



University of  
Zurich<sup>UZH</sup>

Division of Biostatistics

---

# Introduction to Biostatistics

As part of Cancer Biology PhD Program

Kaspar Rufibach

Division of Biostatistics

Institute of Social and Preventive Medicine

University of Zurich

# Agenda

What is statistics?

Types of data, descriptive statistics, effect measures

Distribution, Normal distribution

Uncertainty and confidence intervals

Hypothesis testing

Multiple testing

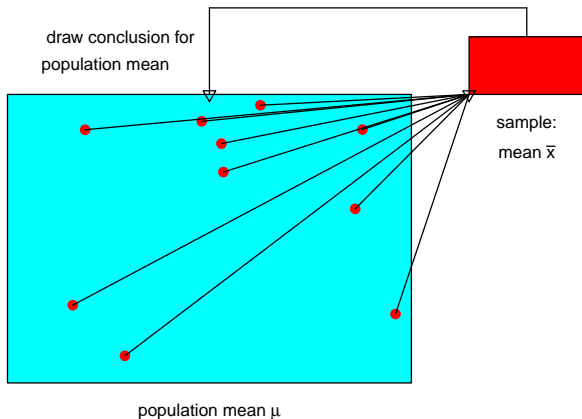
Study design and sample size computation

## Fish data

- Randomized **therapy study** in 308 heart patients in Germany.
- Regimens: normal, normal + fish (3×/week), reduced, reduced + fish.
- Primary endpoint: change in cholesterol level from randomization to four weeks later, **comparison** of groups: “Does a fish diet decrease cholesterol level?”

patnr	geschl	alter	groesse	gew0	gew28	chol0	chol28	fisch	kalorie	gruppe
1	m	44	180	84.4	84.1	246	246	1	0	normal + fish
1	m	58	180	91.3	81.3	450	323	1	1	reduced + fish
2	f	56	163	65.5	64.2	226	216	1	0	normal + fish
3	m	69	171	59.5	59.2	275	245	1	0	normal + fish
3	m	61	182	86.4	82.6	261	206	1	1	reduced + fish
4	m	80	164	67.0	68.2	219	225	0	0	normal
4	m	59	176	75.9	76.1	298	225	0	1	reduced
5	m	44	183	101.3	96.6	279	217	0	1	reduced
5	m	49	170	79.4	78.0	217	222	0	0	normal
6	f	60	156	81.2	78.6	302	221	0	1	reduced

# Population and sample



⇒ Draw conclusions about population from **random** sample.

# Terminology

**Population:** subjects of interest for given question.

All patients with heart disease in Germany.

**Sample:** **randomly** drawn subset of population.

Patients in fish study.

**Variable:** quantity of interest.

Sex, age, height, cholesterol level.

**Parameter:** value of variable in population.

True mean height in population, i.e. of all persons in Germany with heart disease (mean or some other quantity of interest, e.g. standard deviation).

**Estimate:** value of variable in sample.

Mean height in fish study.

# Terminology

## Population:

- Quantities of interest (e.g. mean, standard deviation) are **fixed**.
- **Parameter, true** or **theoretical** value.

## Sample:

- Quantities of interest are **random**, since the sample is **randomly drawn**.
- Different sample – different value of quantity of interest.
- We compute **estimates** that bear **UNCERTAINTY**.

## Descriptive statistics:

- Describe data in your sample, no conclusions about population.
- Tables, graphs, quantification of treatment effect.

## Inferential statistics:

- Draw inference from sample about population: From fish study about all heart disease patients in Germany.
- Quantify uncertainty in estimation: Statistical tests and confidence intervals.

# Randomly drawing a sample

Goal: sample should allow for inference about population, “representative”.

Approach: draw sample **randomly**. Distribution of characteristics approximately as in population.

Statistics vs. Mathematics:

- Statistics = field of mathematics, “mathematics of uncertainty”.
- **Quantification of uncertainty** using mathematical methods.
- “Inability” to provide  $\mu$ : intrinsic to the problem.

# Types of data

**Discrete, nominal:** Observations are in a given set, e.g. {green, blue, red}, {0, 1, 2, ...}, {0, A, B, AB}.

Binary or dichotomous: {male, female}.

Ordered: performance status, nodal status.

**Continuous:** any value of observation possible (measurements), e.g. tumor size, laboratory parameters, temperature.

**Survival data:** quantitative measurement, e.g. time from diagnosis to death. But: At time of analysis, some patients are still alive.

Methods for

- description and
- analysis

of data **depend on its type!**

# Descriptive statistics

## Goals of descriptive statistics:

- Summarize and identify main patterns in **sample**.
- Assess properties of distributions.
- Find strange values (“outliers”).

## Discrete data

**Absolute** (counts) and **relative** (percentages) frequencies of categories.

# patients in treatment groups:

	normal	normal + fish	reduced	reduced + fish	total
female	18	30	19	25	92
male	79	68	31	38	216
total	97	98	50	63	308

Barplots, sector diagrams.

Percentages: What is 100%? Count missings or not?

# Descriptive statistics for continuous data

## Location:

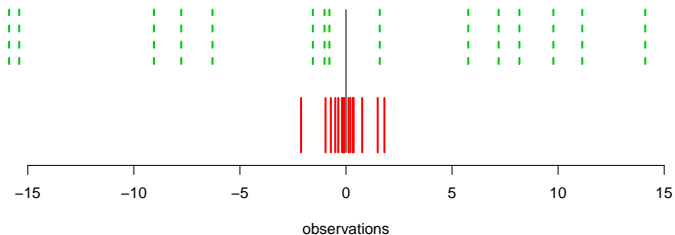
- Goal: describe “location”, “center” of the data.
- Mean (“average”): describes average location, not robust to extreme observations.
- Median: measure of data center (“50% point”), “typical” value, **robust** to extreme observations.

## Use **mean or median**?

- Which aspect should be described?
- How does the distribution look?
- Consistency with inferential methods?

Quantiles: the  $\alpha$ -quantile is the value  $q_\alpha$  so that  $100 \cdot q_\alpha\%$  of the data are smaller than  $q_\alpha$ .

## Descriptive statistics for continuous data



⇒ Same mean, different variability.

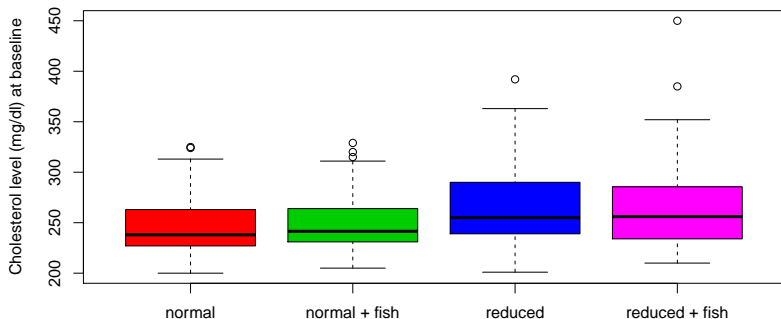
### Measures of variability:

- Standard deviation: measure of spread.
- Variance: squared standard deviation.
- Interquartile range: 3rd - 1st quartile.
- Range: maximum - minimum.

