then receives 2 hours of mandatory guidance with instructor to clarify unsolved issues
- Guidance approx. every 3. week (KLMED8004: week no. 39, 44 and 47, KLH3004: week no. 38, 41 and 46).
- Joint guidance for two groups:
  - Group 1,2: Monday 1415 -1600
  - 3, 4: Wednesday 1215 -1400 (In English)
  - 5, 6: Wednesday 1415 -1600
  - 7, 8: Wednesday 1515 -1700
  - 9,10: Wednesday 1315 -1500
  - 11,12,13,14: Thursday 0815- 1000

Practical...

- The instructor will ...
  record attendance, receive the written assignment (i.e. exercise), correct and comment it, eventually approve, and return it to the group contact
- Every course participant shall possess an approved copy of every assignment

Course assessment

- Course approval requires completed and approved assignments
- Examination *used to* consist of presenting the exercises (15-20 min):
  - The complete set of approved exercises must be brought and displayed
  - One exercise is drawn at random for detailed discussion
  - Supplementary questions in relation to this
- This term possibly written examination (4 hours)

What is statistics?

"Statistics" means:

- Numerical summaries of information, as commonly understood
  (tables, census, graphs etc.)
- A branch of mathematics with its own terminology and methods
- Numerical quantities calculated from a sample
  (mean, median, sample variance, range etc.)

  (*– ikke noe helt godt norsk ord for dette. "observator")*

What is statistics?

«The science of collecting, displaying and analysing data»

Oxford Dictionary of Statistics, 2004

What is statistics?

«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

What is statistics?

«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.“

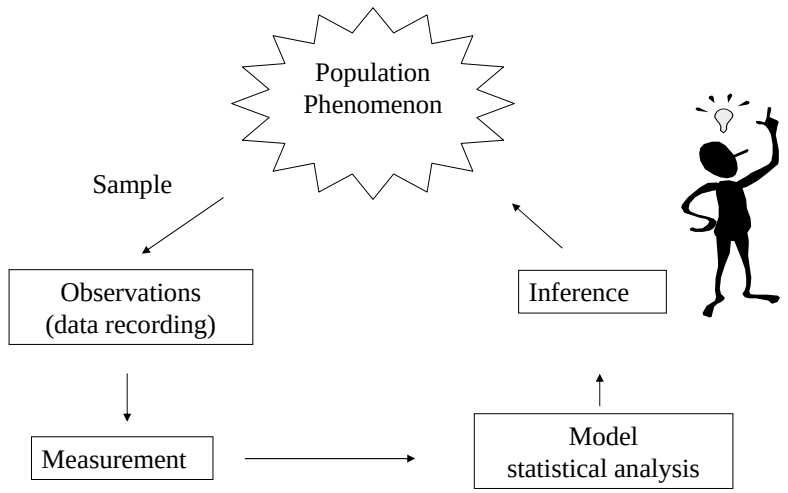
This misunderstanding of statistics as a body of *seemingly arbitrary dogma* leads many (…) to perceive violations even when the research has not actually been harmed.»

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

Statistical terminology

- Population
- Sample
- Variable
- Probability distribution
- Parameter

Population and sample



Variable

Definition:
An observation which is not constant, but may take on different values

Numerical variable

- Continuous: may take on all possible values on the real line

  Examples:
  - Weight (g)
  - Height (cm)
  - income (kr)
- Discrete: a countable set

  Examples:
  - Cells per microscopy field
  - No. of deaths over a time period
  - The results from a dice roll

Variable

Numerical variable

Other aspects.

- Natural zero point?

  Example: In the physical sense, 200 K is twice 100 K, but 30 °C is not "twice as hot" as 15 °C
- Censored observation?
  We only know whether the observation exceeds or is less than some limit.

  Examples:
  - CRP < 5 mg/l (limit of detection)
  - Survival time > 120 days (last appointment)

Variable

Categorical variable

- Ordinal: rank, but distance is undefined ("3 is more than 2, but how much more?")

  Examples:
  - APGAR score (0-10),
  - NYHA classification (I-IV)
- Nominal: no ranking

  Examples:
  - Philosophy/ religion (Jew, Muslim, Christian, Atheist, Buddhist)
  - gender (boy, girl)
  - Place of living (city, town, village)

Variable

Stochastic (random) variable

- May take on different values. Mathematically described with a probability distribution (Normal-, χ^2, etc.)

