Faculty of Health Sciences

# Day 1: Essential statistical concepts

Paul Blanche

Section of Biostatistics, University of Copenhagen

April 17, 2023



## Outline/Intended Learning Outcome (ILOs)







## Overall (course) intended learning objectives

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 & topics

| Date | Day | Room (8:00-15:00) | Topics | Teachers |
|---|---|---|---|---|
| 17 April 2023 | Monday | CSS-7.0.40 | Overview, data, descriptive statistics, concept of statistical inference, confidence intervals | Paul Blanche, Carolin Herrmann |
| 19 April 2023 | Wednesday | CSS-7.0.40 | Hypothesis testing, tests for continuous outcomes, multiple testing | Paul Blanche, Carolin Herrmann |
| 24 April 2023 | Monday | CSS-7.0.40 | Univariate linear regression, correlation, regression to the mean | Paul Blanche, Zehao Su |
| 26 April 2023 | Wednesday | CSS-7.0.40 | Analysis of Variance (One-way and Two-way ANOVA) | Paul Blanche, Zehao Su |
| 3 May 2023 | Wednesday | CSS-7.0.40 | 2x2 tables, odds ratio, two sample tests for binary responses | Paul Blanche, Zehao Su |
| 8 May 2023 | Monday | CSS-7.0.40 | Logistic regression | Paul Blanche, Zehao Su |
| 10 May 2023 | Wednesday | CSS-7.0.40 | Multiple linear regression, confounding, interaction | Paul Blanche, Alessandra Meddis |
| 15 May 2023 | Monday | CSS-7.0.40 | Repeated measurements | Brice Ozenne, Alessandra Meddis |
| 17 May 2023 | Wednesday | CSS-7.0.40 | Survival analysis | Paul Blanche, Alessandra Meddis |
| 24 May 2023 | Wednesday | CSS-7.0.40, CSS-7.0.06 | Presentation and discussion of homework assignments | Paul Blanche, Brice Ozenne |

## Formalities

1. Lecture start at 8:00, computer exercises around 12:00.
2. A homework assignment is handed out after lecture 5.
3. The homework assignment is turned in after lecture 9.
4. This highly ambitious course gives **8 ECTS** points, **if**
   - you attend 80% of all teaching units (we count the signatures)
   - 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/BasicStat2023.html`





## Challenges for students and teachers

We start with he 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.





## Sorry!

We are continuously working to update and improve the course.

There might lead to:

- Last minute adjustments to the course material.
- Typos.
- Exercises that are too hard or too easy.
- etc.

**Your feedback is much appreciated.**





## Textbooks

This course is not based on a single textbook, but the following book is an excellent reference. It covers most of the topics of the course (often in more details) and has been written for a similar audience.

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

There are many other books on "elementary" statistics appropriate for this course. Those are 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.

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

Links are provided via the webpage.



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Statistics is not only about data analysis

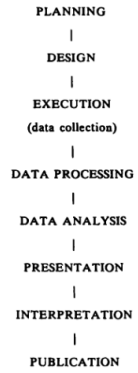



**Figure 1.1** General sequence of steps in a research project.

*"Statistical thinking can contribute to every stage"* [1].

 [1] *Practical statistics for medical research*, Altman (1990).

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## What is statistics?

*"Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data."*[2]

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

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

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

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

 [2] https://en.wikipedia.org/wiki/Statistics

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Word of caution

When used well, statistics turns data into knowledge.

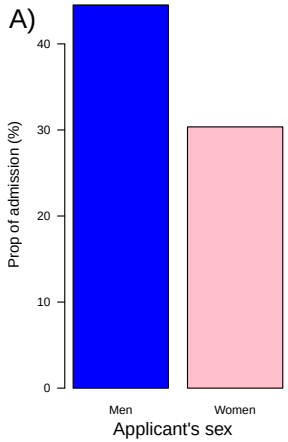
When misused, statistics can be seriously misleading and counterproductive to advance science.

Statistics is not easy to learn and developing good skills usually takes time and is hard work.

Rigorous statistical thinking is not always intuitive! "*Are humans good intuitive statisticians?*", not really. [3]

 [3] see e.g. *'Thinking, Fast and Slow'*, by Daniel Kahneman (Nobel prize).