# Descriptive statistics for continuous data

## Graphs:

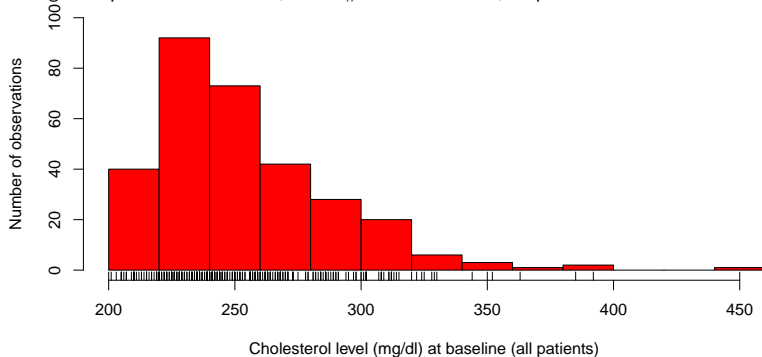
- Assess distribution: boxplot, histogram, frequency polygon, density estimates.
- Relation between two variables: scatterplot.



Differences in location, skewed distributions, identify strange values.

## Descriptive statistics for continuous data

Histogram: split data into classes, count #obs in each class, barplot.

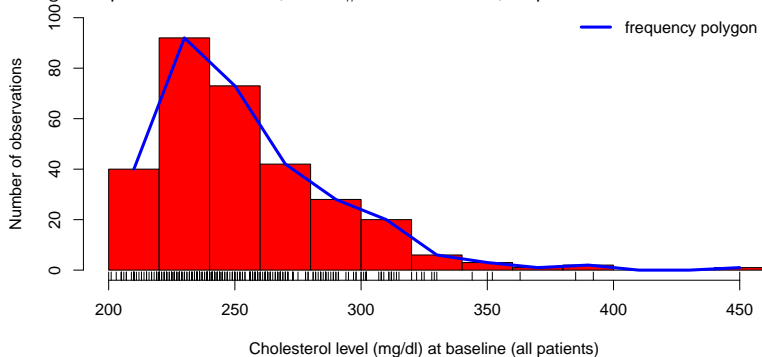


Class width? Try different widths.

Compare different histograms: probability scaled.

# Descriptive statistics for continuous data

Histogram: split data into classes, count #obs in each class, barplot

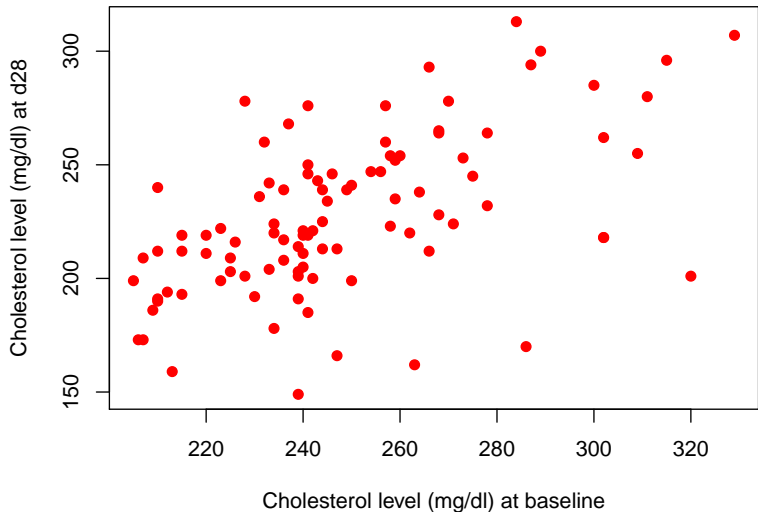


Class width? Try different widths.

Compare different histograms: probability scaled

## Descriptive statistics for continuous data

Scatterplot (Diet: normal + fish)



## Effect measures for binary and nominal data

Are vaccinated patients less susceptible to flu compared to placebo?

	no flu	flu	total
placebo	150	70	220
vaccinated	200	35	235
total	350	105	455

Proportions of flu patients:

- placebo:  $\hat{p}_1 = x_1/n_1 = 70/220 = 0.32$
- vaccination:  $\hat{p}_2 = x_2/n_2 = 35/235 = 0.15$

Binary data. How to **quantify** this effect?

# Effect measures for binary and nominal data

Four measures:

- **risk difference:**  $RD = p_1 - p_2 = 0.32 - 0.15 = 0.17$ .
- **relative risk:**  $RR = p_1/p_2 = 2.14$ .
- **odds ratio:**  $OR = (p_1/(1 - p_1))/(p_2/(1 - p_2)) = 2.67$ .
- **number needed to treat:**  $NNT = 1/|RD| = 5.91$ .

No difference  $p_1 = p_2$ :  $RD = 0$ ,  $RR = OR = 1$ .

OR often preferred, since:

- If  $p_1, p_2$  small  $\Rightarrow$  interpretation as RR.
- Consider event or non-event  $\Rightarrow$  only OR does not care.
- Many methods formulated using odds (e.g. logistic regression)  $\Rightarrow$  natural interpretation with OR.
- Case-control study: disease OR only interpretable quantity.

## Distribution (informally)

Given a sample  $x_1, \dots, x_n$ , distribution gives for any interval  $I$  number of values in  $I$ .

**Symmetric:** similar pattern left and right of data center.

**Right-skewed:** extreme observations tend to be on the right side.

Examples: blood pressure, lab parameters as cholesterol (see histogram on previous slide).

# Normal distribution

Properties of Normal distribution:

- **Symmetry**: Deviations to left/right are equally often and of equal size.
- Large **deviations** rarely occur.
- Characterized through **mean** and **standard deviation**.

For a normally distributed sample we have:

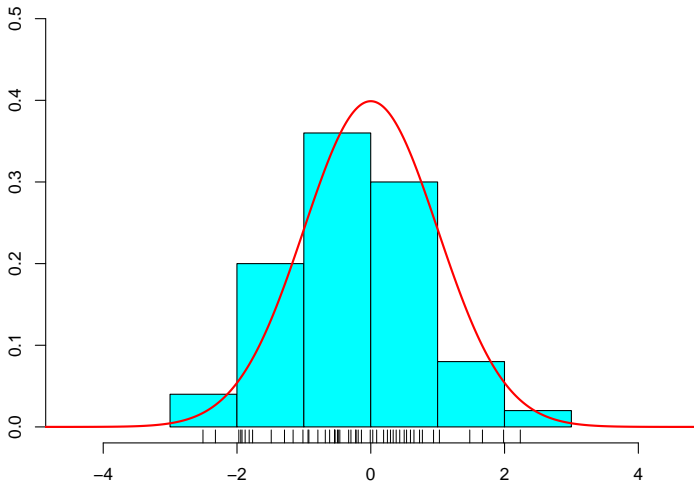
- $\approx 2/3$  (68.2%) of data lie in mean  $\pm$  standard deviation ( $\bar{x} \pm s$ ).
- $\approx 95\%$  of data lie in mean  $\pm 2 \times$  standard deviations ( $\bar{x} \pm 2s$ ).
- Publications:  $\bar{x}$  und  $s$  provided as descriptive statistics.