Dependent variable, outcome variable

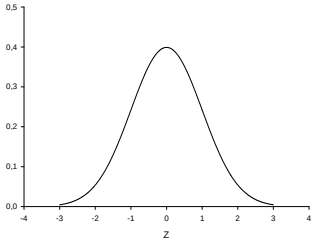
- stochastic, "the one we measure and have focus on"

Independent variable

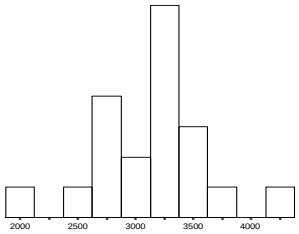
- known, deterministic, (usually) not stochastic, e.g. a grouping variable (then known as a "factor")

Probability distribution

- Describes which values the stochastic variable may take on, and the theoretical probability of each value →



- A histogram describes the observed frequency →



Parameter

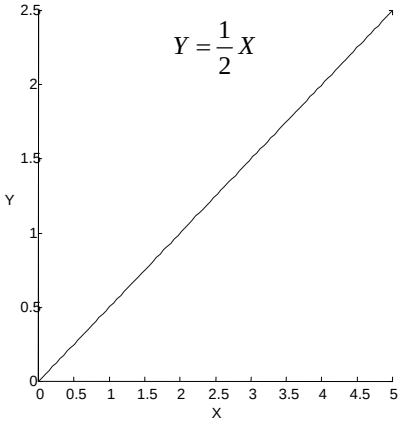
- A numerical constant
- Defines the properties (location and shape) of a probability distribution
  - Example: the variable "height" is Normally distributed with two parameters: expected value (μ) and variance (σ^2)
- Usually needs to be estimated (calculated) from data

Parameter ≠ variable !

Variable vs. parameter

Equation of a straight line:

$Y = \alpha + \beta X$



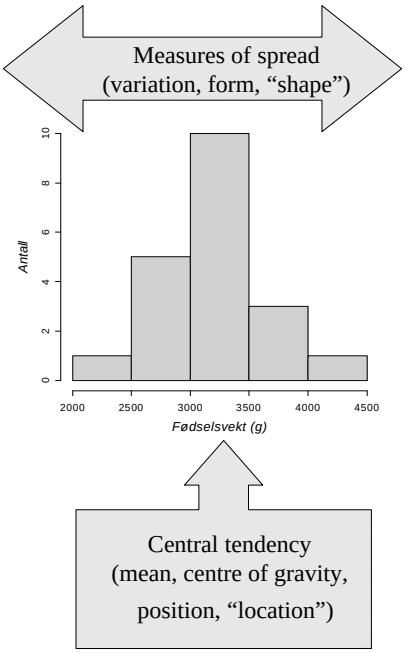
•Dependent and independent variable?

• Parameter (s)?

Descriptive statistics

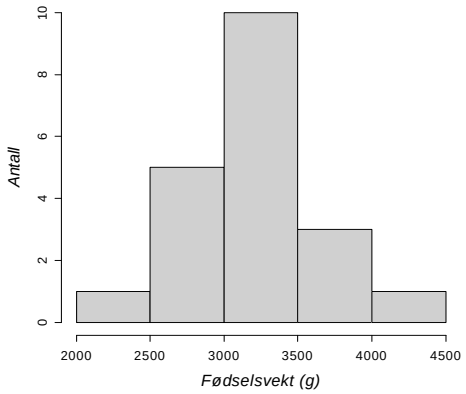
Summarizing data using a few useful measures (numbers):

- Central tendency (often mean value)
- Measures of spread (often standard deviation)



Measures of Central Tendency

- Arithmetic mean
- Trimmed) mean
- Median
- Mode
- (Midrange)
- (Geometric mean)



Central tendency
(mean, centre of gravity,
position, "location"

Arithmetic mean

Synonyms

- Average
- Mean value
- Sample mean

Definition (Rosner def. 2.1, s. 9)

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

Arithmetic mean

Properties

- All observations must be known
- Observations need not be ranked by value
- Mathematically favourable
- Sensitive to “outliers”: extreme, untypical observations

Definition (Rosner def. 2.1, s. 9)

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

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

Example (Rosner expl. 2.4, p. 11)

(from Rosner table 2.1, p. 10)

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

If $x_1 = 500$ g rather than 3265 g ...

