



Faculty of Health Sciences

# Day 1: Essential statistical concepts

Paul Blanche

Section of Biostatistics, University of Copenhagen

February 16, 2026

# Outline/Intended Learning Outcome (ILOs)

## About the course

ILO: to know what to expect about/from the course

## Statistical inference

ILO: to outline the principles of statistical inference

## Data, distribution and descriptive statistics

ILO: to identify different types of data and contrast different options to summarize them

## Statistical uncertainty

ILO: to express and exemplify statistical uncertainty

ILO: to contrast confidence and prediction intervals



## Overall ILOs for this course

1. to identify the role of statistics for your research.
2. to present data and descriptive statistics.
3. to understand uncertainty and its impact for statistical conclusions.
4. to recognize what can and cannot be done with statistics.
5. to become autonomous to plan, perform and report “Basic” statistical analyses.
6. to know when to seek help / need to learn more.



# Calendar, teachers & topics

Date	Day	Room (8:00-15:00)	Topics	Teachers
16 February 2026	Monday	CSS-7.0.22	Overview, data, descriptive statistics, concept of statistical inference, confidence intervals	Paul Blanche, Gabriele Pittarello
18 February 2026	Wednesday	CSS-2.0.22	Hypothesis testing, tests for continuous outcomes, multiple testing	Paul Blanche, Gabriele Pittarello
23 February 2026	Monday	CSS-7.0.22	Univariate linear regression, correlation, regression to the mean	Paul Blanche, Gabriele Pittarello
25 February 2026	Wednesday	CSS-7.0.22	Analysis of Variance (One-way and Two-way ANOVA)	Paul Blanche, Gabriele Pittarello
2 March 2026	Monday	CSS-7.0.22	2x2 tables, odds ratio, two sample tests for binary responses	Karl Bang Christensen, Gabriele Pittarello
4 March 2026	Wednesday	CSS-7.0.22	Logistic regression	Paul Blanche, Gabriele Pittarello
9 March 2026	Monday	CSS-7.0.22	Multiple linear regression, confounding, interaction	Paul Blanche, Gabriele Pittarello
11 March 2026	Wednesday	CSS-7.0.22	Repeated measurements	Paul Blanche, Gabriele Pittarello
16 March 2026	Monday	CSS-7.0.22	Survival analysis	Paul Blanche, Gabriele Pittarello
18 March 2026	Wednesday	CSS-7.0.22	Optional: feedback/advice and work on homework assignment	Paul Blanche, Karl Bang Christensen
25 March 2026	Wednesday	CSS-7.0.22, CSS-7.0.06	Presentation and discussion of homework assignments	Paul Blanche, Karl Bang Christensen



# Formalities

1. Lecture start at 8:00, computer exercises around 12:00.
2. A homework assignment is handed out after lecture 6 (March 4).
3. The **homework assignment** is turned in on Friday March 20 (two days after the optional day to receive feedback/advice needed).
4. This highly **ambitious** course gives **7.5 ECTS** points, **if**
  - ▶ you attend 80% of all teaching units (we count the signatures; optional day on 18/3 does not count)
  - ▶ you present your homework on the last course day.
5. **We work with R** and we planned the teaching assuming that you have the **prerequisites stated on the course description**.
6. Material is available at the course Homepage:

<http://paulblanche.com/files/BasicStat2026.html>



# Homework assignment

- ▶ website (screenshot):

## Homework assignment

A homework assignment will be handed out in due time. Participants work with their own data or data related to their own research provided by their PhD supervisor. The homework assignment is turned in during the second last week of the course (**deadline** is Friday March 20). Before the deadline you can attend one day (non mandatory, on March 18), during which you can work on the assignment and receive feedback from the teachers.

More details will be provided later.

- ▶ Start thinking about which data you will use; discuss with your supervisor(s).
- ▶ Roughly, you will have to perform a statistical analysis of your choice<sup>1</sup> using your own data, defend the rationale for the method you chose to use and interpret the results. Alternatively, you can perform thought-through sample size or power calculations related to your research. You will have to present your work and constructively discuss that of others.



# Challenges for students and teachers

We start with the most basic concepts:

- ▶ But we **move fast** to cover many ("basic") topics needed by most health researchers.
- ▶ Hence, you need to **work hard**.

Nearly all students have some pre-knowledge:

- ▶ But **you differ a lot in level** of expertise/experience.
- ▶ Hence not all students should have the same ambitions. **You** are not all expected to solve all exercises entirely.<sup>2</sup>



# Textbooks

This course is not based on a single textbook, but the following book is excellent and covers most of the topics of the course (often in more details).

- ▶ *Regression with linear predictors*, by Per Kragh Andersen and Lene Theil Skovgaard (Springer, 2010).

There are many books on “elementary” statistics interesting to complement the course material. Those have been popular:

- ▶ *Practical statistics for medical research*, by Douglas G. Altman (CRC press, 1990).
- ▶ *Essential medical statistics, 2nd edition*, by Kirkwood & Sterne (Wiley, 2003)



## Supplementary teaching material

To complement the lectures and practicals, for each course day we recommend:

- ▶ short videos
- ▶ or short papers

to watch/read before or after each course day (see links on webpage).

This is to help you **to come better prepared or to learn further** on the topic of each day.



# Recommended use of AI / Chatbots (during the course)

- ▶ Do not use AI / Chatbots to ask questions/learn about statistical concepts and reasoning. Ask the teachers instead!

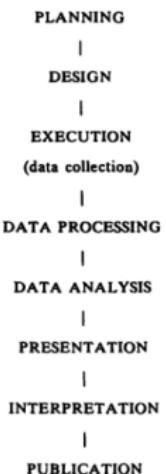
**Rationale:** experience showed that widespread misunderstandings are repeated by the chatbots and/or important details are overlooked.

- ▶ Consider using them (especially outside of the classroom) to help you coding in R, especially for tasks for which you can check the correctness of the result.

**Example:** ask for help to perform specific data management tasks or make better plots.



# Statistics is not only about data analysis



**Figure 1.1** General sequence of steps in a research project.

*“Statistical thinking can contribute to every stage”<sup>3</sup>.*



# What is statistics?

*"Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data."<sup>4</sup>*

For **summarizing** data we usually use **descriptive statistics**.

- ▶ Summary statistics (key figures).
- ▶ Statistical graphs (plots).

For **analyzing** data we usually use **statistical inference**.

- ▶ Estimates and confidence intervals.
- ▶ Hypothesis testing and p-values.



## Word of caution

When used well, statistics turns data into knowledge.

When misused, statistics can be seriously misleading and counterproductive to advance science.



## Word of caution

When used well, statistics turns data into knowledge.

When misused, statistics can be seriously misleading and counterproductive to advance science.

Statistics is not easy to learn and developing good skills usually takes time and is hard work. You will have to learn the theory and acquiring hands-on experience.



## Word of caution

When used well, statistics turns data into knowledge.

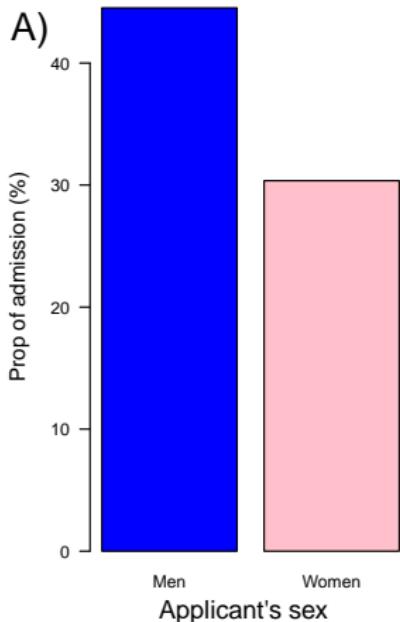
When misused, statistics can be seriously misleading and counterproductive to advance science.

Statistics is not easy to learn and developing good skills usually takes time and is hard work. You will have to learn the theory and acquiring hands-on experience.

Rigorous statistical thinking is not always intuitive! “Are humans good intuitive statisticians?”, not really.<sup>5</sup>



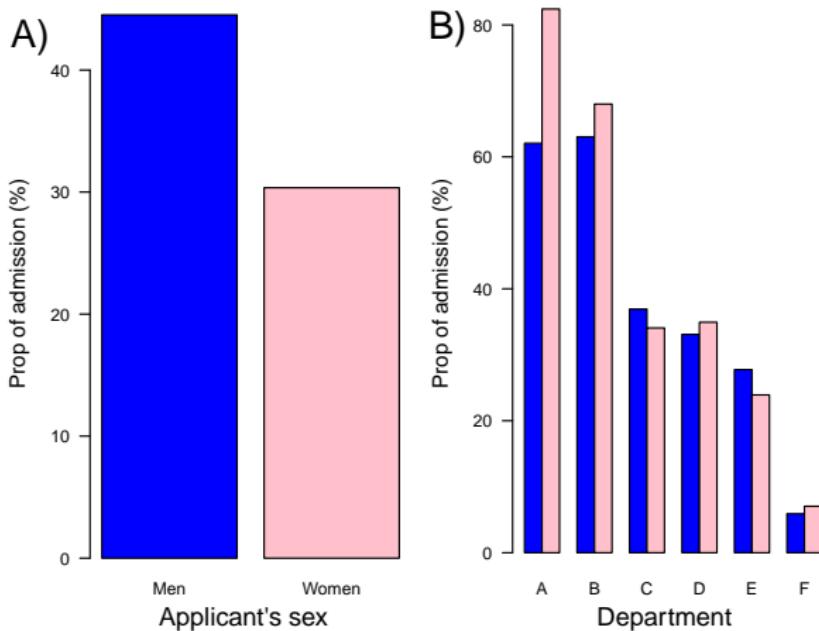
# Simpson paradox: Berkeley admission (1973)<sup>6</sup>



