Biophysics 2.
for dentistry students

Lecture 14th:

Biostatistics III.

2021. May 20.

Dániel Veres

Schedule

I. Repetition

(variability, variables, estimations, normal dist., hypothesis tests)

II. Multiplicity

III. Correlation and regression

VARIABILITY

„varietas delectat”

To describe and get to know:

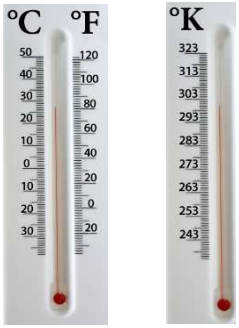
Different **Way of Thinking**

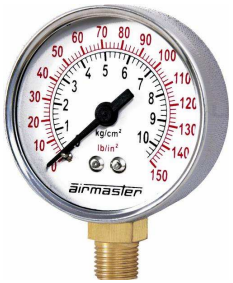
new **Nomenclature**

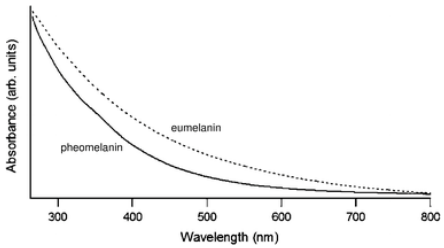
small **math**

Variables, outcomes

Could be measured or observed







Variable Types:
Levels of Measurement

qualitative (categorical)

quantitative (numerical)

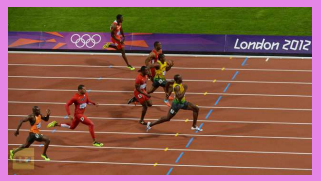
	nominal "scale"	ordinal scale	interval scale	ratio scale
	= ≠	= ≠ < >	= ≠ < > + −	= ≠ < > + − × ÷

natural order or ranking

known distance

natural zero value

(QUASI-) DISCRETE



(QUASI-) CONTINUOUS

Population and Sample

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

Therefore, only a subset of the population is examined. That is what we call a **sample**.

Population and Sample

Population

RANDOMNESS!

Sample

The size of the **population** usually does not allow the examination of all of its elements.

Therefore, only a subset of the population is examined. That is what we call a **sample**.

UNCERTAINTY!
(P and E)

Characteristics of the sample can be used to draw conclusions on the population.

We carry out measurements on the sample elements, then this data set (which is also called **sample**) will be characterized by graphs and numbers

Estimation

Population		Sample
Real value	Estimation	Estimate
Probability		Relative frequency
Expected value (mean of population)		Mean of the sample
Theoretical variance		(Empirical) sample variance
Difference of 2 expected value		Difference between 2 sample mean

Error

Population

- Real value (estimand)