Why is Normal distribution important?

- Distribution of mean **approximately normal**.
- Many variables normal: Body height  $\Rightarrow$  "mean" of many small effects.
- Statistical methods (tests, confidence intervals, regression) that rely on means  $\Rightarrow$  Normal distribution appears at some point.

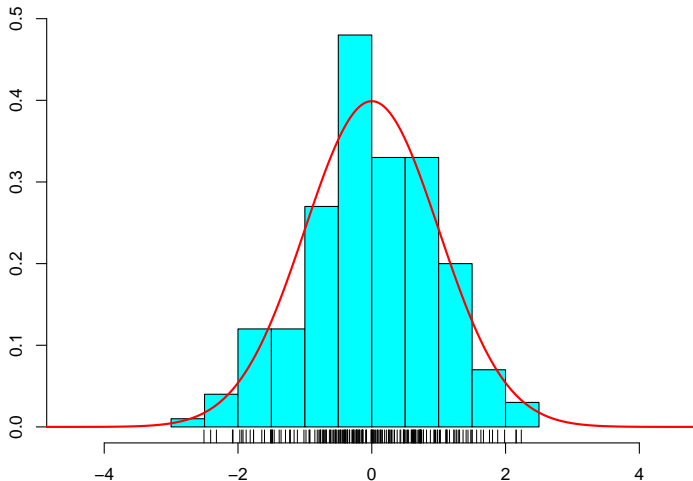
# Illustration Normal distribution

Normal sample,  $n = 50$



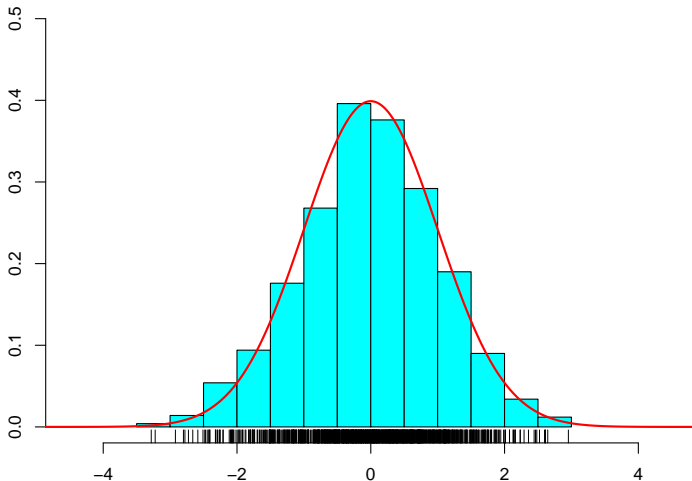
# Illustration Normal distribution

Normal sample,  $n = 200$



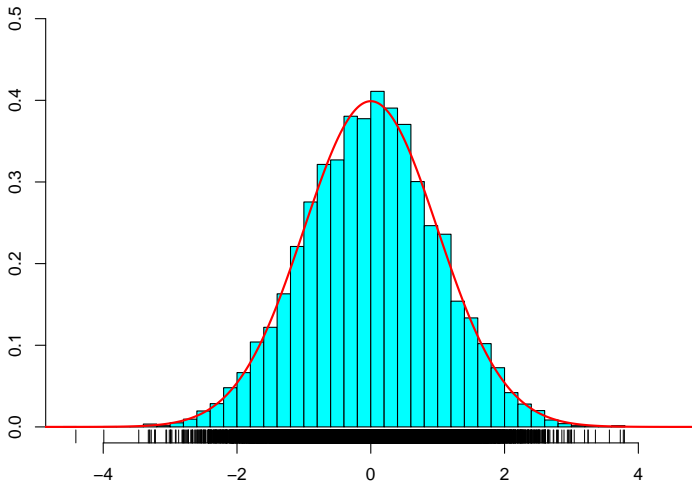
# Illustration Normal distribution

Normal sample,  $n = 1000$



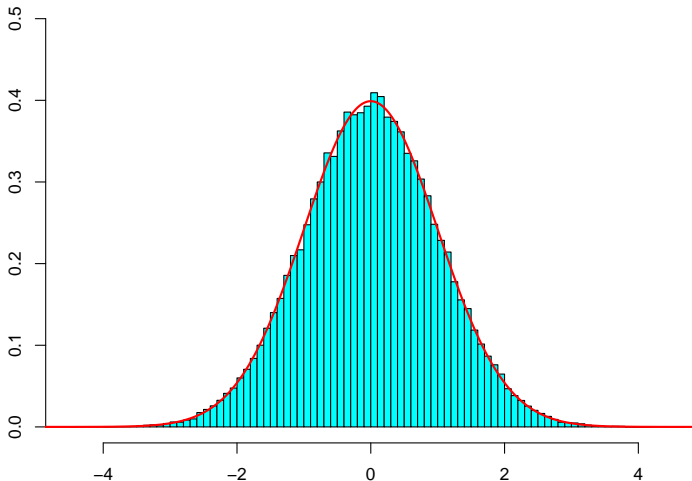
# Illustration Normal distribution

Normal sample,  $n = 10000$



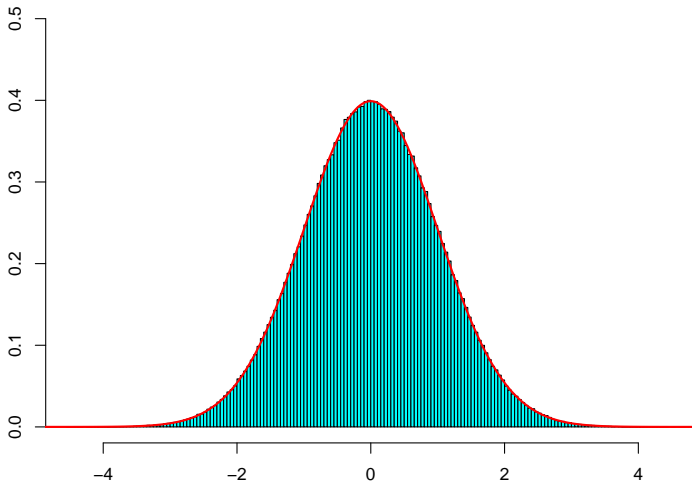
# Illustration Normal distribution

Normal sample,  $n = 100000$



# Illustration Normal distribution

Normal sample,  $n = 1000000$



# “Outlier”

No **unanimous** definition of outlier.

“Remarkable observations”:

- Possibly contain a lot of information, **do not just omit!**
- Outlier: Value that is **not compatible with Normal** distribution (boxplot).
- Many statistics and methods heavily depend on large values.
  - ▶ Verify, that remarkable observations are correct.
  - ▶ Apply **robust** methods.

# Uncertainty

Samples are drawn **randomly**  $\Rightarrow$  different samples, different estimates.

How can we **quantify this uncertainty**?