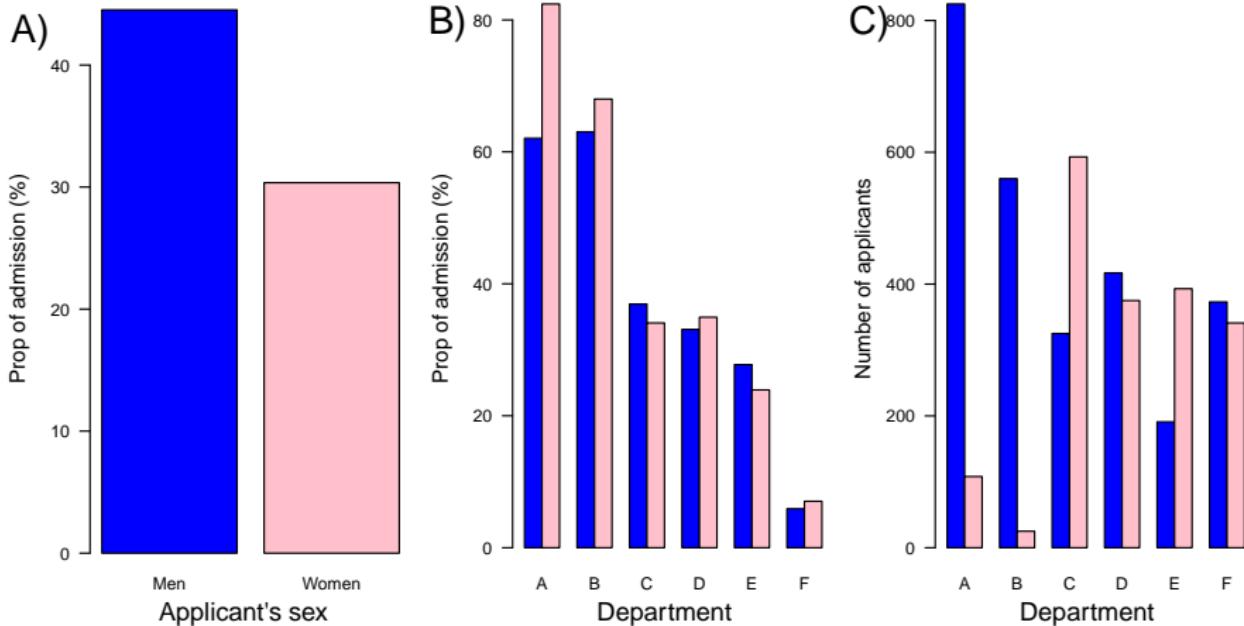
14 / 68 <sup>6</sup> Bickel et al. (1975), Science, 187, 398–403.



# Simpson paradox: Berkeley admission (1973)<sup>6</sup>



# Simpson paradox: Berkeley admission (1973)<sup>6</sup>



# Outline/Intended Learning Outcome (ILOs)

## About the course

ILO: to know what to expect about/from the course

## Statistical inference

ILO: to outline the principles of statistical inference

## Data, distribution and descriptive statistics

ILO: to identify different types of data and contrast different options to summarize them

## Statistical uncertainty

ILO: to express and exemplify statistical uncertainty

ILO: to contrast confidence and prediction intervals



# The general principle of statistical inference

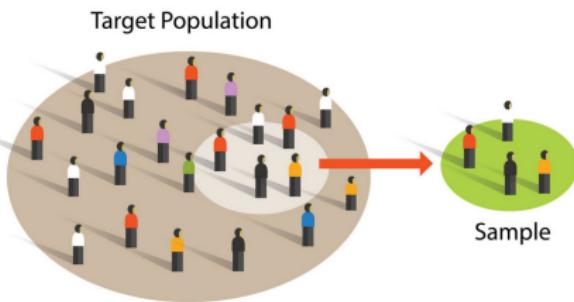
1. Research **aims** and hypotheses are formulated.
2. The study population is defined and data are collected (measured) on a **random subset**<sup>7</sup> of all samples/individuals in the population
3. The sampled data are analyzed but conclusions are drawn about the **full study population**.

---

<sup>7</sup>It is crucial for the interpretation that the sampled samples/individuals are “representative” for the full study population.



**Descriptive statistics** aims to describe the data that you have seen (here among the 4 persons in the small oval).



**Statistical inference** aims to answer research questions about the entire population (here among all in the big oval).



# Example: COVID-19 in Denmark (SSI publication May 20, 2020)

## Descriptive statistics

Out of 2,600 randomly selected people, 1,071 (41.2%) have agreed to be tested for antibodies.<sup>8</sup>

## Statistical inference

The prevalence/proportion (95% confidence limits) of persons who have had COVID-19 infection is estimated to 1.1% (0.6%–1.9%).<sup>9</sup>

---

<sup>8</sup> <https://www.ssi.dk/aktuelt/nyheder/2020/de-forste-foreløbige-resultater-af-undersøgelsen-for-covid-19-i-befolkningen-er-nu klar> (last accessed Feb 4, 2025)

<sup>9</sup> Note: there were 12 tested positive out of 1,071.



# Not exactly a random subset of the target population

**Target population:** the entire Danish population.

**Research question:** “How many Danish residents have been infected?”

Ideally, the sample would have been sampled from the entire Danish population. However, it was (initially) impossible and nicely acknowledged in the report of the results:

*“The preliminary results presented in this note should be interpreted with caution. So far, it is only possible to perform antibody tests in the five tents located in Copenhagen, Aarhus, Aalborg, Næstved and Odense”.*



Often, the sample is not “perfectly representative” of the full study population of interest. But we can:

- ▶ acknowledge the potential biases and sometimes discuss how big we expect them to be.
- ▶ adapt the interpretation accordingly, i.e. clarifying what defines the study population (e.g. inhabitants of the five cities).



# Data driven research questions

We can sometimes define/update the research question after having looked at the data and still get valid statistical inference, but not always. **Great caution is needed!**

In short, it is often fine to look at the data on exposure and other “baseline” variables (e.g. gender, age, medical history before inclusion). <sup>10</sup> However, looking at the outcome data, especially at associations between the outcome and other variables, will generally weaken the validity of any subsequent statistical analysis, possibly dramatically.

---

<sup>10</sup> Actually, this is often helpful in observational study, to better understand what can and cannot be studied using the available data (lecture 7).



# Prespecification matters!

In this course, the interpretation of the statistical results (confidence intervals, p-values) will be given assuming that the **research questions and statistical analyses are formulated before seeing the data**. That is, in the context of **prespecified** research questions and analyses.

This is because statistical inference is all about understanding and quantifying **uncertainty**, which is very difficult without **prespecification**.

This is also because we should prespecify to follow simple rules for **good research practice**.<sup>11</sup>



# Outline/Intended Learning Outcome (ILOs)

## About the course

ILO: to know what to expect about/from the course

## Statistical inference

ILO: to outline the principles of statistical inference

## Data, distribution and descriptive statistics

ILO: to identify different types of data and contrast different options to summarize them

## Statistical uncertainty

ILO: to express and exemplify statistical uncertainty

ILO: to contrast confidence and prediction intervals



# Dataset

Before starting statistical analyses, we need to organize the data (like in a spreadsheet):

- ▶ Columns=variables
- ▶ Rows=records

**Example:** 2,600 records (randomly selected people) and 5 variables (tested: yes/no, test result: positive/negative , city, age, gender).

## What roles do the variables have to play?

- ▶ Outcome/Dependent variable/Phenotype: presumed “effect”.
- ▶ Explanatory variable/Independent variable/Predictor variable/Feature/Covariate: presumed “causes”.
- ▶ Other variables (e.g. CPR number).



# Data type I: quantitative data

**Quantitative data** measure an amount or quantity either on a **continuous** or a **discrete** (i.e. integer) scale.

## Examples of continuous variables:

- ▶ Duration, age, concentration, volume, gene expression level.

## Examples of discrete variables:

- ▶ Number of metastases, number of pups in a litter (i.e. usually counts).

**Note:** Discrete data are often treated as continuous when they have a wide range of values. Otherwise, they are sometimes categorized, e.g. as 0, 1, 2 or more.



## Data type II: categorical data

**Categorical data** are about classification into groups. These can be either **ordinal** (ordered) or **nominal** (unordered).

### Examples of ordinal variables:

- ▶ Disease stage, scores such as none/mild/moderate/severe.

### Examples of nominal variables:

- ▶ Genotype, diagnosis, treatment.
- ▶ Hospital center, patient id<sup>12</sup>.

**Categorical variables with only two groups are termed **binary**.**

- ▶ male/female, dead/alive, diseased/healthy



## Data type III: Troublesome data

### Censored data:

- ▶ Concentration below the limit of detection.
- ▶ Survival time beyond maximal follow-up duration (lecture 9).

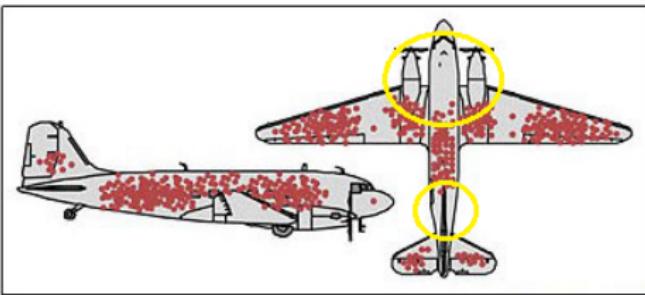
### Missing data: (may occur for any type of data)

- ▶ Failed, lost or cancelled measurements.
- ▶ Patients not showing up for blood sample (lecture 8).

If not accounted for in the statistical analyses, censored or missing data may lead to severe bias.