Estimation

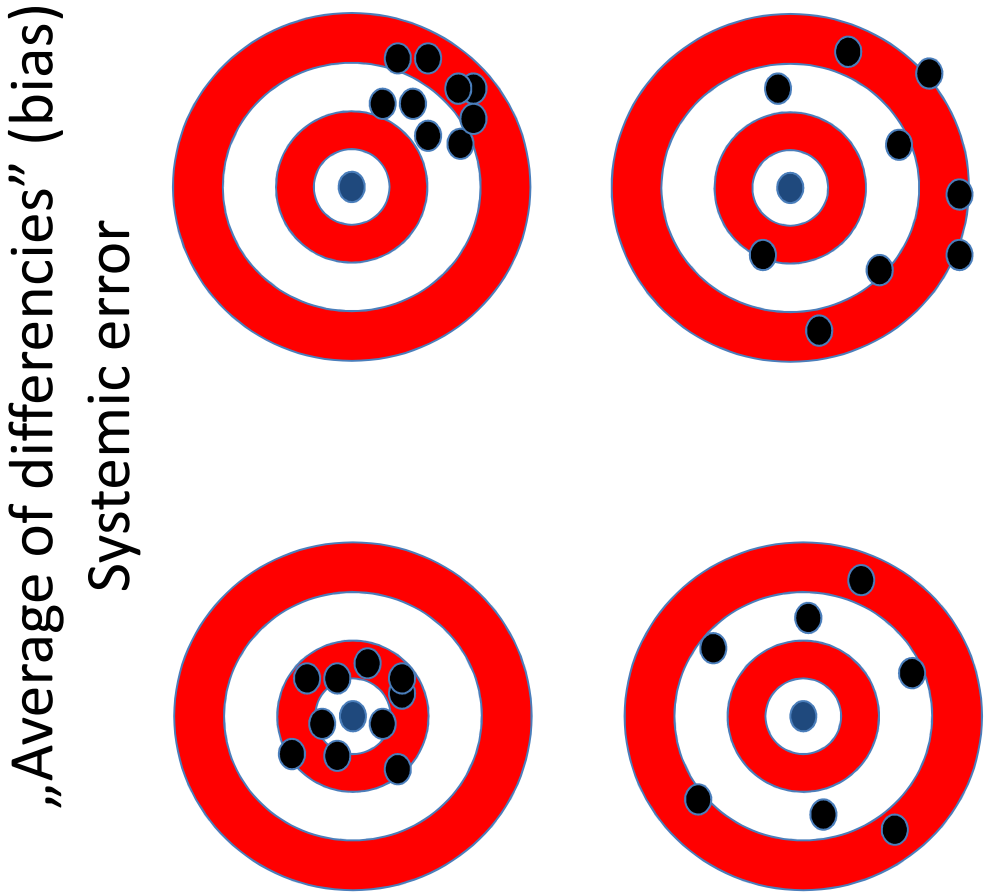
Sample

- Estimates

Error

Imagine more random samples – more estimates

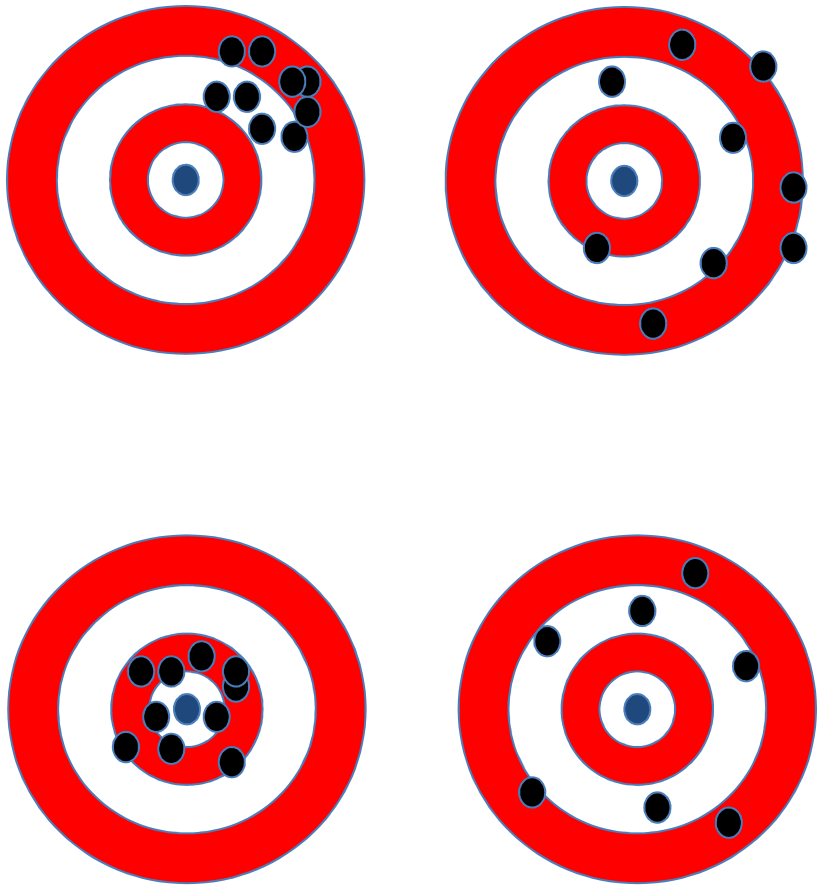
Error – 2 dimension



1. Variability of estimates
2. Difference between the real value and the „center” of the estimates

Error – 2 dimension

„Average of differencies” (bias)
Systemic error



„Variability” (standard deviation);
Random error

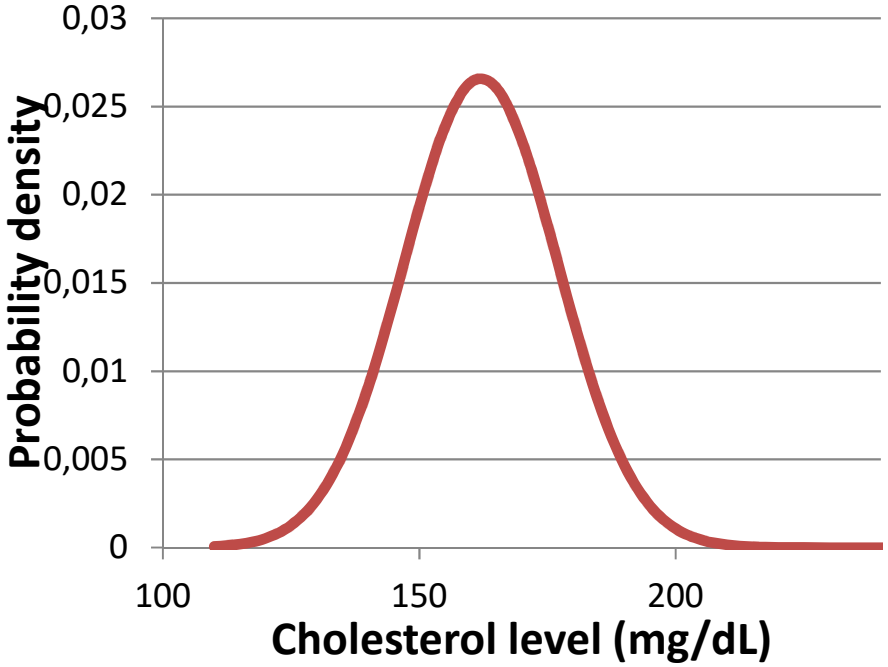
Good estimation, if:

