

Basic Statistical Concepts

José Sánchez 2020-01-24

Contents

- Bias and variability
- Confounding and interaction
- Descriptive and inferential statistics
- Hypothesis testing and p-values

Bias and variability

$$E[\hat{\mu}] - \mu$$

Bias: Systemtic deviation from the true value

Design, Conduct, Analysis, Evaluation

Bias and variability

Larger study does not decrease bias

$$\hat{\mu}_n \rightarrow \mu + \phi ; n \rightarrow \infty$$

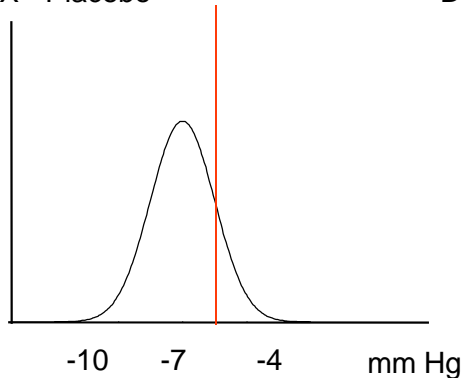
μ Population mean

ϕ bias

Distribution of sample means:

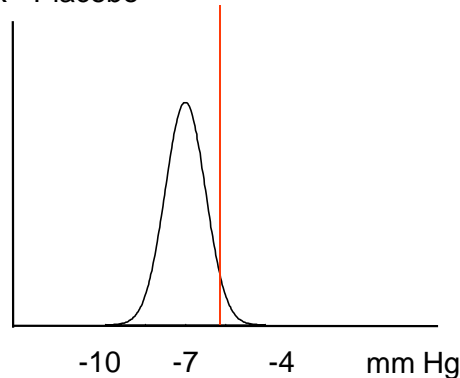
— = population mean

Drug X - Placebo



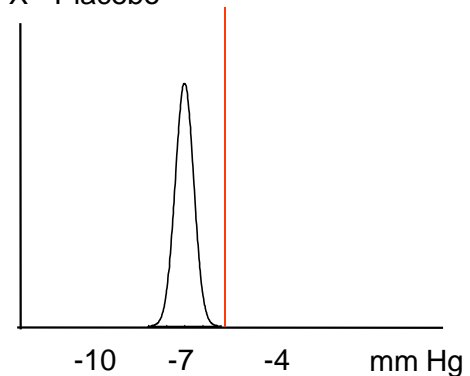
$n=40$

Drug X - Placebo



$n=200$

Drug X - Placebo



$N=2000$

Bias and variability

There is a multitude of sources for bias

Publication bias	Positive results tend to be published while negative or inconclusive results tend to not be published
Selection bias	The outcome is correlated with the exposure. As an example, treatments tend to be prescribed to those thought to benefit from them. Can be controlled by randomization
Exposure bias	Differences in exposure e.g. compliance to treatment could be associated with the outcome, e.g. patients with side effects stop taking their treatment
Detection bias	The outcome is observed with different intensity depending on the exposure. Can be controlled by blinding investigators and patients
Analysis bias	Essentially the I error, but also bias caused by model misspecifications and choice of estimation technique
Interpretation bias	Strong preconceived views can influence how analysis results are interpreted.

Bias and variability

Amount of difference between observations

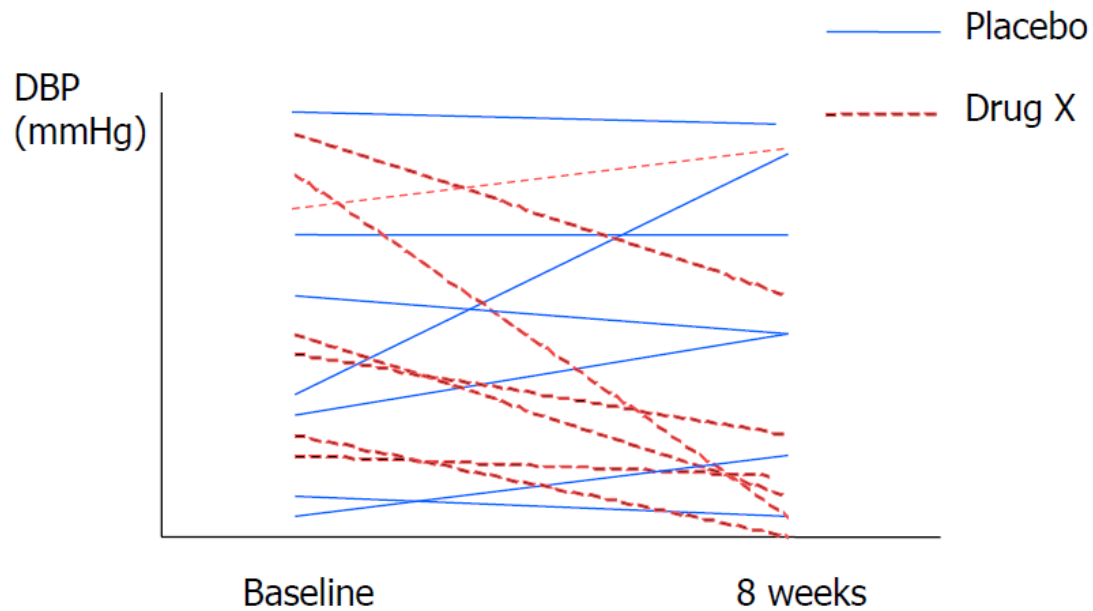
True biological: Variation between subject due to biological factors (covariates) including the treatment.

Temporal: Variation over time (and space)
Often within subjects.