# Beware of missing values!



Credit: Cameron Moll

Gentlemen, you need to put more armour-plate  
where the holes aren't because that's where the holes  
were on the airplanes that didn't return - Abraham  
Wald 1942.

Missing data can be informative!



# What are descriptive statistics good for?

- ▶ Getting an **overview** of your data.
- ▶ **Describing** the representativeness of the sample (“Table I” or “Baseline table”).<sup>13</sup>
- ▶ Identifying outliers and correcting errors.
- ▶ Identifying and visualizing trends.
- ▶ Supporting the interpretation of your (main) statistical analyses.
- ▶ Checking model assumptions.
- ▶ etc.

**But generally not good for making strong conclusions.**

- ▶ risk of overinterpreting a random trend in the data or neglecting a substantial difference.



# Which descriptive statistics? When ?

The datatype determines what descriptive statistics can be used.

Common in “Statistical methods” section:<sup>14</sup>

*Quantitative normally distributed data was summarized by means and standard deviations, non-normally distributed data by medians and quartiles, and categorical data by numbers and percentages....*

**But:** keep study aims and sample size in mind. Use common sense!

- ▶ What information about the data is **interesting**?
- ▶ Some methods are not suited for very small datasets.

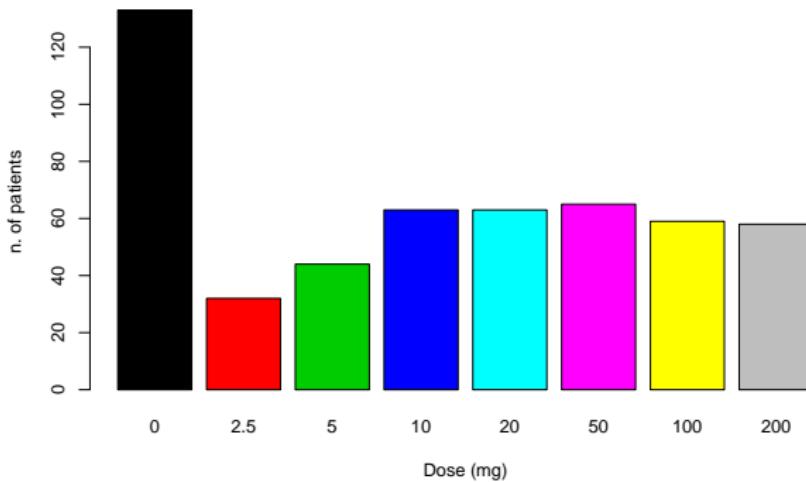
---

<sup>14</sup> BTW, this is usually not very interesting/needed... consider using the limited available space for describing more important information!



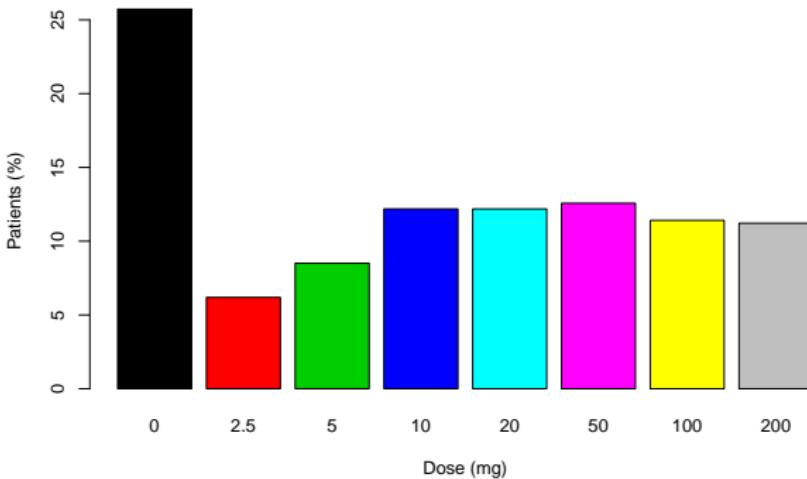
## Categorical data: easy case

Categorical data can be summarized in group totals and percentages. Use **barplots** for graphical presentation.



## Categorical data: easy case

Categorical data can be summarized in group totals and percentages. Use **barplots** for graphical presentation.



Often the percentages, i.e. the distribution of the data, are the most interesting, but not always.



# Quantitative data: more challenging

To summarize with numbers we provide measures of:

Location The average or most typical measurement, e.g. the sample mean or median, sometimes the mode.

Variation A range of most typical measurements, e.g. the standard deviation, normal range or the inter quartile range (IQR).

Beware: different measures of location/variation are most appropriate for approximately normally distributed data or not.

For graphical presentation, options include:

- ▶ Dotplot (aka Stripchart), boxplot, or histogram.<sup>15</sup>
- ▶ QQ-plot<sup>16</sup> to check an assumed normal distribution.

---

<sup>15</sup> preferably the first for small samples.

<sup>16</sup>i.e. Quantile-quantile-plot.



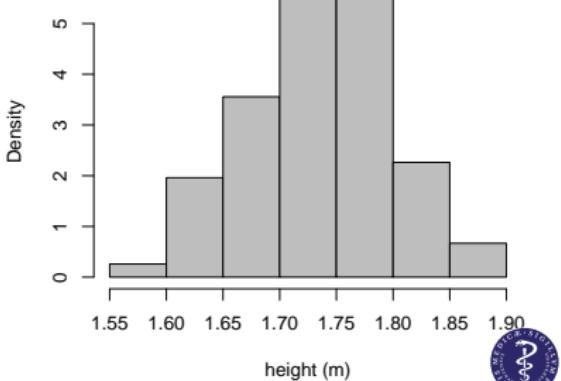
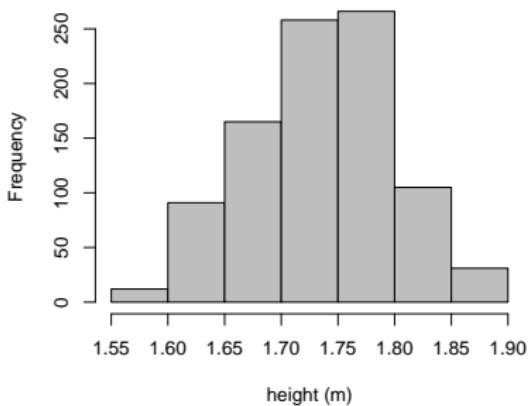
# Histogram

(Galton data (1886))

**Two versions:** same shape different y-axis.

Frequency: numbers of observations in interval (left).

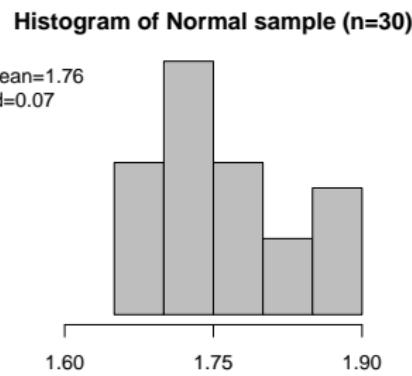
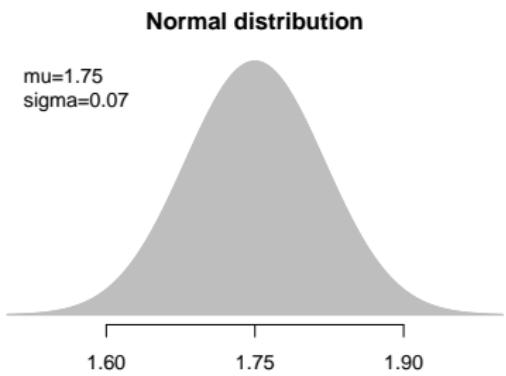
Density: frequencies are normalized so that the areas of the bars add to one (probability distribution; facilitate comparisons between groups).



# The normal distribution (1/2)

**A model** of the distribution of a variable (e.g. outcome) in the population.

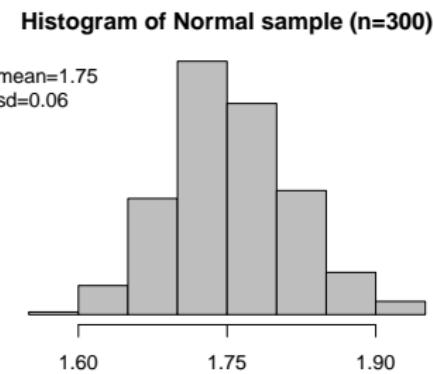
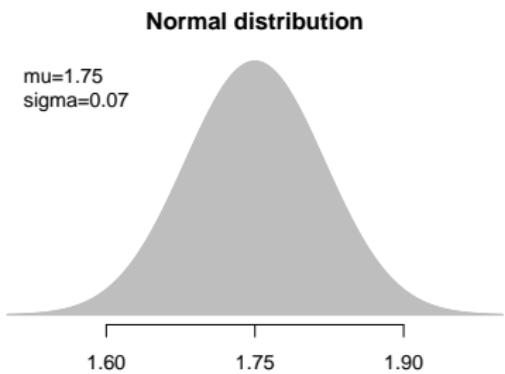
- ▶  $\mu$  is the population mean
- ▶  $\sigma$  the population standard deviation
- ▶ the distribution is symmetric and unimodal