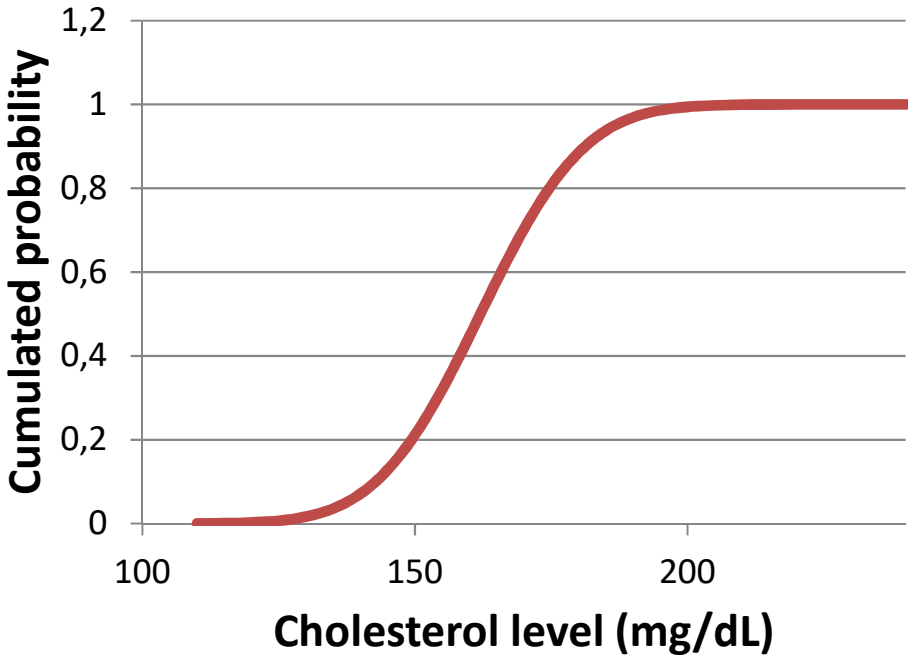
Unbiased:
the expected value of the estimates is equal with the estimand

Effective:
the variability is small

(Consistent:
the variability of estimates decreases with increased sample size)

Normal (Gaussian) Distribution I.

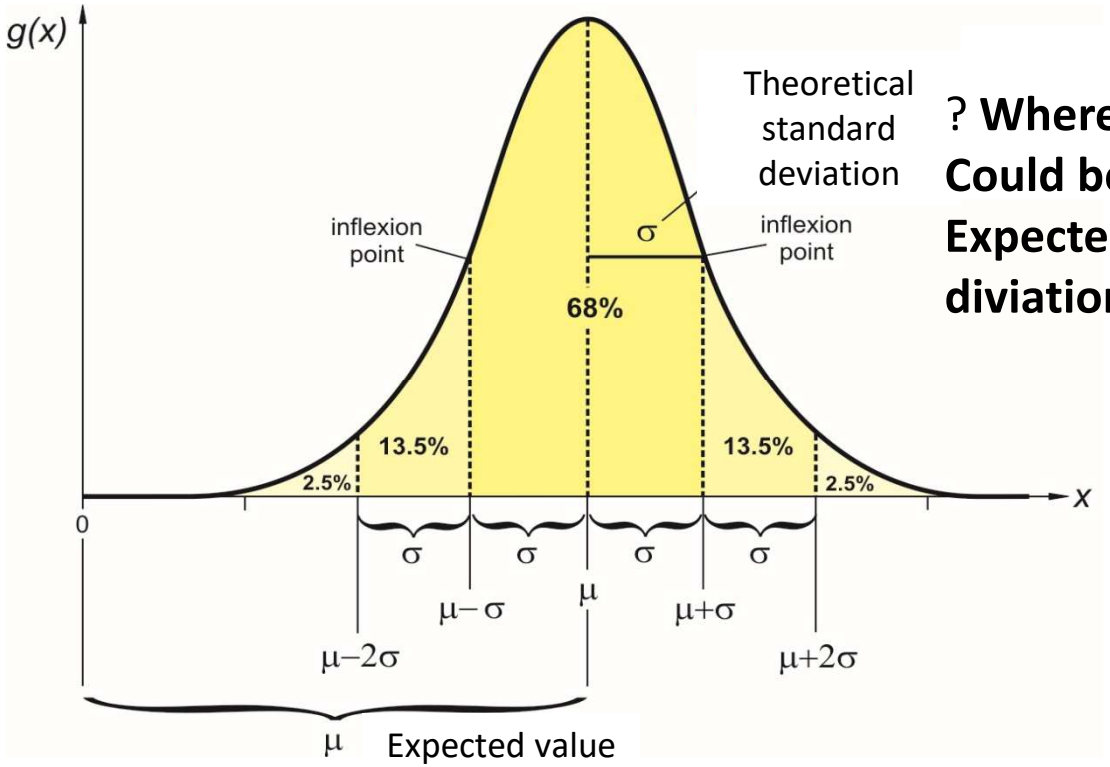




Cholesterol level, glucose level, jaw joint angle……..

Central limit theorem (on variables): for given conditions, adding a large number of independent variables yields a normally distributed variable.

Prediction intervals



? Where is X% of the data in the population?
Could be calculated if:
Expected value and theoretical standard diviation (sd) is known

$$eg\ :95\ \%\ :\mu \pm\ \sim\ 2 * \sigma$$

Based on the sample?

Population	**<----**	**sample**
Expected value (μ)	<----	mean ($\bar{x}$)
Theoretical sd (σ)	<----	sample sd (s)
Normal dist.	<----	t-dist. (more „precise”...)

$$eg\ :95\ \%\ :\bar{x} \pm\ \sim\ 2 * s$$

$$95\ \%\ :\bar{x} \pm t * s$$

Estimation of the mean

From the sample, the estimate of the expected value based on the average is also uncertain!