Measurement error: Related to instruments or observers

Design, Conduct, Analysis, Evaluation

Raw Blood pressure data



Subset of plotted data

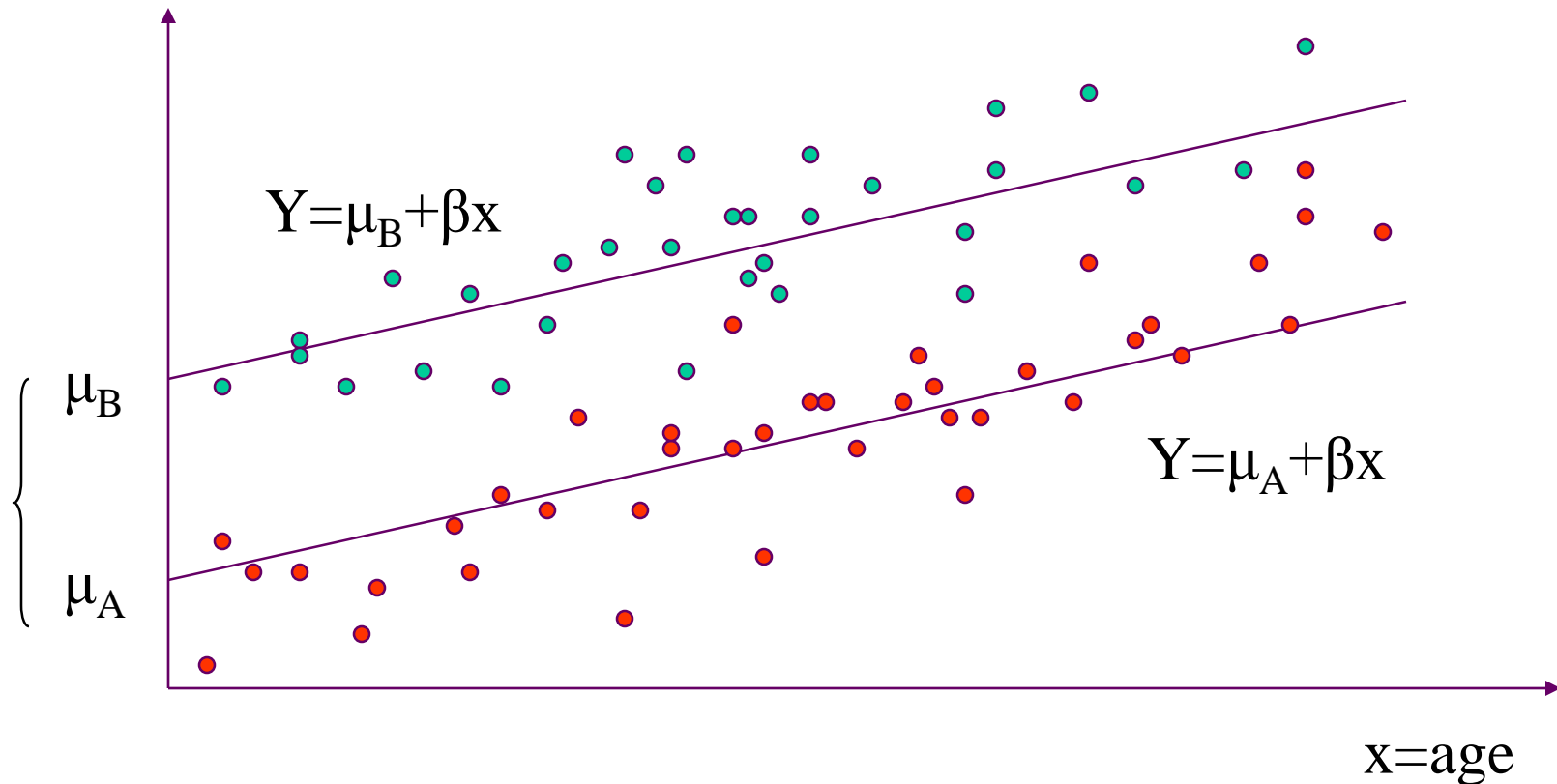
Bias and variability

$$Y = X\beta + \varepsilon$$

Variation in observations = Explained variation + Unexplained variation

Bias and variability

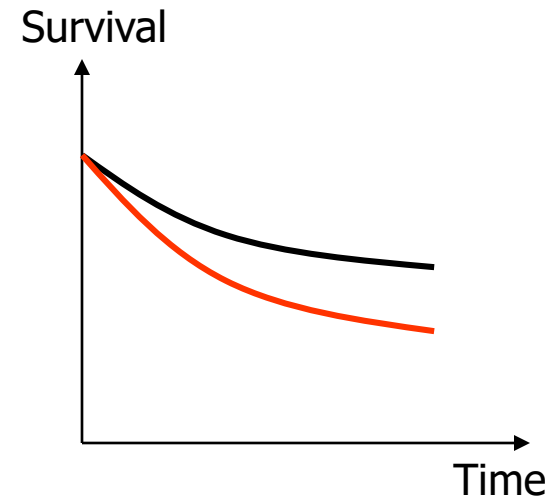
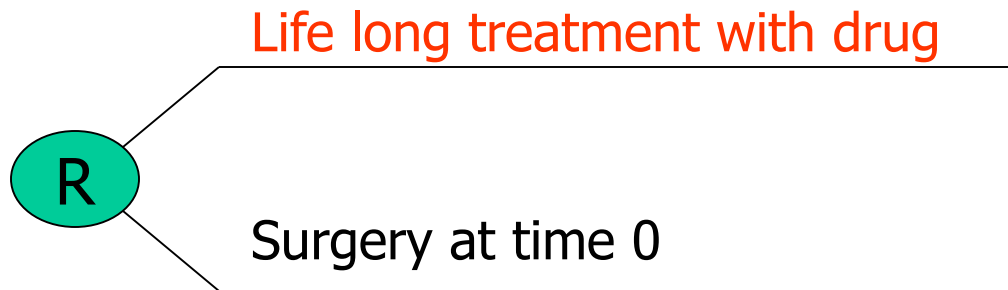
$$\text{Model: } Y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij}$$



Confounding

The effect of two or more factors can not be separated

Example: Compare survival for surgery and drug



Looks ok but:

- Surgery only if healthy enough
- Patients in the surgery arm may take drug
- Compliance in the drug arm may be poor

Example

Smoking Cigaretts is not so bad but watch out for Cigars or Pipes (at least in Canada)

Variable	Non smokers	Cigarette smokers	Cigar or pipe smokers
Mortality rate*	20.2	20.5	35.5

*) per 1000 person-years %

Cochran, Biometrics 1968

Example

Smoking Cigaretts is not so bad but watch out for Cigars or Pipes (at least in Canada)

Variable	Non smokers	Cigarette smokers	Cigar or pipe smokers
Mortality rate*	20.2	20.5	35.5
Average age	54.9	50.5	65.9

*) per 1000 person-years %

Cochran, Biometrics 1968

Example

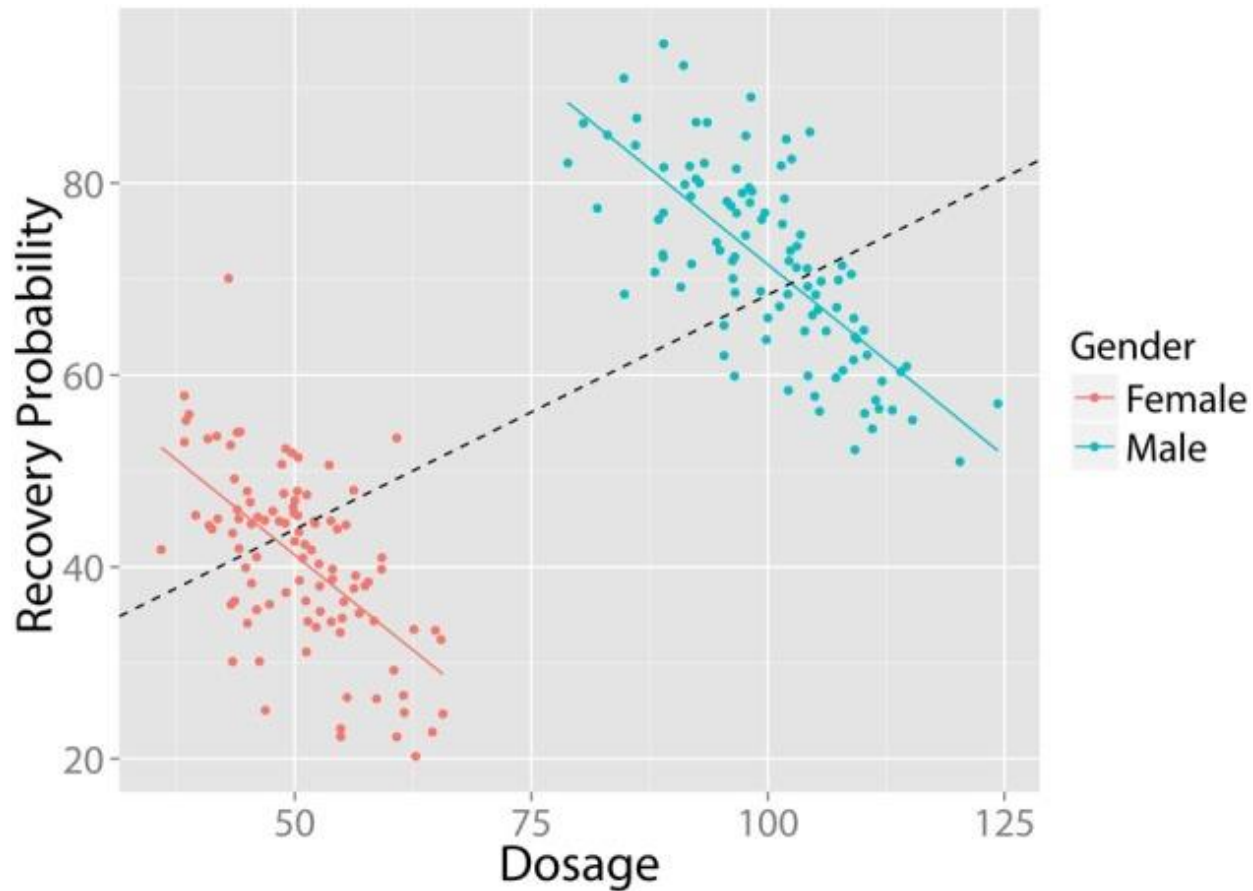
Smoking Cigaretts is not so bad but watch out for Cigars or Pipes (at least in Canada)