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

... a relatively large change due to a single observation

Arithmetic mean

Properties

- Not defined for censored observations
- Sample means may be combined: →

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

- "In the long run" (i.e. as n → infinity) the sample mean will converge towards the Expected value of the stochastic variable

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

Example. Combining sample means:

$\bar{x}_1 = 18,4 \quad n_1 = 22$
$\bar{x}_2 = 17,9 \quad n_2 = 11$
$\bar{x}_3 = 9,5 \quad n_3 = 16$
$\bar{x}_4 = 15,1 \quad n_4 = 5 \qquad$ (i.e. $k = 4$)

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

$$= \frac{22 \cdot 18,4 + 11 \cdot 17,9 + 16 \cdot 9,5 + 5 \cdot 15,1}{22 + 11 + 16 + 5}$$

$$= \frac{404,8 + 196,9 + 152 + 75,5}{54} = \frac{829,2}{54} = 15,36$$

Trimmed mean

- Arithmetic mean based on the central 90 - 95 % of observations; the “tails” (5 - 10 %) have been trimmed
- Less sensitive to outliers

Median

Synonyms
- 50 percentile
- sample median

Properties
- Observations are ranked by value, in increasing order
- “The median is the value that cuts the data into two pieces”
- Insensitive to outliers

Transformation:

$$x_i \xrightarrow[ordered]{} x_{(i)}$$

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

Eksemple (Rosner expl. 2.5, p. 11)

(from Rosner table 2.1, p. 10)

$x_1 = 3265$

$x_2 = 3260$

$x_3 = 3245$

$x_4 = 3484$

...

Observations ranked in increasing order:

$x_{(1)} = 2069$

$x_{(2)} = 2581$

$x_{(3)} = 2759$

$x_{(4)} = 2834$

...

Median

Properties

- Up to half of observations may be censored
- Mathematically less favourable
- Sample medians *cannot* be combined

Definition (Rosner def. 2.2, p. 11)

(1) Uneven no. of observations ($n = 3, 5, 7...$)

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

(2) Even no. of observations ($n = 2, 4, 6$ ...)

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

(other definitions may occur)

Median

TABLE 2.2 Sample of admission white-blood counts (× 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		

Example (Rosner expl. 2.6, p.11)

(from Rosner table 2.2, p. 12)

$n = 9$, we use definition (1):

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

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	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

Eksample (Rosner expl. 2.5, p.11)

(from Rosner table 2.1, p. 10)

$n = 20$, we use definition (2):

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

Mode

Rosner def. 2.3
The most frequently observed value

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

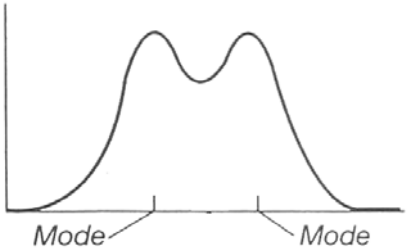
Rosner expl. 2.7, p. 13
Mode = 28

Mode

- Not in much use. Mathematically unfavorable.
- More often used in an abstract sense: unimodal distribution (single peak), bimodal (two peaks) etc.

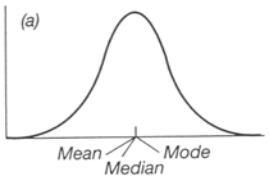
Examples of a bimodal probability distribution

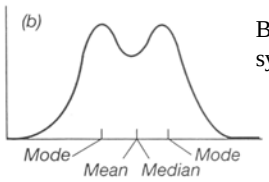
(Zar 1999, fig. 3.2 b)



Measures of Central tendency - summary

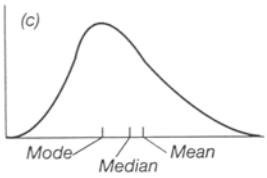
Unimodal, symmetrical

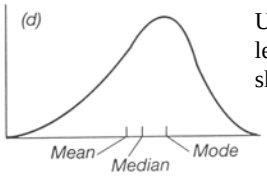




Bimodal, symmetrical

Unimodal, right (pos.) skewed



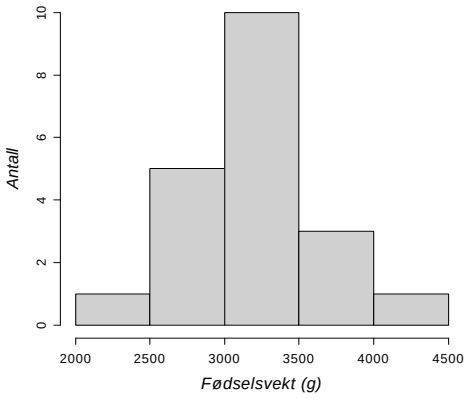


Unimodal, left (neg.) skewed

Measures of spread

Measure of spread, (variation, form, "shape")

- Range
- Quantiles, percentiles: interquartile range
- Variance
- Standard deviation
- Coefficient of variation



Range

Synonym
- Variasjonsbredde (Norw.)