# The normal distribution (1/2)

**A model** of the distribution of a variable (e.g. outcome) in the population.

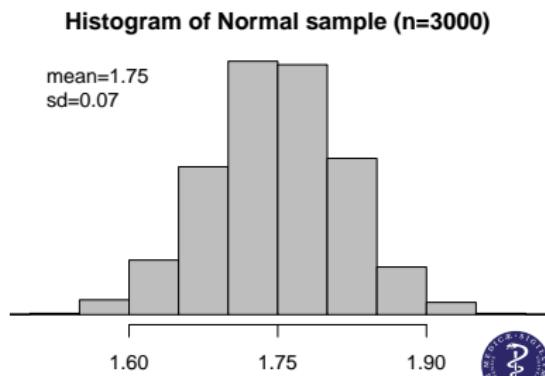
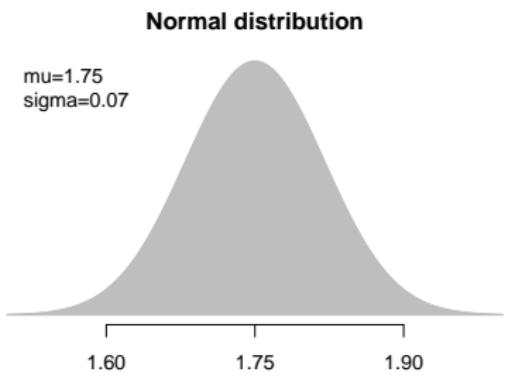
- ▶  $\mu$  is the population mean
- ▶  $\sigma$  the population standard deviation
- ▶ the distribution is symmetric and unimodal



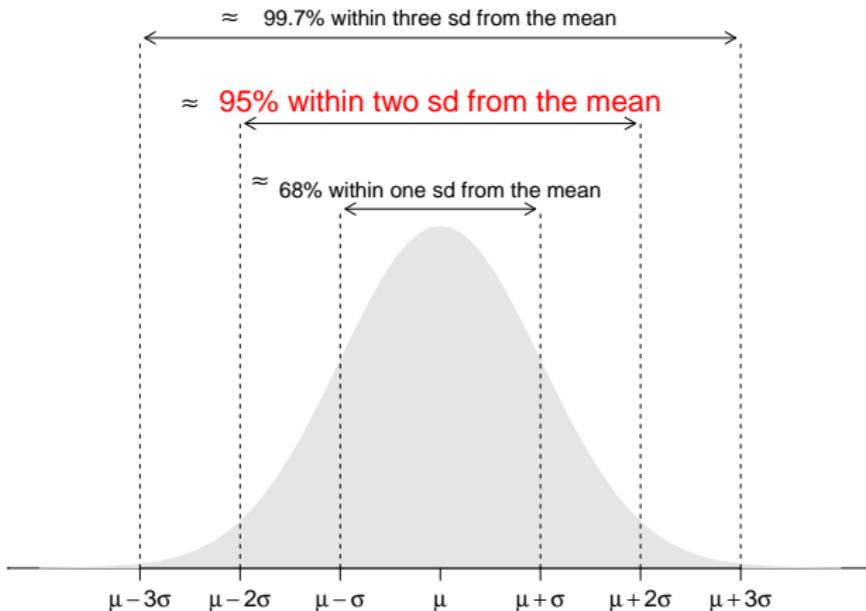
# The normal distribution (1/2)

**A model** of the distribution of a variable (e.g. outcome) in the population.

- ▶  $\mu$  is the population mean
- ▶  $\sigma$  the population standard deviation
- ▶ the distribution is symmetric and unimodal



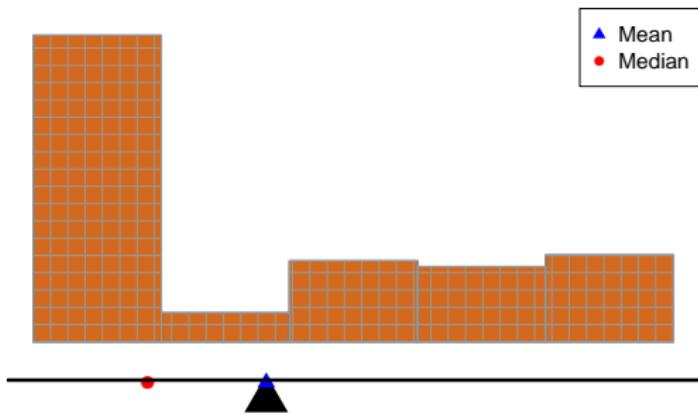
# The normal distribution (2/2)



The 95% normal range is  $\mu \pm 1.96 \cdot \sigma$  (of course  $1.96 \approx 2$ )



# Mean vs Median



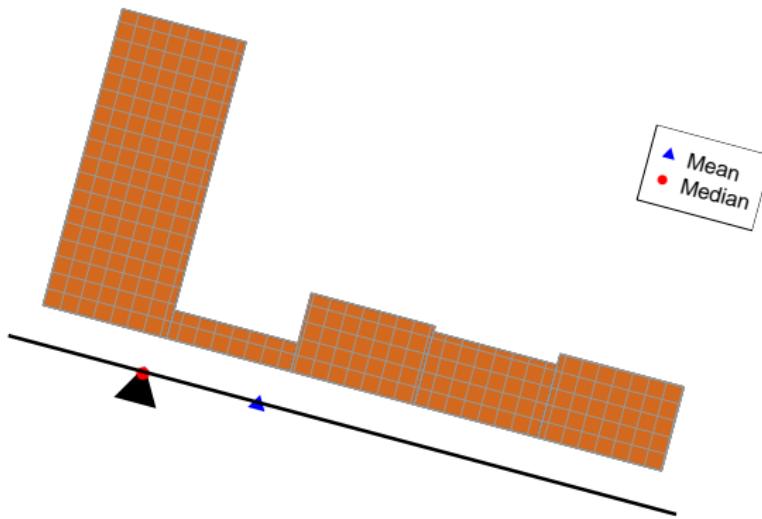
Mean: center of gravity<sup>17</sup>.

Median: obs. are located 50% above, 50% below<sup>18</sup>.

<sup>17</sup>bars of histogram=piles of bricks, x-axis=seesaw, mean=fulcrum.  
36 / 68 <sup>18</sup>50% of the bricks are located to the right and to 50% the left.



# Mean vs Median



Mean: center of gravity<sup>17</sup>.

Median: obs. are located 50% above, 50% below<sup>18</sup>.

<sup>17</sup> bars of histogram=piles of bricks, x-axis=seesaw, mean=fulcrum.  
<sup>18</sup> 50% of the bricks are located to the right and to 50% the left.



# The standard deviation (sd)

The sample standard deviation is given by

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}} \quad \text{with the mean} \quad \bar{x} = \frac{1}{n} \sum_i x_i$$

We divide by  $n - 1$ , the “degrees of freedom”, for technical reasons.

**Interpretation:** mostly meaningful if the distribution is normal.

If that is the case, we can approximate the 95% normal range with

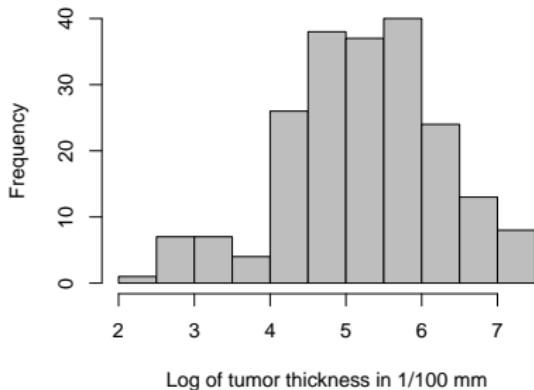
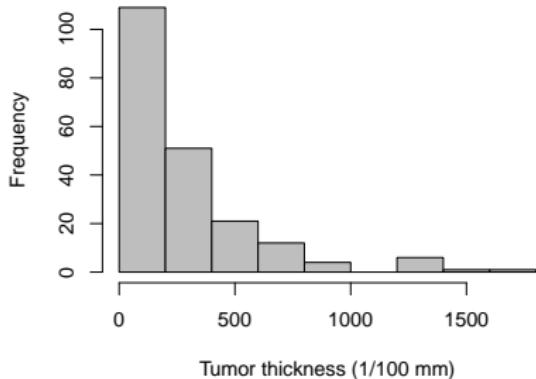
$$\bar{x} \pm 2 \cdot s$$

and this can be used to predict individual responses with 95% certainty if the sample size is large.<sup>19</sup>

<sup>19</sup> but whatever the (often unknown) distribution of the data, the normal range cannot contain less than 75% of the data ([Chebyshev's inequality](#)). 

## Case: log-normal data

Tumor thickness at baseline<sup>20</sup> ( $n=205$ ) has a **skewed** (i.e. asymmetric) distribution.



But the log-transformed data are “approximately” normal.



# Case: interpretation of summary statistics

## Without transformation:

- ▶ Mean  $\pm$  of thickness (in 1/100 mm):  $292.0 \pm 295.4$
- ▶ Normal range  $\approx -299.9\text{--}883.9$  : negative thickness “normal”??

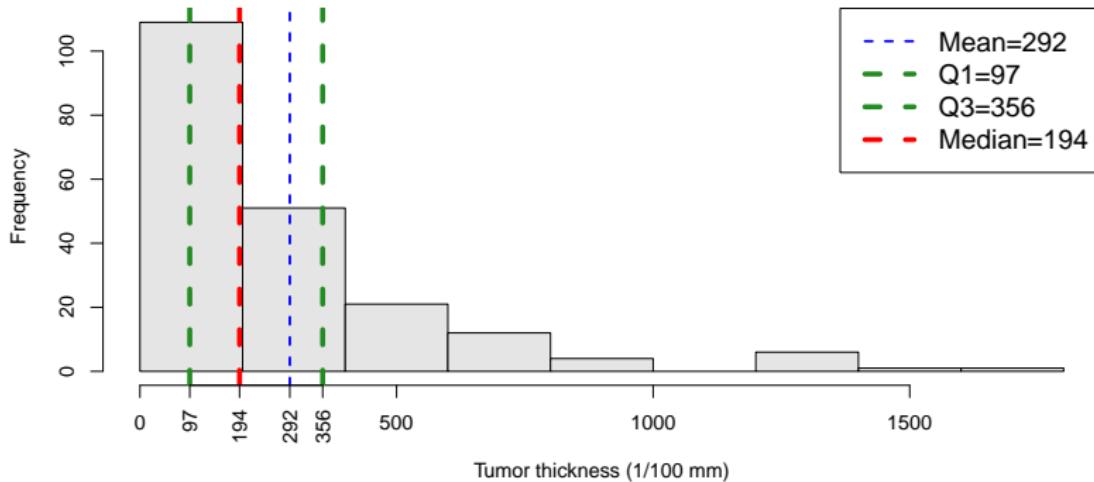
## With log-transformation:

- ▶ Mean  $\pm$  of log thickness:  $5.22 \pm 1.01$
- ▶ Normal range of log thickness:  $\approx 3.20\text{--}7.25$ .
- ▶ **Back-transformed:**  $\exp(3.20)$  to  $\exp(7.25)$ , i.e.  $24.5\text{--}1408$ .