Variable	Non smokers	Cigarette smokers	Cigar or pipe smokers
Mortality rate*	20.2	20.5	35.5
Average age	54.9	50.5	65.9
Adjusted mortality rate*	20.2	26.4	24.0

*) per 1000 person-years %

Cochran, Biometrics 1968

Simpson's paradox



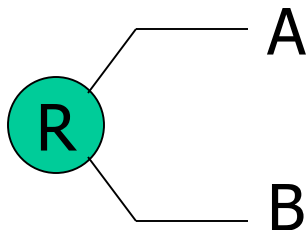
Confounding

Can be sometimes be handled in the design

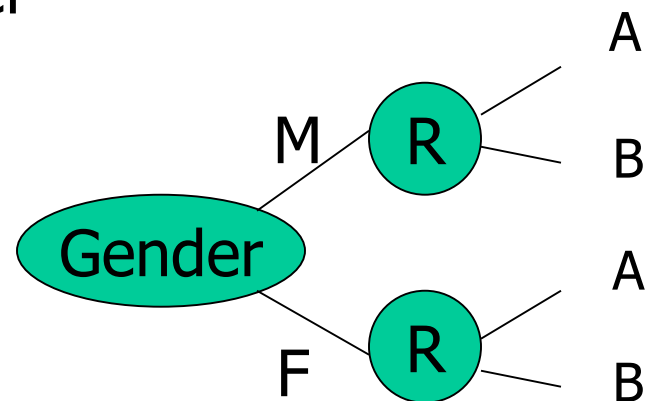
Example: Different effects in males and females

Imbalance between genders affects result

Stratify by gender



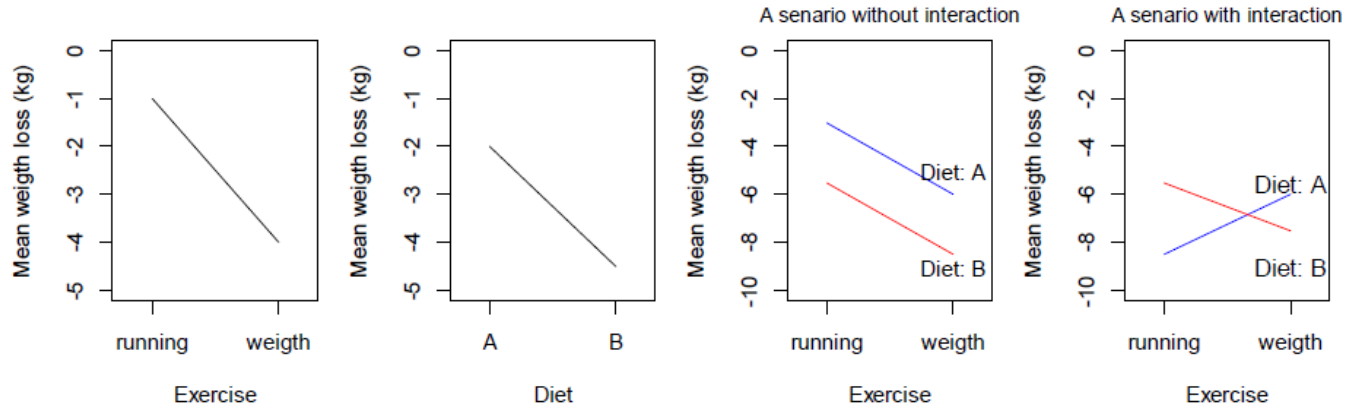
Balance on average



Always balanced

Interaction

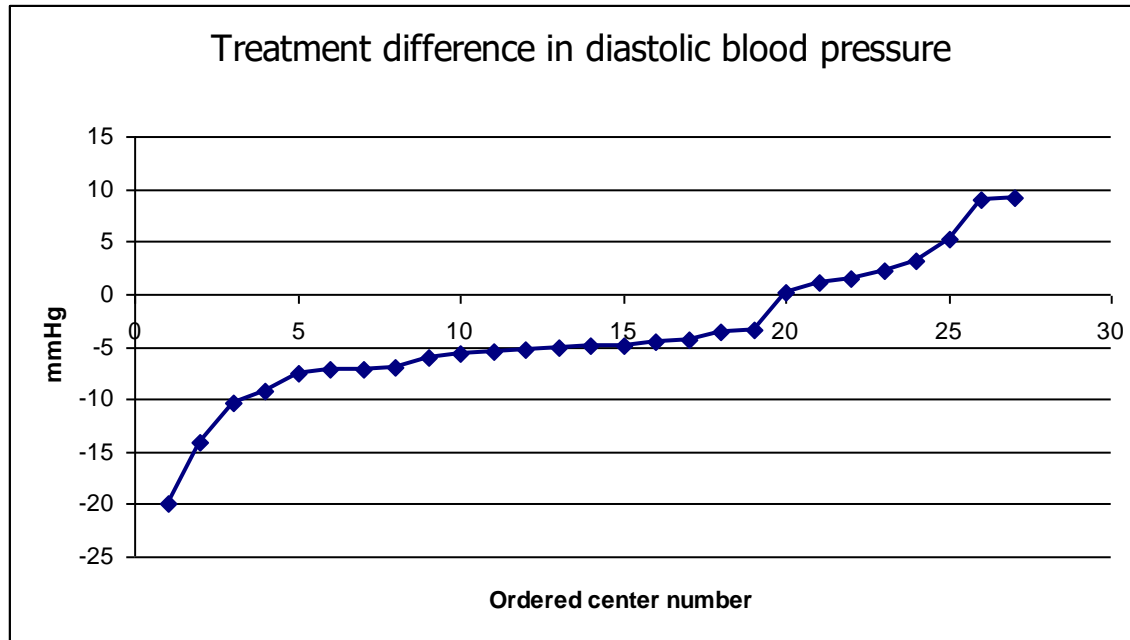
Exercise and Diet



The outcome on one variable depends on the value of another variable.

Interaction

Example: Treatment by center interaction



Average treatment effect: -4.39 [-6.3, -2.4] mmHg

Treatment by center: $p=0.01$

What can be said about the treatment effect?

Descriptive and inferential statistics

The presentation of the results from a clinical trial can be split in three categories:

- Descriptive statistics
- Inferential statistics
- Explorative statistics

Descriptive and inferential statistics

Descriptive statistics aims to describe various aspects of the data obtained in the study.

- Listings.
- Summary statistics (Mean, Standard Deviation...).
- Graphics.

Descriptive and inferential statistics

Inferential statistics forms a basis for a conclusion regarding a prespecified objective addressing the underlying population.