Illustration via mean of height in fish study:

- Consider **fish study** the population:  $n = 308$ ,  $\mu = 169.50$ ,  $\sigma = 8.59$ .
- Goal: learn from **random sample** about population mean  $\mu$ .
- Draw sample of size  $n = 10$ . Compute  $\bar{x} = (x_1 + \dots + x_{10})/10 = 170.65$ .
- How **representative** is the estimate  $\bar{x} = 170.65$  for  $\mu = 169.50$ ?
- $\bar{x}$  is, like the sample itself, random  $\Rightarrow$  different sample, different value for  $\mu \Rightarrow$  **sample variation**.

# Distribution of the mean

We now draw

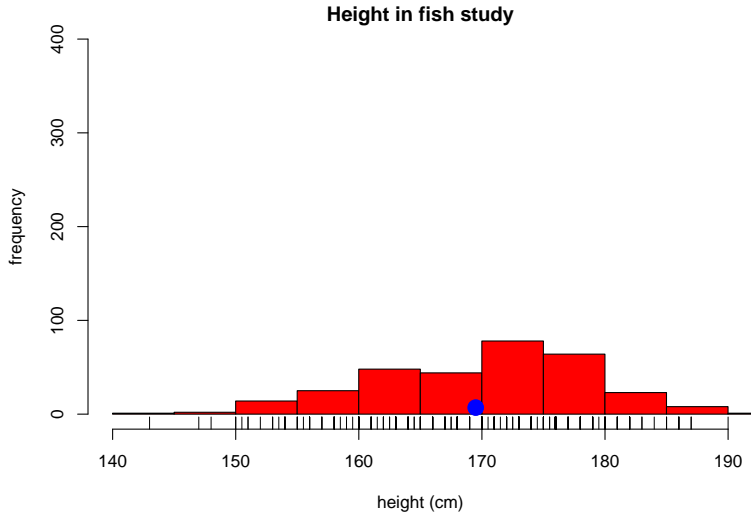
- $r = 2000$  samples of size 10, 30, 50,
- compute the mean in each of these samples,
- and look at a histogram of these means.

First 10 means: 170.1, 166.9, 166.45, 168.5, 167.55, 174.3, 171.65, 172, 166.8, 170.1, . . .

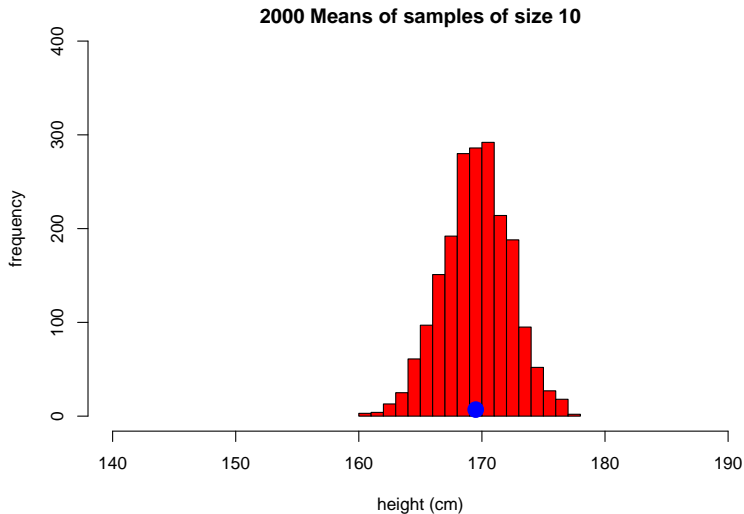
Mean of the 2000 means: 169.53  $\Rightarrow$  close to  $\mu = 169.50$ .

Standard deviation of the 2000 means: 2.73  $\Rightarrow$  remarkably smaller than the population standard deviation:  $\sigma = 8.59$ .

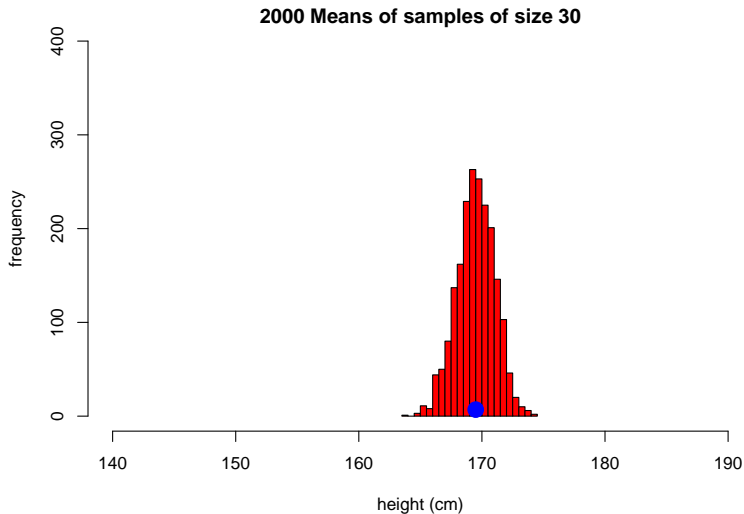
# Illustration distribution of the mean



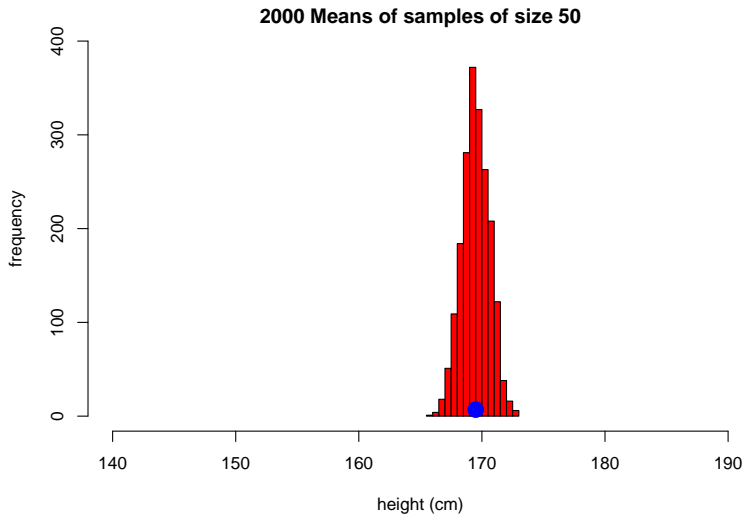
## Illustration distribution of the mean



## Illustration distribution of the mean



## Illustration distribution of the mean



## Observations from simulation

- Population mean:  $\mu = 169.50$ .
- Population standard deviation:  $\sigma = 8.59$ .
- Variability of  $\bar{x}$  **decreases** with growing  $n$ .
- Histograms all “nice and normal”.
- Deviation of a sample mean  $\bar{x}$  from  $\mu$  can be **quantified**:

number of means	n	$\bar{x}_{\text{simul}}$ in cm	$s_{\text{simul}}$ in cm	$\sigma/\sqrt{n}$ in cm
2000	10	169.53	2.73	2.72
2000	30	169.49	1.53	1.57
2000	50	169.49	1.11	1.21

## ~ Central Limit Theorem

If...

