Elements of experimental design

… is it really that important that the experiment is well planned?

Gdańsk, May 13, 2023

Peter Grešner, PhD.

Centre of Biostatistics and Bioinformatics Analyses
Division of Translational Oncology
Medical University of Gdansk

peter.gresner@gumed.edu.pl

Author’s background…

1997 programming class high school (Slovak Republic)

2002 biomedical physics (MSc)
Faculty of Mathematics, Physics and Informatics,
Comenius University, Bratislava, Slovak Republic

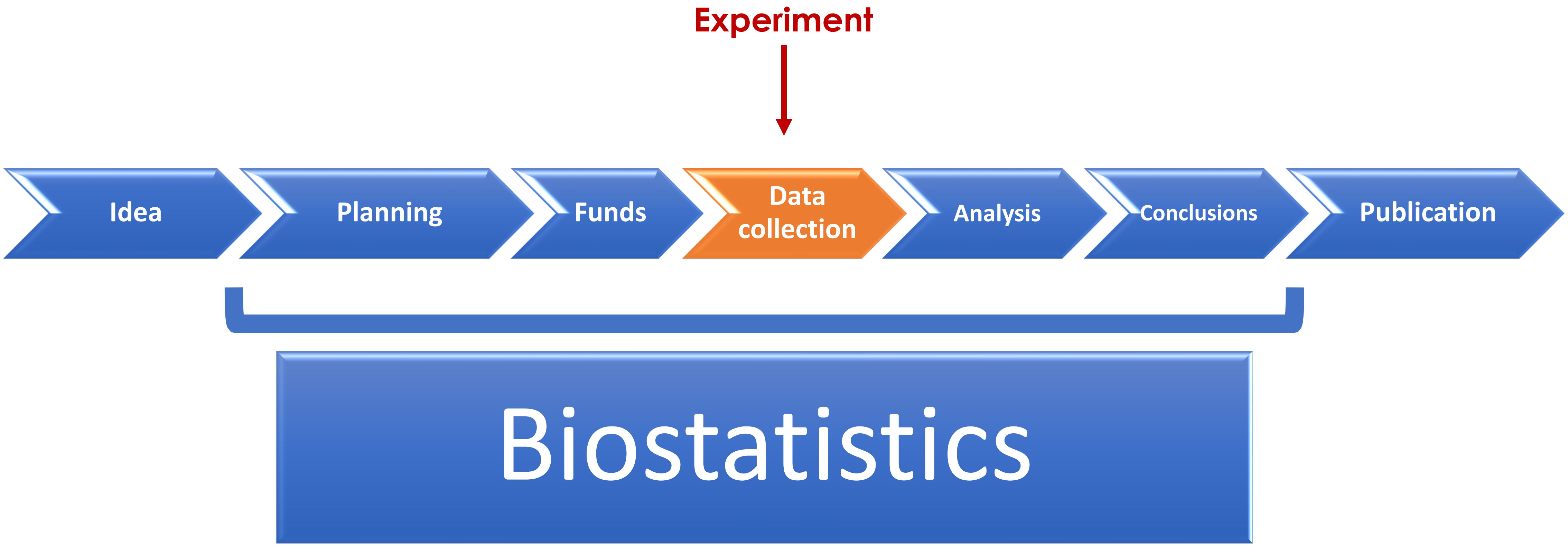
2006 biophysics (PhD)
Faculty of Mathematics, Physics and Informatics,
Comenius University, Bratislava, Slovak Republic

2006 medical statistician
Military Teaching Hospital No.2, Medical University of Lodz,
Poland

2006 – 2020 postdoctoral fellow […] adjunct associate professor
Nofer Institute of Occupational Medicine, Lodz, Poland

2020 – … senior biostatistician
Medical University of Gdansk, Poland

What exactly is (bio)statistics?



Experiment

Idea

Planning

Funds

Data collection

Analysis

Conclusions

Publication

Biostatistics

Mostly, it's not done properly...

- approximately 60% of original papers in the field of biomed/pharm contain statistical errors(De Muth 1999)
  - improper experimental design and **planning**
  - poorly formulated research **hypothesis**
  - incorrectly estimated **sample size** (or no estimation at all)
  - misused **mean & SD**
  - wrong selection of **parametric/non-parametric** tests
  - incorrect use of **paired/unpaired** tests
  - using standard error (**SE**) instead of standard deviation (**SD**)

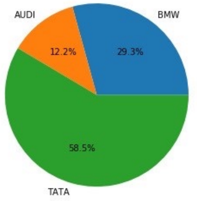
  - using multiple t-tests as an extension of the analysis of variance method
  - using the chi^2 test and Fisher's exact test

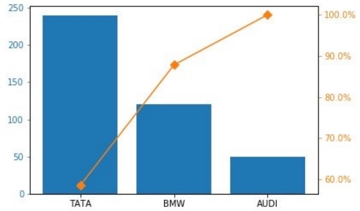
Watała: Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych.
α-medica press, Bielsko-Biała, 2002.

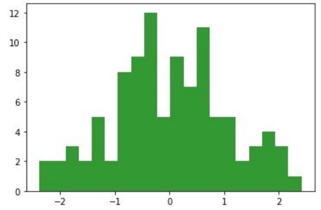
Biostatistics – what are our options?

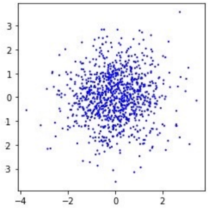
Descriptive statistics

- characteristics of collected dataset
- answers questions:
  - how to describe/present the collected data?
  - what is the most representative value?
  - what are the extremes?
  - what is the spread of data?
  - how to compare data to those from similar sets?
- not going "beyond the scope" of the collected data



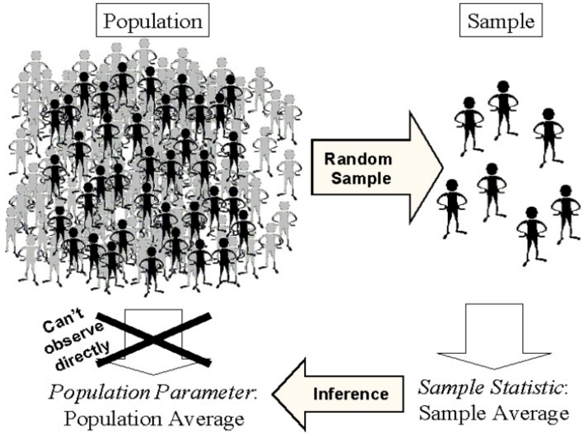






Statistical inference

- a way of "generalizing"
- drawing conclusions about a population from which ONLY a small part (sample) has been analyzed



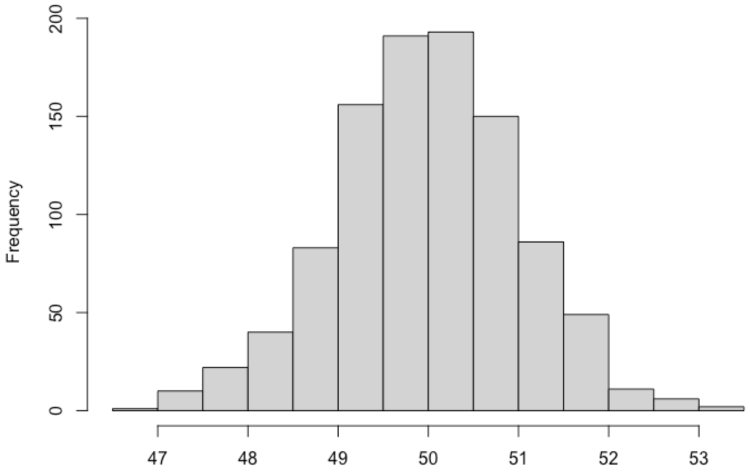
Descriptive statistics

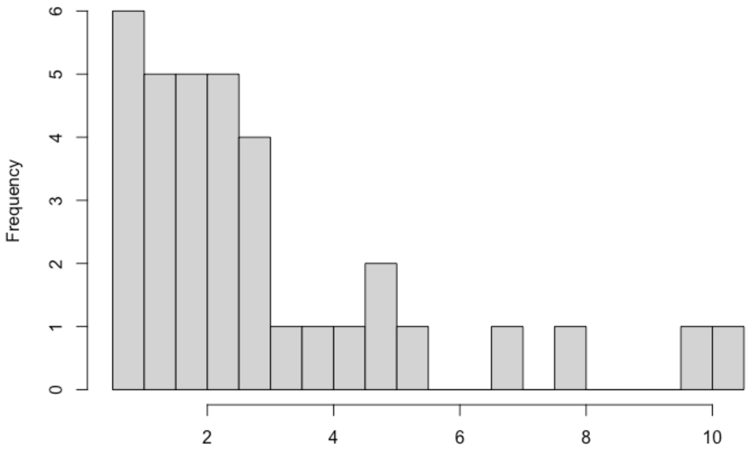
Histogram

Measures of central tendency
mode, mean ($\boldsymbol{\mu}$), median

Measures of spread
range, quartiles (**Q**)
variance (**s**), standard deviation ($\boldsymbol{\sigma}$; ***SD****; 68%; 95%*)
interquartile range (**IQR=Q3-Q1**; middle 50%)

coefficient of variation [%] ($\boldsymbol{\sigma}$/ $\boldsymbol{\mu}$)
quartile coefficient of dispersion [%]
(**IQR/(Q3+Q1)**)





Median vs. Mean

median

- the middle value in distribution
- 50% left / 50% right
- less sensitive to outliers/extreme
- only numerical data

mean (μ; $\bar{y}$)

- sum of all values divided by their number
- sensitive to outliers/extremes
- only numerical data

Normal distribution

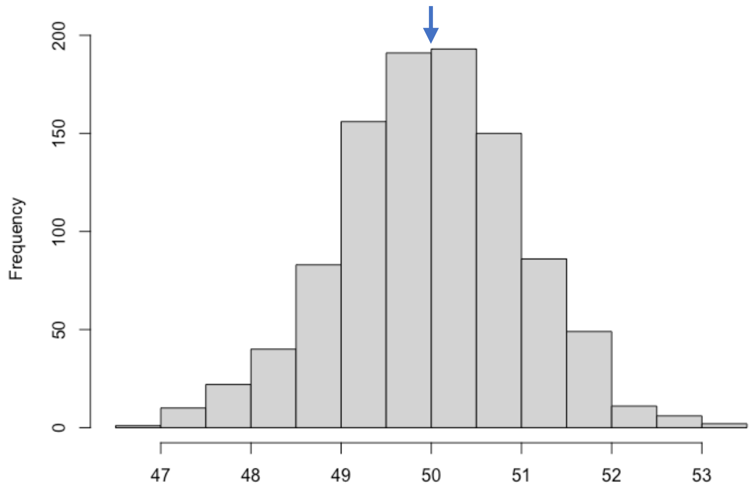
mean & SD

median & IQR

Non-normal distribution

median & IQR (only!)

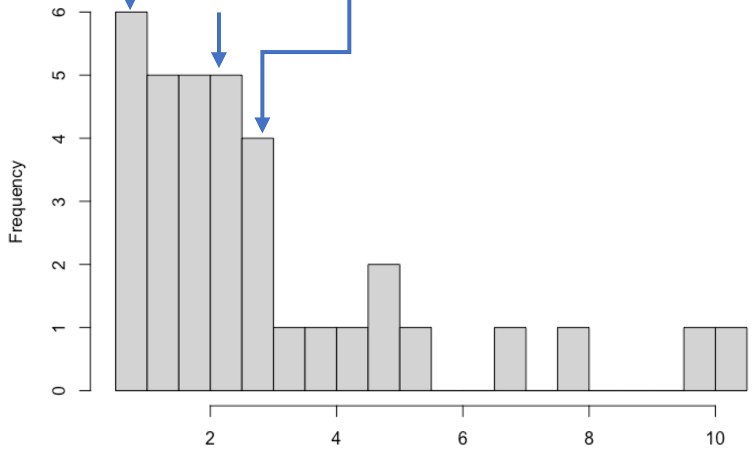
mode, mean, median



mode

median

mean

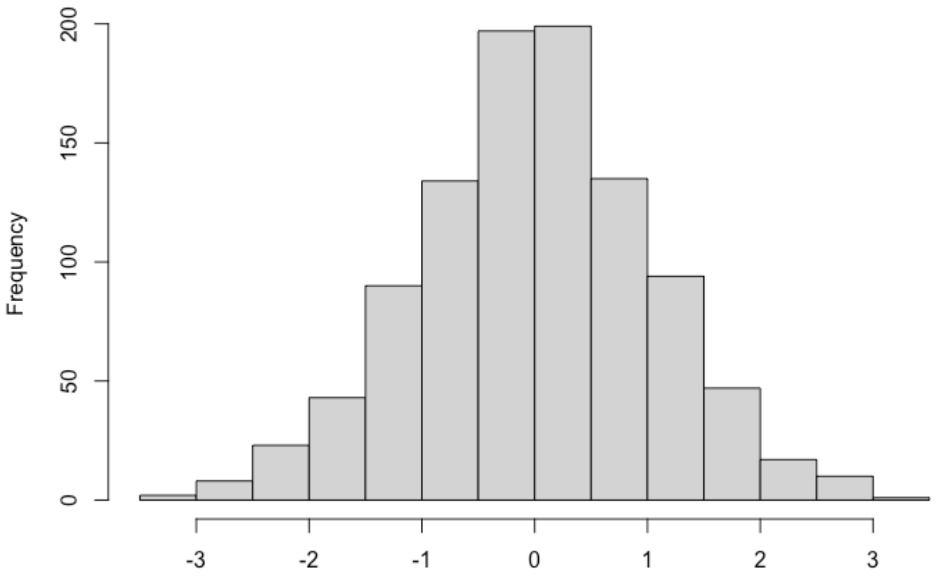


Normal (Gaussian) distribution

- the "bell"-shape, symmetrical around mean
- mathematically **fully** characterized by mean (μ) and standard deviation ($\boldsymbol{\sigma}$)
- $\mu = 0$; $\boldsymbol{\sigma} = 1$; AUC = 1;

Significance of normal distribution

- many biomedical parameters present normal distribution
- many statistical tests thus **assume normally distributed data** (if not met, the logic of analysis fails)



Central limit theorem

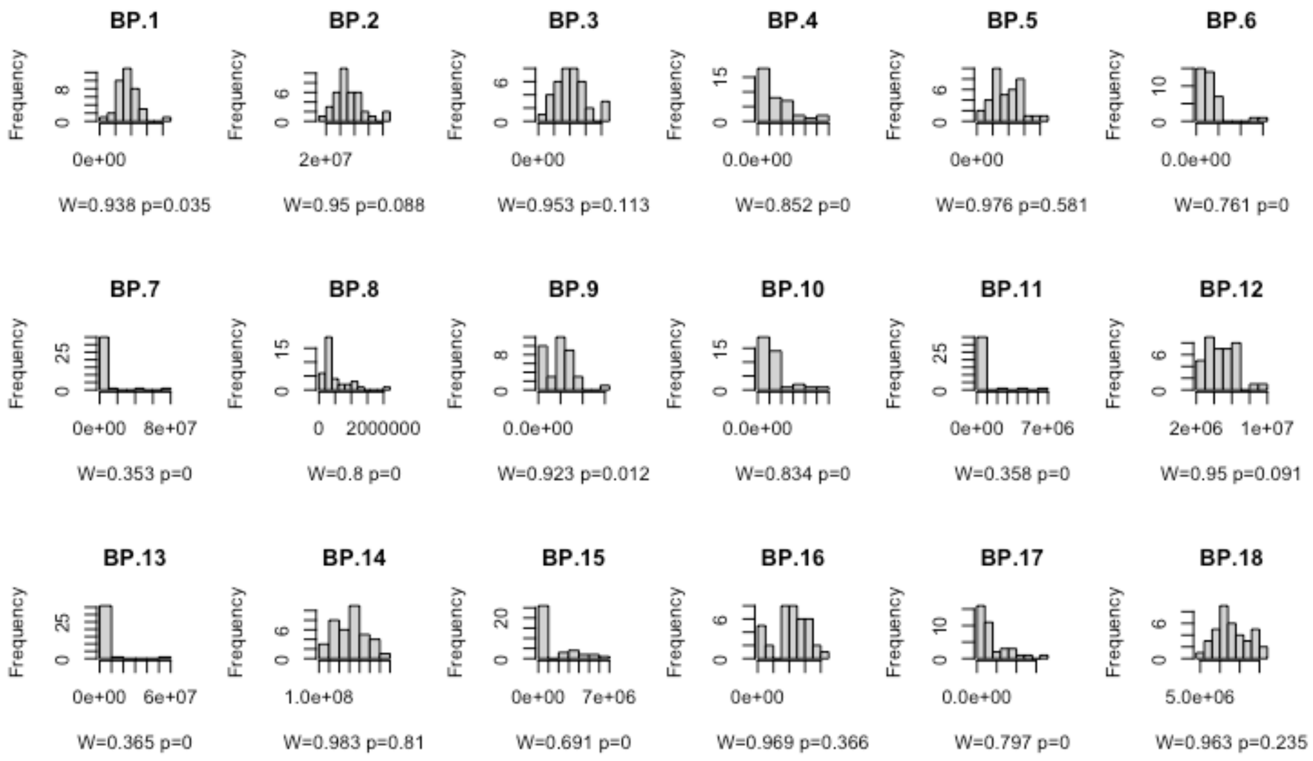
histogram of means from many non-gaussian samples will present normal distribution

How to choose proper measures?

Always check your data first!

- visual inspection of histograms

- statistical tests for testing data normality
    - Shapiro-Wilk W test
    - Kolmogorov-Smirnov test
    - Lilliefors test



Statistical inference

Research hypothesis
A statement specifying the existence of some relationship, difference, mechanism, process, etc.

Statistical hypothesis
Redefinition of research hypothesis into a measurable form.

Hypothesis testing
A sequence of steps allowing us to either accept or reject the hypothesis.

In order to do so, one has to follow **the rules**…

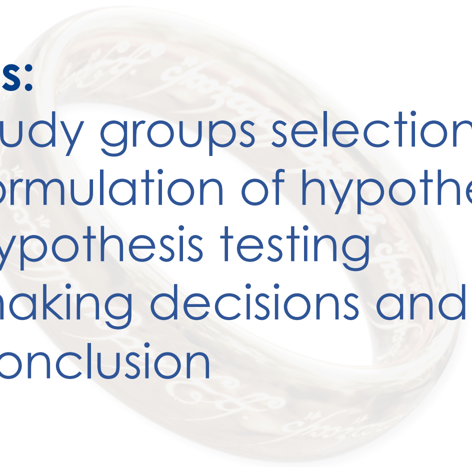
One ring to rule them all…
…and in the darkness bind them.