# Median and quartiles

**Non-normal data** are not well described by mean & SD. Instead, we prefer **median** and **IQR<sup>21</sup>**.



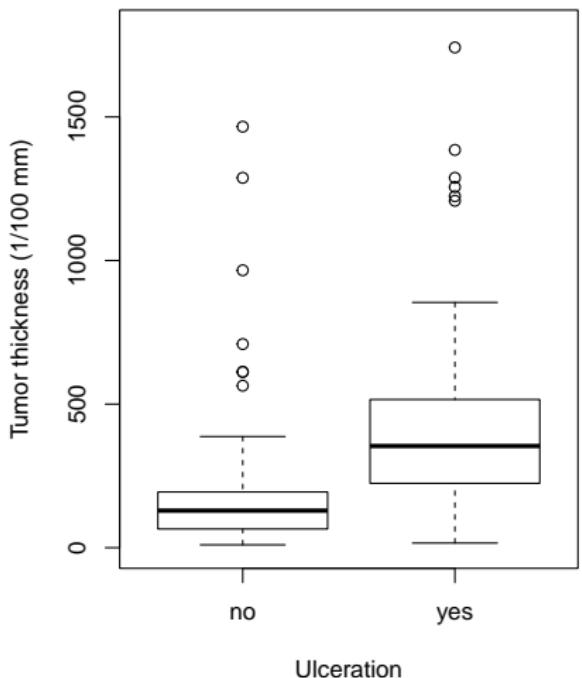
**Median** and **quartiles** divide the data into four equal proportions. We often call the interval  $[Q_1; Q_3]$  the **interquartile range**.<sup>22</sup>

<sup>21</sup>This can be your “default” choice, as it is always easy to interpret and informative.

<sup>22</sup>Especially in medical research, but the difference  $Q_3 - Q_1$  is also often called the interquartile range.



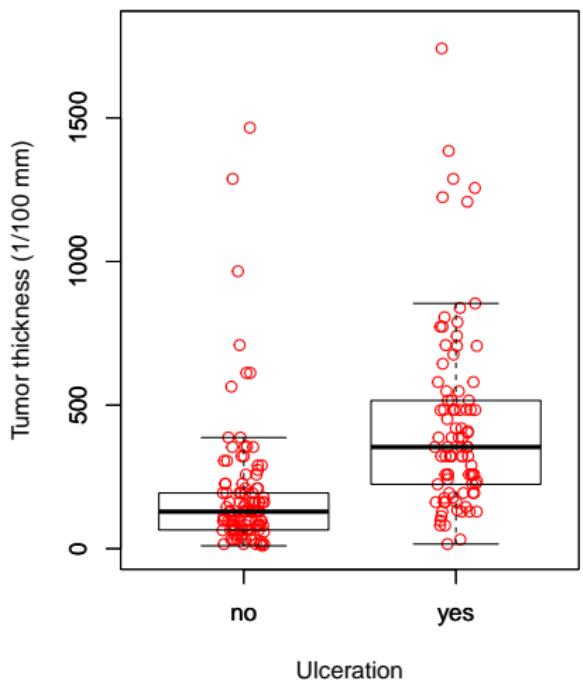
# Boxplot



- ▶ The box shows the **median** and the **quartiles (always)**.
- ▶ The **whiskers** extend to the most extreme data point which is no more than 1.5 times the length of the box (**R default**).
- ▶ Observations beyond are shown with dots (**R default**).



# Boxplot



- ▶ The box shows the **median** and the **quartiles (always)**.
- ▶ The **whiskers** extend to the most extreme data point which is no more than 1.5 times the length of the box (**R default**).
- ▶ Observations beyond are shown with dots (**R default**).
- ▶ (now showing all dots; not only outliers, with “jitter”)



# Quantiles

A  $\gamma\%$ -quantile has  $\gamma\%$  of data below and  $100-\gamma\%$  above.

The 50%-quantile is the **median**.

- ▶ It splits data in the lower and upper half.
- ▶ The middle observation (uneven sample size), or the average of the two mid ones (even sample size).

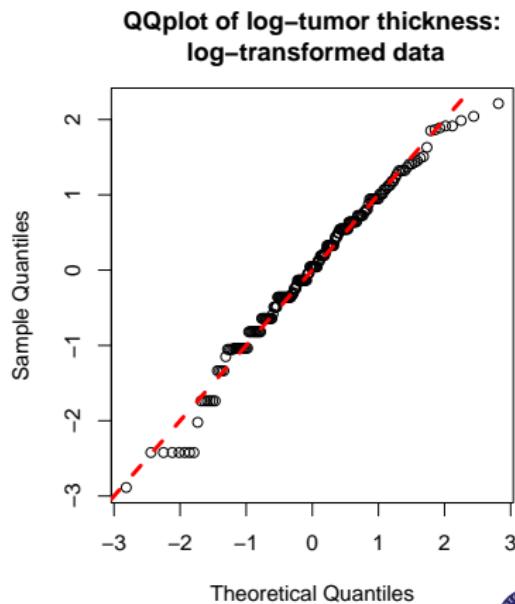
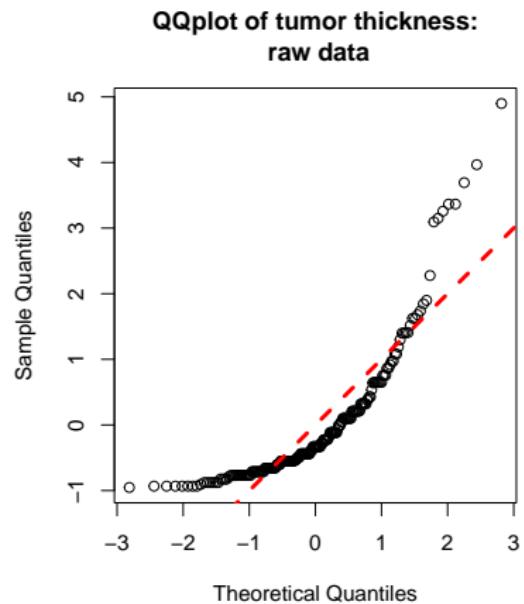
The 25%--, 50%--, and 75% quantiles form the **quartiles**.

- ▶ **Q1**: the 25%-quantile is the **lower quartile** (25% below ...)
- ▶ **Q2**: is the **median**.
- ▶ **Q3**: the 75%-quantile is the **upper quartile** (75% below ...)



# QQplots for checking normality

Compare the ordered (standardized) values from the data (smallest to largest) with the corresponding quantiles of the normal distribution.

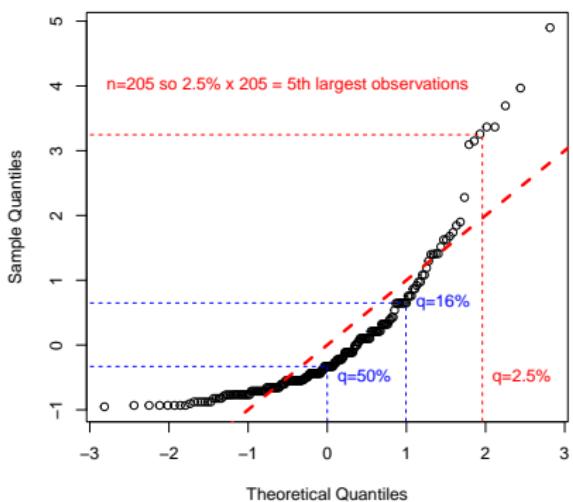


**Interpretation:** the closer the dots to the (red) diagonal the better (i.e. the closer to what is expected from truly normally distributed data the better).