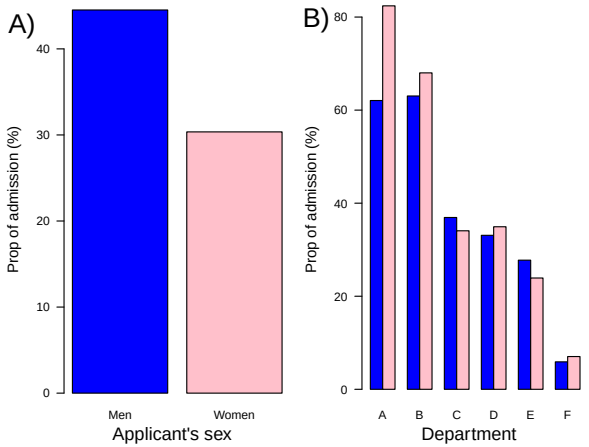
# Simpson paradox: Berkeley admission (1973) [4]




 [4] Bickel et al. (1975), Science, 187, 398–403.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

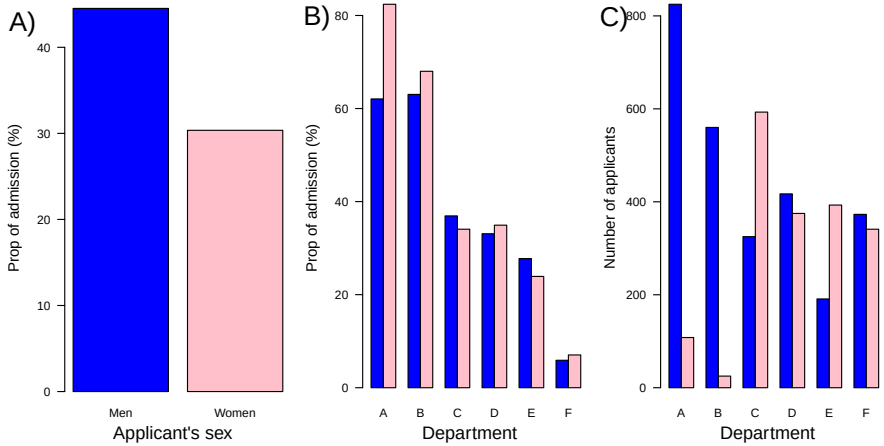
# Simpson paradox: Berkeley admission (1973) [4]




 [4] Bickel et al. (1975), Science, 187, 398–403.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

# Simpson paradox: Berkeley admission (1973) [4]




 [4] Bickel et al. (1975), Science, 187, 398–403.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

# 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

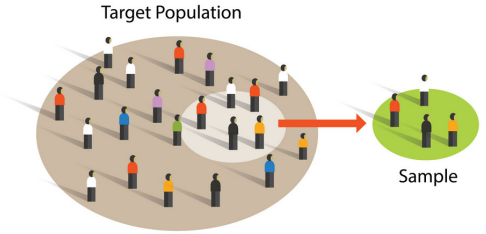
# The general principle of statistical inference

1. Research aims and hypotheses are formulated.
2. The study population is defined and data are collected (measured) on a random subset[5] of all samples/individuals in the population
3. The sampled data are analyzed but conclusions are drawn about the full study population.

[5] 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 on May 20, 2020)

## Descriptive statistics

Out of 2,600 randomly selected people, 1,071 (41.2%) have agreed to be tested for antibodies. [6]

## 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%).[7]

[6] https://www.ssi.dk/aktuelt/nyheder/2020/de-forste-forelobige-resultater-af-undersogelsen-for-covid-19-i-befolkningen-er-nu-klar

[7] 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 question

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). [8] 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.

[8] Actually, this is often helpful in observational study, to better understand what can and cannot be studied using the available data.





## Prespecification matters

In this course, the interpretation of the statistical results (confidence intervals, p-values) will be given assuming than 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.[9]

[9] Schwab et al. (2022) Ten simple rules for good research practice. PLoS Comput Biol 18(6): e1010139.



## 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 spreadsheat):

- 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 [10].

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

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

[10] This might be important when you have repeated/clustered measurements.





## Data type III: Troublesome data

**Censored data:**

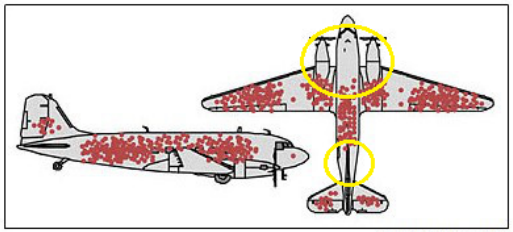
- Concentration below the limit of detection.
- Survival time beyond maximal follow-up (lecture 9).

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

- Failed, lost or cancelled measurements.
- Patients not showing up for blood sample.

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").
- Identifying outliers and correcting errors.
- Identifying and visualizing trends.
- Supporting the interpretaion 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:**

*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...*

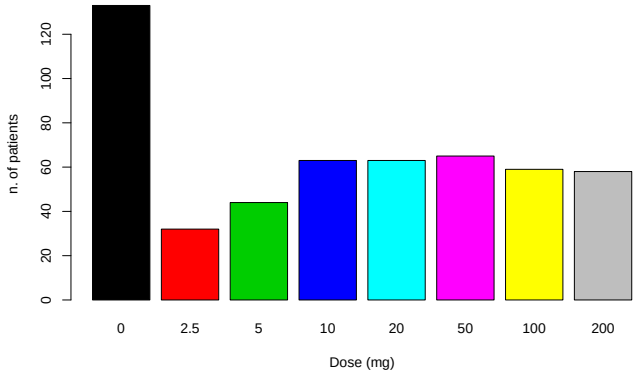
**But:** keep study aims and sample size in mind. Use common sense!