Statistical inference

Sir Karl Raimund Popper (1902 – 1994)

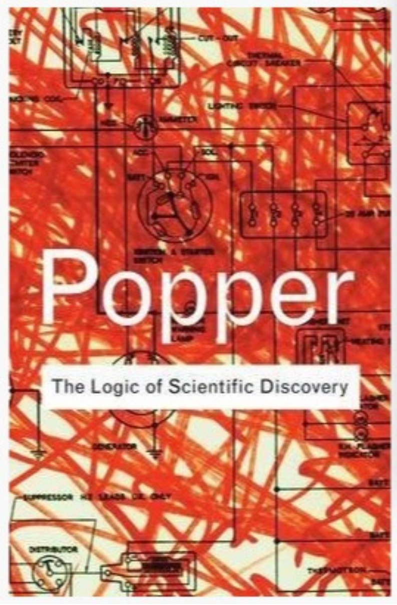
Criterion of falsifiability

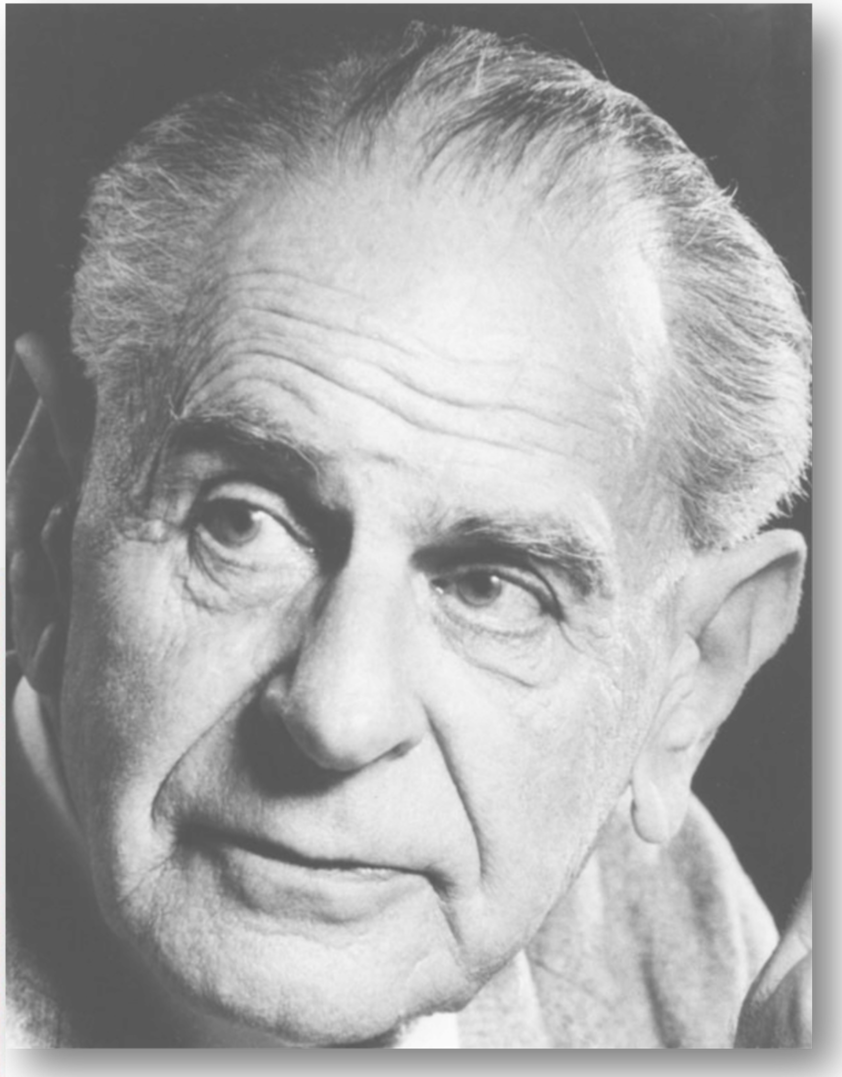
- the main scientific criterion
- In order to prove something, try to reject the negation thereof
- critical rationalism („popperism”)
- philosophy of science
- historical context of Eastern Block 😵‍💫



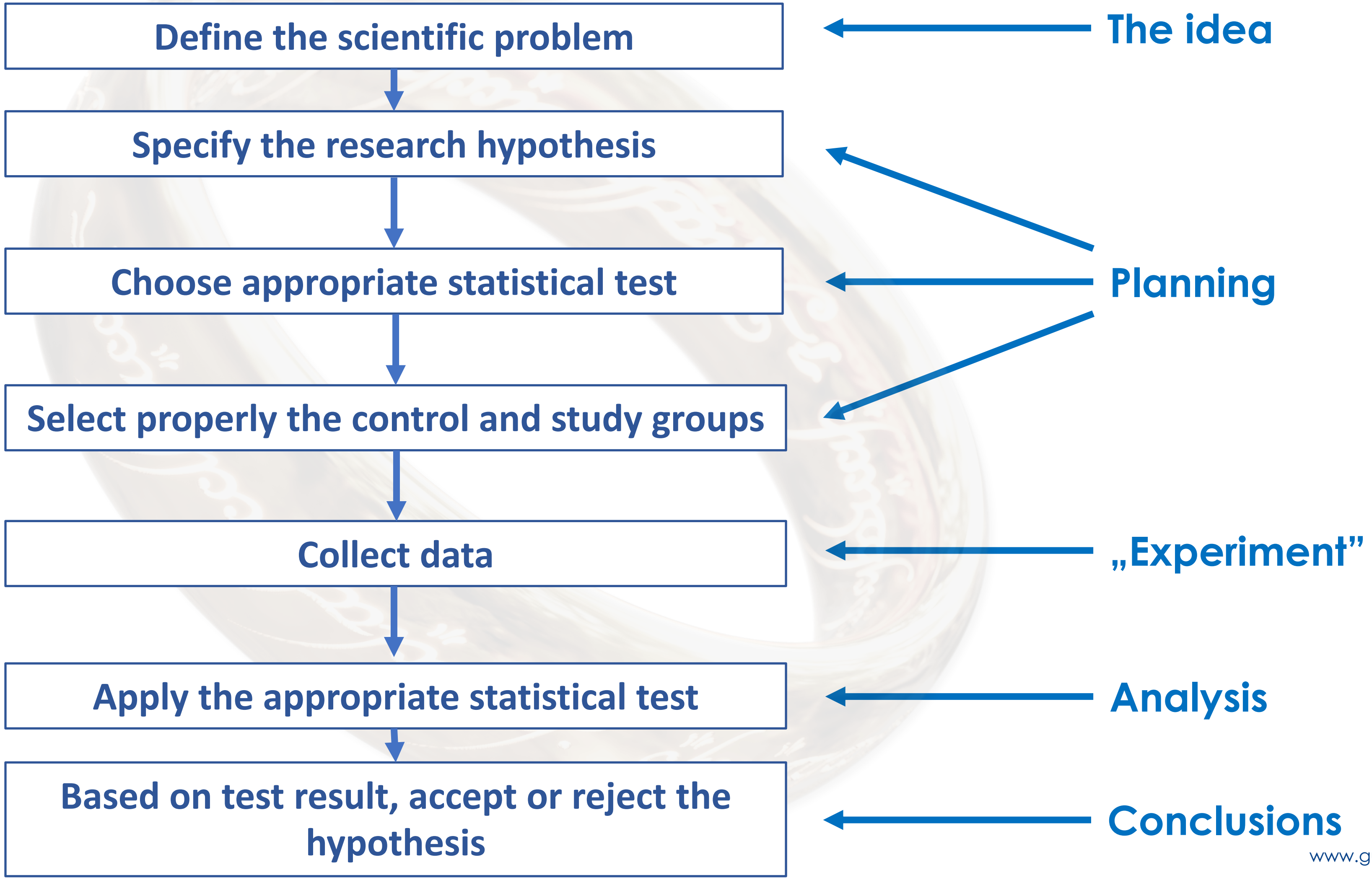
Rules:

- study groups selection
- formulation of hypotheses
- hypothesis testing
- making decisions and drawing conclusion





Statistical inference - workflow



Define the scientific problem

Specify the research hypothesis

Choose appropriate statistical test

Select properly the control and study groups

Collect data

Apply the appropriate statistical test

Based on test result, accept or reject the hypothesis

The idea

Planning

„Experiment”

Analysis

Conclusions

Scientific problem & Hypotheses

Scientific problem

"Does the XY disease affect the patients' IQ?"

Statistical hypotheses

we want to check whether our data allows us to reject this hypothesis as untrue

The "null" hypothesis (H_0)

The average IQ of people with the XY disease does not differ from the one among healthy people (without the XY disease).

$$H_0\colon \overline{IQ_{XY}} = \overline{IQ_{healthy}}$$

The alternative hypothesis (H_A)

The average IQ of people with the XY disease differs from the one among healthy people (without the XY disease).

$$H_A\colon \overline{IQ_{XY}} \neq \overline{IQ_{healthy}}$$

Choosing appropriate statistical test

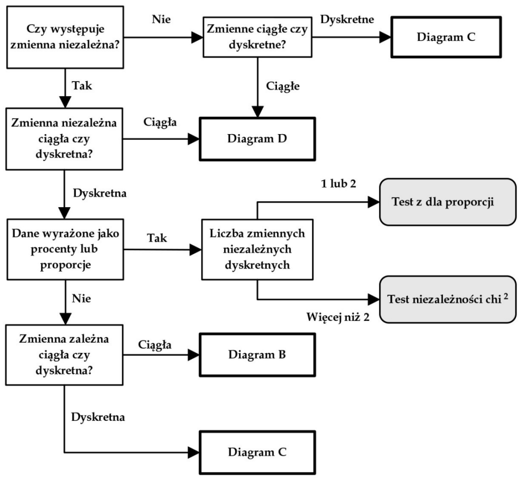
Not a trivial problem, many aspects need to be considered

- type of data: numerical (continuous, discrete) / categorical / proportions
- normality of the data
- dependent/independent variable
- data pairing
- data censoring (right censoring – survival analysis)
- number of levels of categorical variables
- number of compared groups
- additional (confounding) factors

- **experimental design**
  - hierarchical design
  - balanced/unbalanced design
  - fixed/random effects
  - etc.

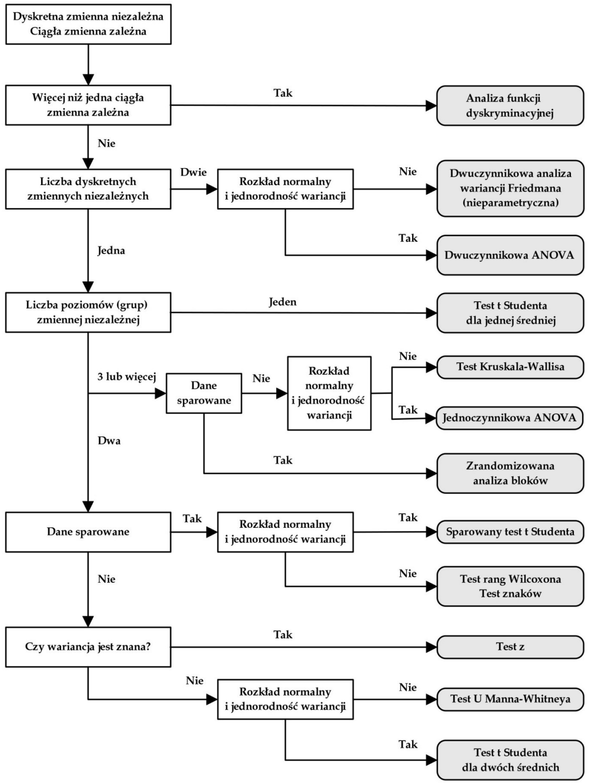
Choosing appropriate statistical test

Diagram A



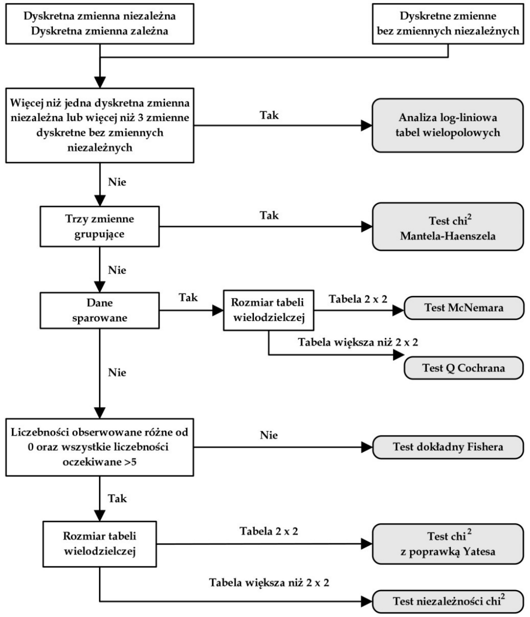
Czy występuje zmienna niezależna?
Nie
Zmienne ciągłe czy dyskretne?
Dyskretne
Diagram C
Ciągłe
Diagram D
Tak
Zmienna niezależna ciągła czy dyskretna?
Ciągła
Diagram D
Dyskretna
Dane wyrażone jako procenty lub proporcje
Tak
Liczba zmiennych niezależnych dyskretnych
1 lub 2
Test z dla proporcji
Więcej niż 2
Test niezależności chi^2
Nie
Zmienna zależna ciągła czy dyskretna?
Ciągła
Diagram B
Dyskretna
Diagram C

Diagram B



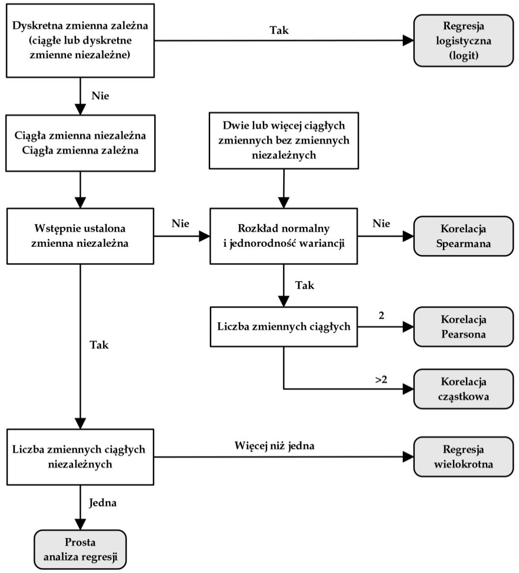
Dyskretna zmienna niezależna
Ciągła zmienna zależna
Więcej niż jedna ciągła zmienna zależna
Tak
Analiza funkcji dyskryminacyjnej
Nie
Liczba dyskretnych zmiennych niezależnych
Dwie
Rozkład normalny i jednorodność wariancji
Nie
Dwuczynnikowa analiza wariancji Friedmana (nieparametryczna)
Tak
Dwuczynnikowa ANOVA
Jedna
Liczba poziomów (grup) zmiennej niezależnej
Jeden
Test t Studenta dla jednej średniej
3 lub więcej
Dane sparowane
Nie
Rozkład normalny i jednorodność wariancji
Nie
Test Kruskala-Wallisa
Tak
Jednoczynnikowa ANOVA
Tak
Zrandomizowana analiza bloków
Dwa
Dane sparowane
Tak
Rozkład normalny i jednorodność wariancji
Tak
Sparowany test t Studenta
Nie
Test rang Wilcoxona
Test znaków
Nie
Czy wariancja jest znana?
Tak
Test z
Nie
Rozkład normalny i jednorodność wariancji
Nie
Test U Manna-Whitneya
Tak
Test t Studenta dla dwóch średnich

Diagram C



Dyskretna zmienna niezależna
Dyskretna zmienna zależna
Dyskretne zmienne bez zmiennych niezależnych
Więcej niż jedna dyskretna zmienna niezależna lub więcej niż 3 zmienne dyskretne bez zmiennych niezależnych
Tak
Analiza log-liniowa tabel wielopolowych
Nie
Trzy zmienne grupujące
Tak
Test chi^2 Mantela-Haenszela
Nie
Dane sparowane
Tak
Rozmiar tabeli wielodzielczej
Tabela 2 x 2
Test McNemara
Tabela większa niż 2 x 2
Test Q Cochrana
Nie
Liczebności obserwowane różne od 0 oraz wszystkie liczebności oczekiwane >5
Nie
Test dokładny Fishera
Tak
Rozmiar tabeli wielodzielczej
Tabela 2 x 2
Test chi^2 z poprawką Yatesa
Tabela większa niż 2 x 2
Test niezależności chi^2

Diagram D



Dyskretna zmienna zależna (ciągłe lub dyskretne zmienne niezależne)
Tak
Regresja logistyczna (logit)
Nie
Ciągła zmienna niezależna
Ciągła zmienna zależna
Wstępnie ustalona zmienna niezależna
Nie
Dwie lub więcej ciągłych zmiennych bez zmiennych niezależnych
Rozkład normalny i jednorodność wariancji
Nie
Korelacja Spearmana
Tak
Liczba zmiennych ciągłych
2
Korelacja Pearsona
>2
Korelacja cząstkowa
Tak
Liczba zmiennych ciągłych niezależnych
Więcej niż jedna
Regresja wielokrotna
Jedna
Prosta analiza regresji

Watała: Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych.
α-medica press, Bielsko-Biała, 2002.

Basic test characteristics

Parametric tests

- make several assumptions concerning the distribution of data
- if not met, their logic fails and provide unreliable results
- limited use but very “powerful” (general linear model, ANOVA, …)

Nonparametric tests

- less or no assumptions concerning the data distribution and other characteristics
- more “universal”

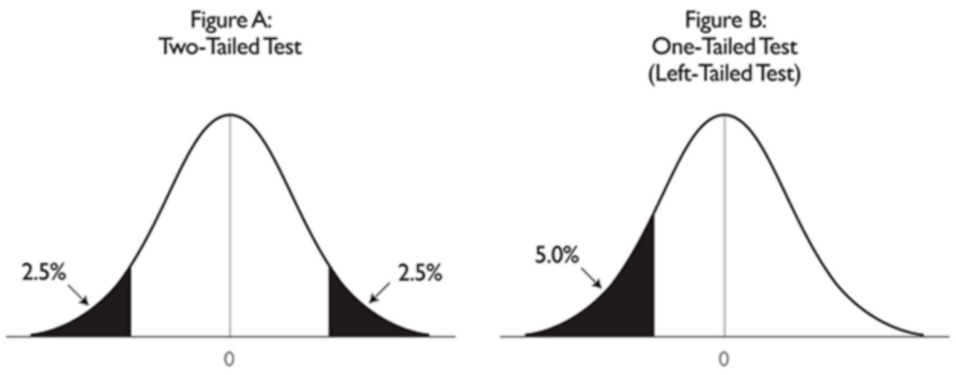
One-tailed tests

- assume one specific change (decrease | increase)

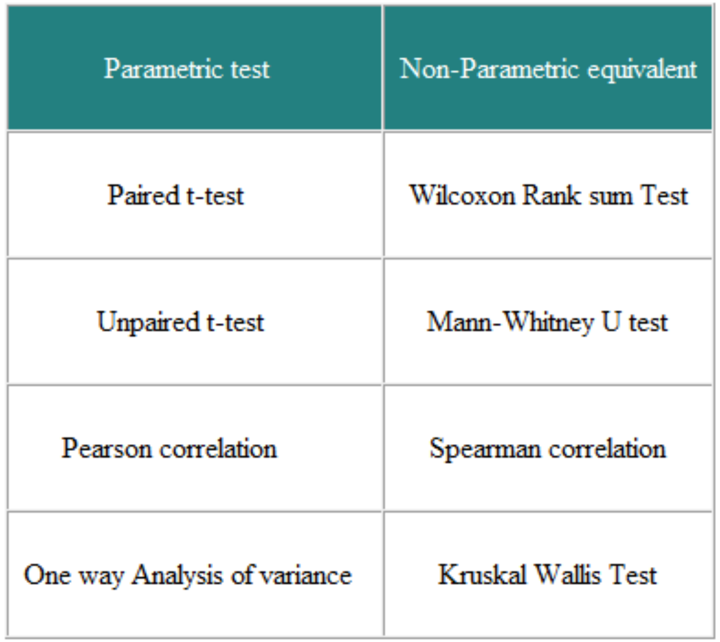
Two-tailed tests

- assume change in general (decrease & increase)