# Appendix: QQplot details

QQplot of tumor thickness:  
raw data

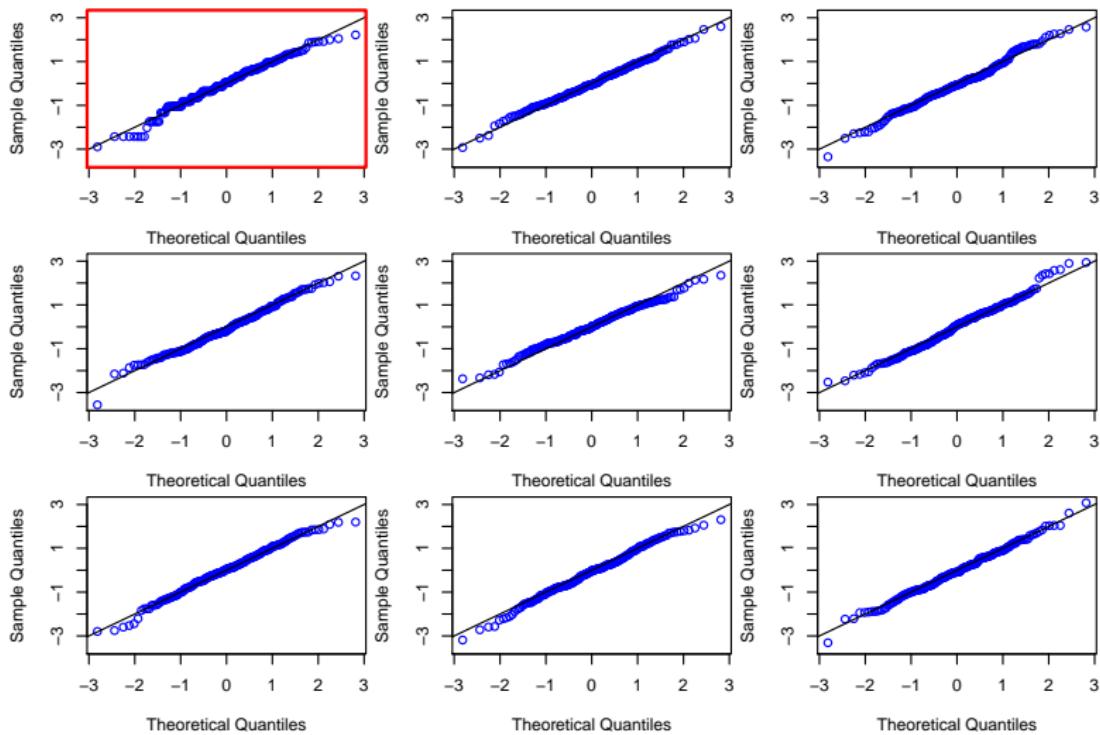


- ▶ **y-axis:** standardized values, i.e.,  $[x - \text{mean}(x)]/\text{sd}(x)$
- ▶ **x-axis:** normal distribution quantiles, e.g.:  
 $1.96 = (100 - 2.5\%)\text{-quantile}$   
 $1 = (100 - 16\%)\text{-quantile}$   
 $0 = (100 - 50\%)\text{-quantile (median)}$
- ▶ Observed  $(100 - 2.5\%)$ -quantile is larger than expected if normal distribution because the data are skewed (to the right), as seen before.

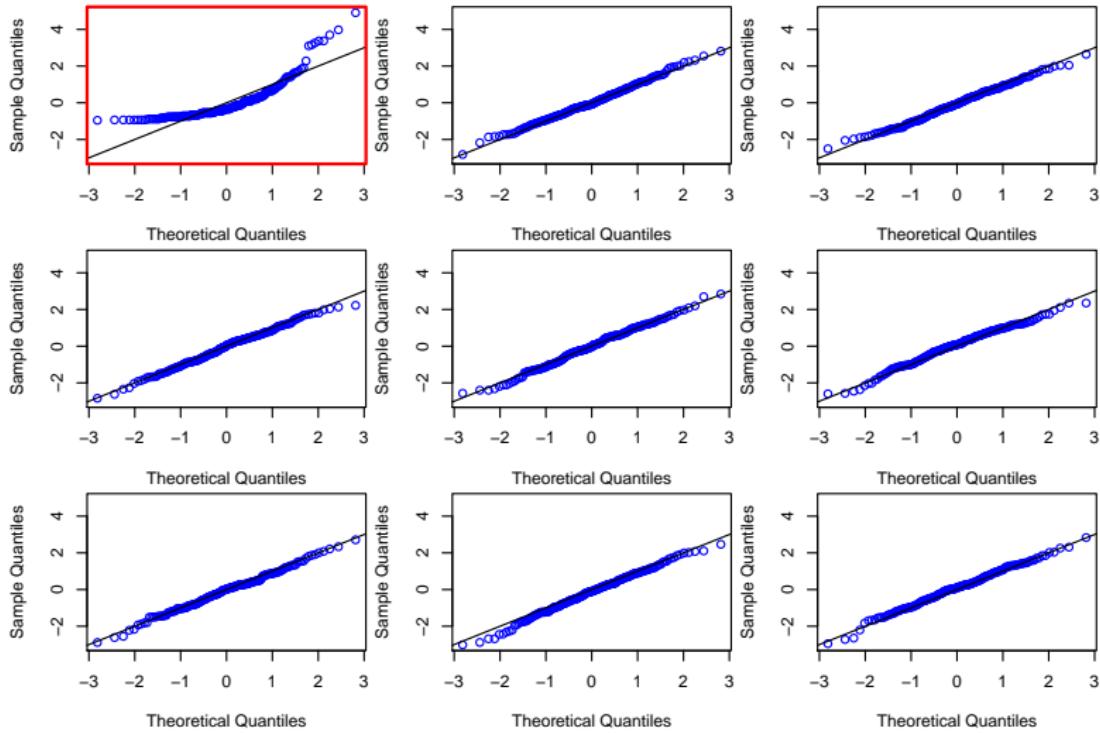


# Wally plot to “visualize” random sampling

Does it look different from what is typically observed with normally distributed data?



# Without log-transformation

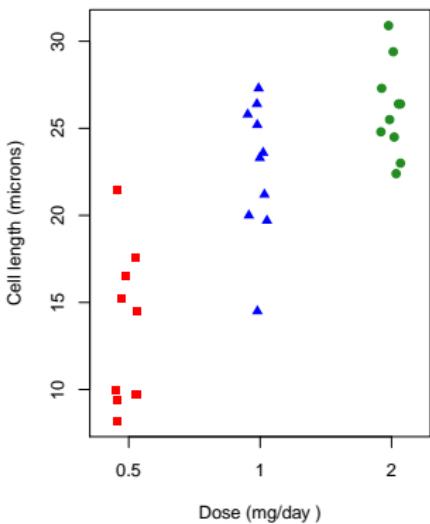


# Small data: what should we do?

The usual descriptive statistics and plots are usually not sensible when the sample size is very small (e.g. with  $n = 3$ ).<sup>23</sup>

## Recommendations:

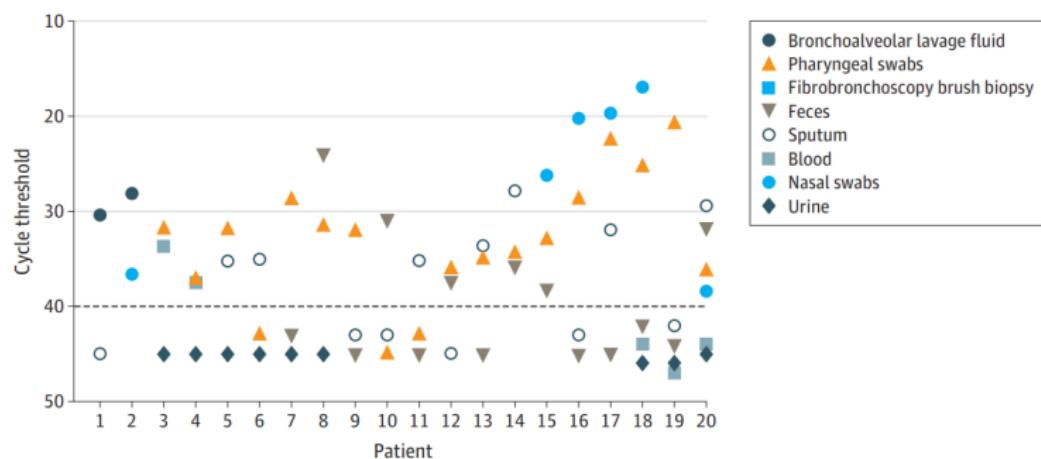
- ▶ **Don't** present summary statistics when you can instead present the original data.
- ▶ **Don't** make a picture of the summary statistics if you can make a nicer one of the original data!
- ▶ **Do** consider presenting a dotplot (aka **stripchart**; use “jitter”).



# Good plots can be very informative...

... and potentially much more than tables/descriptive statistics: **be creative!**

Figure. Severe Acute Respiratory Syndrome Coronavirus 2 Distribution and Shedding Patterns Among 20 Hospitalized Patients



(Wang et al. "Detection of SARS-CoV-2 in different types of clinical specimens." Jama 323.18 (2020): 1843-1844)



# Outline/Intended Learning Outcome (ILOs)

## About the course

ILO: to know what to expect about/from the course

## Statistical inference

ILO: to outline the principles of statistical inference

## Data, distribution and descriptive statistics

ILO: to identify different types of data and contrast different options to summarize them

## Statistical uncertainty

ILO: to express and exemplify statistical uncertainty

ILO: to contrast confidence and prediction intervals



## Reminder: the general principle of statistical inference

1. Research aims and hypotheses are formulated.
2. The study population is defined and data are collected (measured) on a **random subset** of all samples/individuals in the population.
3. The sampled data are analyzed but conclusions are drawn about the **full study population**.

But, which conclusions? With which **confidence level**?



# Accepting uncertainty

**Key idea in statistics:** use probability calculations to **quantify sampling variation** in the results from replicated experiments.

**Case (Covid-19):** 12 tested positive out of 1,071, i.e. 1.1%.

Can we **conclude**:

- ▶ that **exactly** 1.1% have been infected? **No!**
- ▶ that “**approximately**” 1.1% have been infected? **Yes**

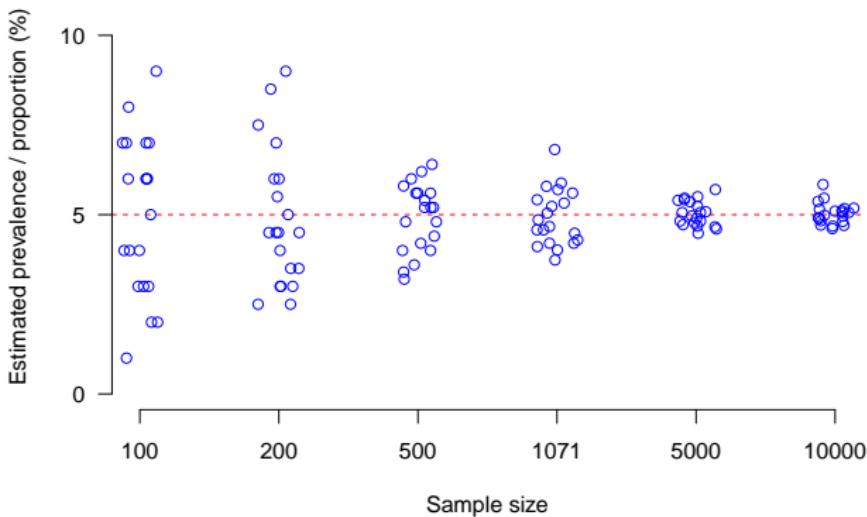
Can we find a **range of values** (e.g. 0.6%–1.9%) that “**likely**” includes the (true) proportion of infected among all Danish residents?