Properties
- Same unit as observations
- Employs only two observations. Inefficient.
- Sensitive to outliers
- Increases with sample size

Definisjon (BR def. 2.5, s. 18)

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

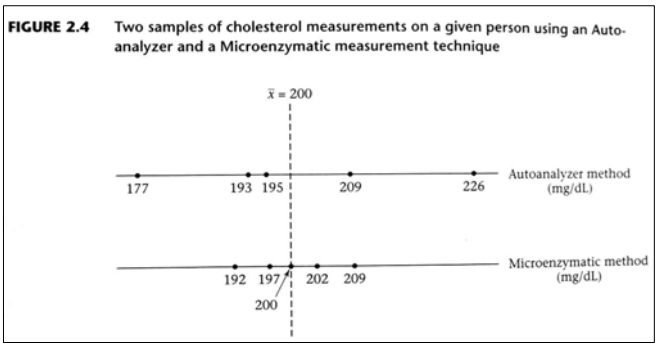
Eksempel (BR expl. 2.14, s. 18)

(fra BR table 2.1, s. 10)

$Range = x_{(20)} - x_{(1)} = 4146 \text{ g} - 2069 \text{ g}$

$= 2077 \text{ g}$

Range



Eksample (Rosner expl. 2.15, p. 18)

(from Rosner Fig 2.4, p. 18)

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

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

Quantiles and percentiles

Definition (Rosner def 2.6)

"The sample value V_p ,which exceeds a given proportion (*p)* of the ordered observations", $0 < p < 1$

(One observation corresponds to 1/*n*'th of the data. If $n = 100$, then one observation equals 0.01 or 1 %. Thus, observation no. 20 (ordered) corresponds to the 0.2 quantile or the 20th percentile.

Formally

n ordered observations: $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(k)}, X_{(k+1)}, \ldots, X_{(n)}$

(1) $V_p = X_{(k+1)}$ if np is not an integer", $k < np < k+1$

(2) $V_p = \frac{X_{(np)} + X_{(np+1)}}{2}$ if np is an integer

Quantiles and percentiles

Example (Rosner, expl. 2.16)
Find $V_{0,1}$ and $V_{0,9}$ (10th and 90th percentile) of the birth weight data in table 2.1.

$p = 0{,}1$ and $np = 2$ (integer) and we employ (2):

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

$p = 0{,}9$ and $np = 18$ (also an integer) and we use (2):

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

The central 80 % of children thus weigh between 2670 g and 3629 g, i.e. a spread of about 1000 g.

Quantiles and percentiles

$V_{0,5}$ = median = 50th percentile. The definitions must agree…

Example (from Rosner table 2.1, p. 10)

$np = 10$, so definition 2.6 (2) is used:

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

Interquartile range (IQR)

Properties

- Same unit as observations
- Contains the central 50 % of (ordered) observations
- Less sensitive to outliers
- Corresponds to "the box" of a box plot

Definition

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

$V_{0,75}$: upper, or 3.rd quartile

$V_{0,25}$: lower, or 1.st quartile

Interquartile range (IQR)

Example (from Rosner table 2.1, p. 9)

Find IQR for birth weight

IQR $= V_{0,75} - V_{0,25}$ (np = 15 og 5, respectively)

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

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

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

Other possible measures of spread...

Observations:

$x_1, x_2, \ldots, x_n$

-Mean deviation?

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

Rather unuseful! Sums to 0 (zero).

-Mean, absolute deviation (MAD)?

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

OK, but mathematically less favourable.

-Mean, *squared* deviation?

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

Best!

Sample variance

Definition (Rosner Def.2.7)

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

Comment

(*n -1*) is used rather than *n* for the denominator, as we already have made use of the sample mean value.
Thus one piece of information (one Degree of Freedom, DF) from the sample has been ”expended”.
This yields a slightly larger estimate of S^2. For large *n*, the difference is trivial.

Sample variance

Example (Rosner Expl. 2.19, p 21-22)

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

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

$$= \frac{1}{4}\cdot\left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]$$

$$= \frac{1}{4}\cdot(529+49+25+81+676) = \frac{1360}{4} = 340$$

Micro – enzymatic (same $\bar{x}$)

$$S^2 = \frac{1}{4}\cdot\left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]$$

$$= \frac{1}{4}\cdot(64+9+0+4+81) = \frac{158}{4} = 39,5$$

Sample variance

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

Properties

- Always positive
- Extreme observations contribute more (squaring)
- Favorable mathematical properties
- *But* cannot be interpreted on original scale

(kg → kg^2, mg/dl → mg^2/dl^2, cm → cm^2 osv.)