How can this uncertainty be estimated?

Central limit theorem (on sampling): for given conditions, sampling with large sample size (n) the **distribution of the sample means** is normal with:

$$\mathrm{Var}_{\mathrm{means}} = \frac{\mathrm{Var}_{\mathrm{sample}}}{\mathrm{n}}$$

Confidence intervals

- If we take samples, then if we look at the averages of the samples, it also follows a normal distribution.
- In this case, the *standard deviation of the distribution of the means is the standard error* (also known as the standard deviation of the mean).
- The prediction interval for the means is called the *confidence interval*.

Therefore we can estimate the 95% confidence interval as follows:

$$\sim \bar{x} \pm 2 * SEM \qquad SEM = \frac{SD}{\sqrt{n}}$$

We can construct confidence intervals for other estimates too!
The CI shows the value of the estimate, its error and its confidence together.

Sampling („random”) error

Aim of hypothesis tests: **Statistical answer on YES/NO question**

Starting point: create a specific statistical question and answers:

H_0: null hypothesis – „random” error only

H_a (or H_1): alternative hypothesis – not H_0

Decision is based on: role of „randomness” if H_0 true (sampling error)

A sample is that could contradict H_0

		In population (in reality) the null hypothesis is:	
		True	False
Decision on null hypothesis:	Accepting (Not rejecting)	Good decision	Error (**type II**) (β) (false negative result)
	Rejecting	Error (**type I**) (α) (false positive result)	Good decision (power) (1-β)

An example - last steps

What is the question : The probability of six-throw is different from 1/6, even bigger?

Null hypothesis: H_0: The probability of rolling 6 is 1/6..

Significance level:10%

Sample: 6 times 6 out of 24 rollings.

Is the difference important at all?– relevant : 1/4 probability, that **1,5** times higher than 1/6 – YES it is

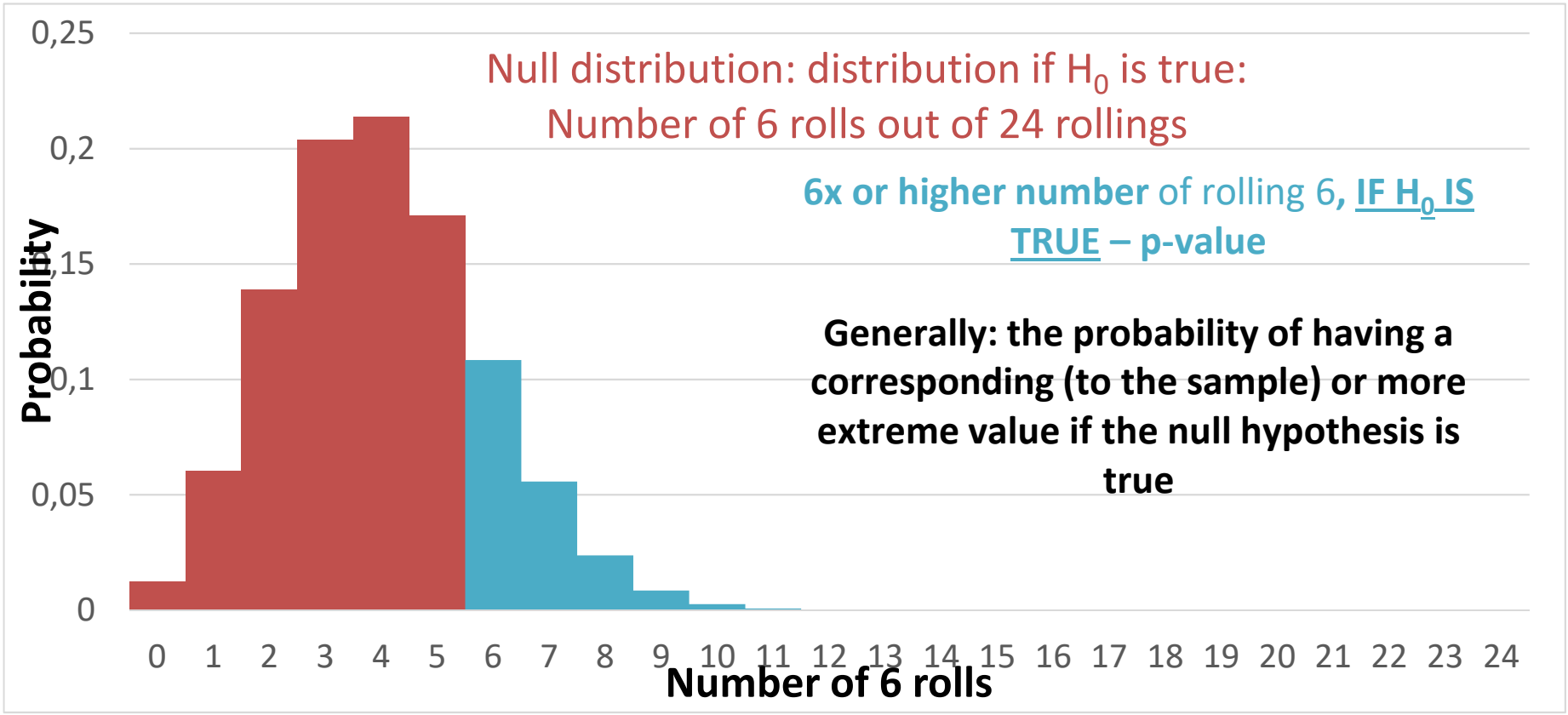
How much evidence? – p-value: 0,1995

Decision: there is not enough evidence for reject H_o– accept H_0

		In population (in reality) the null hypothesis is:	
		True	False
Decision on null hypothesis:	Accepting (Not rejecting)	Good decision	Error (**type II**) (β) (false negative result)
	Rejecting	Error (**type I**) (α) (false positive result)	Good decision (power) (1-β)

Calculation in the „background”

Using binomial test!