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



## Categorical data: easy case

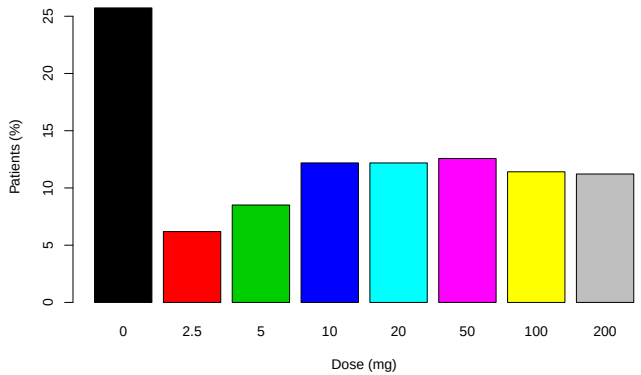
Categorical data can be summarized in group totals and percentages. Use barplots for graphical presentation.



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## 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 normally and non-normally distributed data.

**For graphical presentation, options include:**

- Dotplot (aka Stripchart), boxplot, or histogram.[11]
- QQ-plot[12] to check an assumed normal distribution.

[11] preferably the first for small samples.
[12] i.e. Quantile-quantile-plot.

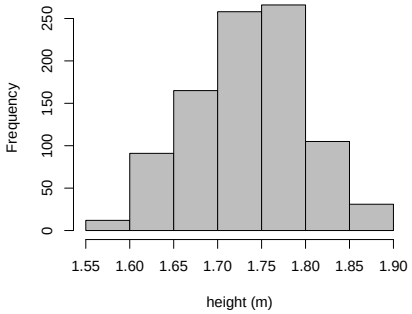


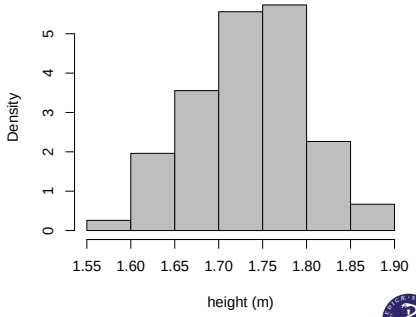
## Histogram (Galton data (1886))

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

Frequency: numbers of observations in interval (to the left).

Density: frequencies are normalized so that the areas of the bars add to one (probabilty distribution).



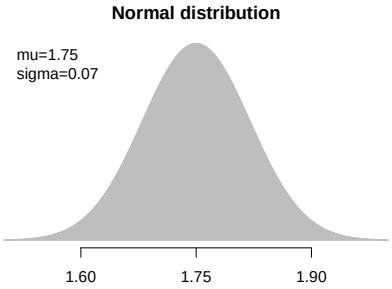



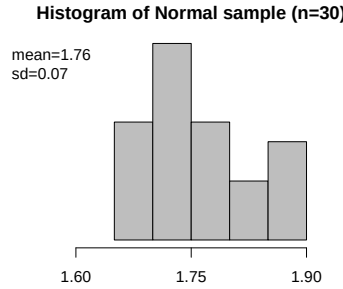



## The normal distribution (1/2)

**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.

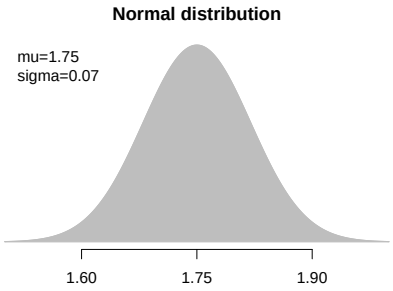


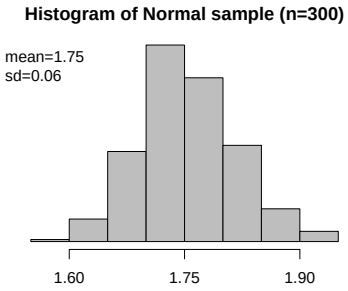



# The normal distribution (1/2)

**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.



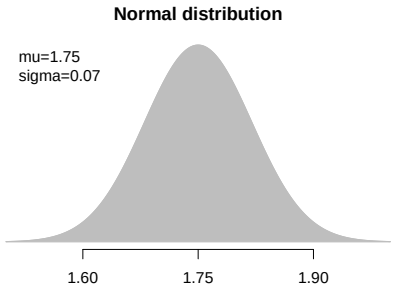



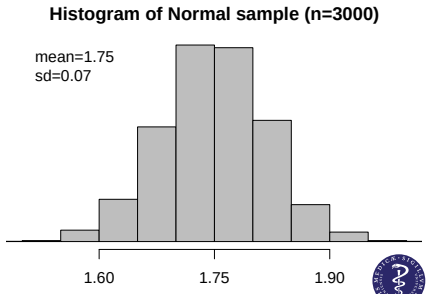



The normal distribution (1/2)