Sample standard deviation (SD)

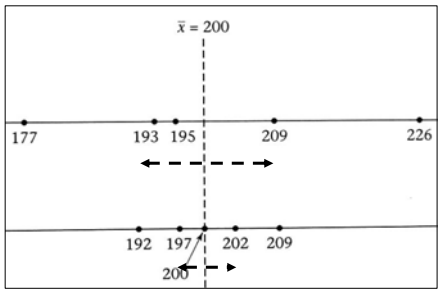
Definition (Rosner Def.2.8)

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

Example (Rosner Expl. 2.19, p. 21)

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

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



Sample standard deviation (SD)

- Approximately 95 % of the population are found within:

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

  … which approximately corresponds to the 2,5 and 97,5 percentiles
  - Follows from the Normal distribution (Gauss)
  - Requires reasonable symmetry and unimodal distribution

- A “quick & dirty” calculation:

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

  - May prove useful if you have to make a rough estimate of SD, in for example sample size determination

Sample standard deviation (SD)

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

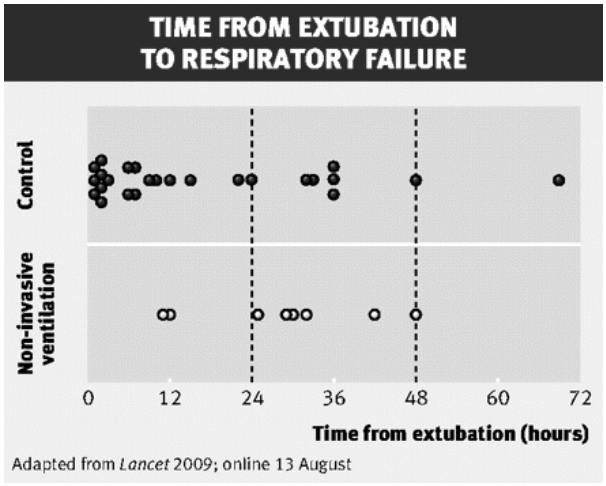
Properties

- Always positive
- Extreme observations contribute more (squaring)
- Favorable mathematical properties
- May be interpreted on original scale!

Graphics: dotplot

"A good figure says more than 1000 t-tests!"

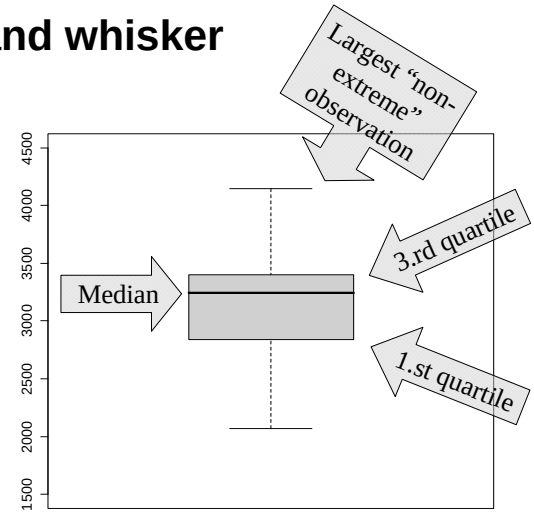
- Plot individual observations. A view of "everything".
- Group comparison is easy
- Useful with a low number of observations

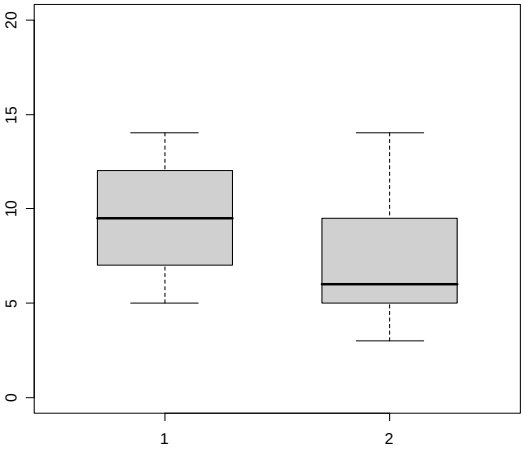


Graphics: Box-and whisker plot (boxplot)

Properties

- Quick view of central tendency and spread
- Key elements:
  - Median (2.nd quartile)
  - 1.st and 3.rd quartile
  - Non-extreme observations
- Useful for "moderately many" observations
- Easy to compare groups





Histogram

Properties

- An empirical "probability distribution"
- Shows the general shape
- Difficult to compare groups
- Useful with a large number of observations