One sample Student t-test

What I'm curious about

Expected value of the sample is equal with a known population mean

Type of variable

1 numerical and continuous

Assumption

Independent observations

distribution of means is normal:

normally distributed sample or large sample size (CLT)

Notes: Calculation:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Paired Student t-test

What I'm curious about

Two expected values in two groups are equals – in paired groups

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations in the groups, paired groups

the distribution of the difference of means is normal

normally distributed differences or large sample size

Notes:

paired test usually has higher power

2 sample Student t-test

What I'm curious about

Two expected values in two groups are equals

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations between and within groups

distribution of means is normal in each group:

distribution is normal in each group or large sample size

distribution of standard deviations are the same

Notes:

use Welch test instead. (typically we do not know the SDs)

Welch test

What I'm curious about

Two expected values in two groups are equals

Type of variable

1 numerical and continuous, 1 binary („groups")

Assumptions

Independent observations between and within groups

distribution of means is normal in each group:

distribution is normal in each group or large sample size

Notes:

suitable to compare other location parameters (quantiles)

not sensitive for different variances (robust for variance differences)

Multiplicity

you do NOT need to know for the exam!

…Chocolate Helps Weight Loss.

„Slim by Chocolate!" the headlines blared. A team of German researchers had found that people on a low-carb diet lost weight 10 percent faster if they ate a chocolate bar every day. It made the front page of Bild, Europe's largest daily newspaper."

„*the statistically significant benefits of chocolate that we reported are based on the actual data*"

But how??

…Chocolate Helps Weight Loss.

„…Frank randomly assigned the subjects to one of three diet groups. One group followed a low-carbohydrate diet. Another followed the same low-carb diet plus a daily 1.5 oz. bar of dark chocolate. And the rest, a control group, were instructed to make no changes to their current diet.”

„Our study included **18 different measurements**—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.—**from 15 people**.”

…Chocolate Helps Weight Loss.

The conventional cutoff for being “significant” is 0.05, which means that there is just a 5% chance that your result is a random fluctuation.

„Error probability” at 1 test:	p
No error probability at 1 test:	$1-p$
No error for k independent tests:	$(1-p)^k$
At least 1 error for k independent tests:	$1 - (1-p)^k$

„With our 18 measurements, **we had a 60% chance** of getting some“significant” result with $p < 0.05$. (The measurements weren’t independent, so it could be even higher.)”

This problem is called multiplicity.

I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.

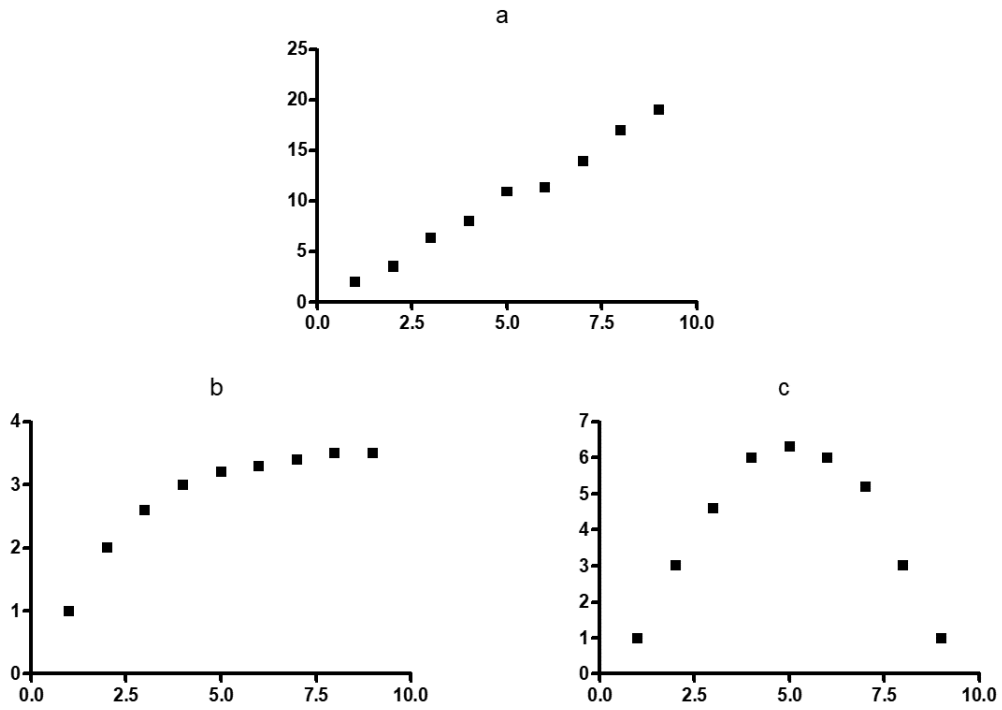
- Multiplicity
  - ☺*eg: Chocolate Helps Weight Loss*
  - *https://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800*