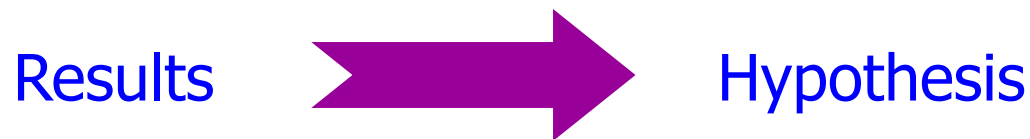
Confirmatory analysis:



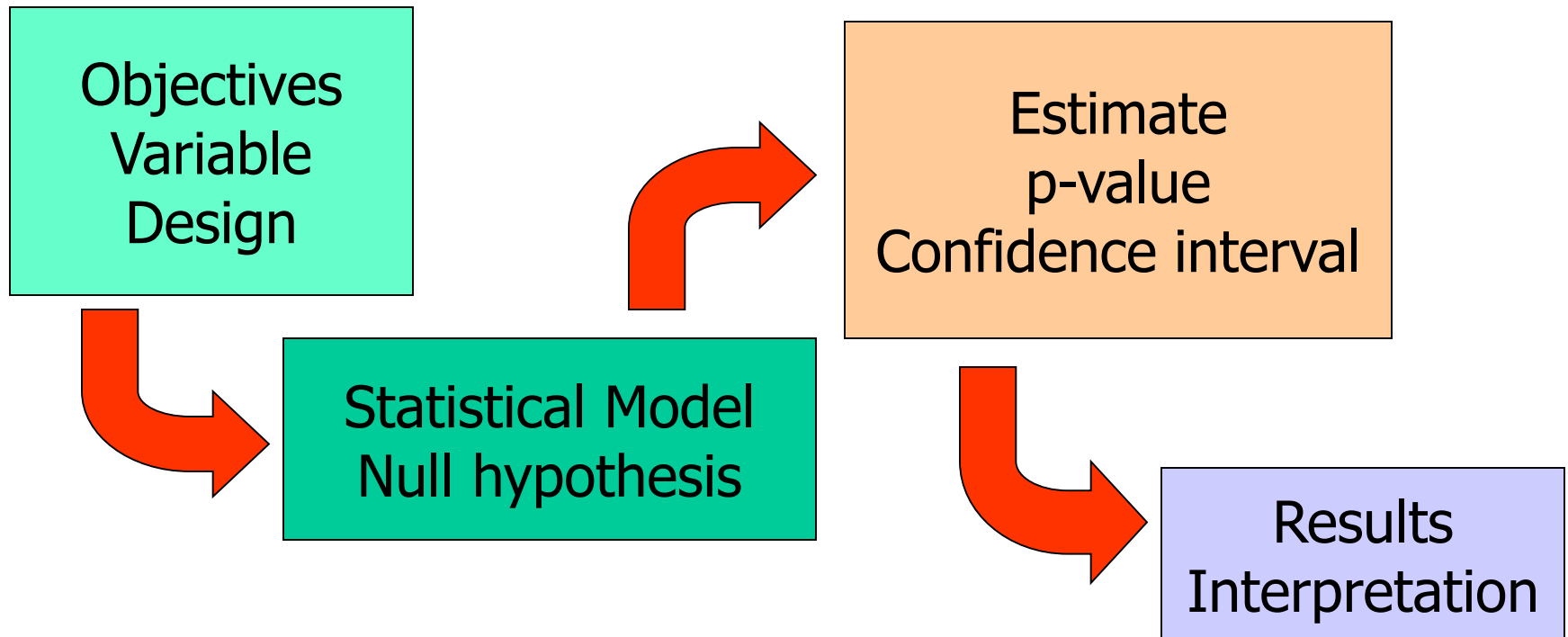
Descriptive and inferential statistics

Explorative statistics aims to find interesting results that can be used to formulate new objectives/hypothesis for further investigation in future studies.

Explorative analysis:



Hypothesis testing, p-values and confidence intervals



What is statistical hypothesis testing?

A statistical hypothesis test is a method of making decisions using data from a scientific study.

Decisions:

Conclude that substance is likely to be efficacious in a target population.

Scientific study:

Collect data measuring efficacy of a substance on a random sample of patients representative of the target population

Formalizing hypothesis testing

As null hypothesis, H_0 , we shall choose the hypothesis we want to reject

$$H_0 : \mu_X = \mu_{\text{placebo}}$$

As alternative hypothesis, H_1 , we shall choose the hypothesis we want to prove.

$$H_1 : \mu_X \neq \mu_{\text{placebo}}$$

(two sided alternative hypothesis)

μ is a parameter that characterizes the efficacy in the target population, for example the mean.


Hypothesis testing, p-values

Statistical model: Observations $\mathbf{X} = (X_1, \dots, X_n) \in R^n$
from a class of distribution functions

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

Hypothesis test: Set up a null hypothesis: $H_0: \theta \in \Theta_0$
and an alternative $H_1: \theta \in \Theta_1$

Reject H_0 if $\mathbf{X} \in S_c \subseteq R^n$ $P(\mathbf{X} \in S_c \mid \theta \in \Theta_0) < \alpha$



Rejection region Significance level

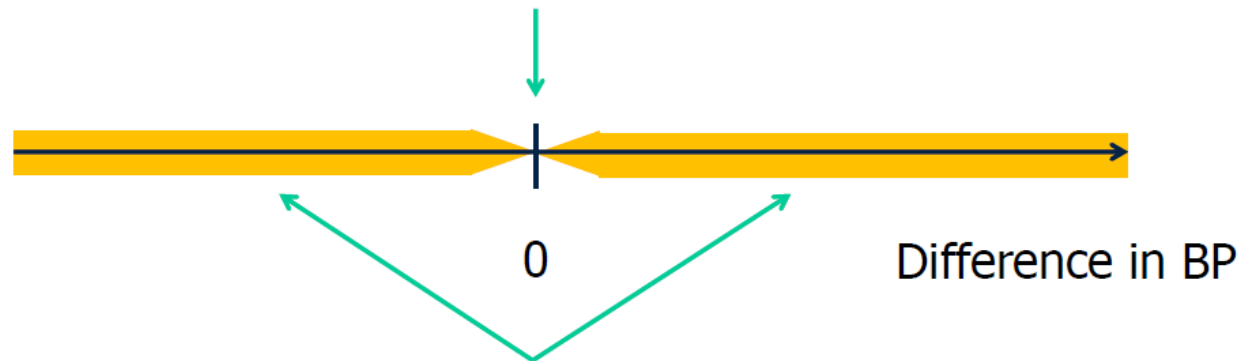
p-value: The smallest significance level for which the null hypothesis can be rejected.

What can go wrong?

	H_0 true	H_0 false
Fail to reject	No error ($1-\alpha$)	Type II error (β)
Reject	Type I error (α)	No error ($1-\beta$)

Example, two sided H_1 :

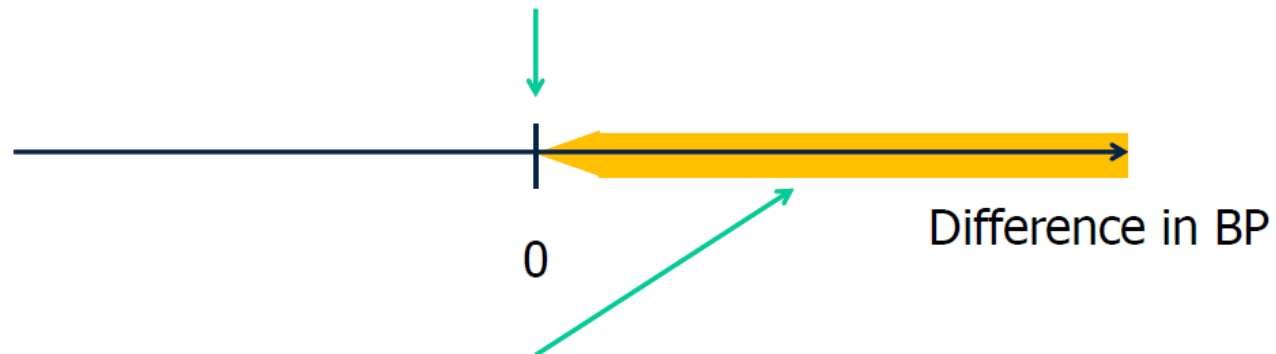
H_0 : Treatment with substance X has no effect on blood pressure compared to placebo



H_1 : Treatment with substance X influence blood pressure compared to placebo

Example: one sided H_1

H_0 : Treatment with substance X has no effect on blood pressure compared to placebo



H_1 : Treatment with substance X decrease the blood pressure more than placebo

Structure of a hypothesis test

Given our H_0 and H_1

Assume that H_0 is true

H_0 : *Treatment with substance X has no effect on blood pressure compared to placebo ($\mu_X = \mu_{placebo}$)*

- 1) We collect data from a scientific study.
- 2) Calculate the likelihood/probability/chance of data if H_0 was true.