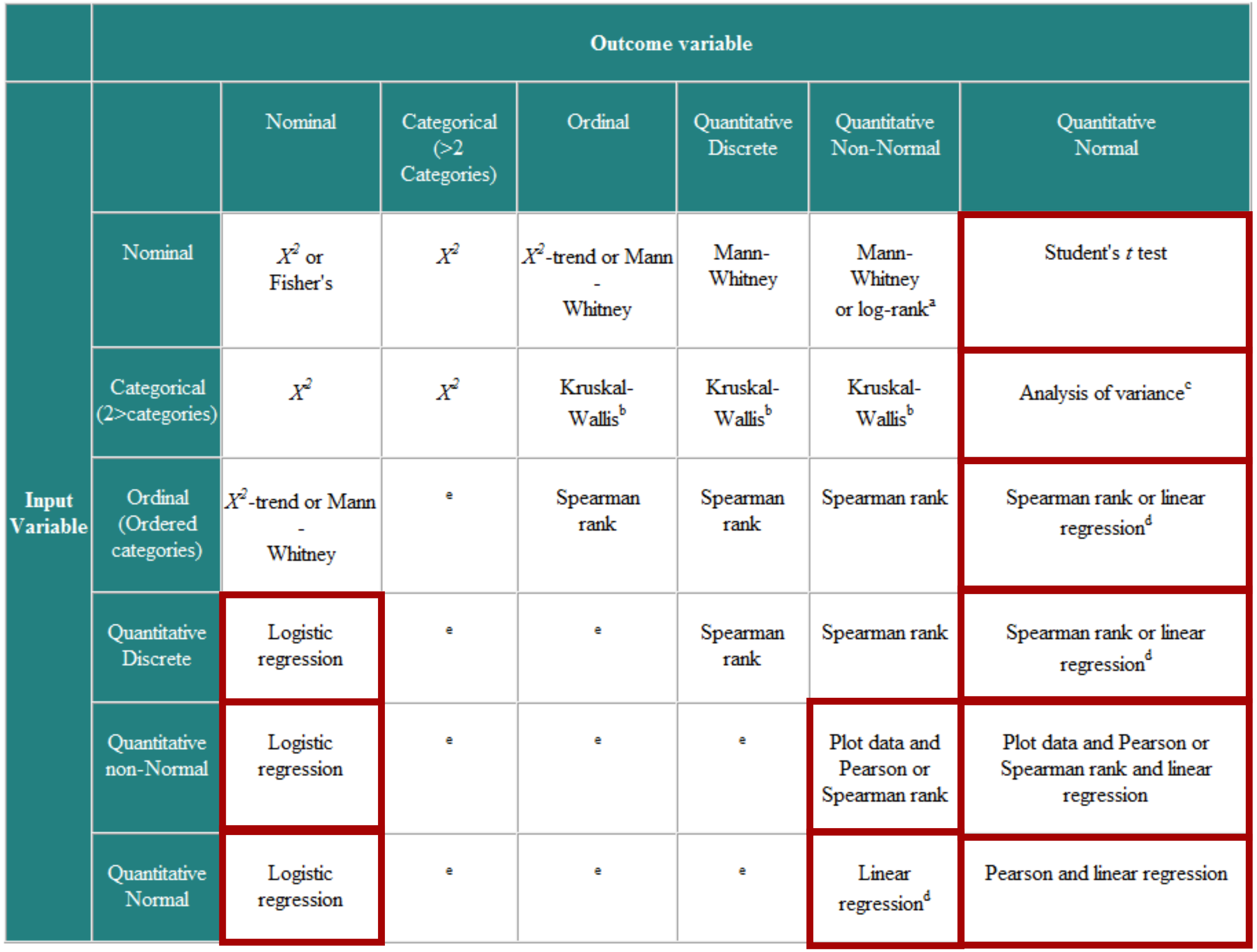
Two-Tailed Versus One-Tailed Hyphothesis Tests



Parametric vs. non-parametric tests



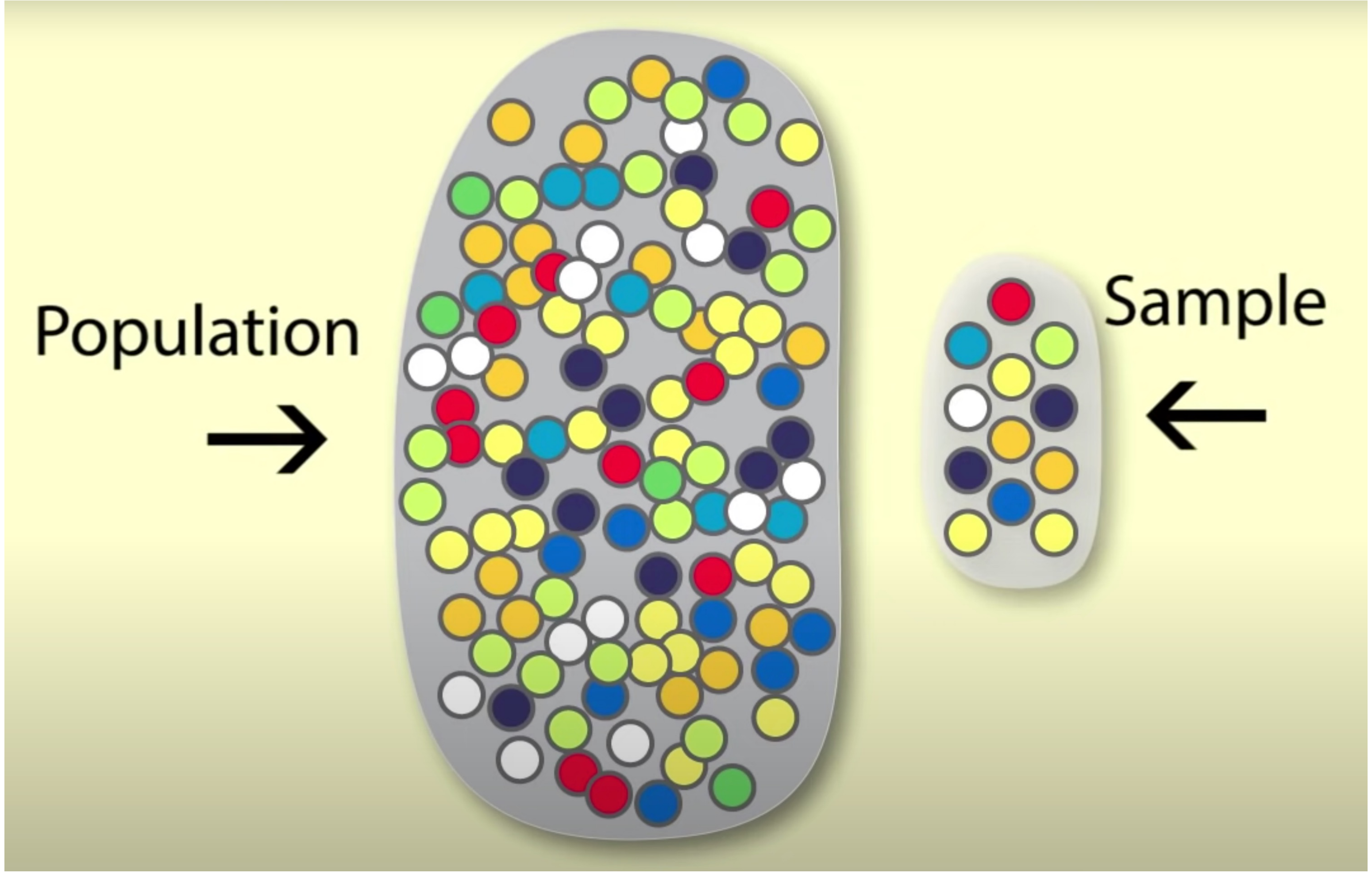
Parametric test	Non-Parametric equivalent
Paired t-test	Wilcoxon Rank sum Test
Unpaired t-test	Mann-Whitney U test
Pearson correlation	Spearman correlation
One way Analysis of variance	Kruskal Wallis Test



	Outcome variable						
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Input Variable	Nominal	X^2 or Fisher's	X^2	X^2-trend or Mann - Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	Student's *t* test
	Categorical (2>categories)	X^2	X^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Analysis of variance[c]
	Ordinal (Ordered categories)	X^2-trend or Mann - Whitney	[e]	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative Discrete	Logistic regression	[e]	[e]	Spearman rank	Spearman rank	Spearman rank or linear regression[d]
	Quantitative non-Normal	Logistic regression	[e]	[e]	[e]	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	[e]	[e]	[e]	Linear regression[d]	Pearson and linear regression

https://www.healthknowledge.org.uk

Selection of study groups



Dr Nic's Maths & Stats YT Channel (https://www.youtube.com/channel/UCG32MfGLit1pcqCRXyy9cAg)

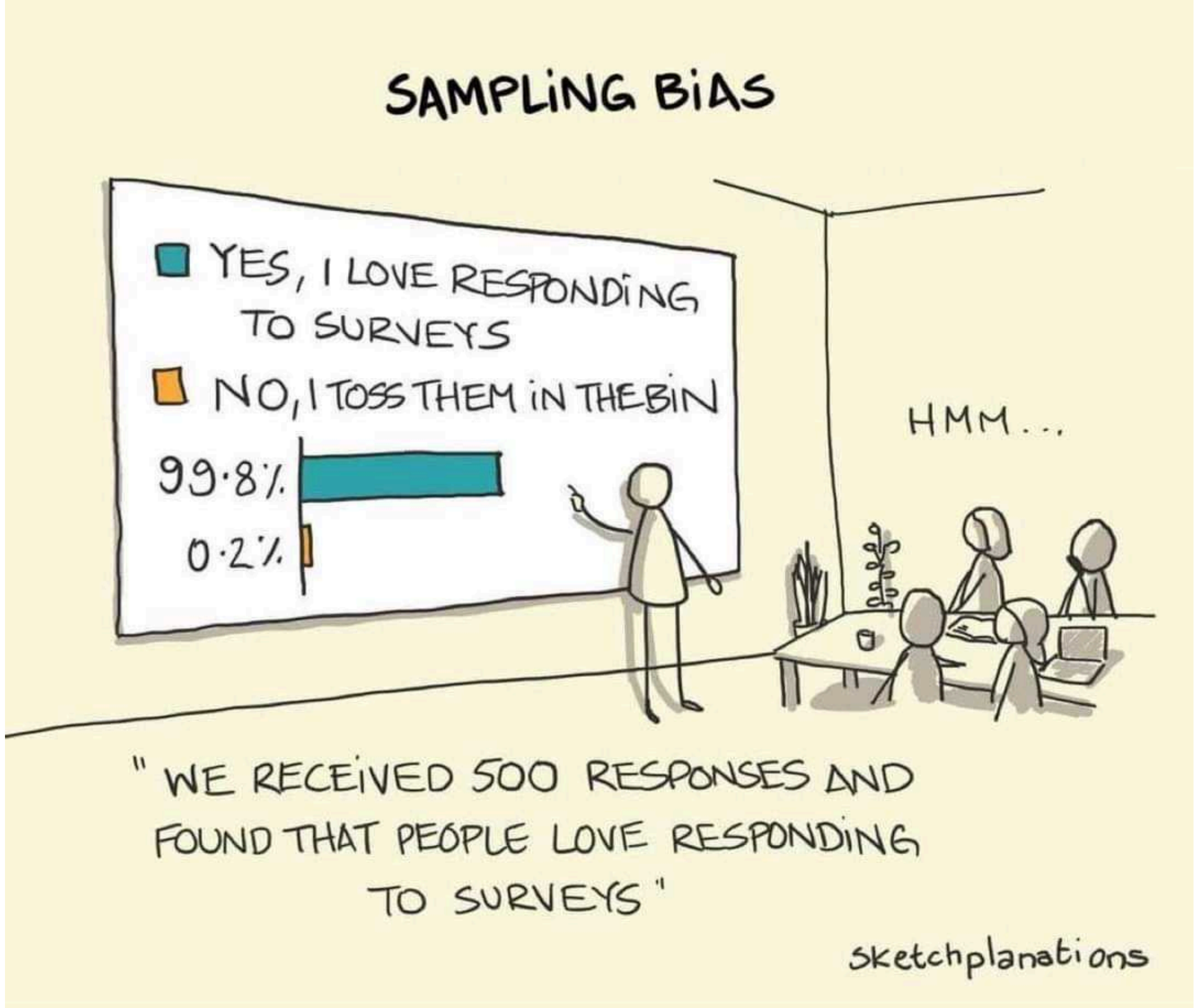
- representative and appropriately selected groups
- inclusion/exclusion criteria (diseases, medications)
- certain degree of uncertainty

- **sampling bias**

Study size estimation

- how large should the study/control group be?
- ethical, financial, temporal, logistical aspects

Sampling bias


SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

Source: Internet

- sampling strategy, in which a certain members of population have higher or lower sampling probability

- if not accounted for during data analysis, it’s effect can be erroneously attributed to the phenomenon under study

- **healthy user bias** (overestimating health of general population)

- **Berkson’s fallacy** (underestimating health of general population)

Study groups selection precautions

- completely randomly selected volunteers; is it even feasible? (control group searched within specific occupation, "healthy" hospital visitors)

- how to match the study and control groups in terms of age, when the examined disease appears only in certain age groups? Will it then be possible to exclude other factors influencing the examined parameters in the compared control group?? (centenarians)

- how to perform randomization if subjects representing the study group are only rarely encountered?

- how to conduct drug effect testing (placebo, single/double blind study, the "noninferiority" problem)

Data collection

Various types of experiments:

- basic research
- population studies
    - observational
        - cohort / case-control studies
        - cross-sectional / monitoring (*longitudinal*: *prospective* / *retrospective*)
    - interventional (experimental)
- clinical trials
- questionnaire studies

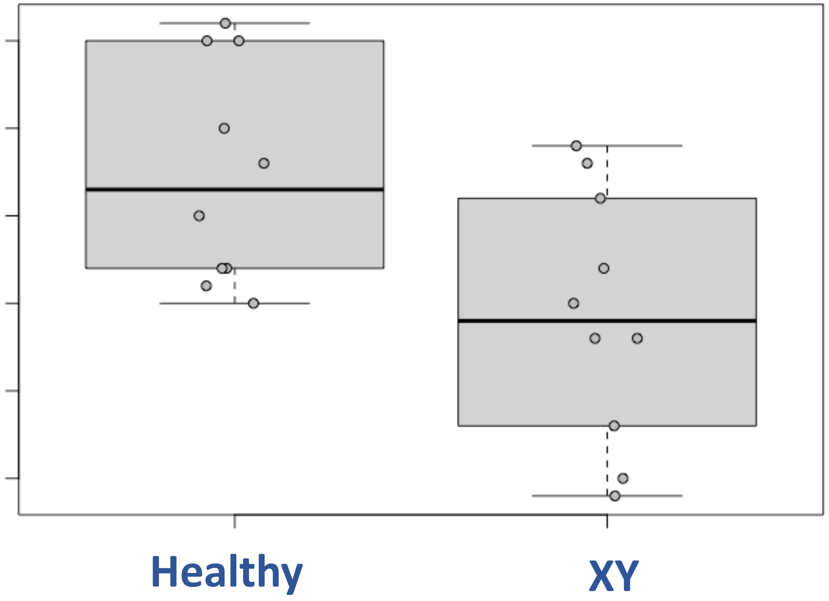
- ***independence, randomness of all observations***

Hypothesis testing

statistical tests

- observed differences between the groups are just by chance or rather indicate a kind of regularity (pattern)
- transform observed differences into *statistics*

remember to check the data distribution first!



$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{(1/n_1 + 1/n_2)}}$$

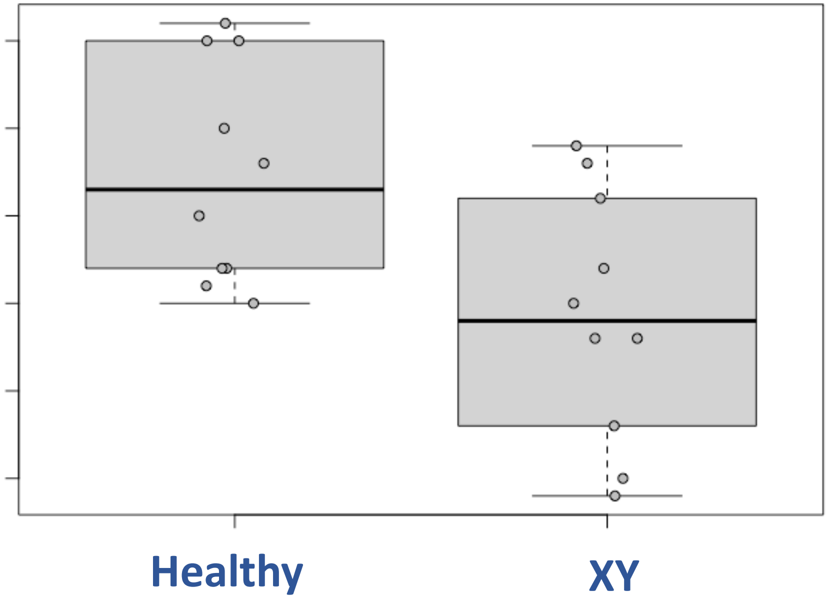
$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right]}$$

Hypothesis testing

Why calculate *t statistic* out of real difference?

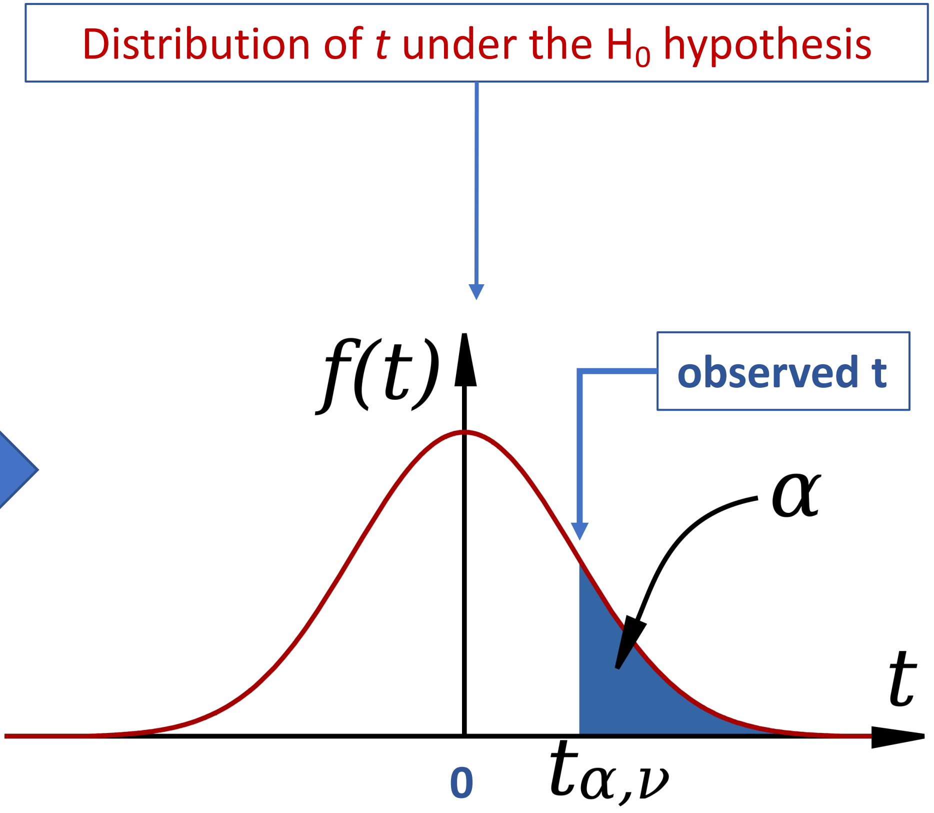
Uncertainty

- exact distribution of IQ is unknown
- the representativeness of groups is unknown



$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{(1/n_1 + 1/n_2)}}$$

$$s = \sqrt{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}\right]}$$



Making decisions

	Real world	
Test (experiment) result	**$x_1 = x_2$** (H_0 is true)	**$x_1 \neq x_2$** (H_A is true)
Reject H_0 **$x_1 \neq x_2$**	**Type I error** (significance) α, p	**Correct decision** (test power)
Do not reject H_0 **$x_1 = x_2$**	**Correct decision** (1- significance)	**Type II error** (1- test power)

Life Sciences
$\boldsymbol{\alpha}$ = 0.05; $\boldsymbol{\alpha}$ = 0.01

Physical sciences
$\boldsymbol{\alpha}$ = 0.0000003 (Higg's boson)

The higher the test power, the better
The lower the significance level, the better (?!?)