- ▶ **Yes**, by computing a **confidence interval**.



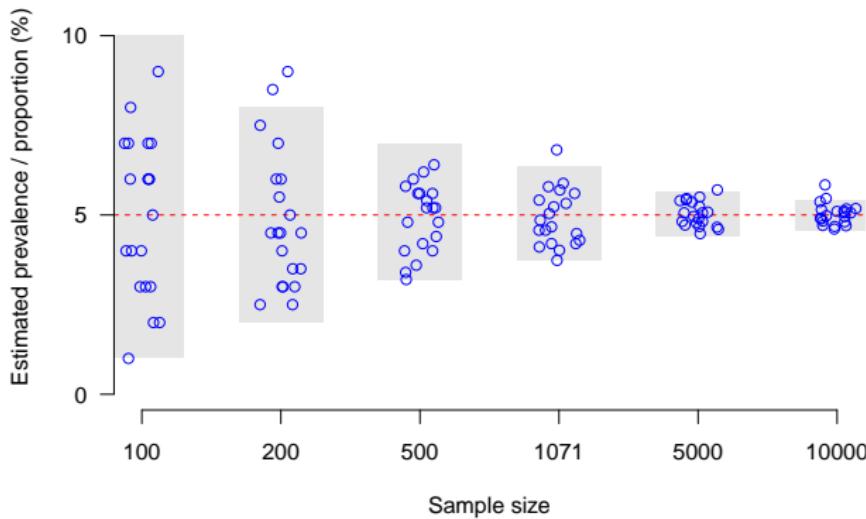
# Understanding uncertainty

Assume that indeed 5% are infected. What sample proportions will we observe from a random sample? Let's do look at 20 such samples, for different sample sizes.



# The role of probability calculation

Using probability calculation, we can actually calculate that, 95% of the observations will be in the grey zones (if we look at results from many samples).



Note: doubling the sample size reduces the height of the grey box by 30% ("central limit theorem":  $1/\sqrt{2} \approx 0.7$ )



# Statistical reasoning and probability calculation

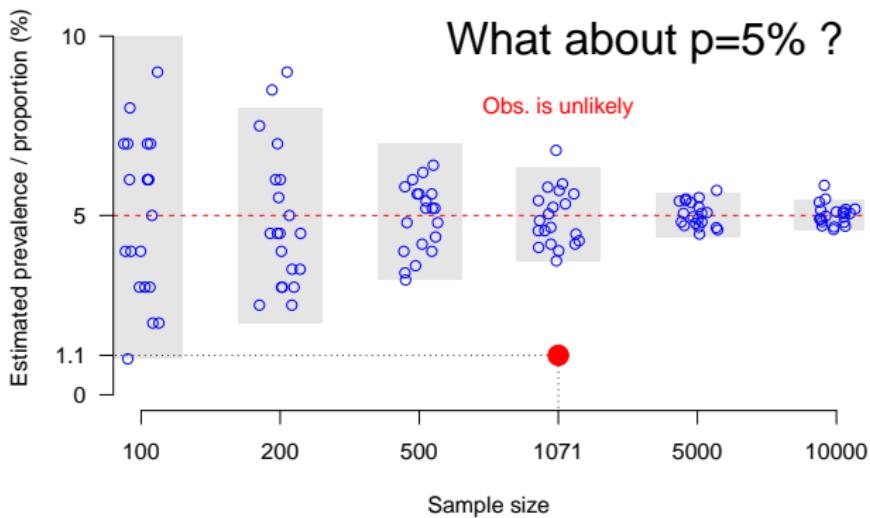
Because we know how to calculate what to expect after assuming a specific value for the (true) population parameter (e.g. a prevalence of 5%), we can “reverse” the reasoning and know which values could “likely” lead to the observation of our sample estimate.

These values will form a **confidence interval**.

The confidence interval gives a range of “plausible” values for the parameter that we are trying to estimate.



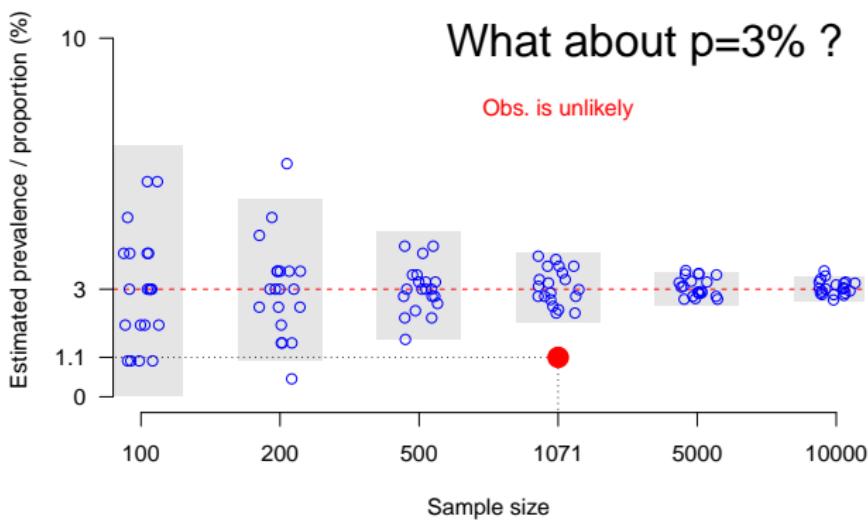
# Statistical reasoning and probability calculation



- observation / estimate is  $\hat{p} = 1.1\%$ , from  $n = 1,071$  (red dot).



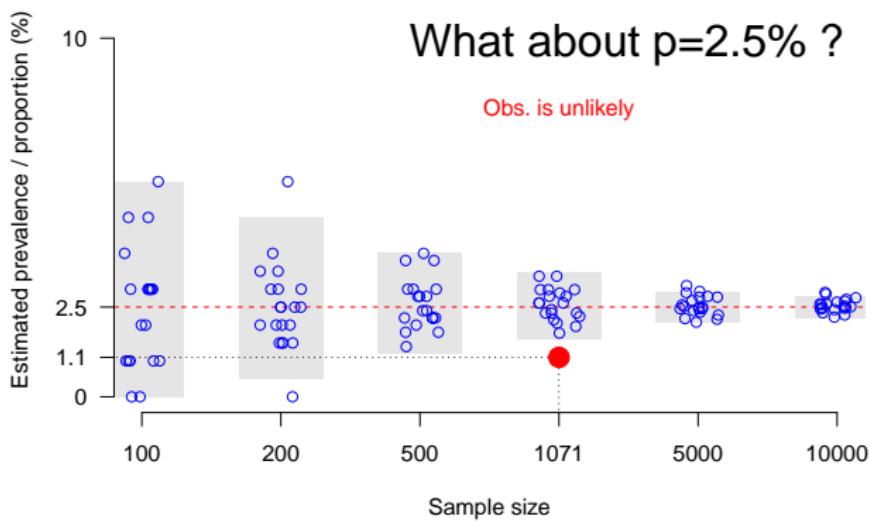
# Statistical reasoning and probability calculation



- ▶ observation / estimate is  $\hat{p} = 1.1\%$ , from  $n = 1,071$  (red dot).



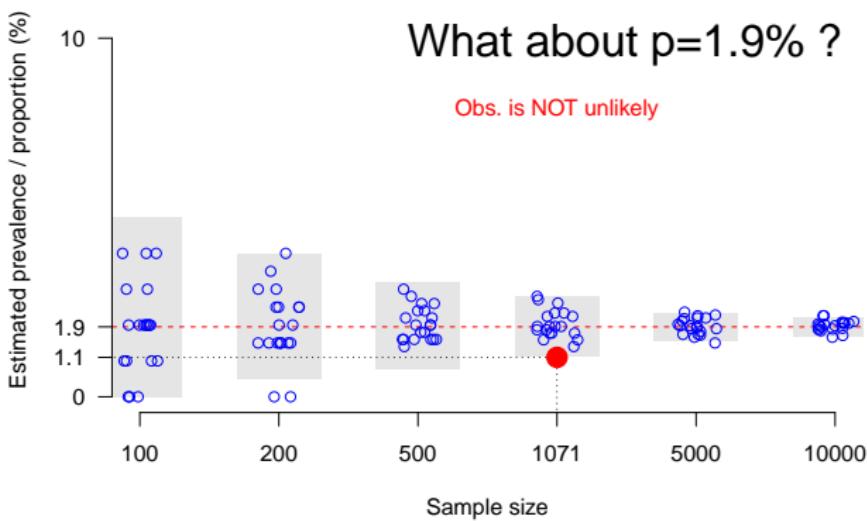
# Statistical reasoning and probability calculation



- observation / estimate is  $\hat{p} = 1.1\%$ , from  $n = 1,071$  (red dot).



# Statistical reasoning and probability calculation



- ▶ observation / estimate is  $\hat{p} = 1.1\%$ , from  $n = 1,071$  (red dot).
- ▶ Hence, the upper limit of the 95% confidence interval is 1.9%.
- ▶ **Note:** the height of the grey bars depends on the sample size, hence so does our confidence interval.