**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.

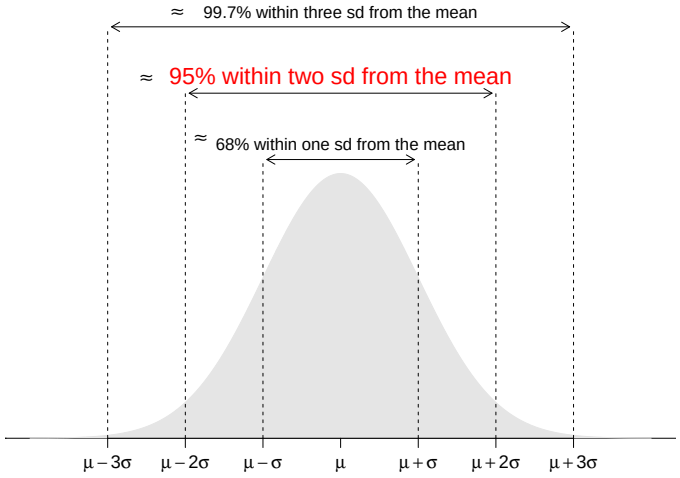



33 / 66

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

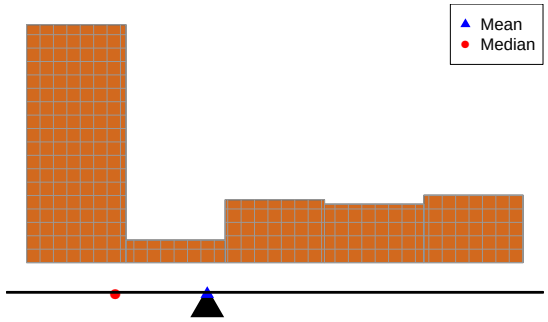
# The normal distribution (2/2)




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



# Mean vs Median




Mean: center of gravity[13] .

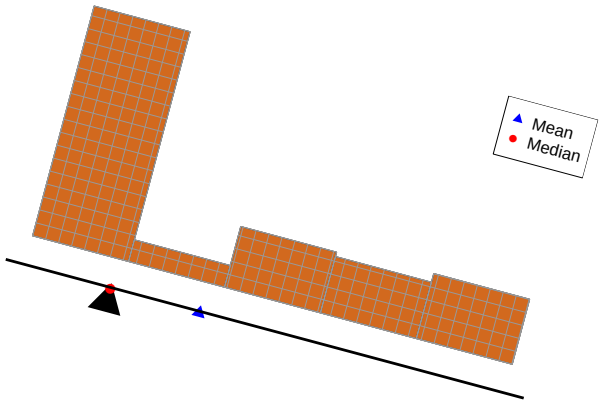
Median: obs. are located 50% above, 50% below[14].

[13] bars of histogram=piles of bricks, x-axis=seesaw, mean=fulcrum.

[14] 50% of the bricks are located to the right and to 50% the left.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Mean vs Median




Mean: center of gravity[13] .

Median: obs. are located 50% above, 50% below[14].

[13] bars of histogram=piles of bricks, x-axis=seesaw, mean=fulcrum.

[14] 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.[15]

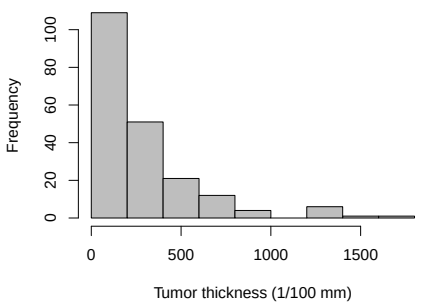
[15] but whatever the (often unknown) distribution of the data, the normal range cannot contain less than 75% of the data (Chebyshev's inequality).



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Case: log-normal data

Tumor thickness at baseline[16] (n=205) has a **skewed** (i.e. asymmetric) distribution.





But the log-transformed data are "approximately" normal.

[16] i.e. at surgery for malignant melanoma. Drzewiecki et al (1980).



## Case: interpretation of summary statistics

**Without transformation:**

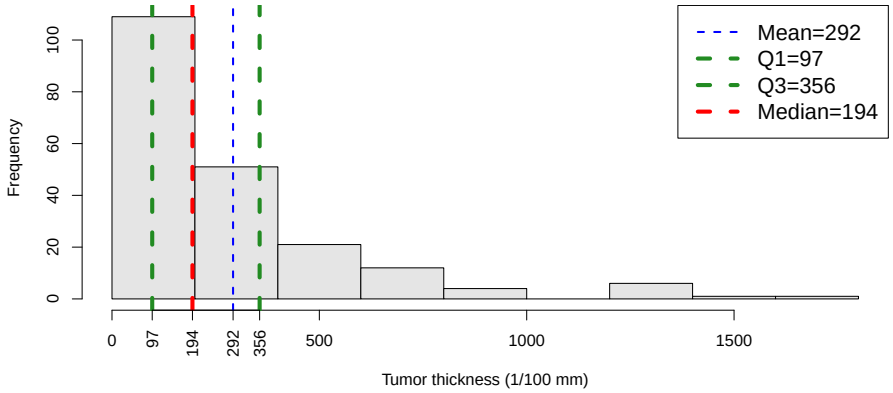
- Mean ± of thickness (in 1/100 mm): 292.0 ± 295.4
- Normal range ≈ -299.9–883.9 : negative thickness "normal"??