Correlation, regression

Relation between 2 numerical variable

Type of relation:

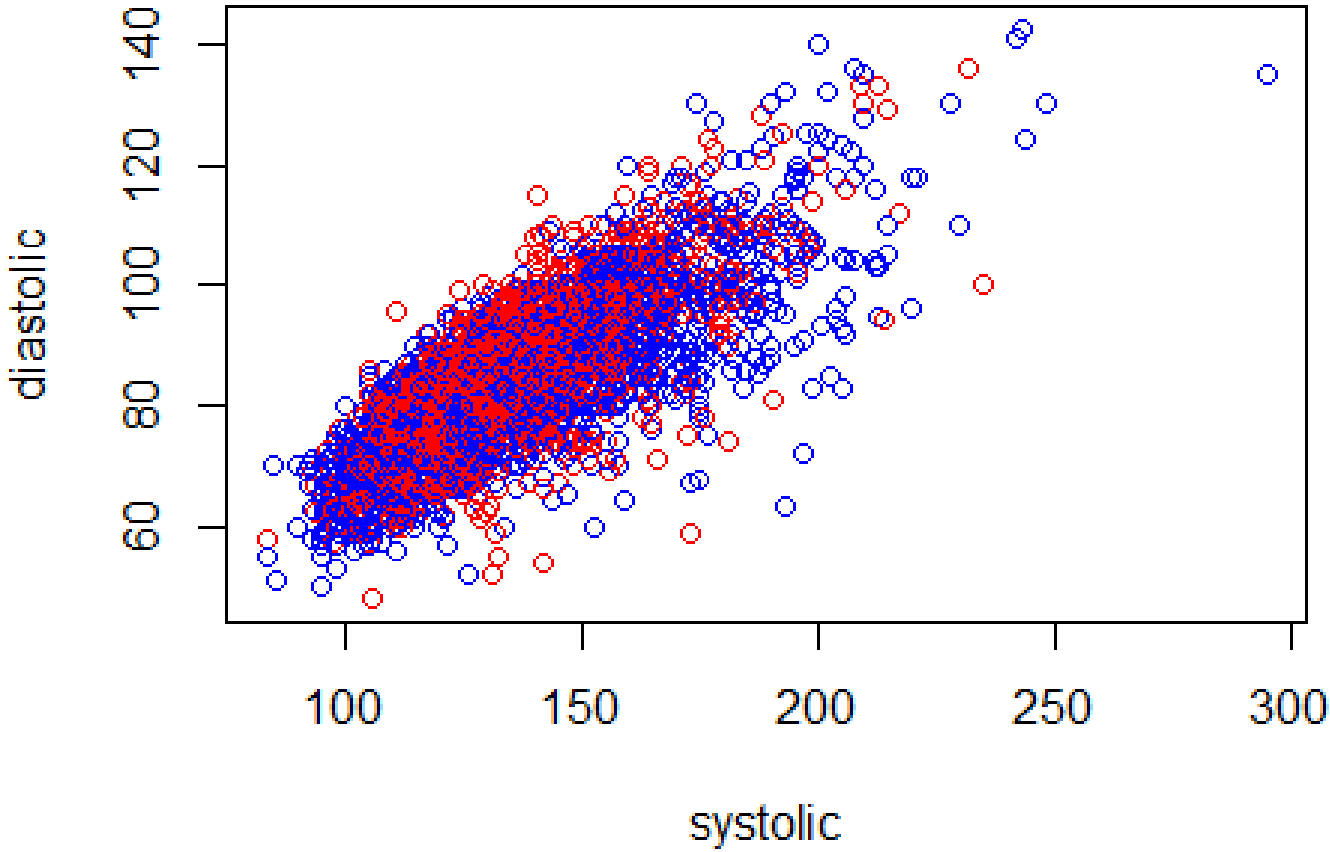
- monotonic
  - positive
  - negative
  - positive linear
  - …
- non monotonic
  - parabolic
  - …
- No relation



Correlation

Monotonic,

symmetric (it cannot be said which one would depend on which)
Relation between **2 random** (random error, "not set") variable.



Correlation

Strength of correlation:

Correlation coefficients (r):

if **linear corelation** is assumed: **Pearson** r

if **monotonic** (not necessarily linear): **Spearman rank** r,

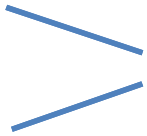
The **value** of the correlation coefficient:

-1 to +1

negative: negative correlation

positive: positive correlation

closer to|1|means more strength



$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

„Distance from the mean" – both for y and x

„Correlation” t-test (on Pearson r)

What I’m curious about

Two variables are (linearly) depends on each other, r?=0

Type of variable

2 numerical variable (X and Y)