**B1** the distribution of a single measurement has mean  $\mu$  and standard deviation  $\sigma$ ,

**B2** we have  $n$  **independent** measurements,

then...

the **mean**  $\bar{x}$  follows approximately a **Normal distribution** with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ .

- Use result to **quantify uncertainty** in estimation of  $\mu$  by  $\bar{x}$  (tests, confidence intervals).
- Approximation the “better” the larger  $n$  is.
- Using  $\bar{x}$  we thus estimate  $\mu$  accurately, in the long run. This estimate varies around  $\mu$ , and that with standard deviation  $\sigma/\sqrt{n}$ .
- Note:  $\bar{x}$  follows a Normal distribution even if single measurements are **not normal** (but fulfill B1)!

# The $\sqrt{n}$ law

$\bar{x}$  becomes **more precise with growing  $n$** .

Standard deviation of  $\bar{x}$  is  $\sigma/\sqrt{n} \Rightarrow$  precision of  $\bar{x}$  (as estimate of  $\mu$ ) is proportional to  $\sqrt{n}$ .

One has to multiply  $n$  by

- **the factor 4** to get double precision,
- **the factor 9** to get triple precision,
- $\delta^2$  to get  $\delta \times$  higher precision.

# Standard error

Standard deviation  $\sigma/\sqrt{n}$  of  $\bar{x}$  (the estimator of  $\mu$ ) depends on  $\sigma \Rightarrow$  unknown.

Replace  $\sigma$  by its estimation  $s \Rightarrow$  **standard error of the mean:**

$$\text{se}(\bar{x}) = s/\sqrt{n}.$$

Standard error: describes the variability of an estimate (not only mean).

## **DISTINGUISH:**

- Standard deviation of the population:  $\sigma$ . Estimated by  $s$ .
- Standard deviation of estimation of the mean:  $\sigma/\sqrt{n}$ . Estimated by SEM  $s/\sqrt{n}$ .

# Confidence interval

Follow up cholesterol level in normal + fish group: 227.4mg/dl.

**Point estimate** of population parameter  $\mu$ .

Different sample – different value of point estimate.

Goal: provide interval  $I$  that contains **plausible values** for  $\mu$ .

Since we know the distribution of  $\bar{x}$  we can provide probability with which  $I$  covers  $\mu$ .

Typical procedure:

- Define **confidence level**  $1 - \alpha$ .
- Compute corresponding  $(1 - \alpha)$ -confidence interval.

Follow-up Cholesterol level normal + fish:

227.4mg/dl with 95% confidence interval [220.6mg/dl, 234.3mg/dl].

# Confidence interval

227.4mg/dl with 95% confidence interval [220.6mg/dl, 234.3mg/dl].

Confidence interval – population parameter: **What is random?**

Formulation: “The interval [220.6mg/dl, 234.3mg/dl] **covers** the true follow-up cholesterol level in normal + fish group with probability not smaller than 95%.”

95% confidence interval for mean:

$$\left[ \bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right],$$

if ***n* large enough**. Here: 98 observations. *z*-confidence interval.

*n* “small” (e.g.  $n \leq 60$ ): 1.96 needs to be increased, based on *t*-distribution.

Example: for  $n = 10$  use 2.262.

# Confidence interval

## A $(1-\alpha)$ -confidence interval

- depends on the sample and is therefore **random**,
- contains **plausible values** for a parameter (the true, unknown effect), those values of the parameter that are compatible with the data.
- **covers** the true parameter with probability  $1 - \alpha \Rightarrow$  i.e. with probability 95% if  $\alpha = 0.05$ .
- is            the larger the uncertainty with regard to the parameter,
- is            the larger the number of observations  $n$ ,
- is            the smaller the standard error.

# Confidence interval

## A $(1-\alpha)$ -confidence interval

- depends on the sample and is therefore **random**,
- contains **plausible values** for a parameter (the true, unknown effect), those values of the parameter that are compatible with the data.
- **covers** the true parameter with probability  $1 - \alpha \Rightarrow$  i.e. with probability 95% if  $\alpha = 0.05$ .
- is **wider** the larger the uncertainty with regard to the parameter,
- is           the larger the number of observations  $n$ ,
- is           the smaller the standard error.

# Confidence interval

## A $(1-\alpha)$ -confidence interval

- depends on the sample and is therefore **random**,
- contains **plausible values** for a parameter (the true, unknown effect), those values of the parameter that are compatible with the data.
- **covers** the true parameter with probability  $1 - \alpha \Rightarrow$  i.e. with probability 95% if  $\alpha = 0.05$ .
- is **wider** the larger the uncertainty with regard to the parameter,
- is **smaller** the larger the number of observations  $n$ ,
- is           the smaller the standard error.

# Confidence interval

## A $(1-\alpha)$ -confidence interval

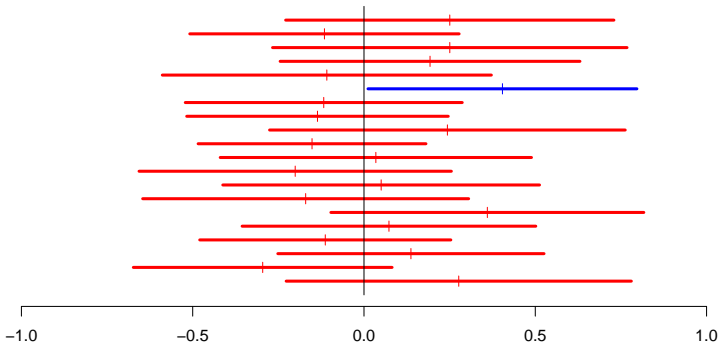
- depends on the sample and is therefore **random**,
- contains **plausible values** for a parameter (the true, unknown effect), those values of the parameter that are compatible with the data.
- **covers** the true parameter with probability  $1 - \alpha \Rightarrow$  i.e. with probability 95% if  $\alpha = 0.05$ .
- is **wider** the larger the uncertainty with regard to the parameter,
- is **smaller** the larger the number of observations  $n$ ,
- is **smaller** the smaller the standard error.

# Confidence interval and coverage probability

No guarantee that  $\mu$  is covered, only high probability!

20 samples of size  $1e + 06 = 20$  from a  $\text{Normal}(0, 1)$ -distribution

$\Rightarrow$  19 intervals cover true  $\mu = 0$ .



# Statistical test

Goal: assess **scientific hypothesis**.

Example: Follow-up cholesterol level different in normal group  $\pm$  fish?