What is the result?

- the effect size (i.e. difference between means)
- level of significance (the degree of our (un)certainty)

Group size estimation

In order to generalize the study/control groups must be of required sizes.

- required study group size (**n**) depends on several parameters
  - statistical power (the ability to reject H_0 when it is false; 80%; $\boldsymbol{z_\beta}$)
  - level of significance (the probability of rejecting H_0 when it is *de facto* true; 5%; $\boldsymbol{z_\alpha}$)
  - effect size (i. e. assumed difference in mean values ($\boldsymbol{\delta}$))
  - assumed spread (e. g. variance; $\boldsymbol{\sigma^2}$)

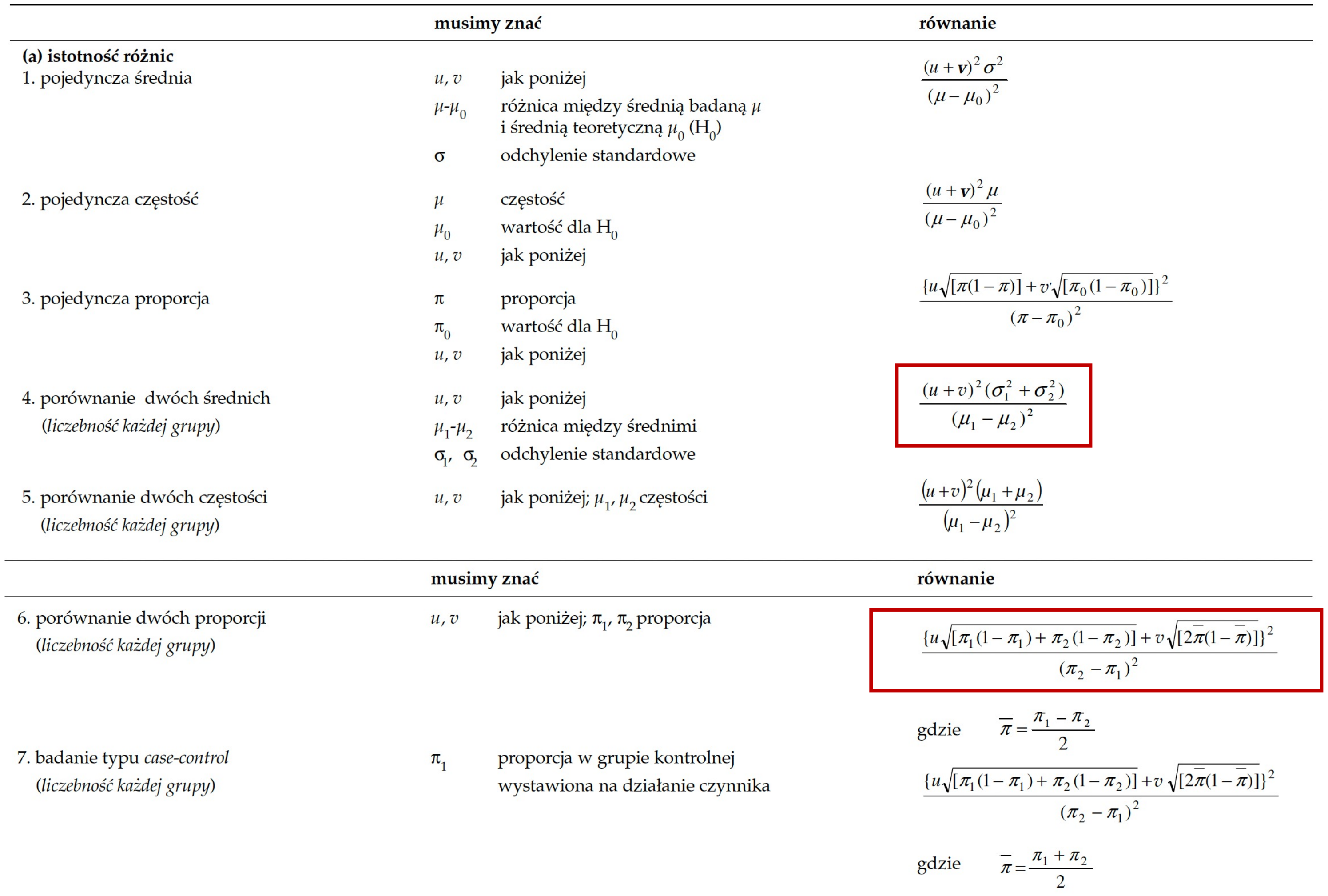
$$z_\beta = \frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}} - z_{\alpha/2} \quad \Rightarrow \quad \boxed{n \approx \frac{\sigma^2\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2}}$$

  - larger the spread ➔ higher ***n***
  - smaller the difference ➔ higher ***n***

Group size estimation

$$n \approx \frac{\sigma^2\left(z_\beta + z_{\alpha/2}\right)^2}{\delta^2}$$

- independent observations
- numerical data
- trying to prove the difference between groups



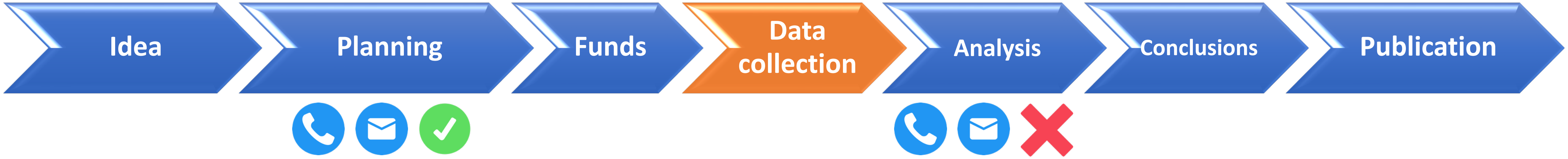
	musimy znać		równanie
(a) istotność różnic			
1. pojedyncza średnia	u, v	jak poniżej	$\frac{(u+\mathbf{v})^2\sigma^2}{(\mu-\mu_0)^2}$
	μ-μ_0	różnica między średnią badaną μ i średnią teoretyczną μ_0 (H_0)	
	σ	odchylenie standardowe	
2. pojedyncza częstość	μ	częstość	$\frac{(u+\mathbf{v})^2\mu}{(\mu-\mu_0)^2}$
	μ_0	wartość dla H_0	
	u, v	jak poniżej	
3. pojedyncza proporcja	π	proporcja	$\frac{\{u\sqrt{[\pi(1-\pi)]}+v'\sqrt{[\pi_0(1-\pi_0)]}\}^2}{(\pi-\pi_0)^2}$
	π_0	wartość dla H_0	
	u, v	jak poniżej	
4. porównanie dwóch średnich	u, v	jak poniżej	$\boxed{\frac{(u+v)^2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}}$
(liczebność każdej grupy)	μ_1-μ_2	różnica między średnimi	
	σ_1, σ_2	odchylenie standardowe	
5. porównanie dwóch częstości	u, v	jak poniżej; μ_1, μ_2 częstości	$\frac{(u+v)^2(\mu_1+\mu_2)}{(\mu_1-\mu_2)^2}$
(liczebność każdej grupy)			
6. porównanie dwóch proporcji	u, v	jak poniżej; π_1, π_2 proporcja	$\boxed{\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2-\pi_1)^2}}$
(liczebność każdej grupy)			gdzie $\bar{\pi}=\frac{\pi_1-\pi_2}{2}$
7. badanie typu *case-control*	π_1	proporcja w grupie kontrolnej wystawiona na działanie czynnika	$\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2-\pi_1)^2}$
(liczebność każdej grupy)			gdzie $\bar{\pi}=\frac{\pi_1+\pi_2}{2}$

Watała: Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych. α-medica press, Bielsko-Biała, 2002.

What should you take from this part?

1. **Plan** your experiments properly (in cooperation with a specialist)
2. **Check** your data prior to analysis (if it meets assumptions of tests)
3. Use **appropriate measures** of centrality and spread
4. Don't be afraid to use **other than parametric tests**
5. The **p-value is not a result** of experiment!
   It is the effect size (difference), while the p-value tells something about how certain you are when it comes to the effect size.

6. Contact **CABiB** if you need help



Idea → Planning → Funds → Data collection → Analysis → Conclusions → Publication

Planning: ☎ ✉ ✓

Analysis: ☎ ✉ ✗

Selected practical issues

Group size estimation

- determine a number of factors ***a priori***
  - the level of significance
  - statistical power of the test
  - the effect size
  - the dispersion measure (variance)

“to start from something is better than starting from nothing”

How to formulate the question concerning the group size?

✗ „ So how many subjects do I need to include in the study to obtain statistically significant difference?”

✓ „Assuming the variability of data at the level of 25%, the level of statistical significance below 0.01, statistical power above 80% and the planned use of one-tailed Student’s t-test, how many patients should be included in the study to show a 15% higher level of biomarker expression in the group subjected to treatment?”

✓ ” If the study and control groups include both 30 subjects per group, then assuming the variability data at the level of 25%, the significance level below 0.01, the statistical power above 80% and the planned use of one-tailed Student's t-test, how large an increase in the biomarker expression level will I be able to evaluate as statistically significant?”

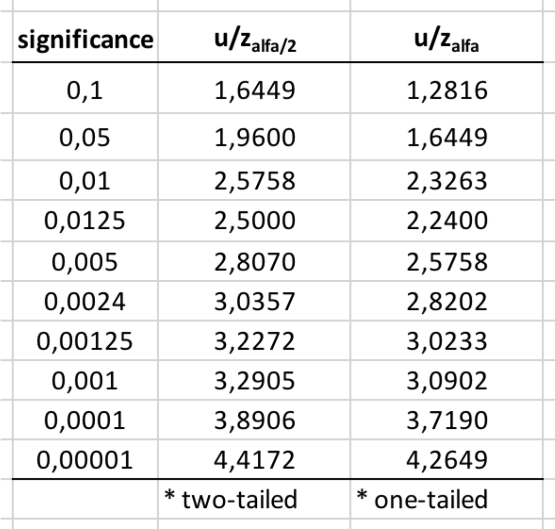
What if we expect a non-Gaussian distribution of the data?

Group size estimation: two means

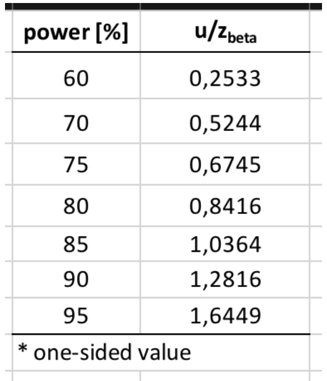
Required information:

- expected means (μ_1; μ_2)
- expected spreads ($\boldsymbol{\sigma}_1$; $\boldsymbol{\sigma}_2$)
- u – a value related to level of significance
- v – a value related to statistical power of test

$$\frac{(u+v)^2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}$$



significance	u/$z_{alfa/2}$	u/z_{alfa}
0,1	1,6449	1,2816
0,05	1,9600	1,6449
0,01	2,5758	2,3263
0,0125	2,5000	2,2400
0,005	2,8070	2,5758
0,0024	3,0357	2,8202
0,00125	3,2272	3,0233
0,001	3,2905	3,0902
0,0001	3,8906	3,7190
0,00001	4,4172	4,2649
	* two-tailed	* one-tailed



power [%]	u/z_{beta}
60	0,2533
70	0,5244
75	0,6745
80	0,8416
85	1,0364
90	1,2816
95	1,6449
* one-sided value	

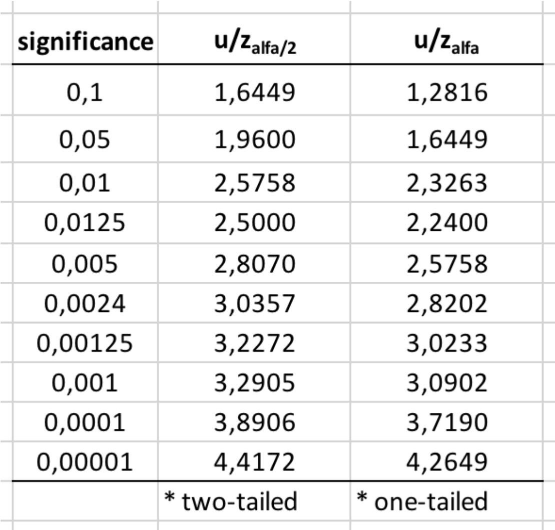
Group size estimation: three means

Required information:

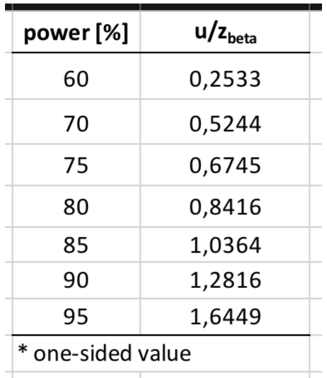
- expected means (μ_1; μ_2 ; μ_3)
- expected spreads ($\boldsymbol{\sigma}_1$; $\boldsymbol{\sigma}_2$; $\boldsymbol{\sigma}_3$)
- u – a value related to level of significance
- v – a value related to statistical power of test

- Bonferroni correction for multiple testing (adjust p)

$$\frac{(u+v)^2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}$$



significance	u/z$_{alfa/2}$	u/z$_{alfa}$
0,1	1,6449	1,2816
0,05	1,9600	1,6449
0,01	2,5758	2,3263
0,0125	2,5000	2,2400
0,005	2,8070	2,5758
0,0024	3,0357	2,8202
0,00125	3,2272	3,0233
0,001	3,2905	3,0902
0,0001	3,8906	3,7190
0,00001	4,4172	4,2649
	* two-tailed	* one-tailed



power [%]	u/z$_{beta}$
60	0,2533
70	0,5244
75	0,6745
80	0,8416
85	1,0364
90	1,2816
95	1,6449
* one-sided value	

Group size estimation: three groups; non-Gaussian distribution

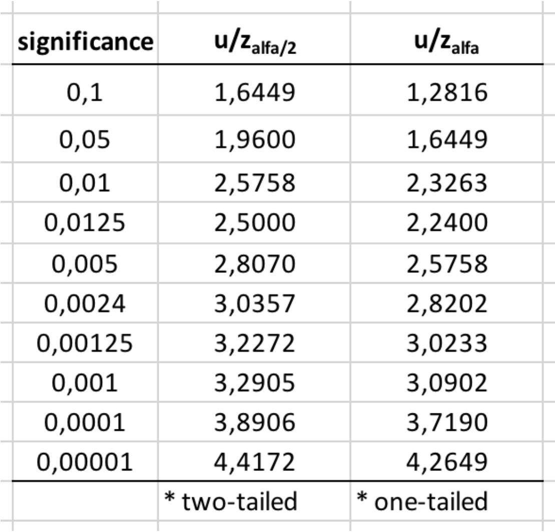
Required information:

- expected means (μ_1; μ_2 ; μ_3)
- expected spreads ($\boldsymbol{\sigma}_1$; $\boldsymbol{\sigma}_2$; $\boldsymbol{\sigma}_3$)
- u – a value related to level of significance
- v – a value related to statistical power of test

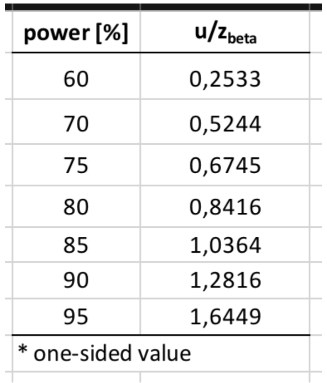
$$\frac{(u+v)^2(\sigma_1^2+\sigma_2^2)}{(\mu_1-\mu_2)^2}$$

- Bonferroni correction for multiple testing (adjust p)

- Add +15% to each group's N



significance	u/z$_{alfa/2}$	u/z$_{alfa}$
0,1	1,6449	1,2816
0,05	1,9600	1,6449
0,01	2,5758	2,3263
0,0125	2,5000	2,2400
0,005	2,8070	2,5758
0,0024	3,0357	2,8202
0,00125	3,2272	3,0233
0,001	3,2905	3,0902
0,0001	3,8906	3,7190
0,00001	4,4172	4,2649
	* two-tailed	* one-tailed



power [%]	u/z$_{beta}$
60	0,2533
70	0,5244
75	0,6745
80	0,8416
85	1,0364
90	1,2816
95	1,6449

* one-sided value

Group size estimation: two proportions

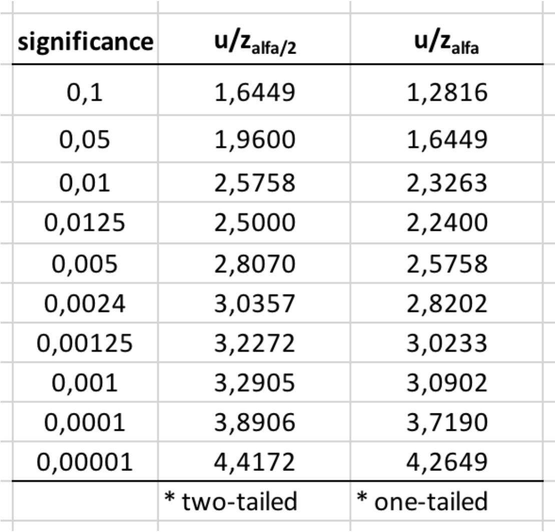
Required information:

- expected proportions (π_1; π_2)

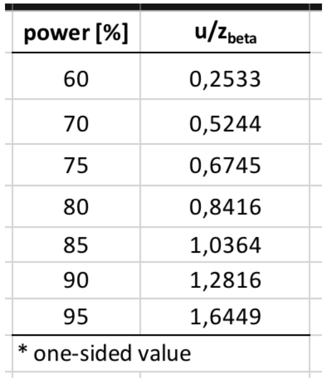
- OR, alternatively:
- expected proportion π_1 and effect size (OR, RR, …)

- u – a value related to level of significance
- v – a value related to statistical power of test

$$\frac{\{u\sqrt{[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}+v\sqrt{[2\bar{\pi}(1-\bar{\pi})]}\}^2}{(\pi_2-\pi_1)^2}$$



significance	$u/z_{alfa/2}$	u/z_{alfa}
0,1	1,6449	1,2816
0,05	1,9600	1,6449
0,01	2,5758	2,3263
0,0125	2,5000	2,2400
0,005	2,8070	2,5758
0,0024	3,0357	2,8202
0,00125	3,2272	3,0233
0,001	3,2905	3,0902
0,0001	3,8906	3,7190
0,00001	4,4172	4,2649
	* two-tailed	* one-tailed