**With log-transformation:**

- Mean ± of log thickness: 5.22 ± 1.01
- Normal range of log thickness: ≈ 3.20–7.25.
- **Back-transformed:** $\exp(3.20)$ to $\exp(7.25)$, i.e. 24.5–1408.

## Median and quartiles

**Non-normal data should not be described by mean (SD)**.
Instead, we prefer median and IQR[17].

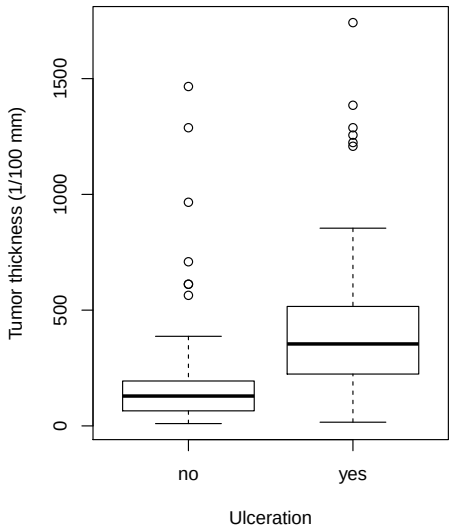



Median and quartiles divide the data into four equal proportions.
We call the interval $[Q1; Q3]$ the interquartile range.

[17] This can be the "default" choice.



## 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).





## 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).

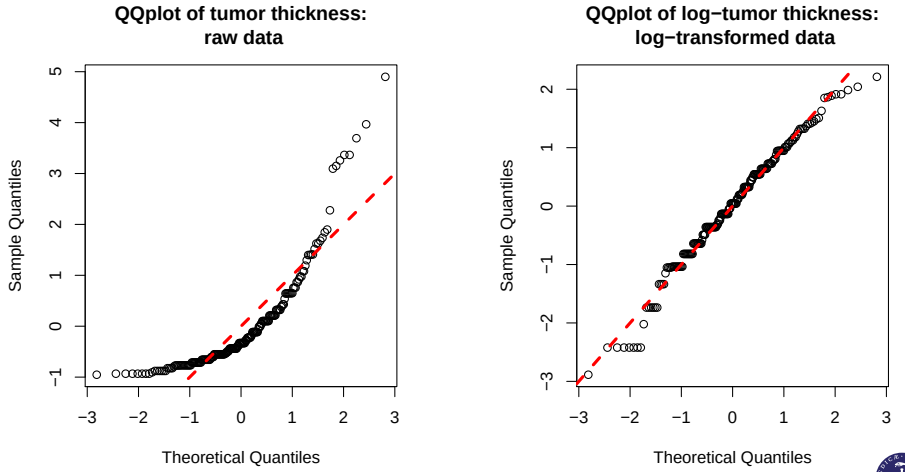
The 25%-, 50%-, and 75% quantiles form the quartiles.

- **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

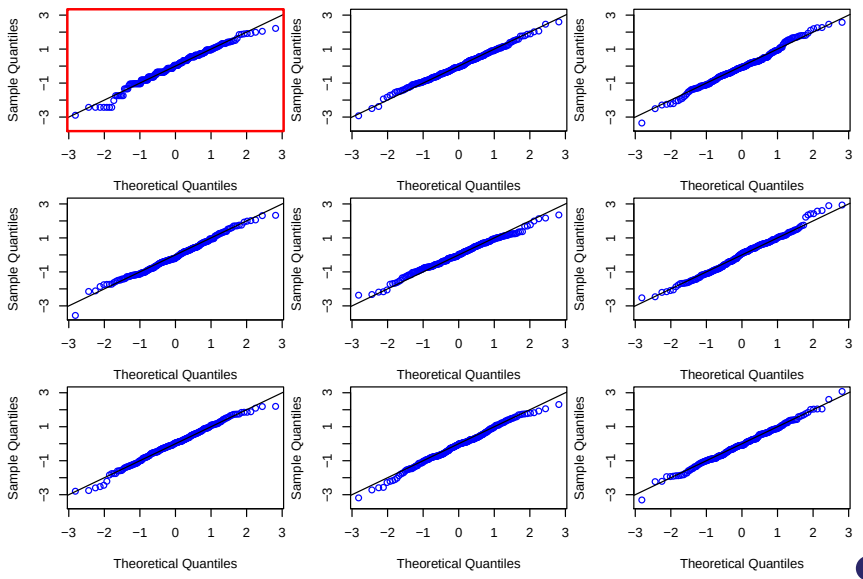
Compare the ordered (standardized) values from the data (smallest to largest) with the corresponding quantiles of the normal distribution.