Group	<i>n</i>	Mean	Standard error	95%-confidence interval
No fish	95	223.5	3.3	[217.1, 229.9]
Fish	98	227.4	3.5	[220.6, 234.3]
<b>Difference</b>		<b>4.0</b>	<b>4.8</b>	<b>[-5.5, 13.4]</b>

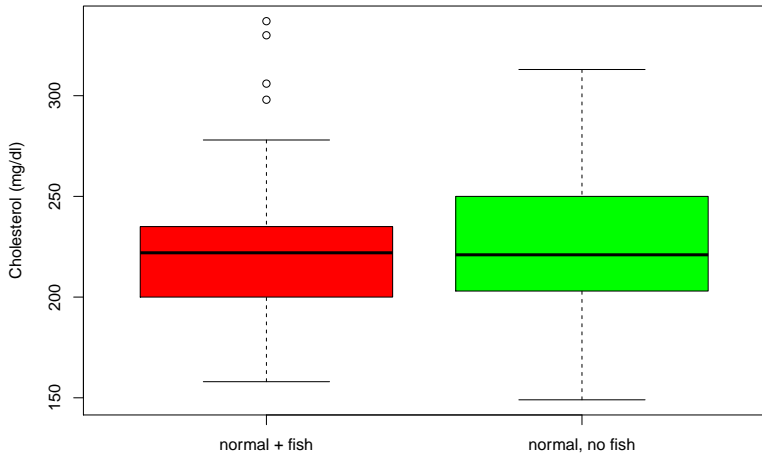
Interpretation:

- 0 mg/dl is covered by the confidence interval for the difference in follow-up values.
- It is plausible that there is **no population difference**, that the observed difference is **simply random**.

Statistical test: We assess whether an observed effect can be **considered random** or whether it is larger than can be expected if there were no underlying population difference.

# Statistical test

Scientific question: Follow-up cholesterol level different in normal group  $\pm$  fish?



# Statistical test

Goal: collect data (=draw a sample), assess scientific hypothesis.

Steps in performing a statistical test:

- Define hypotheses:
  - ▶ **Null hypothesis  $H_0$** : Statement that one seeks to reject.
  - ▶ **Alternative hypothesis  $H_1$** : Scientific hypothesis, what you are interested in.
- Fix significance level  $\alpha$  (see later).
- Compute **test statistic** from data  $\Rightarrow$  quantifies “distance” between estimate and hypothetical value (e.g. mean).

$$t = \frac{\text{estimate} - \text{null value}}{\text{standard error}}.$$

- $|t|$ : How many standard errors is estimate from null value?
- Distribution of test statistic under  $H_0$  must be known.

Samples are **random**  $\Rightarrow$  impossible to **prove**  $H_0$  or  $H_1$ !

## Statistical test - z-test

z-test: Compare means if  **$n$  large enough**.

Scientific question: Follow-up cholesterol level different in normal group  $\pm$  fish?

**Null hypothesis  $H_0$** : what we want to reject:

$$H_0 : \mu_{\text{no fish}} = \mu_{\text{with fish}} \Leftrightarrow \delta = \mu_{\text{no fish}} - \mu_{\text{with fish}} = 0.$$

**Alternative hypothesis  $H_1$** : Scientific question, what you are interested in:

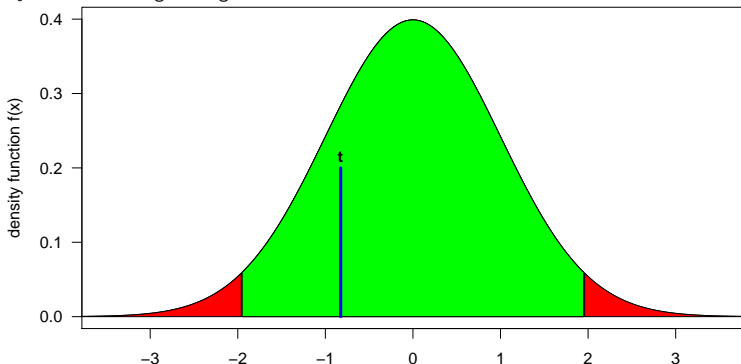
$$H_1 : \delta \neq 0.$$

Compute test statistic:

$$\begin{aligned} t &= \frac{(223.5 - 227.4) - 0}{4.8} \\ &= -0.8. \end{aligned}$$

## Statistical test - z-test

If  $H_0$  holds and  $n$  large enough  $\Rightarrow t$  **standard normal**.



If  $H_0$  holds, i.e. **no effect in population**, then  $t$  is expected in the green area (=95% of area, corresponds to  $1 - \alpha$ ).

$t$  in red area  $\Rightarrow H_0$  is rejected.

## Statistical test - z-test

Limits of green area: -1.96 and 1.96 – **critical values** for z-test.

Here:  $t = -0.8$  inside  $[-1.96, 1.96]$ , so in green area.

- There is no support from data to reject  $H_0$ .
- Test result is **not significant**.
- We **ARE NOT ALLOWED TO SAY THAT  $H_0$  IS TRUE!**
- Two scenarios:
  - ▶  $H_0$  is indeed true, e.g. there is no population effect.
  - ▶ There is an effect, but it is too small to be detected with given  $n$ .

## Errors in a statistical test

	Truth (population)	
	$H_0$ true	$H_1$ true
test not significant	correct decision probability $1 - \alpha$	<b>wrong</b> decision probability $\beta$
test significant	<b>wrong</b> decision probability $\alpha$	correct decision probability $1 - \beta$

We never know the truth, i.e.

- whether and
- which error we possibly make  $\Rightarrow$  uncertainty.

**Significance level  $\alpha$ :**

- Probability to reject  $H_0$  **although**  $H_0$  is true.
- Typically:  $\alpha = 0.05$ . Determines red (and so green) area in plot.

Error probabilities can **not be both bounded** at the same time  $\Rightarrow$  bound  $\alpha$ .

Idea: We want to know how often an ineffective therapy is falsely declared effective.

# Errors in a statistical test

**Error of the first kind**,  $\alpha$ -error, type I error:

- Reject  $H_0$  (based on test) although  $H_0$  is true (in population).
- Probability of  $\alpha$ -error:  $\alpha$ .
- **Want to avoid this error by all means!** Not declare ineffective treatment as effective.
- Determine  $\alpha$  **in advance**, do not change it.
- Does not depend on  $n$ .

**Error of the second kind**,  $\beta$ -error, type II error:

- Do not reject  $H_0$  although  $H_1$  is true (formulation!).
- Probability of  $\beta$ -error:  $\beta$ .
- Power =  $1 - \beta$ : "Ability" to detect an underlying effect.
- The larger  $\alpha$  and/or  $n$ , the smaller  $\beta$ .
- $\beta$  is only known if the true effect known or specified.

# Statistical tests and confidence intervals

A two-sided statistical test at  $\alpha = 5\%$  corresponds to a 95%-confidence interval in the following sense:

- All values **in the confidence interval** would **not be rejected** with the test if considered the null hypothesis.
- All values **not in the confidence interval** would be rejected with the test if considered the null hypothesis.

**Duality** of confidence interval and statistical test.

Similar for  $\alpha = 1\%$  and 99%-confidence interval etc.

Example: all null hypotheses

$$H_0 : \delta = \mu_{\text{with fish}} - \mu_{\text{no fish}}$$

with  $\delta$  in  $[-5.5\text{mg/dl}, 13.4\text{mg/dl}]$  would **not be rejected**. Specifically  $\delta = 0$ .