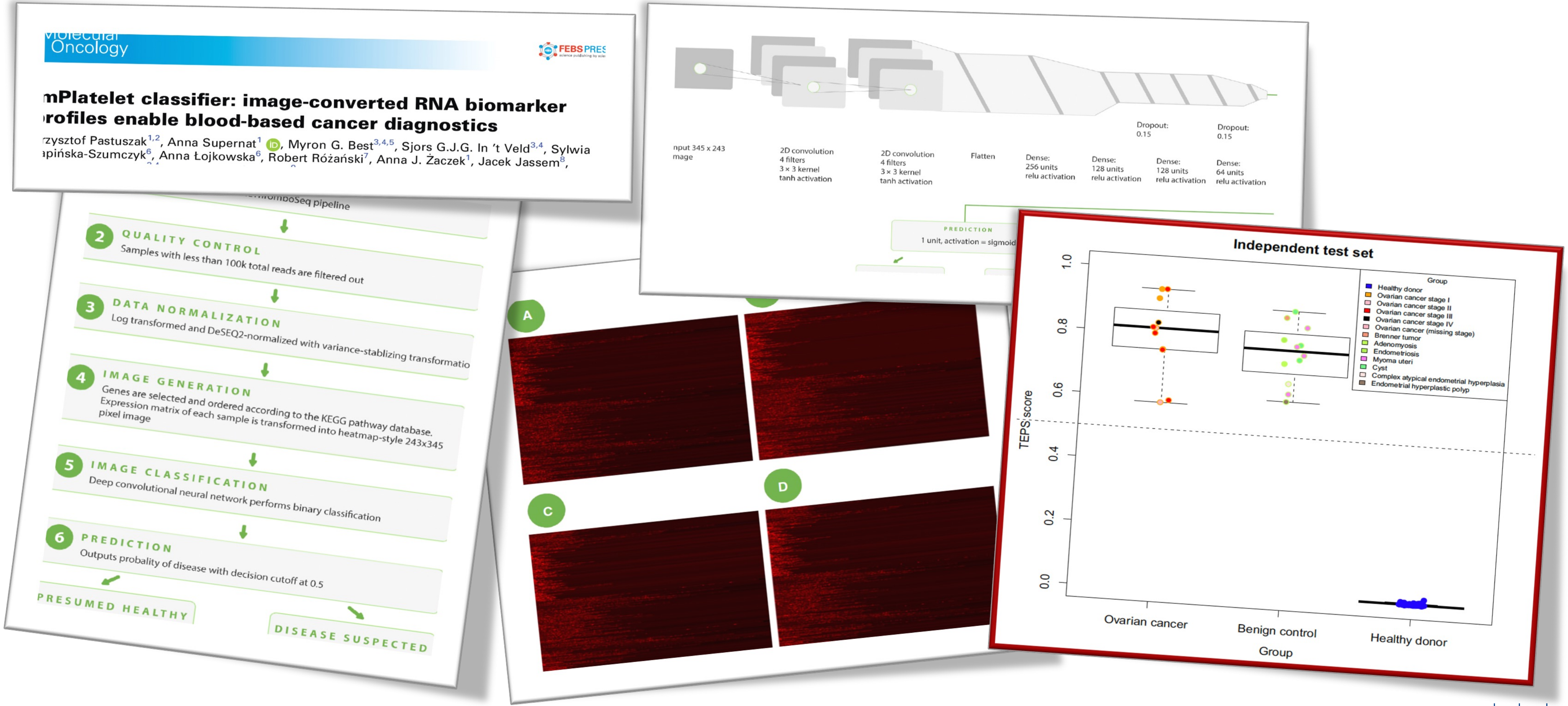
power [%]	u/z_{beta}
60	0,2533
70	0,5244
75	0,6745
80	0,8416
85	1,0364
90	1,2816
95	1,6449

* one-sided value

	Females	*Males*	*Total*
Smoking	**10**	**90**	*100*
Non-smoking	**110**	**30**	*140*
Total	*120*	*120*	*240*

Group size estimation: more complex situation

Calculating the required study size for testing the new diagnostic tool (e.g. AI-based classifier)



Molecular Oncology

FEBS PRESS

mPlatelet classifier: image-converted RNA biomarker profiles enable blood-based cancer diagnostics

zysztof Pastuszak[1,2], Anna Supernat[1], Myron G. Best[3,4,5], Sjors G.J.G. In 't Veld[3,4], Sylwia apińska-Szumczyk[6], Anna Łojkowska[6], Robert Różański[7], Anna J. Żaczek[1], Jacek Jassem[8],

romboSeq pipeline

2 QUALITY CONTROL
Samples with less than 100k total reads are filtered out

3 DATA NORMALIZATION
Log transformed and DeSEQ2-normalized with variance-stablizing transformatio

4 IMAGE GENERATION
Genes are selected and ordered according to the KEGG pathway database. Expression matrix of each sample is transformed into heatmap-style 243x345 pixel image

5 IMAGE CLASSIFICATION
Deep convolutional neural network performs binary classification

6 PREDICTION
Outputs probality of disease with decision cutoff at 0.5

PRESUMED HEALTHY

DISEASE SUSPECTED

Pastuszak. Molecular Oncology (2021) doi: 10.1002/1878-0261.13014

Group size estimation: more complex situation

Calculating the required study size for testing the new diagnostic tool (e.g. AI-based classifier)

- **paired** data
- data as **proportions**
- trying to prove **noninferiority** of the new diagnostic tool (that it is not worse compared to standard test)

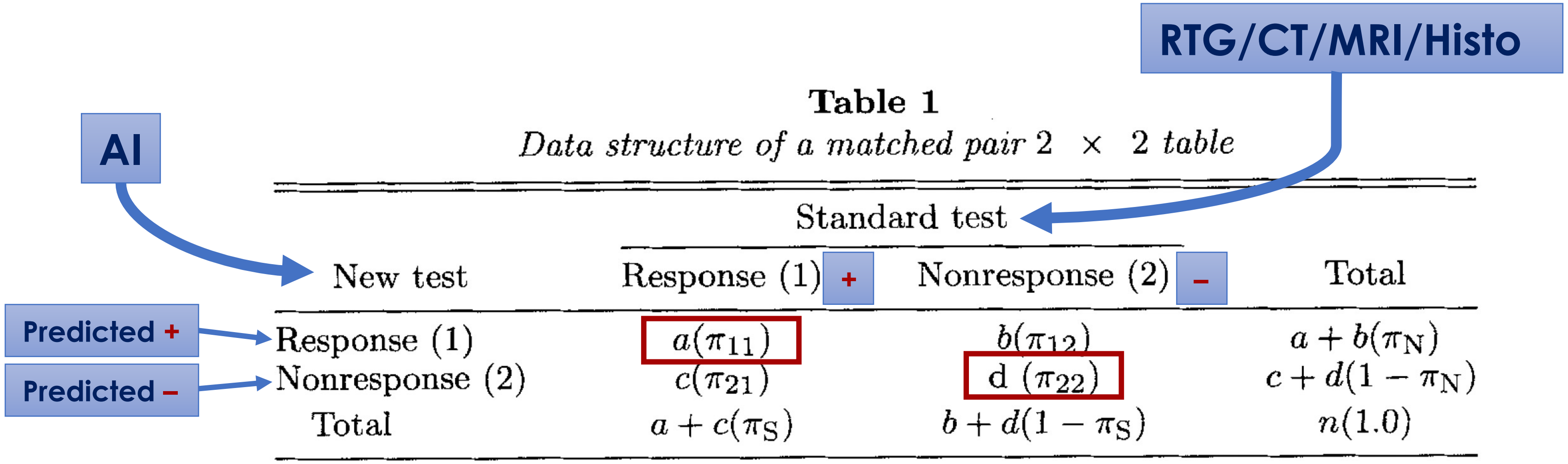

RTG/CT/MRI/Histo
AI
Predicted +
Predicted –
+
–

Table 1

Data structure of a matched pair 2 × 2 table

	Standard test		
New test	Response (1)	Nonresponse (2)	Total
Response (1)	$a(\pi_{11})$	$b(\pi_{12})$	$a + b(\pi_N)$
Nonresponse (2)	$c(\pi_{21})$	d (π_{22})	$c + d(1 - \pi_N)$
Total	$a + c(\pi_S)$	$b + d(1 - \pi_S)$	$n(1.0)$

Group size estimation: more complex situation

$$n_{TS} = \left\{ \frac{z_{(1-\alpha)}\sqrt{(1+\delta_0)\bar{\pi}_{21} + (\delta_1\pi_S + \pi_{21})(\delta_0 - 1)}}{(\delta_1 - \delta_0)\pi_S} \right.$$

$$+ z_{(1-\beta)}\left\{2\delta_0\pi_{21} - \delta_0(1-\delta_0)\pi_S \right.$$

$$\left. + (\delta_1 - \delta_0)\pi_S[1 - (\delta_1 - \delta)\pi_S]\right\}^{1/2}$$

$$\left. \div (\delta_1 - \delta_0)\pi_S \right\}^2 .$$

$$\pi_{21}: \quad \begin{array}{l} \min[(2-\delta_1)\pi_S/2,\ (1+\pi_S)/2 - \delta_1\pi_S] \text{ for } \delta_1 \leq 1 \\ \min[(1-\delta_1\pi_S)/2,\ \pi_S/2] \text{ for } 1 < \delta_1 \leq 1/\pi_S \end{array}$$

BIOMETRICS 58, 957–963
December 2002

Sample Size Determination for Establishing Equivalence/Noninferiority via Ratio of Two Proportions in Matched-Pair Design

Man-Lai Tang,[1,*] Nian-Sheng Tang,[2] Ivan Siu-Fung Chan,[3] and Ben Ping-Shing Chan[4]

Group size estimation: more complex situation

GDAŃSKI UNIWERSYTET MEDYCZNY

Circulating tumour cells mRNA transcriptomics study

Q: how many reads should a differentially-expressed miRNA transcript have to provide reliable results?

Assumptions

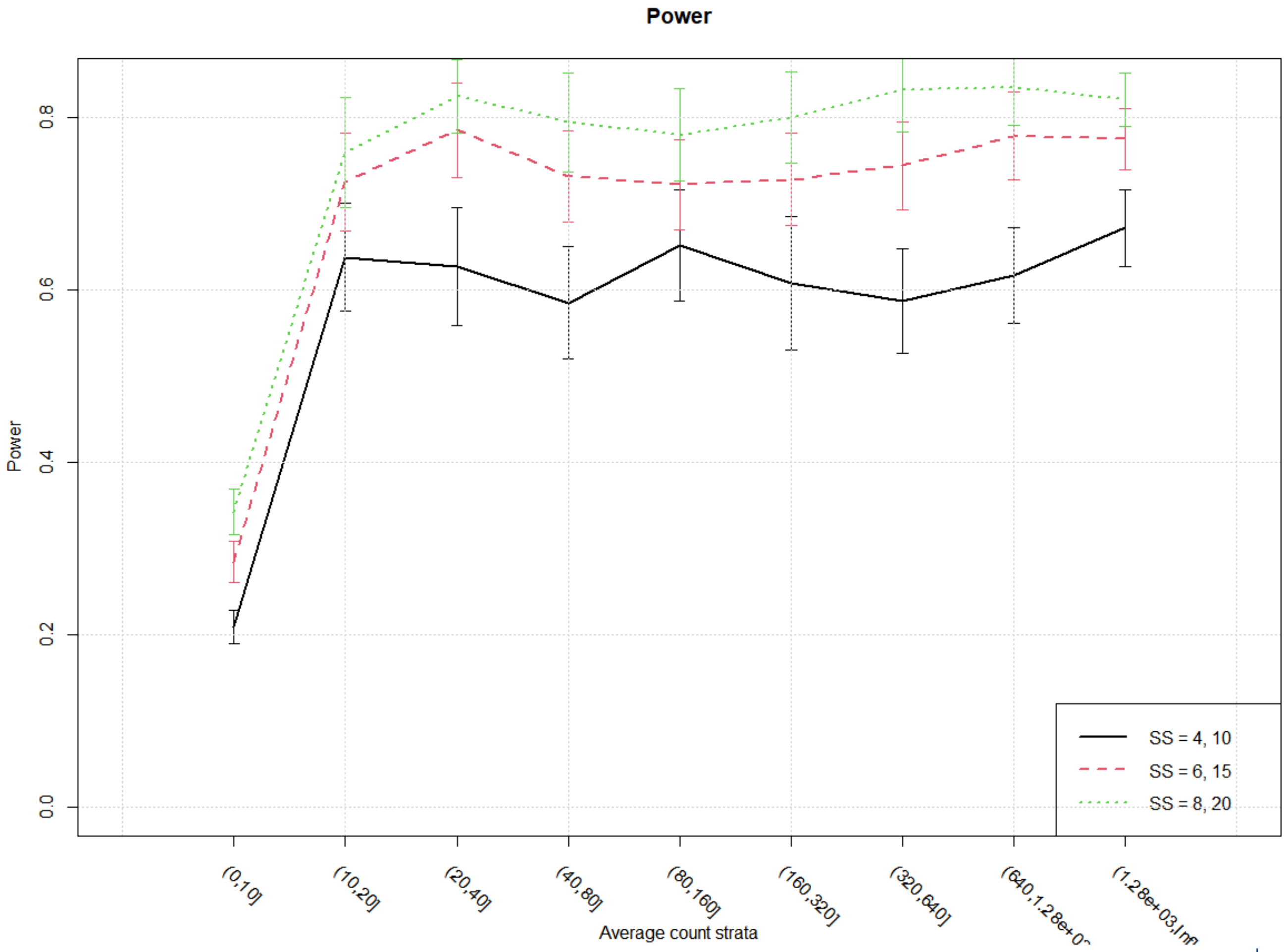
- *N = 5000 transcripts;*
- *DEG = 0.1; FDR = 0.1*
- *log(FC) = 0.5 (~3.16x)*

- *4 and 10 subjects per group*
- *6 and 15 subjects per group*
- *8 and 20 subjects per group*

What is the statistical power?

simulation: 50 iterations

*(**R**; proper package) (free)*



Group size estimation: dynamic processes

GDAŃSKI UNIWERSYTET MEDYCZNY

Clinical trial assessing the effectiveness of stem cell wound dressings against commercially available dressings

Literature

- ***Area = 6.84 * e$^{-0.124 * T}$***
- ***k_T = -0.124 week^{-1}***
- *approx. 50% area reduction each 6 weeks*

Assumptions

- *10% faster healing (**k_T = -0.138 week^{-1}**)*
- *80% statistical power*
- *0.05 level of significance*
- *one-sided tests*

What is the required group size?

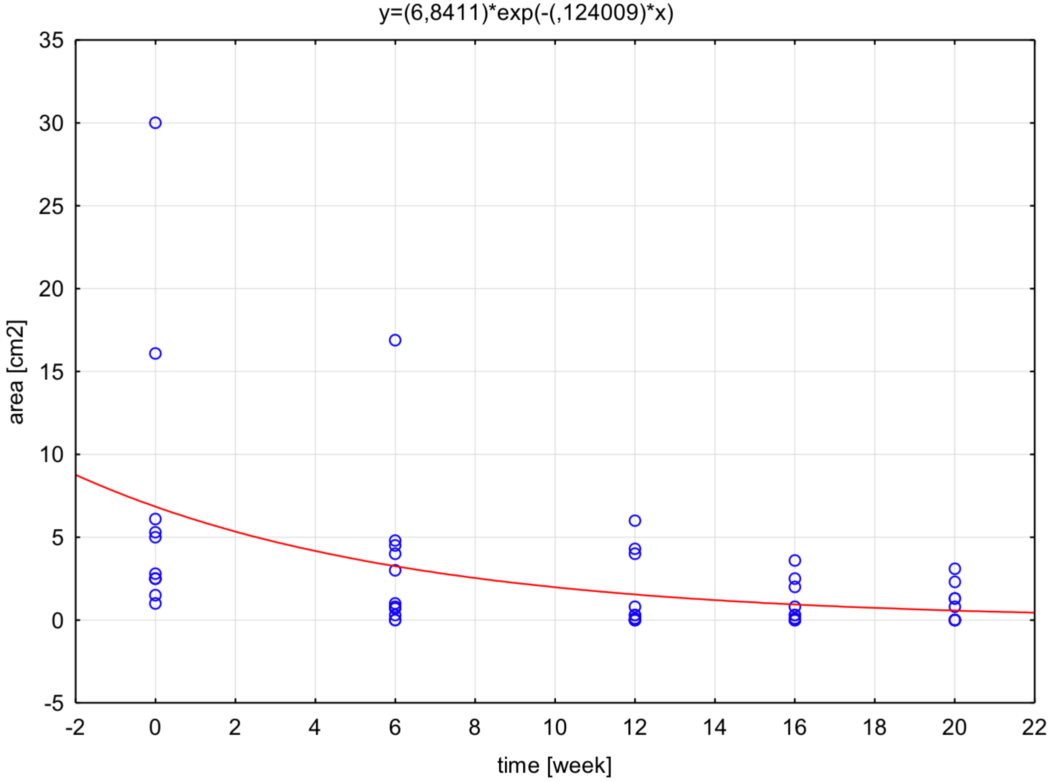
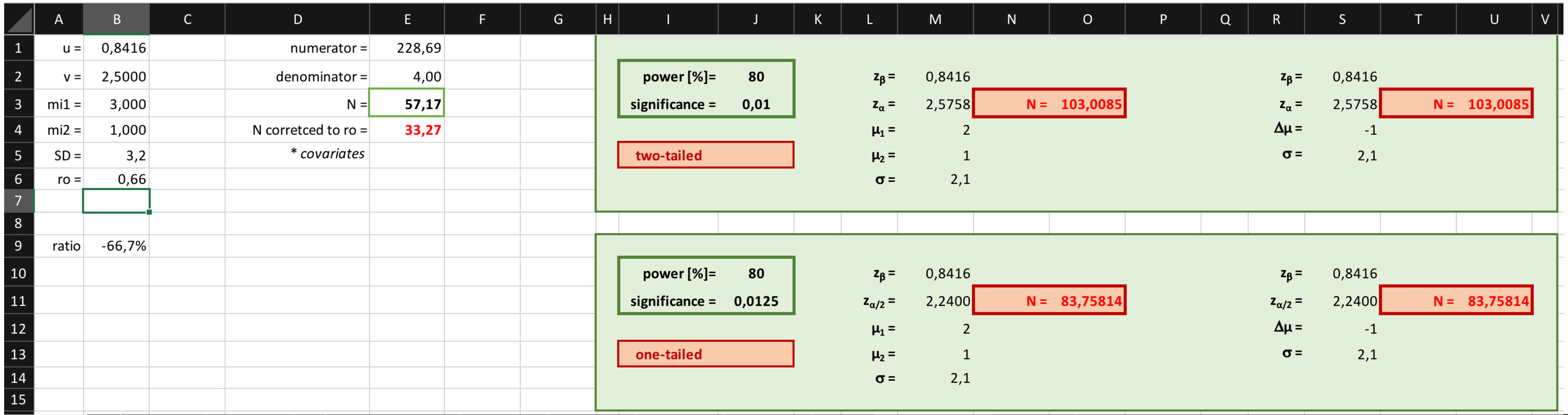


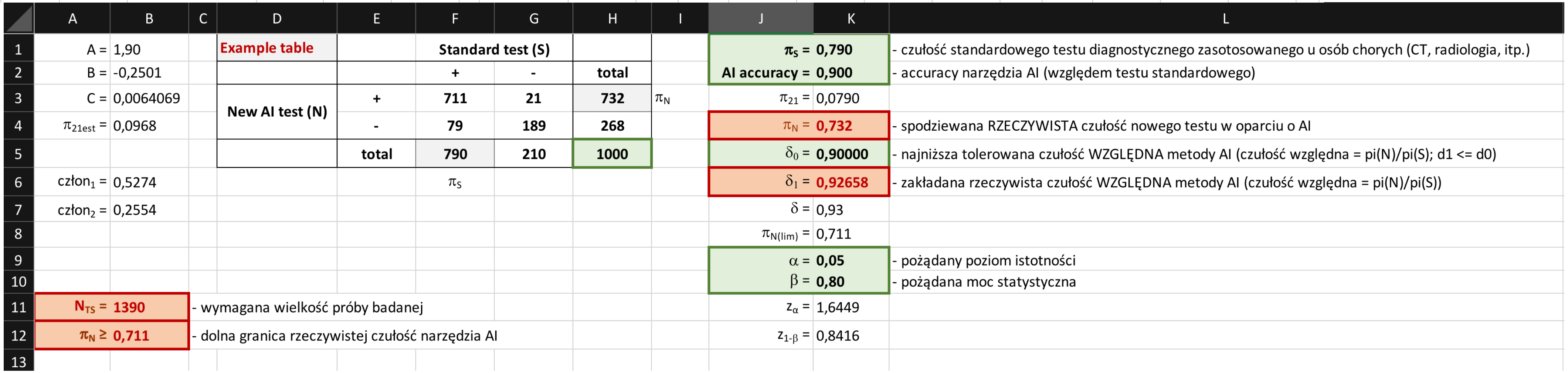
Fig. 1. Wound surface area reduction following the use of commercially available dressing [1]