Assumptions

Independent observations for pairs

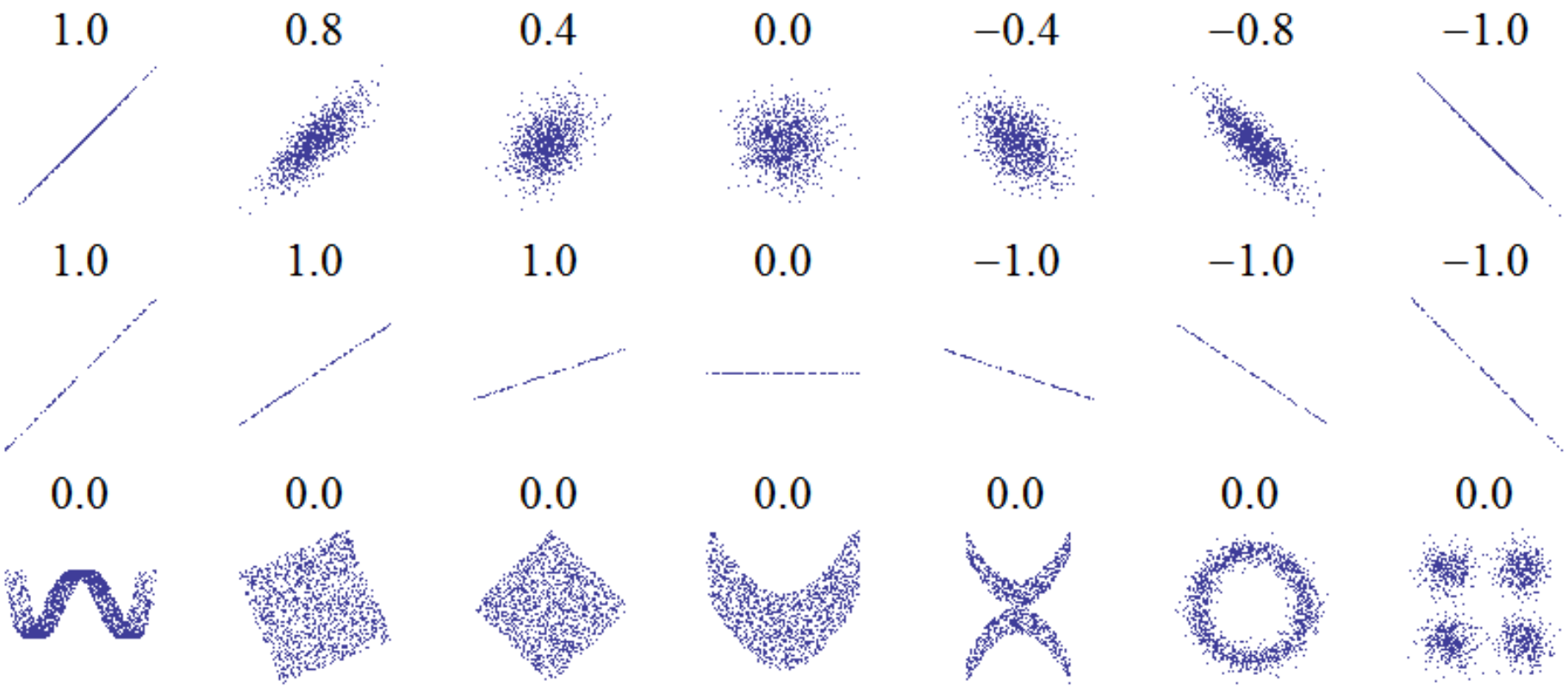
symmetric, linear relation assumed

x and y variables are random variables

Notes

NOTES

- *Always make a graph!;*
- *Correlation does not mean causality!*
  ☺*eg: http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation*



Regression

Function relation (NOT symmetric) between a dependent (outcome, predicted, Y) variable and an independent (predictor, X) variable. [Y is a random variable, X not necessarily random variable]

Y depends on X – this dependency could not cheked by statistics – it should be assumed by **clinical knowledge**.

Questions could be answered:

- Is there a (linear) function relation between the variables?
- What is the value of Y is if X is…? (estimations)
- What is the value of X is if Y is…?(estimations)

Linear regression

Linear function relation is assumed**.**

In the case of 2 variables, the questions and calculations of the regression and correlation can most often be made “equivalent” to each other.