**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).

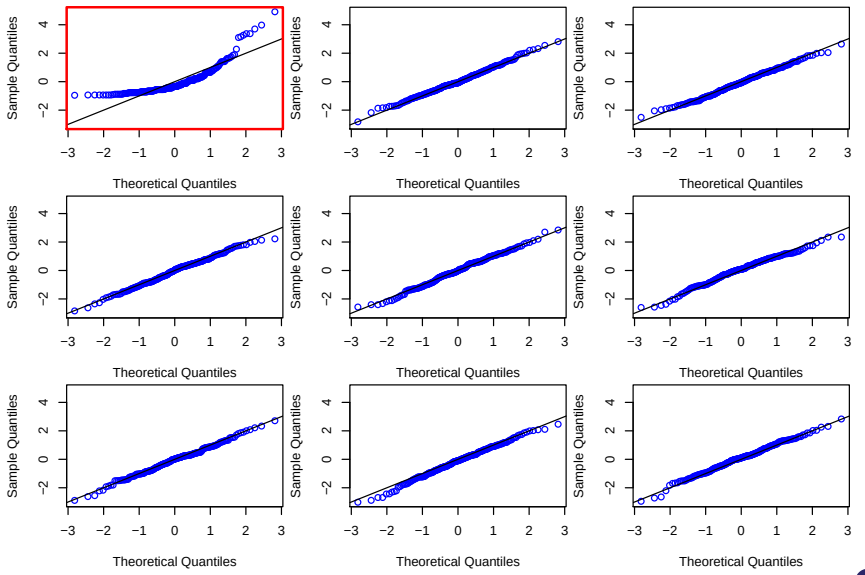
# Wally plot to "visualize" random sampling

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






# Without log-transformation





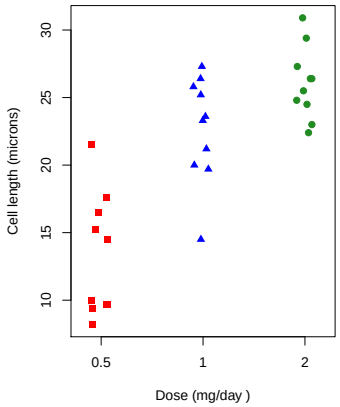



# Small data: what should we do?

The usual descriptive statistics and plots are usually not sensible when the sample size is very small (e.g. with $n = 3$).[18]

**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").

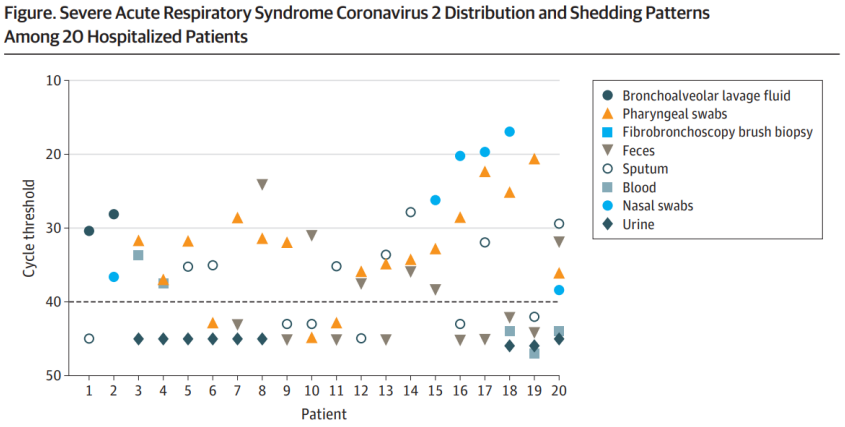



[18] See e.g.: *"Show the data, don't conceal them"* by Drummond & Vowler (2011).



# Good plots can be very informative...

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




(Wang et al. "Detection of SARS-CoV-2 in different types of clinical specimens." Jama 323.18 (2020): 1843-18...

# 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%.

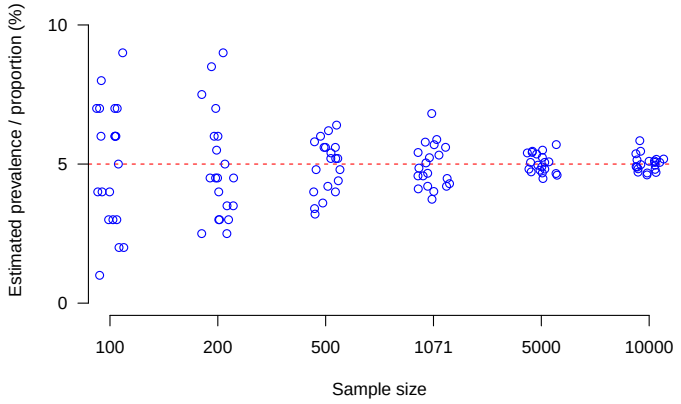
Can we **conclude**:
- that **exactly** 1.1% have been infected? **No!**
- that **"approximately"** 1.1% have been infected? **Yes**

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.