Group size estimation: DIY solutions

GDAŃSKI UNIWERSYTET MEDYCZNY



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	u =	0,8416		numerator =	228,69																	
2	v =	2,5000		denominator =	4,00				power [%]=	80		$z_β$ =	0,8416					$z_β$ =	0,8416			
3	mi1 =	3,000		N =	57,17				significance =	0,01		$z_α$ =	2,5758	N =	103,0085			$z_α$ =	2,5758	N =	103,0085	
4	mi2 =	1,000		N corretced to ro =	33,27							$μ_1$ =	2					Δμ =	-1			
5	SD =	3,2		* *covariates*					two-tailed			$μ_2$ =	1					σ =	2,1			
6	ro =	0,66										σ =	2,1									
7																						
8																						
9	ratio	-66,7%																				
10									power [%]=	80		$z_β$ =	0,8416					$z_β$ =	0,8416			
11									significance =	0,0125		$z_{α/2}$ =	2,2400	N =	83,75814			$z_{α/2}$ =	2,2400	N =	83,75814	
12												$μ_1$ =	2					Δμ =	-1			
13									one-tailed			$μ_2$ =	1					σ =	2,1			
14												σ =	2,1									
15																						

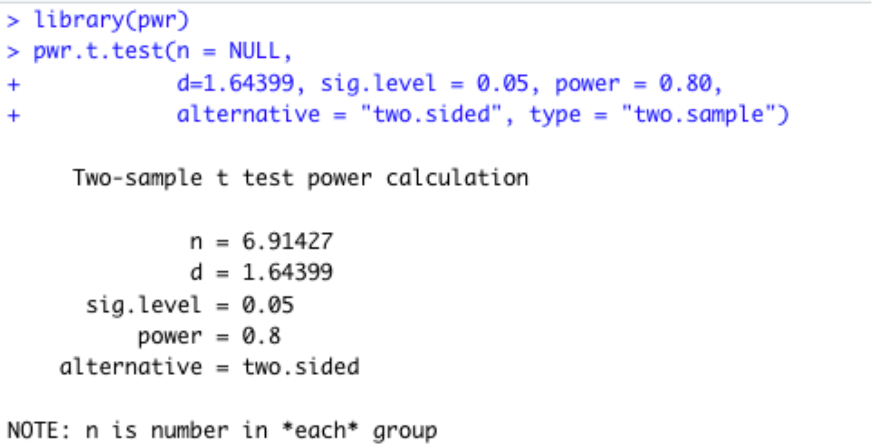


	A	B	C	D	E	F	G	H	I	J	K	L
1	A =	1,90		Example table		Standard test (S)				$π_S$ =	0,790	- czułość standardowego testu diagnostycznego zasotosowanego u osób chorych (CT, radiologia, itp.)
2	B =	-0,2501				+	-	total		AI accuracy =	0,900	- accuracy narzędzia AI (względem testu standardowego)
3	C =	0,0064069		New AI test (N)	+	711	21	732	$π_N$	$π_{21}$ =	0,0790	
4	$π_{21est}$ =	0,0968			-	79	189	268		$π_N$ =	0,732	- spodziewana RZECZYWISTA czułość nowego testu w oparciu o AI
5					total	790	210	1000		$δ_0$ =	0,90000	- najniższa tolerowana czułość WZGLĘDNA metody AI (czułość względna = pi(N)/pi(S); d1 <= d0)
6	człon$_1$ =	0,5274				$π_S$				$δ_1$ =	0,92658	- zakładana rzeczywista czułość WZGLĘDNA metody AI (czułość względna = pi(N)/pi(S))
7	człon$_2$ =	0,2554								δ =	0,93	
8										$π_{N(lim)}$ =	0,711	
9										α =	0,05	- pożądany poziom istotności
10										β =	0,80	- pożądana moc statystyczna
11	N_{TS} =	1390	- wymagana wielkość próby badanej							$z_α$ =	1,6449	
12	$π_N$ ≥	0,711	- dolna granica rzeczywistej czułość narzędzia AI							$z_{1-β}$ =	0,8416	
13												

Group size estimation: available software solutions

- **G*Power** (free)
- **R** (*pwr package*) (free)