P-value:

Probability to get at least as extreme as what we observed given H_0

How to interpret the p-value?

P-value small:

Correct: our data is unlikely given the H_0 .

We don't believe in unlikely things so something is wrong.

Reject H_0 .

P-value large:

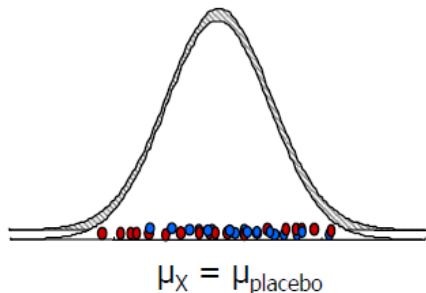
We can't rule out that H_0 is true; we can't reject H_0

A statistical test can **never** prove the null hypothesis!!!!

How to interpret the p-value? Cont.

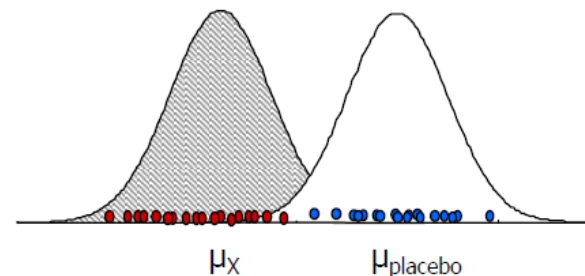
From the assumption that if the treatment does not have effect we calculate: How likely is it to get the data we collected in the study?

Expected outcome of the study if H_0 is true ($\mu_X = \mu_{\text{placebo}}$)



Likely outcome if the treatments are equal:
The p-value is big

Expected outcome of the study for some H_1 ($\mu_X < \mu_{\text{placebo}}$)

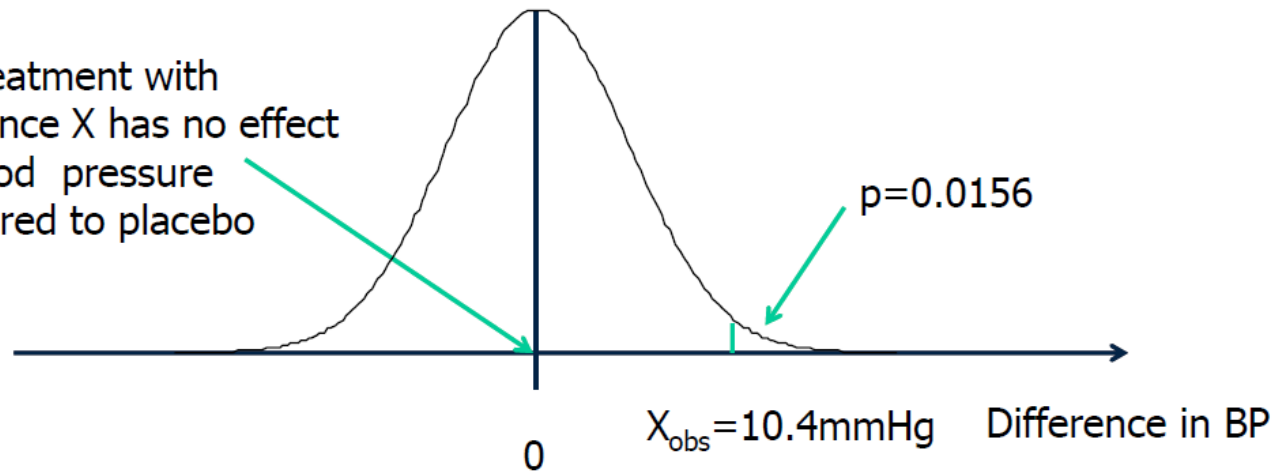


Unlikely outcome if the treatments are equal:
The p-value is small

Example:

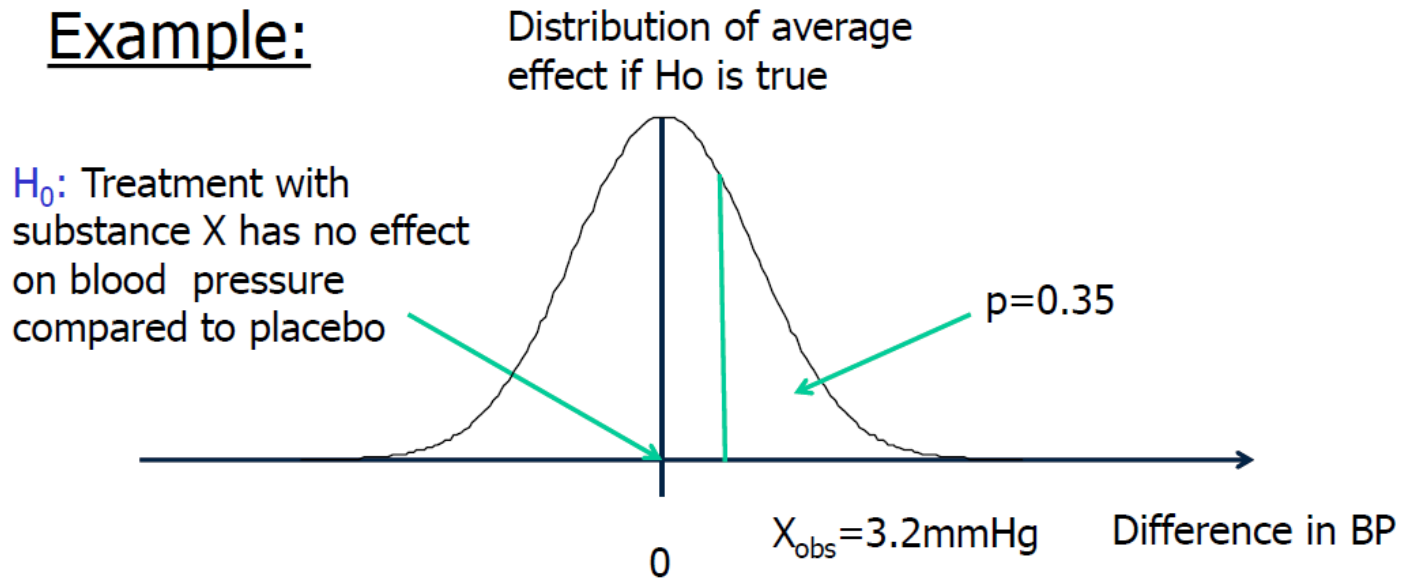
Distribution of average
effect if H_0 is true

H_0 : Treatment with
substance X has no effect
on blood pressure
compared to placebo



- If the null hypothesis is true it is unlikely to observe 10.4
- We don't believe in unlikely things so something is wrong
- The data is "never" wrong
- The null hypothesis must be wrong!

Example:



- If the null hypothesis is true it is not that unlikely to observe 3.2
- Nothing unlikely has happened
- There is no reason to reject the null hypothesis
- This does not prove that the null hypothesis is true!

However....