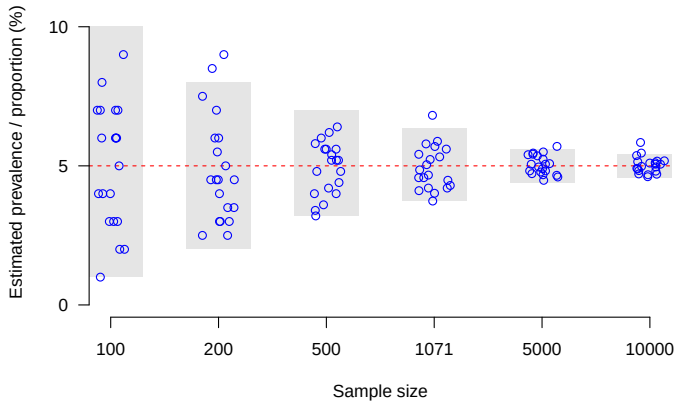
# Understanding uncertainty

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

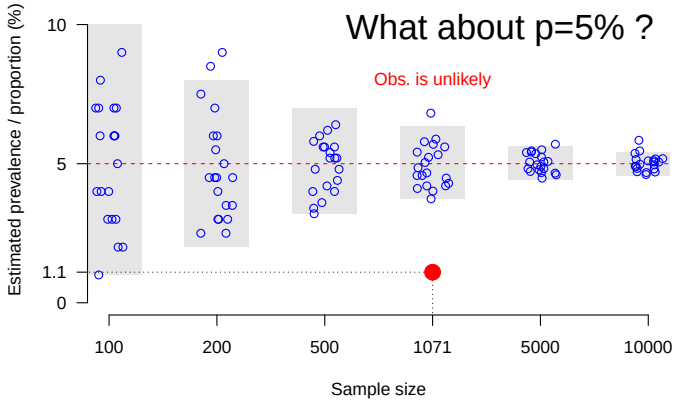
Because we know how to calculate what to expect after assuming a specific value for the (true) population parameter (e.g. a prevalence of 5%), we can "reverse" the reasoning and know which values could "likely" lead to the observation of our sample estimate.

These values will form a confidence interval.

The confidence interval gives a range of "plausible" values for the parameter that we are tying to estimate.



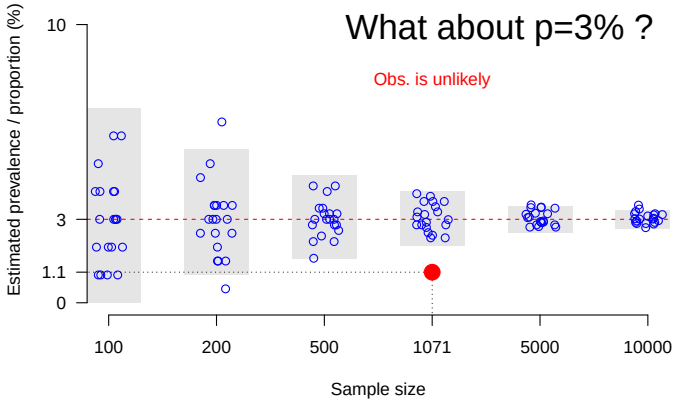
## Statistical reasoning and probability calculation