- *a priori vs. a posteriori*



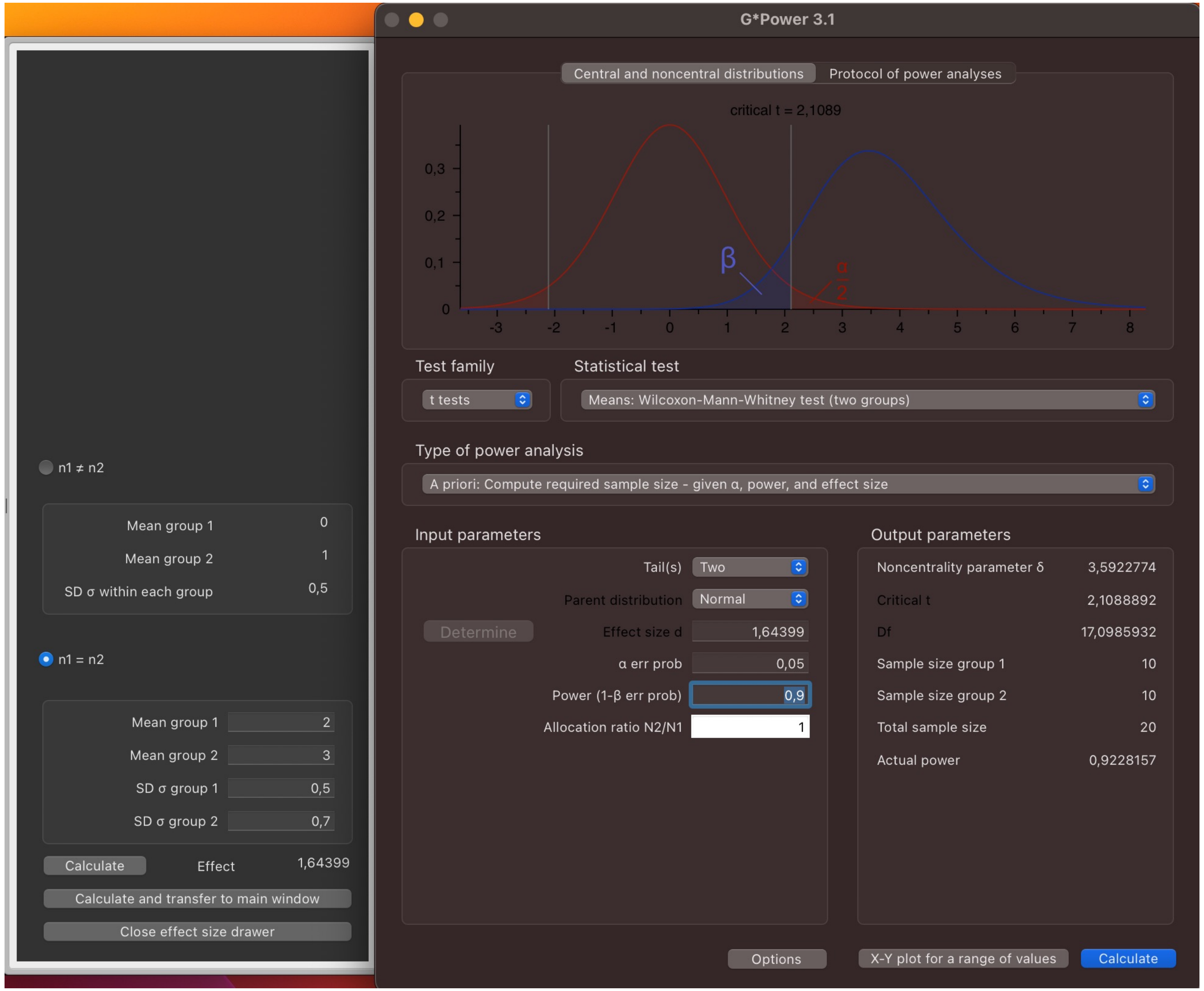
```
> library(pwr)
> pwr.t.test(n = NULL,
+            d=1.64399, sig.level = 0.05, power = 0.80,
+            alternative = "two.sided", type = "two.sample")

     Two-sample t test power calculation

              n = 6.91427
              d = 1.64399
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

pwr-package 2
cohen.ES 3
ES.h 4
ES.w1 5
ES.w2 6
plot.power.htest 7
pwr.2p.test 8
pwr.2p2n.test 9
pwr.anova.test 10
pwr.chisq.test 11
pwr.f2.test 13
pwr.norm.test 14
pwr.p.test 15
pwr.r.test 17
pwr.t.test 18
pwr.t2n.test 20



G*Power 3.1
Central and noncentral distributions
Protocol of power analyses
critical t = 2,1089
0,3
0,2
0,1
0
-3
-2
-1
0
1
2
3
4
5
6
7
8
β
α/2
Determine
Calculate
Calculate and transfer to main window
Close effect size drawer
Options
X-Y plot for a range of values
Calculate

○ n1 ≠ n2

Mean group 1 0
Mean group 2 1
SD σ within each group 0,5

● n1 = n2

Mean group 1 2
Mean group 2 3
SD σ group 1 0,5
SD σ group 2 0,7

Effect 1,64399

Test family: t tests

Statistical test: Means: Wilcoxon-Mann-Whitney test (two groups)

Type of power analysis: A priori: Compute required sample size - given α, power, and effect size

Input parameters

Tail(s) Two
Parent distribution Normal
Effect size d 1,64399
α err prob 0,05
Power (1-β err prob) 0,9
Allocation ratio N2/N1 1

Output parameters

Noncentrality parameter δ 3,5922774
Critical t 2,1088892
Df 17,0985932
Sample size group 1 10
Sample size group 2 10
Total sample size 20
Actual power 0,9228157

Group size estimation: survival analysis

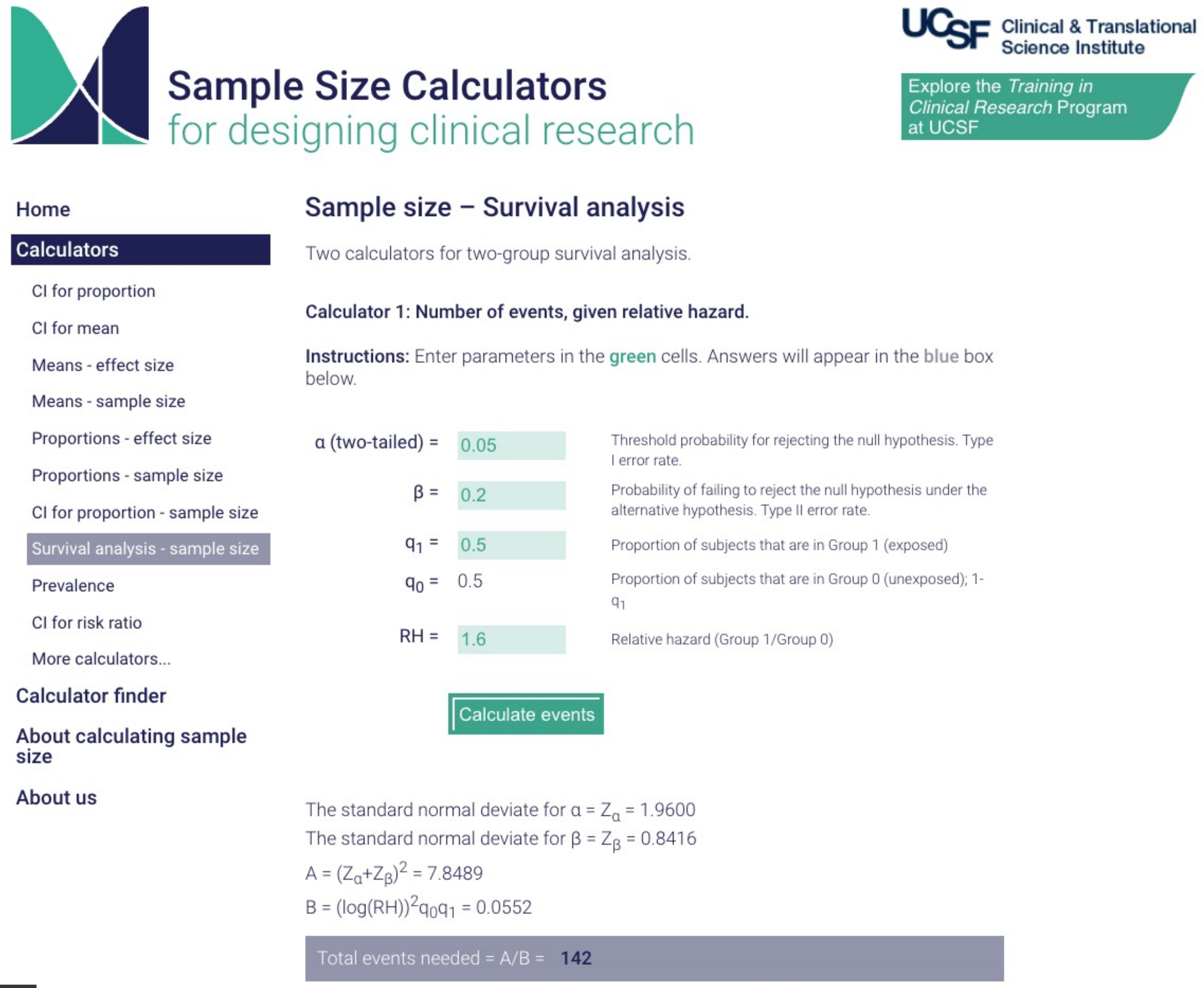
Required information

- expected Hazard Ratio (HR)
- proportion of exposed/unexposed subjects

- *u* – a value related to level of significance
- *v* – a value related to statistical power of test

Output

- required number of events (!!!)


Sample Size Calculators
for designing clinical research
UCSF Clinical & Translational Science Institute
Explore the *Training in Clinical Research* Program at UCSF
Home
Calculators
CI for proportion
CI for mean
Means - effect size
Means - sample size
Proportions - effect size
Proportions - sample size
CI for proportion - sample size
Survival analysis - sample size
Prevalence
CI for risk ratio
More calculators...
Calculator finder
About calculating sample size
About us
Calculate events

Sample size – Survival analysis

Two calculators for two-group survival analysis.

Calculator 1: Number of events, given relative hazard.

Instructions: Enter parameters in the **green** cells. Answers will appear in the blue box below.

α (two-tailed) =	0.05	Threshold probability for rejecting the null hypothesis. Type I error rate.
β =	0.2	Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.
q_1 =	0.5	Proportion of subjects that are in Group 1 (exposed)
q_0 =	0.5	Proportion of subjects that are in Group 0 (unexposed); 1-q_1
RH =	1.6	Relative hazard (Group 1/Group 0)

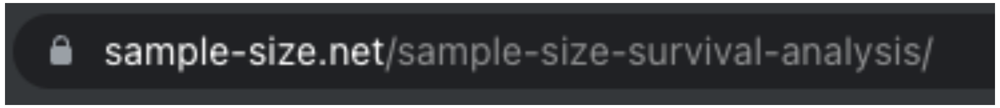
The standard normal deviate for α = Z_α = 1.9600

The standard normal deviate for β = Z_β = 0.8416

A = $(Z_\alpha+Z_\beta)^2$ = 7.8489

B = $(\log(RH))^2 q_0 q_1$ = 0.0552

Total events needed = A/B = **142**

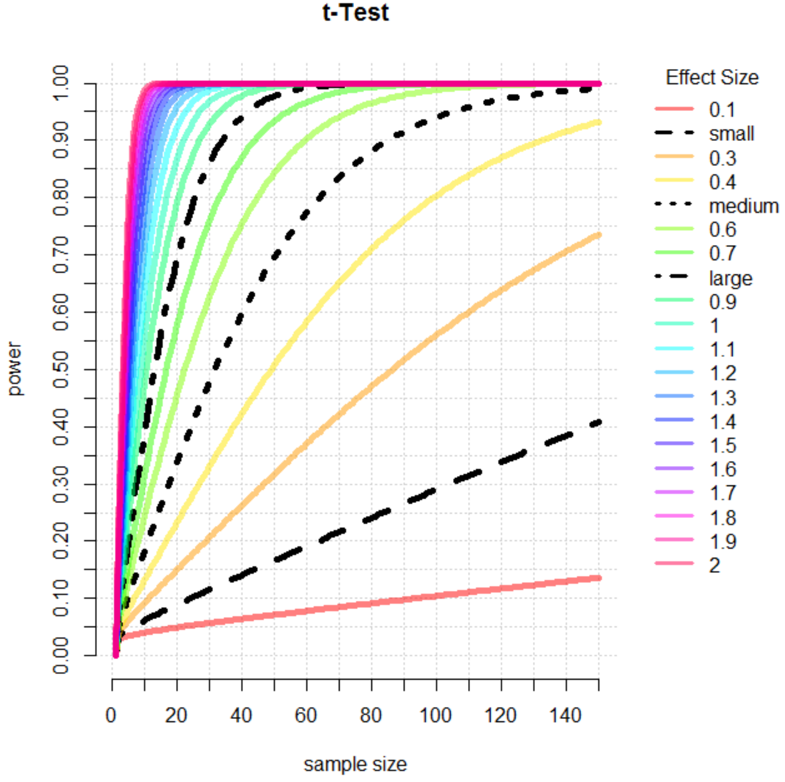


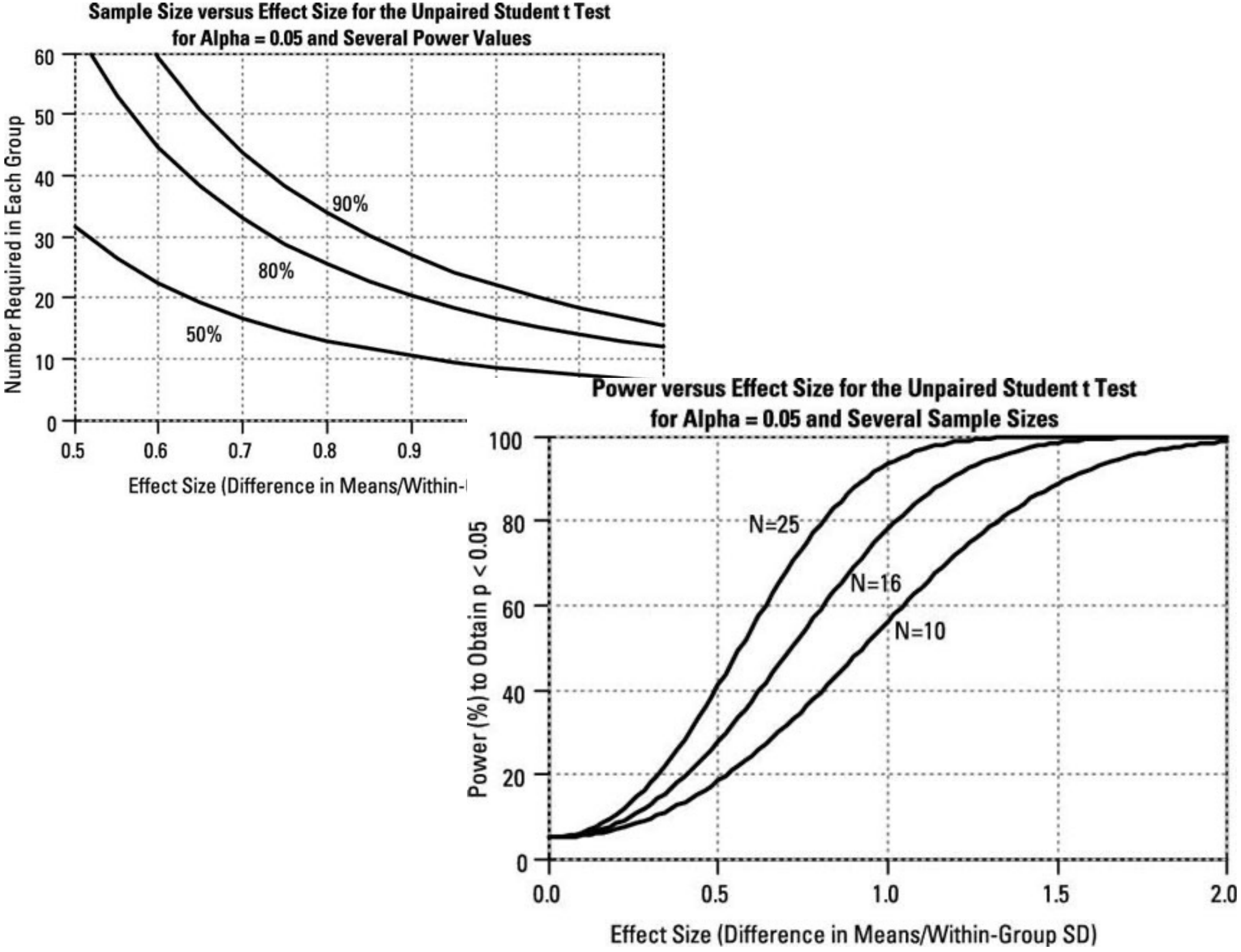
sample-size.net/sample-size-survival-analysis/

Group size estimation: graphical outputs interpretation

- *p*: constant (0.01 or 0.05)
- $\boldsymbol{\sigma}$: beyond our control

N, δ, 1-β



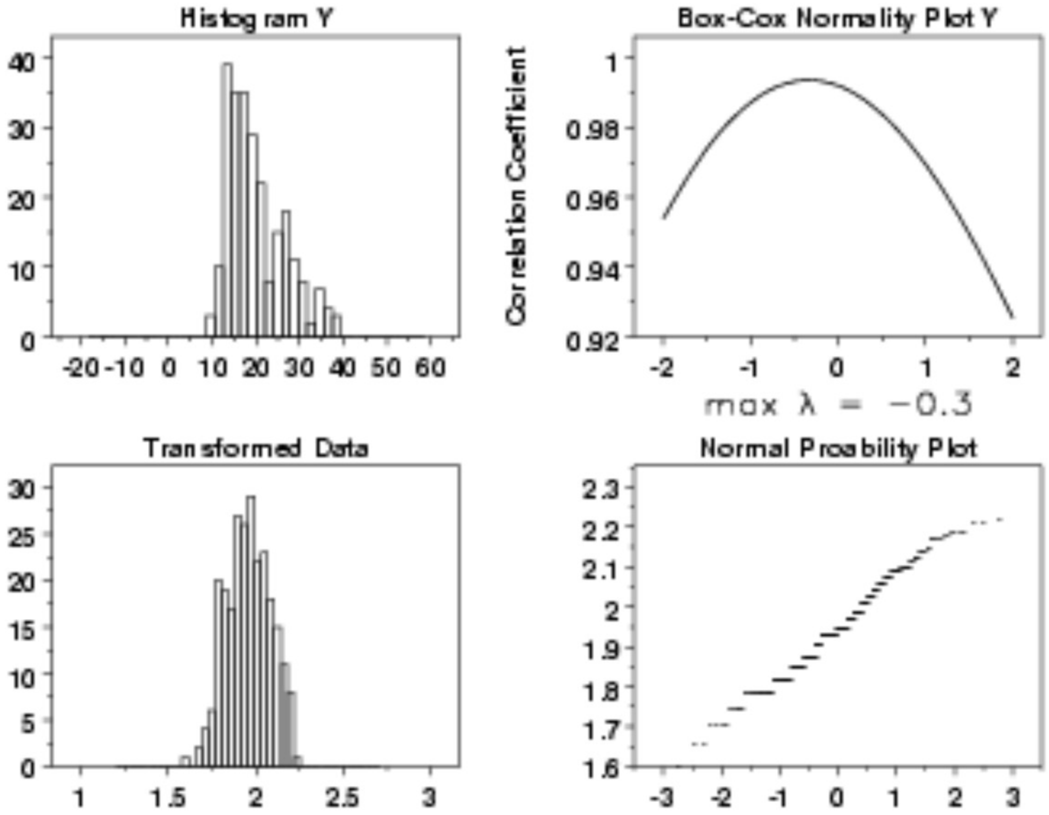


Transformed data or non-parametric?

Situation	Transformation
right-skewed distribution	
lognormal more skewed than lognormal less skewed than lognormal	x' = log(x) x' = 1/x x’ = sqrt(x)
left-skewed distribution	
moderately skewed more skewed	x’ = x^2 x’ = x^3
nonhomogeneous variances	
SD proportional to means SD proportional to means^2 SD proportional to sqrt(means)	x' = log(x) x' = 1/x x’ = sqrt(x)
data as percentages (0-100%)	p’ = arcsin(sqrt(p))
proportions (p, X/n, 0-1)	p’ = arcsin(sqrt(p))

Watała: Biostatystyka - wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych.
α-medica press, Bielsko-Biała, 2002.

Transformed data or non-parametric?



Box-Cox transformation
Rank-transformation

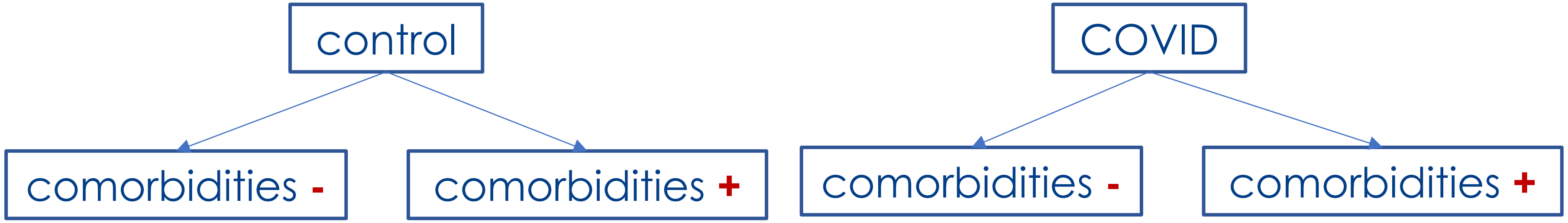
Box, G. and Cox, D. (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26, 211-252.

Experimental design: selection of the control group

✗ **More than just 1 control group**

healthy **(HC)** | healthy + comorbidities **(DC)** | disease **(sick?)** | disease + comorbidities **(sicker?)**

✓



Additional advantages of the model
(possible answers to various pre- or post-planned questions within one analysis)

GDAŃSKI
UNIWERSYTET
MEDYCZNY

Experimental design: multiple comparisons

Bonferroni correction

$$p_0 = 1 - \sqrt[n]{1-p}$$

$$p_0 \sim p/n$$

N = 3	$p_0 \approx 0.016$
N = 10	$p_0 \approx 0.005$
N = 30	$p_0 \approx 0.0017$

Very conservative, restrictive

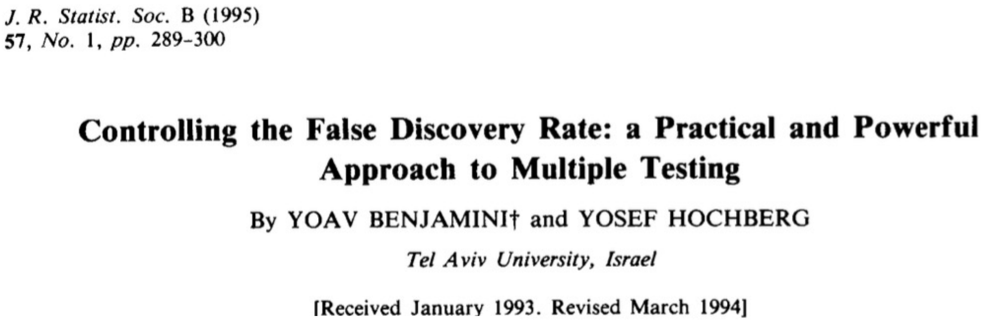
FDR (Benjamini & Hochberg)

different logic

less restrictive

Unjustified use of corrections

relevant answer to question of no interest



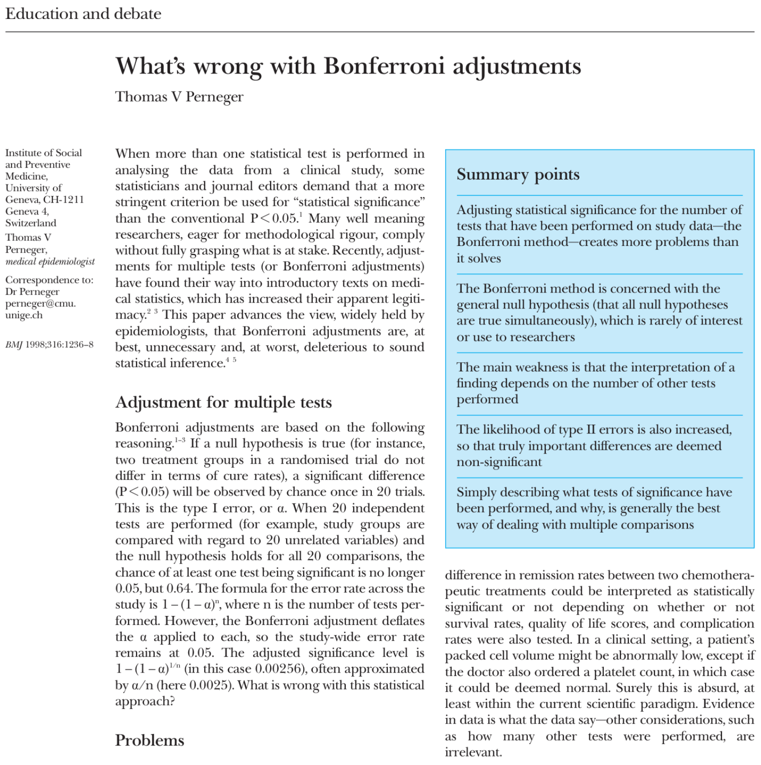
J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]



Education and debate

What's wrong with Bonferroni adjustments

Thomas V Perneger

Institute of Social and Preventive Medicine, University of Geneva, CH-1211 Geneva 4, Switzerland

Thomas V Perneger, *medical epidemiologist*

Correspondence to: Dr Perneger perneger@cmu.unige.ch

BMJ 1998;316:1236–8

When more than one statistical test is performed in analysing the data from a clinical study, some statisticians and journal editors demand that a more stringent criterion be used for "statistical significance" than the conventional P < 0.05.[1] Many well meaning researchers, eager for methodological rigour, comply without fully grasping what is at stake. Recently, adjustments for multiple tests (or Bonferroni adjustments) have found their way into introductory texts on medical statistics, which has increased their apparent legitimacy.[2 3] This paper advances the view, widely held by epidemiologists, that Bonferroni adjustments are, at best, unnecessary and, at worst, deleterious to sound statistical inference.[4 5]

Adjustment for multiple tests

Bonferroni adjustments are based on the following reasoning.[1–3] If a null hypothesis is true (for instance, two treatment groups in a randomised trial do not differ in terms of cure rates), a significant difference (P < 0.05) will be observed by chance once in 20 trials. This is the type I error, or α. When 20 independent tests are performed (for example, study groups are compared with regard to 20 unrelated variables) and the null hypothesis holds for all 20 comparisons, the chance of at least one test being significant is no longer 0.05, but 0.64. The formula for the error rate across the study is $1-(1-\alpha)^n$, where n is the number of tests performed. However, the Bonferroni adjustment deflates the α applied to each, so the study-wide error rate remains at 0.05. The adjusted significance level is $1-(1-\alpha)^{1/n}$ (in this case 0.00256), often approximated by α/n (here 0.0025). What is wrong with this statistical approach?

Problems

Summary points

Adjusting statistical significance for the number of tests that have been performed on study data—the Bonferroni method—creates more problems than it solves