- A high p-value is not evidence that H_0 is true
- P-value is not the probability of the null hypothesis
- P-value says nothing about the *size of the effect*
- A statistical significant difference does not need to be *clinically relevant!*
- Easy to misinterpret and mix things up (p-value, size of test, significance level, etc)

Confidence intervals

Let $\delta(\mathbf{X}, \theta^*) = \begin{cases} 1 & \text{if } H_0 : \theta = \theta^* \text{ rejected} \\ 0 & \text{if } H_0 : \theta = \theta^* \text{ not rejected} \end{cases}$ (critical function)

Confidence set: $C(\mathbf{X}) = \{\theta : \delta(\mathbf{X}, \theta) = 0\}$

The set of parameter values corresponding to hypotheses that can not be rejected.

A confidence set is a random subset $C(\mathbf{X}) \subseteq \Theta$ covering the true parameter value with probability at least $1 - \alpha$.

Example:

- We study the effect of treatment X on change in DBP from baseline to 8 weeks

- The true treatment effect is:

$$d = \mu_X - \mu_{\text{placebo}}$$

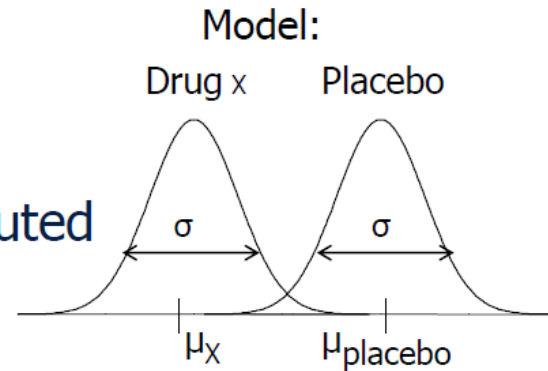
- The estimated treatment effect is:

$$\bar{x}_X - \bar{x}_{\text{placebo}} = -7.5 \text{ mmHg}$$

–How reliable is this estimation?

Example

- The 95% confidence interval is $[-9.5, -4.9]$ mmHg (based on a model assuming normally distributed data)
- With great likelihood (95%), the interval $[-9.5, -4.9]$ contains the true treatment effect $\mu_X - \mu_{\text{placebo}}$
- Provided that our model is correct...

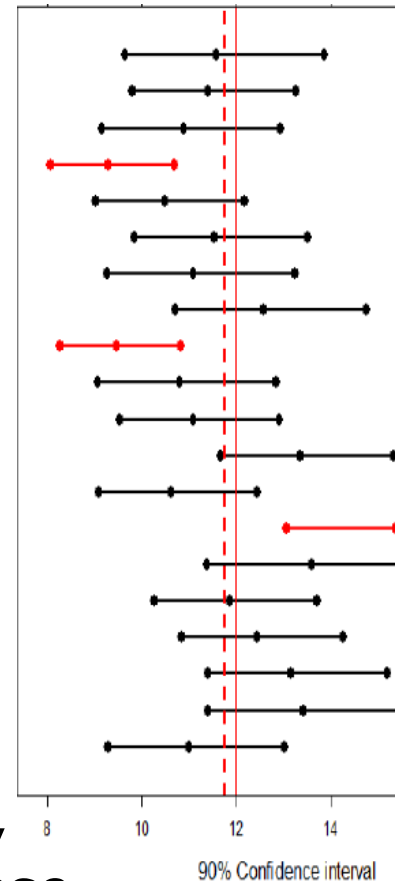


General principle

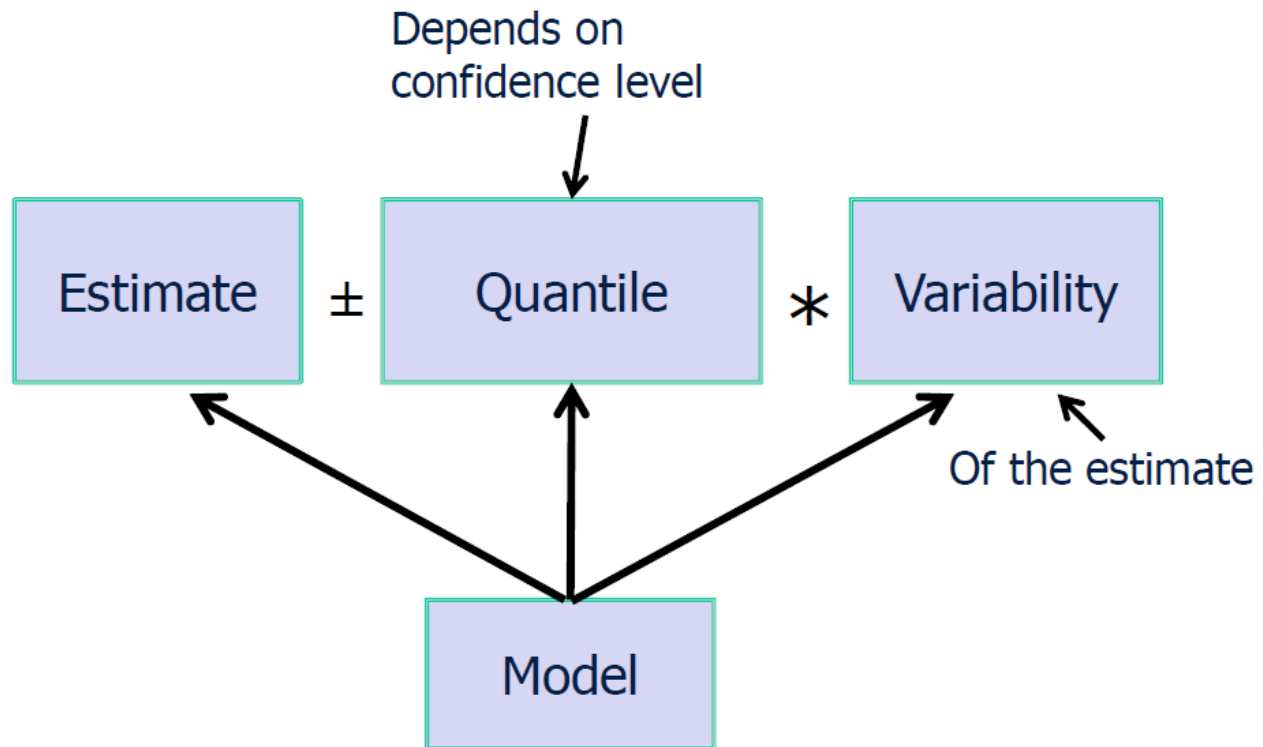
- There is a True fixed value
- We estimate this value using data from a clinical trial
- Estimate deviations from the true value (by an unknown amount)

If repeating the study a number of times, the different estimates would be centered around the true value