## The infamous “error bars”

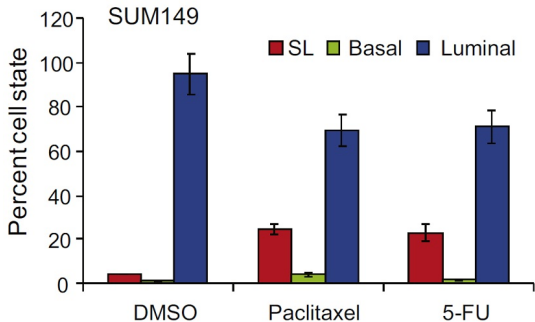


Figure 4. Analysis of Chemical Treatment Sensitivities by Monte Carlo Simulation with the Stochastic Cell-State Transition Model (A) Cell-state proportions after a 6 day treatment with either paclitaxel or 5-FU are shown for SUM159 and SUM149 populations.

Taken from [Gupta et al. \(2011\)](#). No indication of what the bars are (only in other plot of the paper)!

- What do we learn if the bars were the estimated standard deviation  $s$ ?
- What do we learn if the bars were the standard error of the mean  $s/\sqrt{n}$ ?
- What would really be **informative**?

## Significance vs. relevance

**Statistical significance** and **biological relevance** are two different aspects and **must** both be discussed!

Change in cholesterol level baseline - follow-up:

Scenario	Change (mg/dl)	Significant?	Clinical relevant?	Action
A	16	yes	yes	Bingo!
B	0.3	yes	no	Evidence of an irrelevant difference.
C	16	no	yes	Further investigations necessary.
D	0.3	no	no	No evidence of difference.

# Choice of statistical tests

z-Test: **Difference** of **means** of a **continuous** variable of **two large independent** samples.

Further relevant scenarios:

- **Unpaired, independent**: Samples independent. Normal vs. normal + fish.
- **Paired**: Two measurements per unit (e.g. patient). Consider differences. Follow-up level in normal group different from baseline measurement?
- **Small samples**, data **normal**, variances similar: t-test.
- **Large samples** ( $\approx n \geq 100$ ): z-test.
- Tests for other types of data:
  - ▶ **Nominal** variable in  $\geq 2$  samples:  $\chi^2$ - or Fisher's exact test.
  - ▶ **Ordinal** or **continuous** variable in two samples: Wilcoxon-test.

# Nonparametric tests

t-test: normal data, comparable variance, small samples.

z-test: large samples.

What if **non-normal** data and **small** samples? Problem: Distribution of ( $t$ -)test statistic under  $H_0$  not clear.

**Nonparametric** procedures:

- Based on **ranking** the data.
- Easy to perform.
- Decent statistical properties: “Loss of efficiency” through ranking is small.
- Parametric assumptions (normality of data) are often
  - ▶ not sufficiently checked,
  - ▶ ignored if not fulfilled.

Sign test, Wilcoxon test, Friedman test.

# The $p$ -value

Test: decision on  $H_0$  or  $H_1$ .

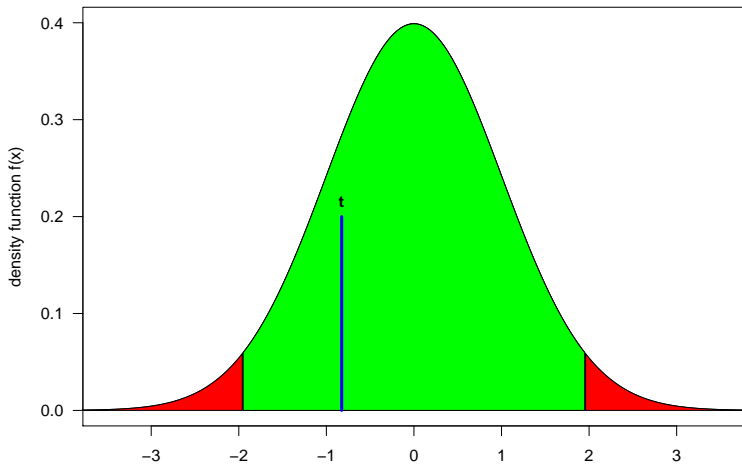
$p$ -value:

- Quantifies **evidence against  $H_0$** .
- Probability to get  $t$  or even **more extreme** values of  $t$ , under  $H_0$ .

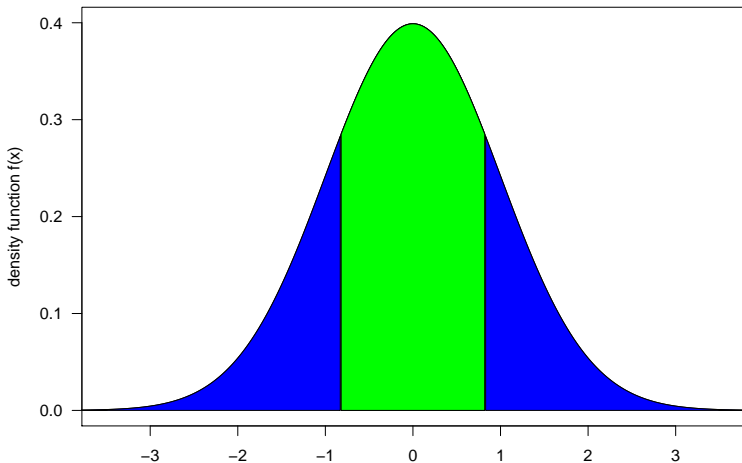
Initial idea (Fisher):  $p$ -value as informal index as a **measure of evidence** against  $H_0$ .

*[Goodman, 1999] Fisher suggested that the  $p$ -value be used as part of the fluid, non-quantifiable process of drawing conclusions from observations, a process that included combining the  $p$ -value in some unspecified way with background information.*

## The $p$ -value



## The $p$ -value



Blue area:  $p = 0.41 \Rightarrow$  not much evidence against  $H_0$ .

# The $p$ -value

Statistical test and  $p$ -value are two entirely **different concepts!**

Connection:

- Both assess a null hypothesis.
- $p \leq \alpha \Rightarrow$  reject  $H_0$ , test is significant.
- $p > \alpha \Rightarrow$  do not reject  $H_0$ , test is not significant.
- $p$ -value can be used to decide on hypotheses.
- Used in this context, the **actual value of the  $p$ -value should not be interpreted.**

Confidence interval: provides estimate of effect size and uncertainty of its estimation.

$p$ -value: no such information.

Recommendation: always report confidence interval, may be supplemented by  $p$ -value.

## Misinterpretations of $p$ -value and $\alpha$

### $p$ -value:

- $p$ -value is **not** probability that  $H_0$  is true!
- Ability to detect a present effect depends on sample size  $\Rightarrow$  also report non-significant results, at best using confidence intervals!
- Do **not overpronounce** significant results!
- **Significance** vs. **relevance**!

Suppose  $\alpha = 0.05$ , 20 tests, no effect in all 20 questions  $\Rightarrow$  on average, one test is **falsely** significant.

$\alpha = 0.05$  **does not** mean that every 20th significant result is false!