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

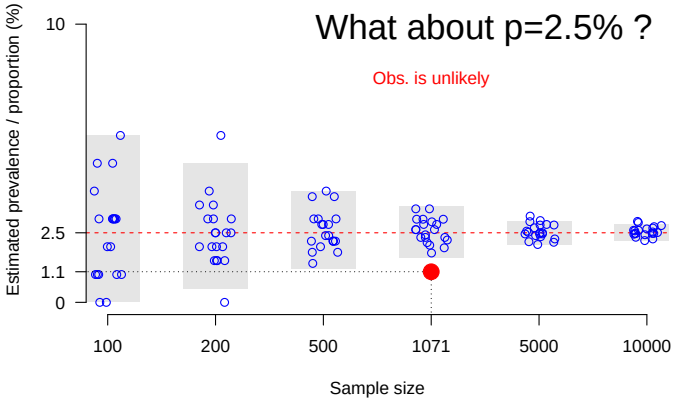


## Statistical reasoning and probability calculation




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

## Statistical reasoning and probability calculation

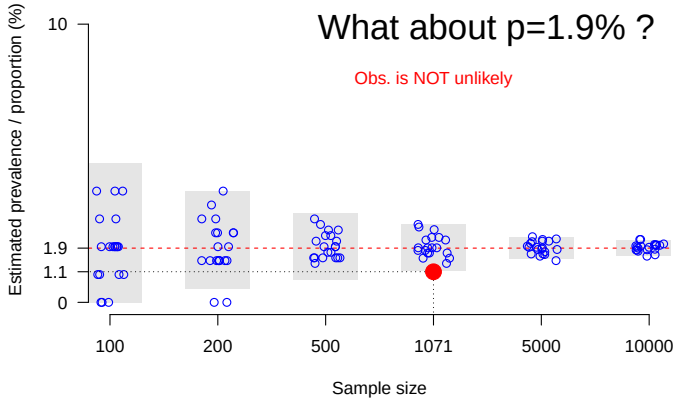



- observation / estimate is $\hat{p} = 1.1\%$, from $n$ =1,071 (red dot).



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## 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.



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Confidence interval: interpretation (1/2)

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

**Interpretation:**[19]

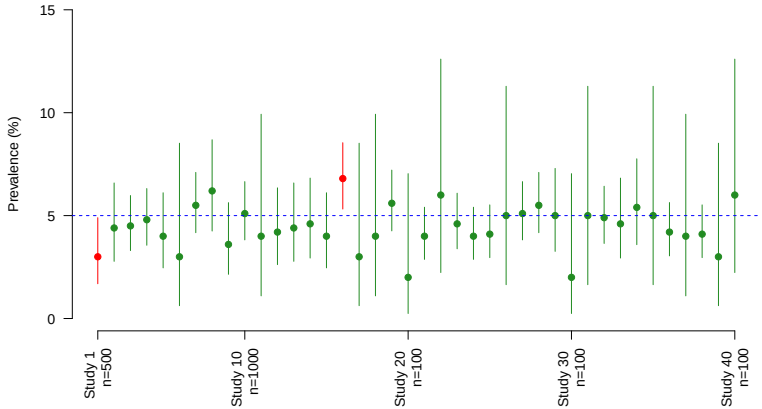
- Loosely speaking, the prevalence in the entire population lies between 0.6% and 1.9%, with 95% probability (at least).
- Formally, 95% confidence intervals are such that if we repeat again and again the process of data collection and data analysis, then (at least) 95% of the confidence intervals catch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

[19] Although there are still some debates on whether the "loosely speaking" interpretation is acceptable, I think it is, as others (See e.g. Goeman (2017), Biom J, 59(5), 884).



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

## Confidence interval: interpretation (2/2)




- prevalence=5%, most of the 95% CI 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 [20]. 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 |

[20] It is the 97.5% quantile of the t-distribution with $n-1$ degrees of freedom, which in R 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"[21]

[21] 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

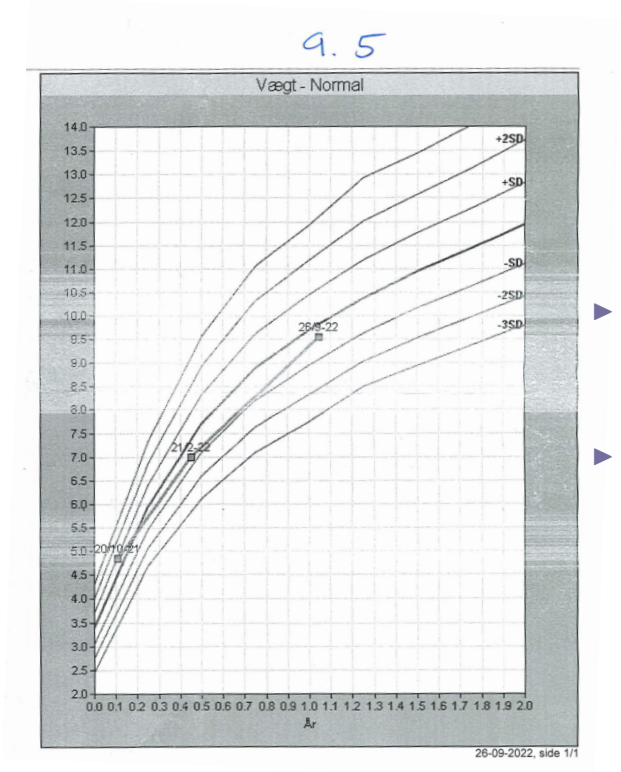
Estimating and comparing population means is often useful, e.g. to estimate average treatment effects. To do so, we use confidence intervals for the mean.

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.




- Growth charts are examples of prediction intervals.
- GP use them to monitor infant's 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





## 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.

**Confidence intervals are more informative than p-values[22].**

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

[22] More on p-values and significance at lecture 2.





## 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)

```
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$ntrt,
        col=1:nrow(migraine),
        names=migraine$dose,
        xlab="Dose (mg)",
        ylab="n. of patients")
```