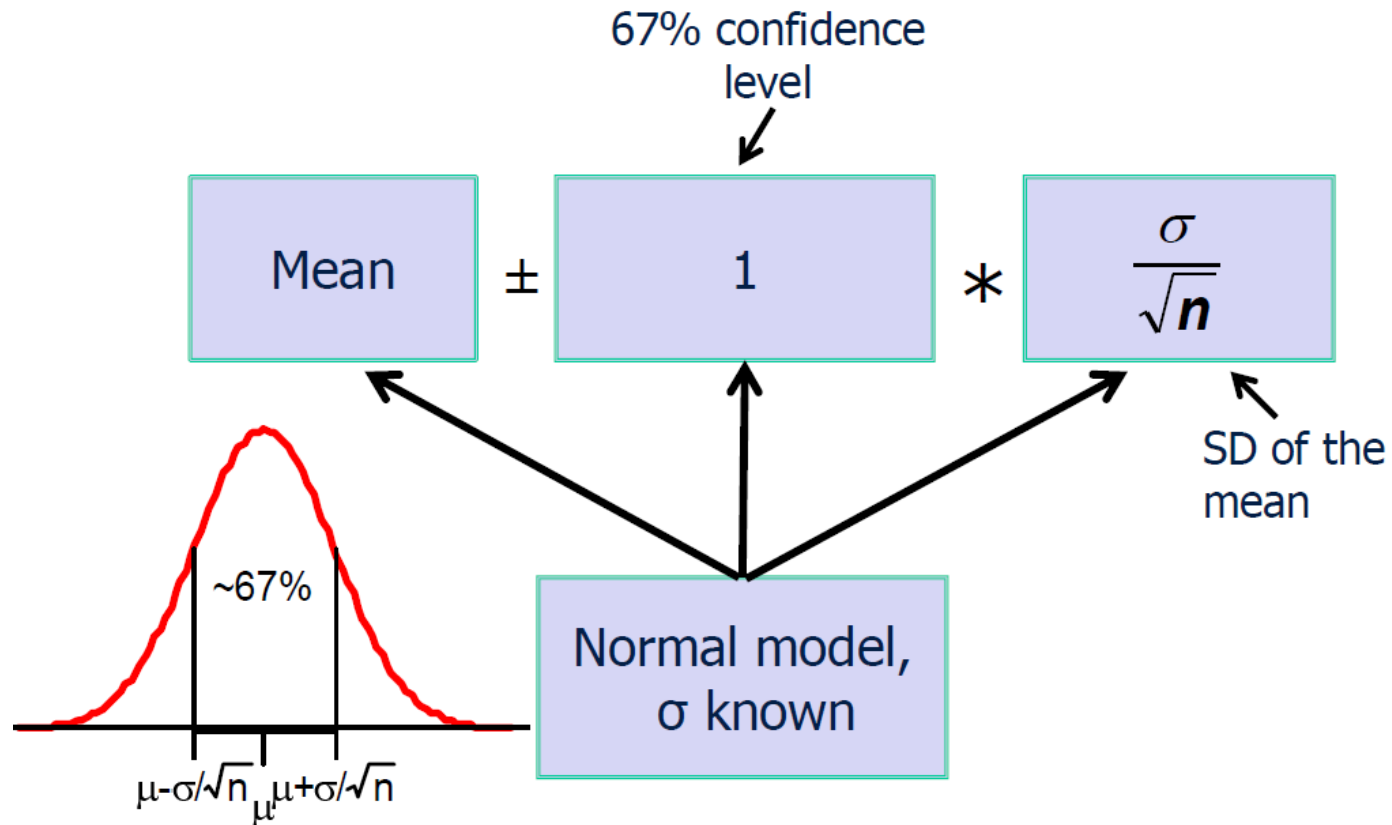
A 95% confidence interval is an interval estimated so that if the study is repeated a number of times, the interval will cover the true value in 95% of these