# Confidence interval: interpretation (1/2)

**Case:** the prevalence (95% confidence limits) is estimated to 1.1% (0.6%–1.9%).

**Interpretation:**<sup>24</sup>

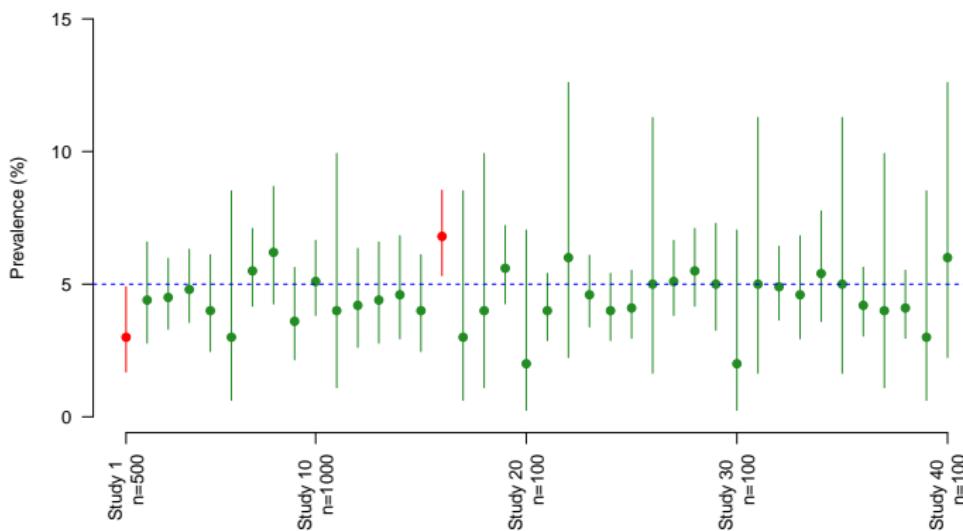
- ▶ **Loosely speaking**, the prevalence in the entire population lies between 0.6% and 1.9%, with 95% probability (at least).
- ▶ **Formally**, 95% confidence intervals are such that if we repeat again and again the process of data collection and data analysis, then (at least) 95% of the confidence intervals catch the parameter (i.e. the prevalence, here).

---

<sup>24</sup> Although there are some debates on whether the “loosely speaking” interpretation is acceptable for practical purpose, I think it is, as others (See e.g. Goeman (2017), Biom J, 59(5), 884).



## Confidence interval: interpretation (2/2)



- ▶ prevalence=5%, most of the 95% CIs catch it, but not all (here  $38/40=95\%$ ).
- ▶ in practice **we never know for sure** whether we catch it or not, but we can be 95% "**confident**" about it.



## 95% confidence interval for the population mean

We can estimate the **population mean** by the **sample mean**  $\bar{x}$  and compute a 95% confidence interval as:

$$\bar{x} \pm t_{n-1} \cdot \frac{s}{\sqrt{n}}$$

- ▶  $s$  is the **standard deviation** (SD) and  $n$  is the sample size,
- ▶  $s/\sqrt{n}$  is the **standard error** (SE) of the mean,
- ▶  $t_{n-1}$  is a specific value that depends on the **degrees of freedom**  $n - 1$ , which can be obtain from statistical tables or software<sup>25</sup>. It is  $\approx 1.96$  ("two") for large  $n$ .

$n$	5	10	20	100	500
$t_{n-1}$	2.78	2.26	2.09	1.98	1.96

---

<sup>25</sup>It is the 97.5% quantile of the t-distribution with  $n - 1$  degrees of freedom, we use e.g. `qt(0.975, df=9)` for  $n = 10$ .



## SE versus SD

"The terms '*standard error*' and '*standard deviation*' are often confused. The contrast between these two terms reflects the **important distinction** between **data description** and **inference**, one that all researchers should appreciate"<sup>26</sup>



---

<sup>26</sup> Altman and Bland. "Standard deviations and standard errors." BMJ (2005).

## Assumptions for CI of the mean

The 95% confidence interval for the mean is **valid** when:

- ▶ **Either** the data are **normally distributed**.
- ▶ **Or** the **sample size is not too small**, some say  $n \geq 15$ .

Otherwise the confidence level can actually be lower than the intended 95% level.



## CI for the mean vs prediction interval

Estimating and comparing population means is often useful, e.g. to estimate average treatment effects. To do so, we use confidence intervals for the mean.

However, individual observations will usually deviate substantially from the mean. Hence, it is sometimes useful to estimate an interval within which most (future) observations can be expected.

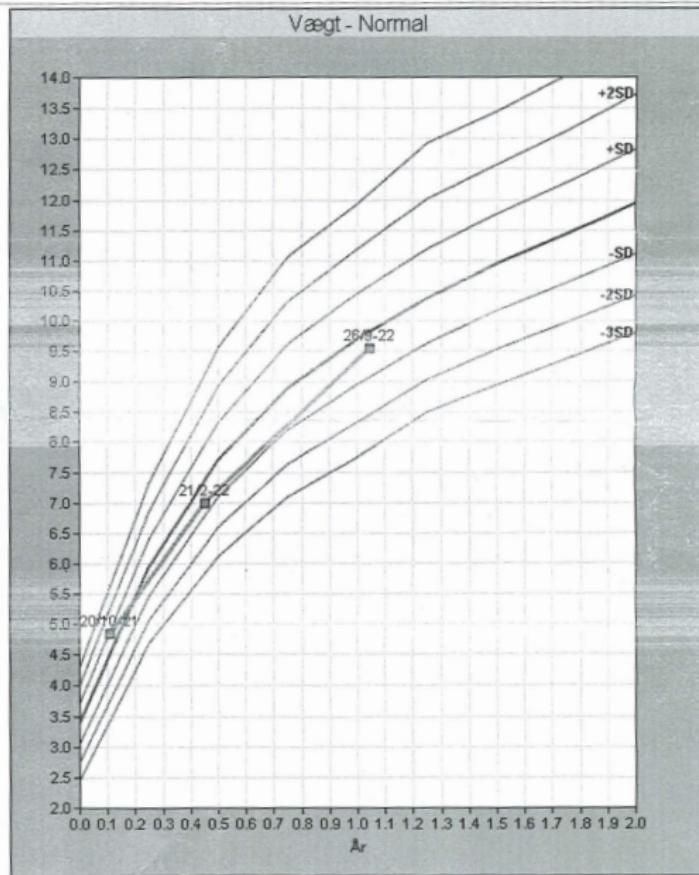
- ▶ **Patient:** Doc, what will likely be my Oxford Knee Score in one year, if I undergo surgery next month?

To provide the patient with a range of likely values, we provide a prediction interval.

- ▶ **Doctor:** About 9 out of 10 patients like you (similar current score, BMI & age) have an improvement between 10 and 38 points.



9. 5



- ▶ Growth charts are examples of prediction intervals.
- ▶ GPs use them to monitor infants' growth.

## Prediction interval

**When the data are normal**, a 95% prediction interval can be computed as

$$\bar{x} \pm t_{n-1} \cdot s \cdot \sqrt{1 + \frac{1}{n}}$$

It catches 95% of the observations in the population.

- ▶ The only difference with the formula of the 95% CI of the mean is that we use  $\sqrt{1 + \frac{1}{n}}$  instead of  $\sqrt{\frac{1}{n}}$ .
- ▶ If  $n$  is large, then approx. “mean  $\pm 2$  SD” (normal range).
- ▶ Does not depend much on the sample size  $n$ .
- ▶ Much wider than the confidence interval (which makes good sense!)



# Tumor thickness example (1/2)

## With log-transformed data:

- ▶  $n = 205$
- ▶  $\bar{x} = 5.22$
- ▶  $s = 1.01$
- ▶  $t_{n-1} = 1.97$
- ▶ Confidence interval for the population mean: 5.08–5.36
- ▶ Prediction interval: 3.22–7.22

## Conclusions (1/2):

- ▶ With 95% certainty we estimate that the population mean of the log of the tumor thickness lies between 5.08 and 5.36.
- ▶ We predict that 95% of the future patients from the same population will have values between 3.22 and 7.22.



## Tumor thickness example (2/2)

### Conclusions (2/2):

- ▶ With 95% certainty we estimate that the population **median** of the tumor thickness lies between

$$\exp(5.08) = 161 \text{ and } \exp(5.36) = 213.$$

- ▶ We predict that **95% of the future patients** from the same population will have tumor thickness between

$$\exp(3.22) = 25 \text{ and } \exp(7.22) = 1371.$$

**Note:** medians (or other quantiles) are “preserved” under monotone transformation, while means and standard deviations are not; **for normally distributed data mean=median.**



# Good reporting practice

Estimates and **confidence intervals** are **important to report** in the results of a statistical analysis.

- ▶ How **large** is the true effect/difference of exposure/treatment? (**effect size**)
- ▶ You can distinguish **lack of effect** from **lack of evidence** (lecture 2).

**Confidence intervals are more informative than p-values<sup>27</sup>.**

- ▶ You can most often tell a **significant** effect/difference from the confidence intervals but **not the other way around!**



# Testing by comparing confidence intervals

Often you can tell whether or not a significant difference in means is found between two groups by comparing the confidence intervals.



**Note:** this works for **independent** samples, not with paired data.



## R-demo ("head" only, see webpage)

```
rm(list=ls()) # clear all objects from R memory

# load relevant packages
library(DoseFinding) # for loading data example 1
library(HistData)    # for loading data example 2
library(timereg)     # for loading data example 3
library(MESS)        # for wally plots

# Load first data example (Data on doses)
data(migraine)

# visualize the first line of the data
head(migraine)

# make a barplot (of counts)
barplot(migraine$ntrt,           # should be a table or counts
         col=1:nrow(migraine), # to color the bars
         names=migraine$dose,  # names under the bars
         xlab="Dose (mg)", 
         ylab="n. of patients")
```