Linear regression

To estimate the linear: OLS (Ordinary Least Square) method

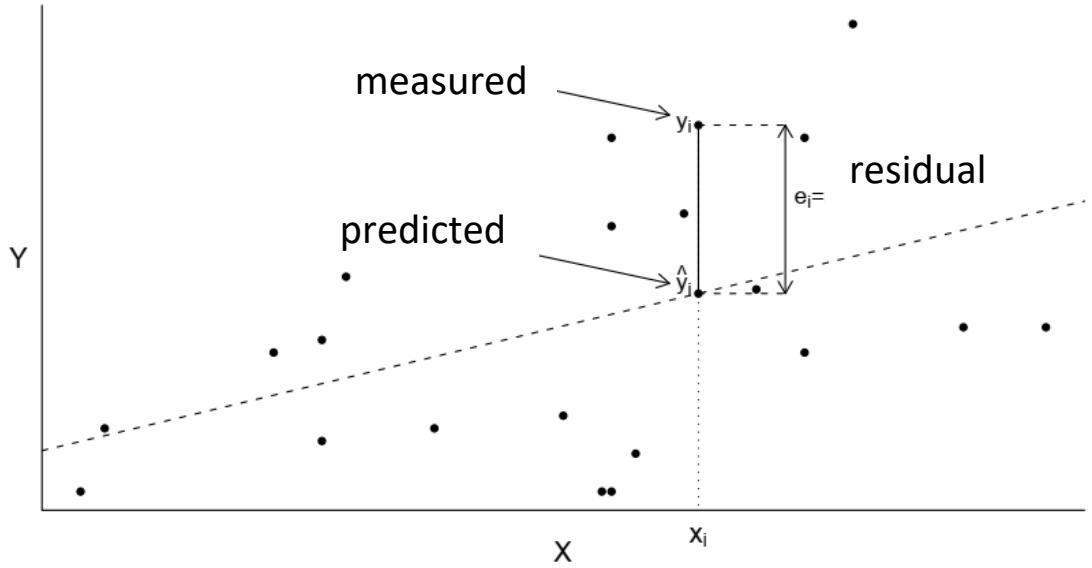
intercept

slope

Linear function: $Y = \beta_0 + \beta_1 * X + \epsilon$

Reziduum (error term): point-linear **vertical** difference
= predicted and measured Y value difference

Based on OLS the best line is where the sum of the **squared residuals** are **minimal**.



„Correlation” t-test (on slope)

What I’m curious about

Y variable is linearly depends on X (slope?=0)

Type of variable

2 numerical variable (X and Y)

Assumptions

Independent observations for pairs

linear relation assumed

x values measured with no error

residuals are normally distribution

residuals have same variance at each x

Notes

Slope and R^2

Slope

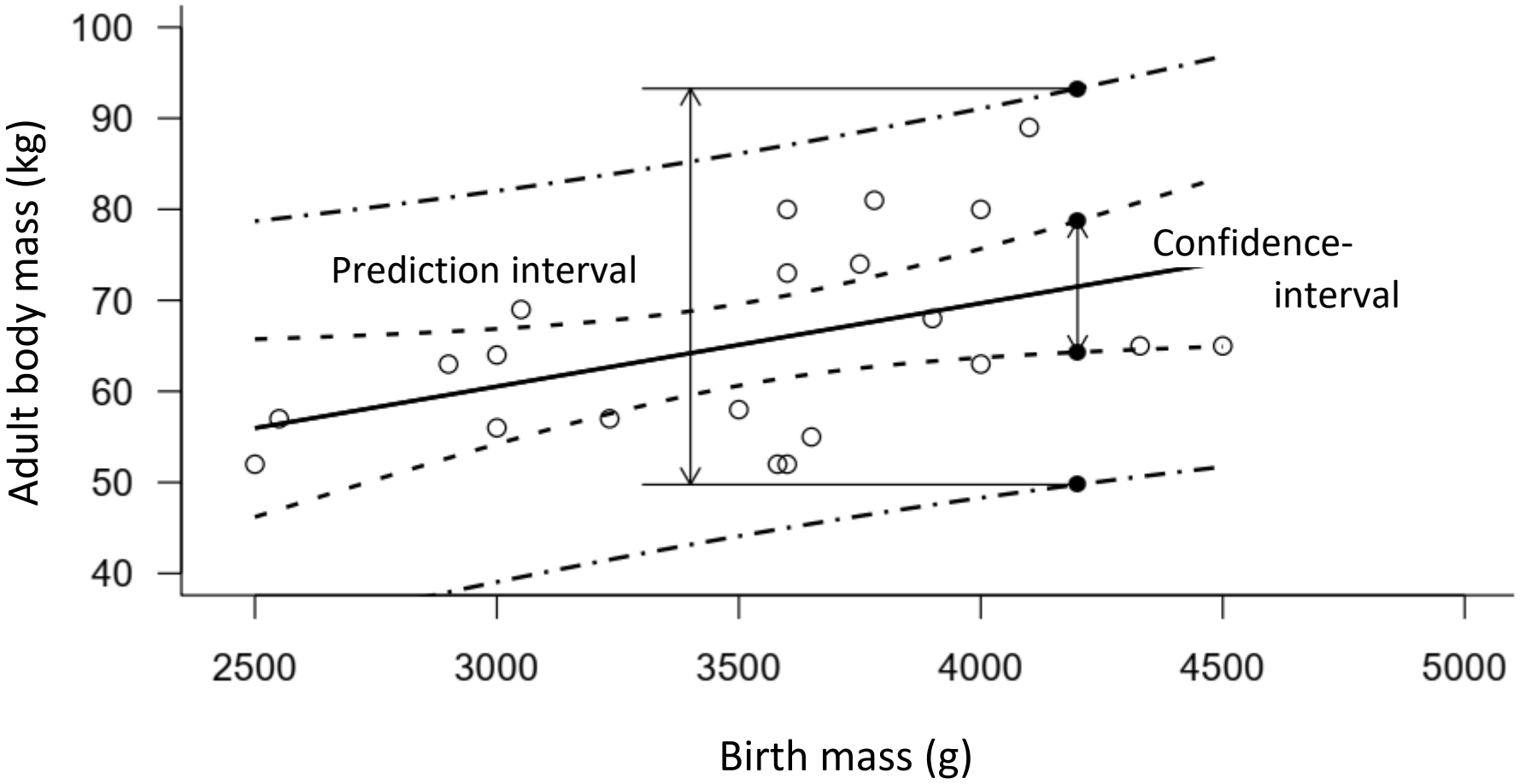
The average change in Y regarding to the 1 unit change in X

R^2 – determination coefficient

square of R

What percentage of the variance (variability) of the variable Y can be explained by the variance (variation) of X

Confidence and prediction intervals



Feedback