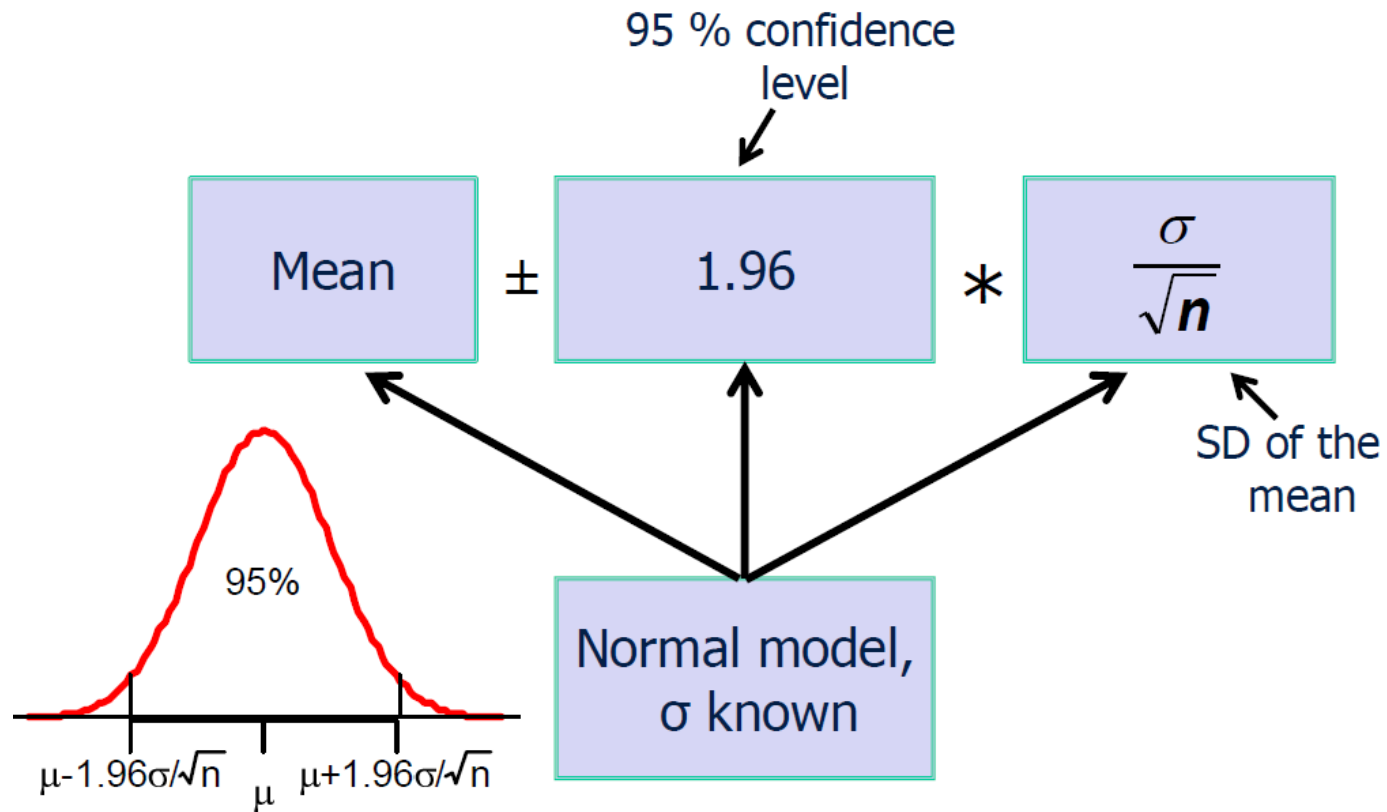
General principle



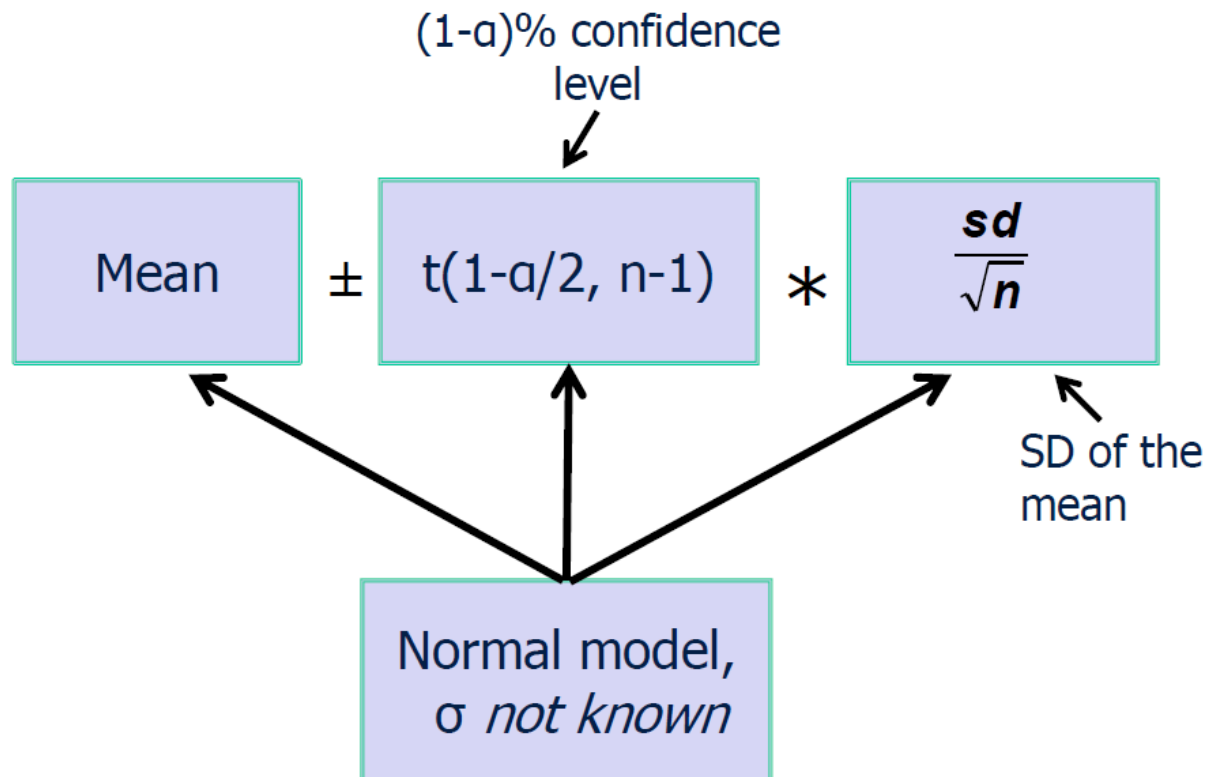
General principle – Normal model



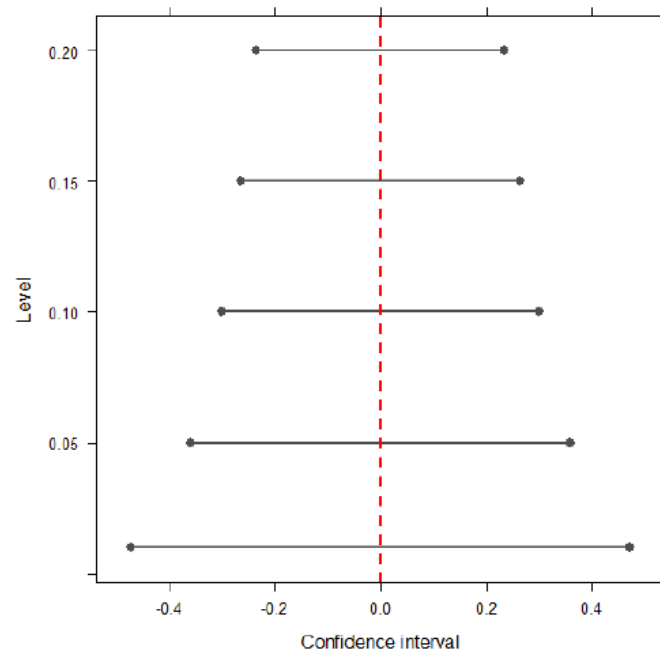
General principle – Normal model



General principle – Normal model



Effect of confidence level



95% → 90%	17% narrower
95% → 80%	36% narrower
95% → 99%	35% wider

(normal model, $n=30$)

“Excuse me, professor. Why a 95% confidence interval and not a 94% or 96% interval?”, asked the student.

“Shut up,” he explained.

Effect of Sample size and SD

Sample size

Double	30% narrower
+50%	20% narrower
+30%	12% narrower

Double SD → double width, etc

Analogy with tests

- For most statistical tests there is a corresponding confidence interval
- If a 95% confidence interval does not cover 'the zero', the corresponding test is to give a p-value < 0.05 .
- If for example your p-value is based on the median and the confidence interval is based on the mean you can get conflicting results = trouble

Example

Objective: To compare sitting diastolic blood pressure (DBP) lowering effect of hypersartan 16 mg, 8 mg and 4 mg

Variable: The change from baseline to end of study in sitting DBP (sitting SBP) will be described with an **ANCOVA** model, with treatment as a factor and baseline blood pressure as a covariate

Model:
$$Y_{ij} = \mu + \tau_i + \beta(x_{ij} - x_{..}) + \varepsilon_{ij}$$

treatment effect
 $i = 1, 2, 3$
 $\{16 \text{ mg}, 8 \text{ mg}, 4 \text{ mg}\}$

Parameter space: R^4
($\tau_1 + \tau_2 + \tau_3 = 0$)

Null hypotheses (subsets of R^4):

$H_{01}: \tau_1 = \tau_2$ (DBP)

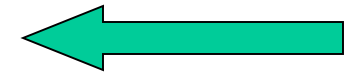
$H_{02}: \tau_1 = \tau_2$ (SBP)

$H_{03}: \tau_2 = \tau_3$ (DBP)

$H_{04}: \tau_2 = \tau_3$ (SBP)

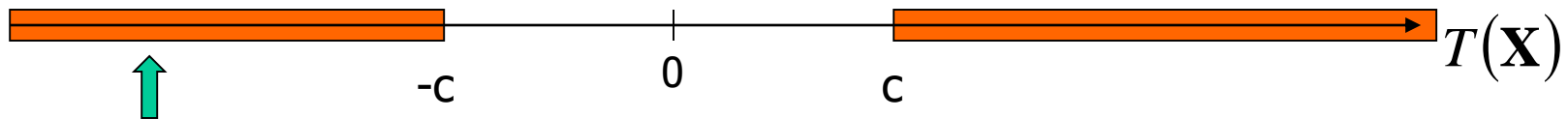
Example continued

Hypothesis	Variable	LS Mean	CI (95%)	p-value
1: 16 mg vs 8 mg	Sitting DBP	-3.7 mmHg	[-4.6, -2.8]	<0.001
2: 16 mg vs 8 mg	Sitting SBP	-7.6 mmHg	[-9.2, -6.1]	<0.001
3: 8 mg vs 4 mg	Sitting DBP	-0.9 mmHg	[-1.8, 0.0]	0.055
4 : 8 mg vs 4 mg	Sitting SBP	-2.1 mmHg	[-3.6, -0.6]	0.005



This is a t-test where the test statistic follows a t-distribution

Rejection region: $\{\mathbf{X} : |T(\mathbf{X})| > c\}$



P-value: The null hypothesis can be rejected at $\alpha < 0.001$



P-value says nothing about the size of the effect!

Example: Simulated data. The difference between treatment and placebo is 0.3 mmHg

No. of patients per group	Estimation of effect	p-value
10	1.94 mmHg	0.376
100	-0.65 mmHg	0.378
1000	0.33 mmHg	0.129
10000	0.28 mmHg	<0.0001
100000	0.30 mmHg	<0.0001

A statistical significant difference does **NOT** need to be clinically relevant!

Statistical and clinical significance

Statistical significance:	Is there any difference between the evaluated treatments?
Clinical significance:	Does this difference have any meaning for the patients?
Health economical relevance:	Is there any economical benefit for the society in using the new treatment?

Statistical and clinical significance

A study comparing gastroprazole 40 mg and mygloprazole 30 mg with respect to healing of erosived eosophagitis after 8 weeks treatment.

Drug	Healing rate
gastroprazole 40 mg	87.6%
mygloprazole 30 mg	84.2%

Cochran Mantel Haenszel p-value = 0.0007

Statistically significant!

Clinically significant?

Health economically relevant?

Chapter 2 Reading instructions

- 2.1 Introduction: Not very important
- 2.2 Uncertainty and probability: Read
- 2.3 Bias and variability: Read
- 2.4 Confounding and interaction: Read
- 2.5 Descriptive and inferential statistics: Repetition
- 2.6 Hypothesis testing and p-values: Repetition*
- 2.7 Clinical significance and clinical equivalence: Read
- 2.8 Reproducibility and generalizability: Read

*) Read extra material