*[Goodman, 2005] In fact, the  $p$ -value is **almost nothing sensible** you can think of.  
I tell students to give up trying.*

# Multiple testing

Ideal situation: clinical trial addresses exactly **one** research question

⇒ one statistical test for **primary endpoint**.

Then:  $p$ -value equal to probability of getting the observed or more extreme data under  $H_0$ .

Routine: more than one (many) tests on same dataset.

Example: perform 200 tests, **no effect** present,  $\alpha = 0.05$

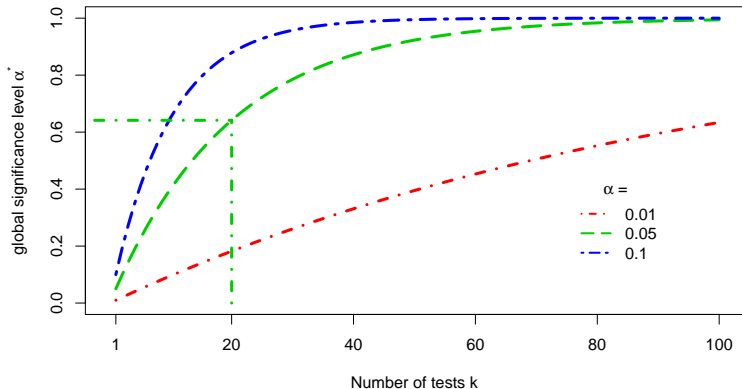
⇒ on average, 10 tests are falsely significant.

May be ok in exploratory, hypothesis generating studies.

Global significance level, familywise error rate: Probability to falsely reject **at least one** of  $H_{0,1}, \dots, H_{0,k}$  although all are valid (no effect).

## Global $\alpha$

Test  $k$  independent hypothesis  $\Rightarrow$  global niveau  $\alpha^*$  is  $\alpha^* = 1 - (1 - \alpha)^k$ .



$\alpha = 0.05$ , 20 tests, no effect  $\Rightarrow$  probability for at least one false positive test is **0.64!**

## How can we keep $\alpha^*$ ?

Decrease  $\alpha$  in single tests.

**Bonferroni:**  $\alpha = \alpha^* / k \Rightarrow p$ -value needs “to be smaller” to be significant. Conservative, i.e. on average “**too many**”  $H_0$ 's can not be rejected.

Less conservative: Bonferroni-Holm.

**Confirmatory analysis:** need to keep global  $\alpha^*$   $\Rightarrow$  “hard” statements, e.g. compare therapies.

**Exploratory analysis:** generate new hypothesis, do not want to miss present effect  $\Rightarrow$  less restrictive on  $\alpha^*$ .

# Study design

First step: **quantify** scientific question!

**Primary endpoint:** quantity we compute sample size for  $\Rightarrow$  "hard" statistical conclusion only valid for primary endpoint!

"hard": maintain  $\alpha$ -error.

**Secondary endpoints:** other quantities of interest, not considered for sample size computation.

Question: What  $\alpha$  do we choose for secondary endpoints  $\Rightarrow$  pre-specify!

# Clinical studies

Assess efficacy and tolerability of new therapy.

Randomization: random allocation of patients to treatments  $\Rightarrow$  control **confounders**.

No randomization: therapy effect may be **biased**. Trial on hypertension: arm A heavier patients than arm B.

- **Therapy study, RCT**: new vs. standard drug, randomization, blinded.
- **Diagnostic study**: assess quality of diagnostic procedure (“healthy” vs. “diseased”).
- **Prognostic study**: find prognostic factors for course of disease.
- **Meta-analysis**: summarize results (effect sizes!) of many studies.

# Epidemiological studies

Distribution of **diseases** and **risk factors** in population.

No randomization possible!

- **Cohort** study: follow-up specific group of persons.
- **Case-control** study: compare **cases** (diseased) to **controls** (healthy). regarding exposition to risk factors. Matching.
- **Cross-sectional** study: assess exposition, disease in population at a given time point.

# Computation of sample size

Why compute sample size?

Treat too few patients: **unethical** since

- can probably not verify scientific hypothesis (not enough power).
- treat patients with possibly ineffective therapy.
- waste of resources.

Treat too many patients: **unethical** since

- more patients than necessary get ineffective therapy.
- delay use of effective therapy.

No sample size computation: properties of statistical test / CI not clear (power).

# Computation of sample size: hypothesis test

Example: Continuous data in two independent groups  $\Rightarrow$  z-test.

Cholesterol change baseline – follow-up for normal vs. fish diet.

Applied researcher – **not statistician!** – needs to specify:

- strength of evidence, i.e. upper bound on prob. type I error,  $\alpha$ ,
- $1 - \beta$ , prob. to detect assumed mean difference, if it is true,
- smallest difference  $\delta$  in Cholesterol change to be detected between groups, for given  $\alpha$  and  $\beta$ ,
- standard deviation  $\sigma$ , of individual measurements of Cholesterol (equal in both groups).

Get  $\delta$  and  $\sigma$  from **literature** or **pilot study**.

# Computation of sample size: hypothesis test

We set

- standard choices (arbitrary):  $\alpha = 0.05$ ,  $\beta = 0.1$ ,
- clinically relevant mean difference:  $\delta = 15$ ,
- standard deviation of Cholesterol measurements (both groups):  $\sigma = 35$ .

We need to include

$$\begin{aligned}n &\geq \frac{2(q_{1-\beta} + q_{1-\alpha/2})^2 \sigma^2}{\delta^2} \\ &= 64.4,\end{aligned}$$

so 65 patients in each group.

# Computation of sample size: confidence interval

Test: provides no effect size  $\Rightarrow$  **confidence interval** for difference of Cholesterol differences.

Applied researcher (not statistician!) needs to specify:

- strength of evidence, i.e. lower bound on confidence level  $1 - \alpha$ ,
- standard deviation of individual measurements of Cholesterol (equal in both groups),
- maximal desired length of confidence interval  $d$ .

We set  $1 - \alpha = 0.95$ ,  $\sigma = 35$ ,  $d = 24$ . Then,

$$\begin{aligned}n &\geq 8\sigma^2 q_{1-\alpha/2}^2 / d^2 \\ &= 65.4.\end{aligned}$$

So we need to include 66 patients in each group

**Thank you for your attention.**

Division of Biostatistics  
Institute of Social- and Preventive Medicine  
University of Zurich  
Hirschengraben 84  
8001 Zurich  
`kaspar.rufibach@ifspm.uzh.ch`

`http://www.biostat.uzh.ch`

download lecture notes: [Held et al. \(2011\)](#).

statistical consulting service

# References

- ▶ Gupta, P., Fillmore, C., Jiang, G., Shapira, S., Tao, K., Kuperwasser, C. and Lander, E. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146** 633 – 644.  
<http://www.sciencedirect.com/science/article/pii/S0092867411008245>
- ▶ Held, L., Rufibach, K. and Seifert, B. (2011). *Einführung in die Biostatistik*. 7th ed. Druckereizentrum Universität Zürich.  
<http://www.biostat.uzh.ch/teaching/lecturenotes/scripts.html>