The Bonferroni method is concerned with the general null hypothesis (that all null hypotheses are true simultaneously), which is rarely of interest or use to researchers

The main weakness is that the interpretation of a finding depends on the number of other tests performed

The likelihood of type II errors is also increased, so that truly important differences are deemed non-significant

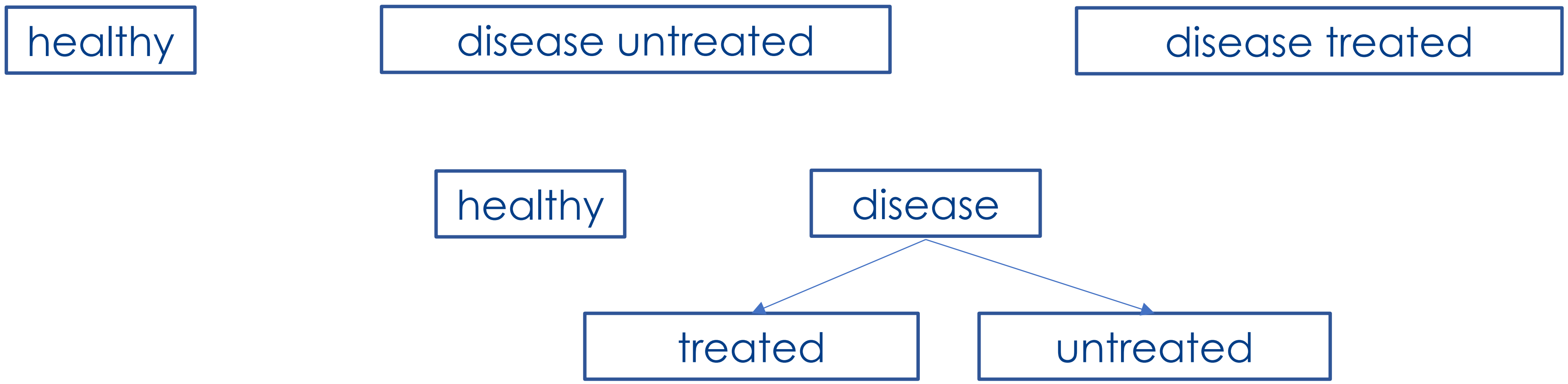
Simply describing what tests of significance have been performed, and why, is generally the best way of dealing with multiple comparisons

difference in remission rates between two chemotherapeutic treatments could be interpreted as statistically significant or not depending on whether or not survival rates, quality of life scores, and complication rates were also tested. In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be deemed normal. Surely this is absurd, at least within the current scientific paradigm. Evidence in data is what the data say—other considerations, such as how many other tests were performed, are irrelevant.

DAVID HITCH

www.gumed.edu.pl

Experimental design: several separate analyses or a model?



healthy

disease untreated

disease treated

✓

healthy

disease

treated

untreated

✗ Breaking down analyses into smaller partial analyses
Multiple comparisons using the simplest possible tests (Student’s *t* test)
Required corrections for multiple testing (Bonferroni; FWER; FDR)

✓ Hierarchical (nested) designs
ANOVA & post-hoc tests
General Linear Model, General Additive Model, …

Lack of appropriate test?

Permutation (randomization) tests

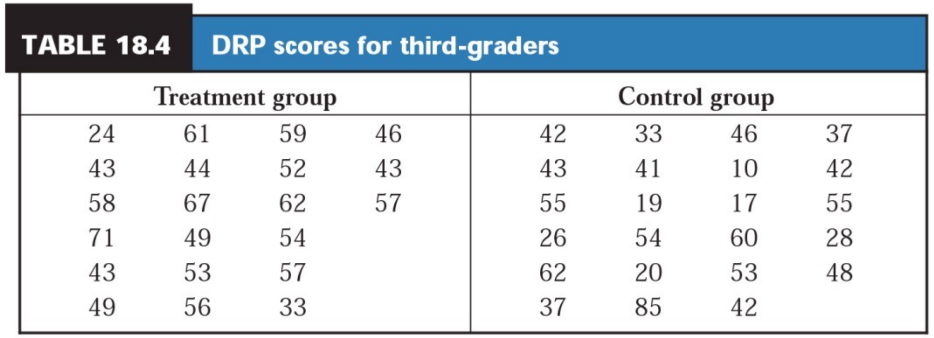
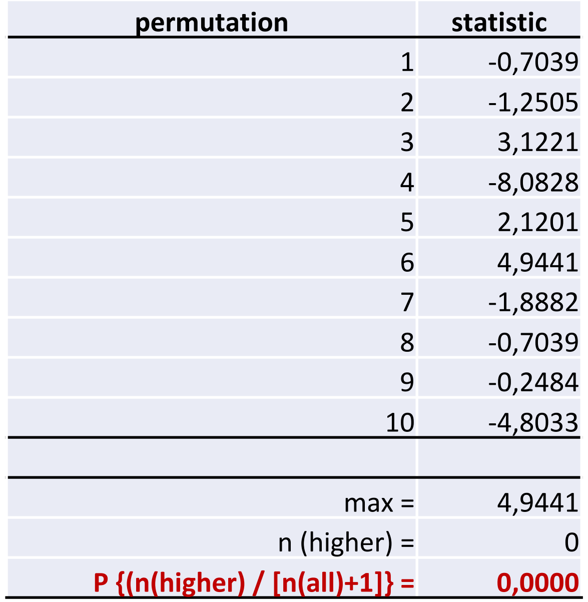


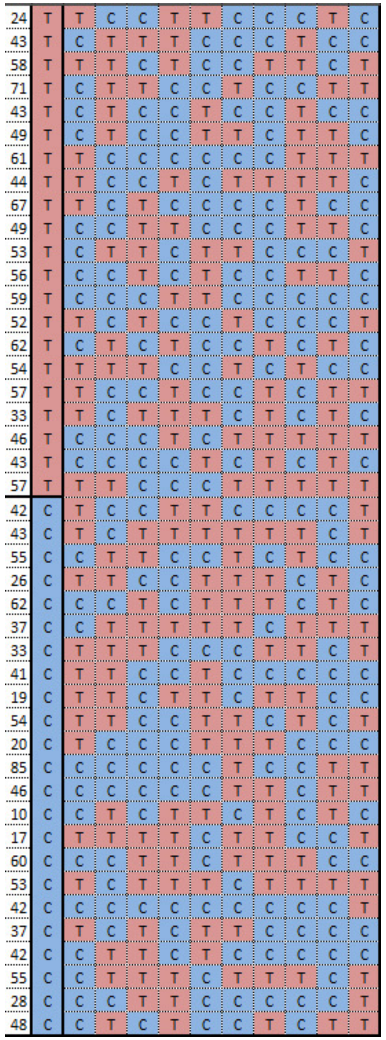
TABLE 18.4 DRP scores for third-graders

Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

$x_T - x_C = 51.472\text{-}41.522 = \mathbf{9.954}$



permutation	statistic
1	-0,7039
2	-1,2505
3	3,1221
4	-8,0828
5	2,1201
6	4,9441
7	-1,8882
8	-0,7039
9	-0,2484
10	-4,8033
max =	4,9441
n (higher) =	0
P {(n(higher) / [n(all)+1]} =	0,0000

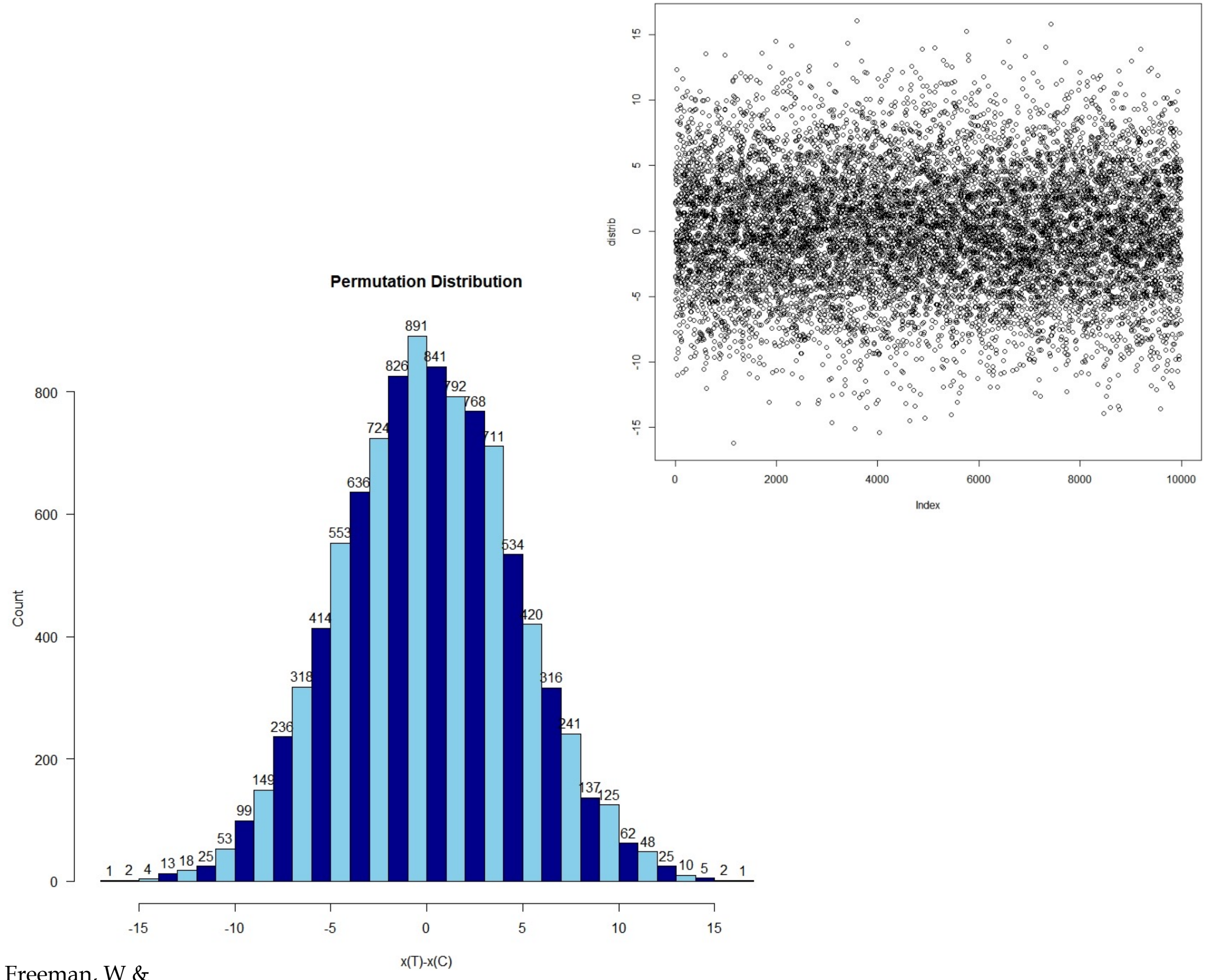


24	T	T	C	C	T	T	C	C	C	T	C
43	T	C	T	T	T	C	C	C	T	C	C
58	T	T	T	C	T	C	C	T	T	C	T
71	T	C	T	T	C	C	T	C	C	T	T
43	T	C	T	C	C	T	C	C	T	C	C
49	T	C	T	C	C	T	T	C	T	T	C
61	T	T	C	C	C	C	C	C	T	T	T
44	T	T	C	C	T	C	T	T	T	T	C
67	T	T	C	T	C	C	C	C	T	C	C
49	T	C	C	T	T	C	C	C	T	T	C
53	T	C	T	T	C	T	T	C	C	C	T
56	T	C	C	T	C	T	C	C	T	T	C
59	T	C	C	C	T	T	C	C	C	C	C
52	T	T	C	T	C	C	T	C	C	C	T
62	T	C	T	C	T	C	C	T	C	T	C
54	T	T	T	T	C	C	T	C	T	C	C
57	T	T	C	C	T	C	C	T	C	T	T
33	T	T	C	T	T	T	C	T	C	T	C
46	T	C	C	C	T	C	T	T	T	T	T
43	T	C	C	C	C	T	C	T	C	T	C
57	T	T	T	C	C	C	T	T	T	T	T
42	C	T	C	C	T	T	C	C	C	C	T
43	C	T	C	T	T	T	T	T	T	C	T
55	C	C	T	T	C	C	T	C	T	C	C
26	C	T	T	C	C	T	T	T	C	T	C
62	C	C	C	T	C	T	T	T	C	T	C
37	C	C	T	T	T	T	T	C	T	T	T
33	C	T	T	T	C	C	C	T	T	C	T
41	C	T	T	C	C	T	C	C	C	C	C
19	C	T	T	C	T	T	C	T	T	C	C
54	C	T	T	C	C	T	T	C	T	C	T
20	C	T	C	C	C	T	T	T	C	C	C
85	C	C	C	C	C	C	T	C	C	T	T
46	C	C	C	C	C	C	T	T	C	T	T
10	C	C	T	C	T	T	C	T	C	T	C
17	C	T	T	T	T	C	T	T	C	C	T
60	C	C	C	T	T	C	T	T	T	C	C
53	C	T	C	T	T	T	C	T	T	T	T
42	C	C	C	C	C	C	C	C	C	C	T
37	C	T	C	T	C	T	T	C	C	C	C
42	C	C	T	T	C	T	C	C	C	C	C
55	C	C	T	T	T	C	T	T	T	C	T
28	C	C	C	T	T	C	C	C	C	C	T
48	C	C	T	C	T	C	C	T	C	T	T

Lack of appropriate test?

Permutation (randomization) tests

no permuations	max stat.	n (higher)	p (n higher / n all)
10	4,9441	0	0,0000
50	13,4162	1	0,0200
100	10,7743	2	0,0200
200	13,5073	3	0,0150
500	16,4224	9	0,0180
1000	12,8696	11	0,0110
10000	16,6957	129	**0,0129**
50000	18,6998	699	**0,0140**
100000	17,6977	1409	**0,0141**
500000	18,6087	6912	**0,0138**
		t-test	**0,0264**
		rang-sum	**0,0127**



Hesterberg T & Monaghan, Shaun & S Moore, David & Clipson, Ashley & Epstein, Rachel & H Freeman, W & New York, Company. (2005). Bootstrap Methods and Permutation Tests. Introduction to the Practice of Statistics. 14.

A permutation test: Fisher’s exact test

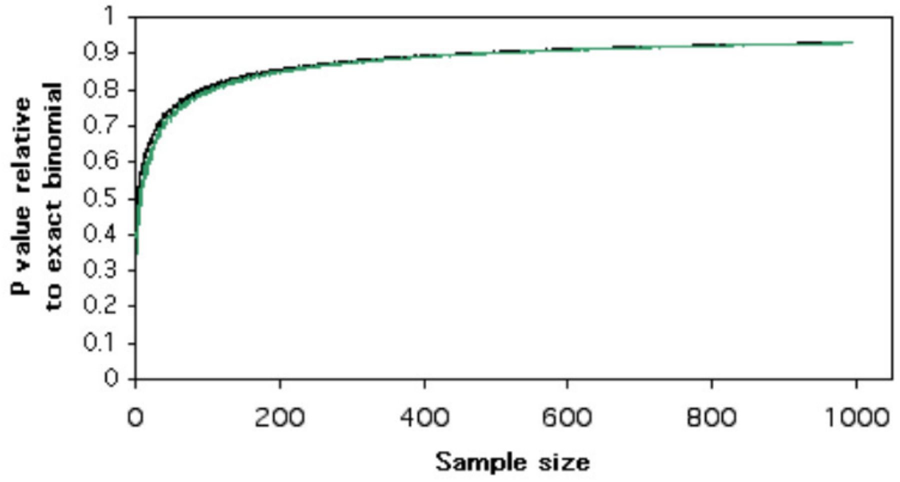
Fisher’s exact test

- an example of commonly used permutation test
- alternative to Pearson’s chi-squared test
- if the count in any cell is below 5
- but…

For every test in the case of which some statistic is being computed (t, z, F, chi2, …) there is also a permutation version thereof

Instead of comparing the calculated value of test statistic against the values in the corresponding tables, one can compared it against the distribution obtained empirically *en route* resampling

	F	*M*	*Total*
Smoking	**1**	**9**	*10*
Non-smoking	**11**	**3**	*14*
Total	*12*	*12*	*24*



***P* values of chi-square and *G*–tests, as a proportion of the *P* value from the exact binomial test.**

www.biostathandbook.com/small.html (5/06/2019)

What is your result?

p-value vs. effect size

p-value is not your result

it only tells how reliable your observed effect size is
it only tells how often you’re about to see the same effect size in repeated experiment

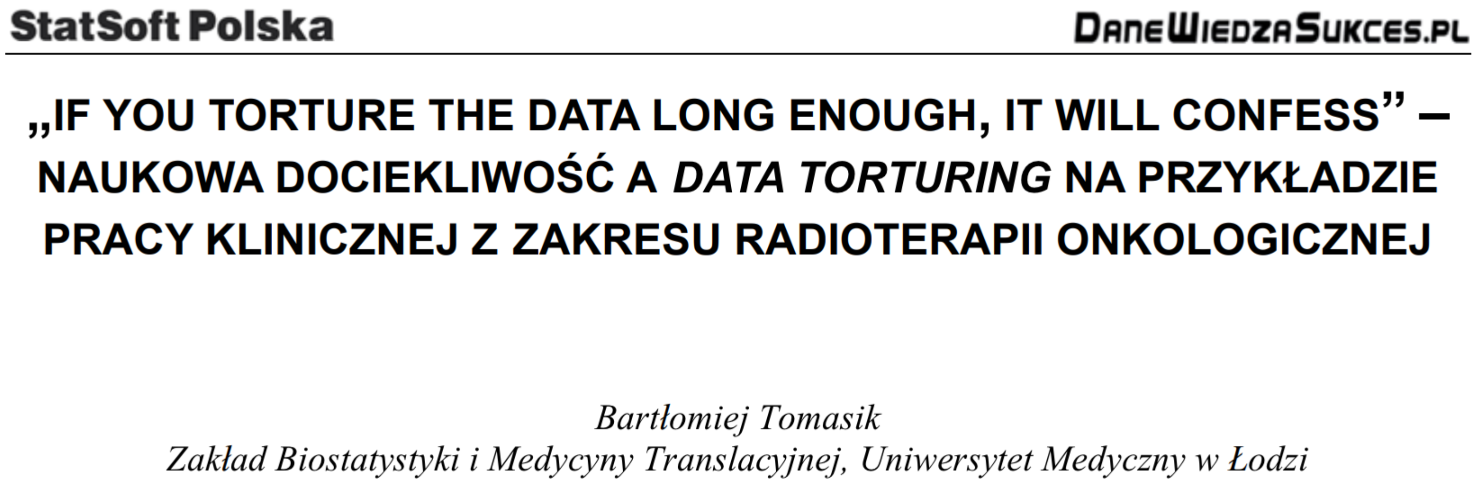
Effect size is your result

are the changes biologically relevant?
are the changes clinically relevant?

Publish or parish

striving for ”p-value”
nonsignificant results are not being published
sufficient power must be shown in order to publish results with p>0.05

An example: Data Torturing

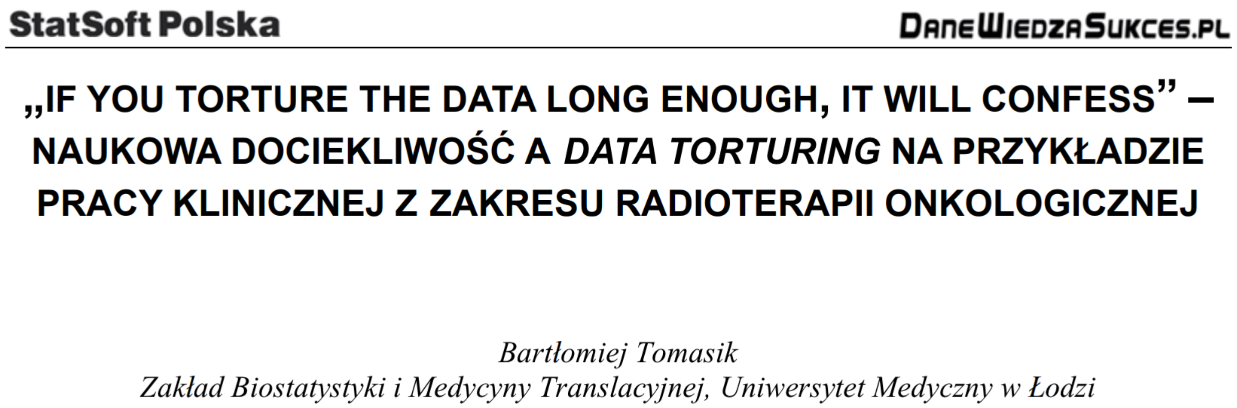


StatSoft Polska DaneWiedzaSukces.pl

„IF YOU TORTURE THE DATA LONG ENOUGH, IT WILL CONFESS" – NAUKOWA DOCIEKLIWOŚĆ A *DATA TORTURING* NA PRZYKŁADZIE PRACY KLINICZNEJ Z ZAKRESU RADIOTERAPII ONKOLOGICZNEJ

Bartłomiej Tomasik
Zakład Biostatystyki i Medycyny Translacyjnej, Uniwersytet Medyczny w Łodzi

Data torturing



StatSoft Polska DaneWiedzaSukces.pl

„IF YOU TORTURE THE DATA LONG ENOUGH, IT WILL CONFESS" – NAUKOWA DOCIEKLIWOŚĆ A *DATA TORTURING* NA PRZYKŁADZIE PRACY KLINICZNEJ Z ZAKRESU RADIOTERAPII ONKOLOGICZNEJ

Bartłomiej Tomasik
Zakład Biostatystyki i Medycyny Translacyjnej, Uniwersytet Medyczny w Łodzi

Aim: Assessment of the impact of IMRT radiotherapy on the development of xerostomia in patients with Head and Neck Cancer

A questionnaire survey + salivary gland scintigraphy
Searching for risk factors of severe xerostomia (grade 3/4) one year after IMRT

The power analysis:
assumed statistical power of 80%
assumed level of significance of 5%

~ 100 subjects per group (control; study)
data collection should take 2 years

GDAŃSKI
UNIWERSYTET
MEDYCZNY

Data torturing

2 years of data collection
- only 53 subjects
- 30 scintigraphic examinations
- 40 questionnaires

a posteriori power analysis:
< 20 % (max **!**)

So what about it now?
Analyse?
Continue collecting data?

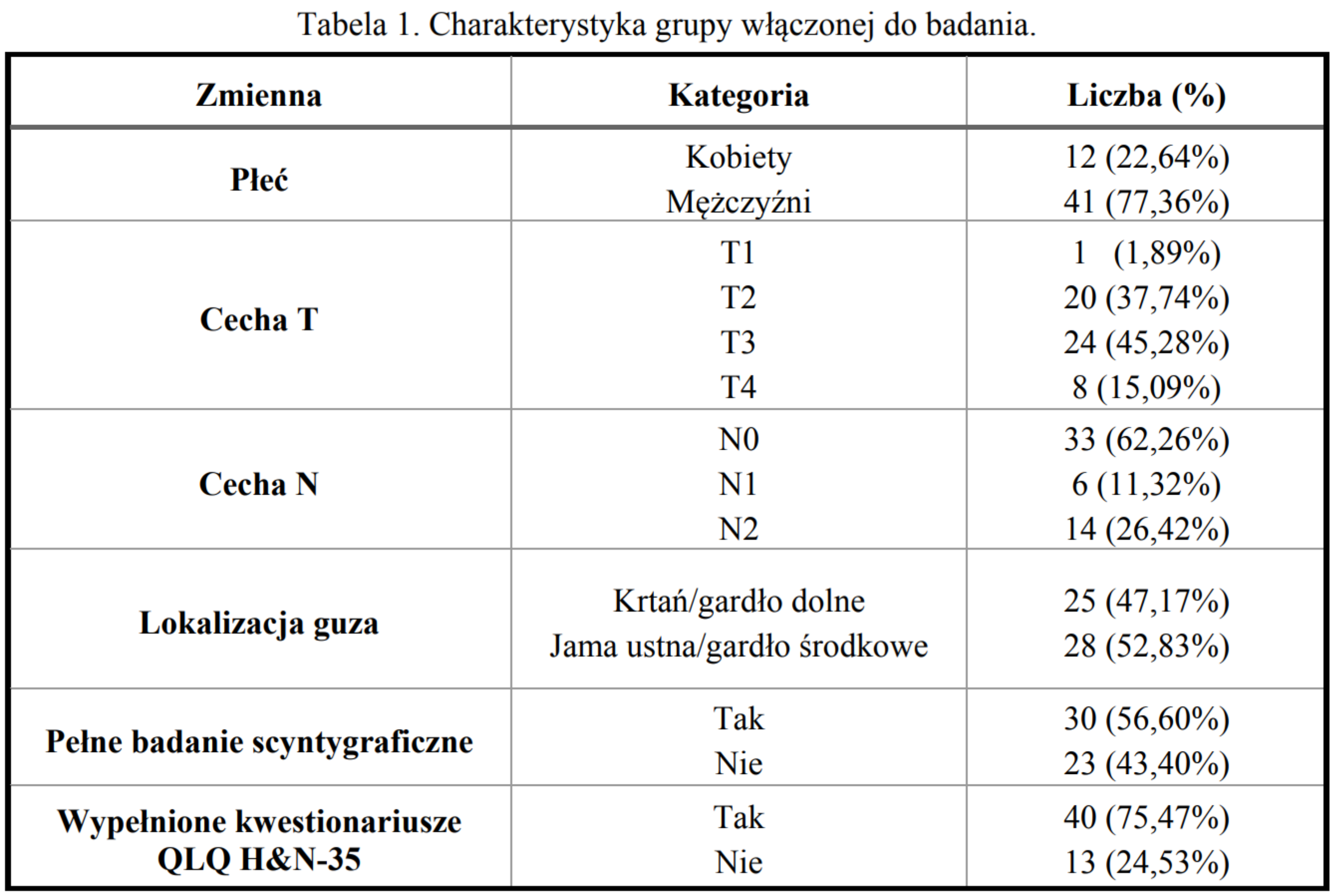


Tabela 1. Charakterystyka grupy włączonej do badania.

Zmienna	Kategoria	Liczba (%)
Płeć	Kobiety Mężczyźni	12 (22,64%) 41 (77,36%)
Cecha T	T1 T2 T3 T4	1 (1,89%) 20 (37,74%) 24 (45,28%) 8 (15,09%)
Cecha N	N0 N1 N2	33 (62,26%) 6 (11,32%) 14 (26,42%)
Lokalizacja guza	Krtań/gardło dolne Jama ustna/gardło środkowe	25 (47,17%) 28 (52,83%)
Pełne badanie scyntygraficzne	Tak Nie	30 (56,60%) 23 (43,40%)
Wypełnione kwestionariusze QLQ H&N-35	Tak Nie	40 (75,47%) 13 (24,53%)

Data torturing

Xerostomia severity (degree) vs. risk group (location)

- Fisher's exact test
- p=0.048 (**!**)

- one-sided test (**!**)

- trying to find an argument for usage of one-sided test *post factum*
- it was not taken into account while planning the experiment

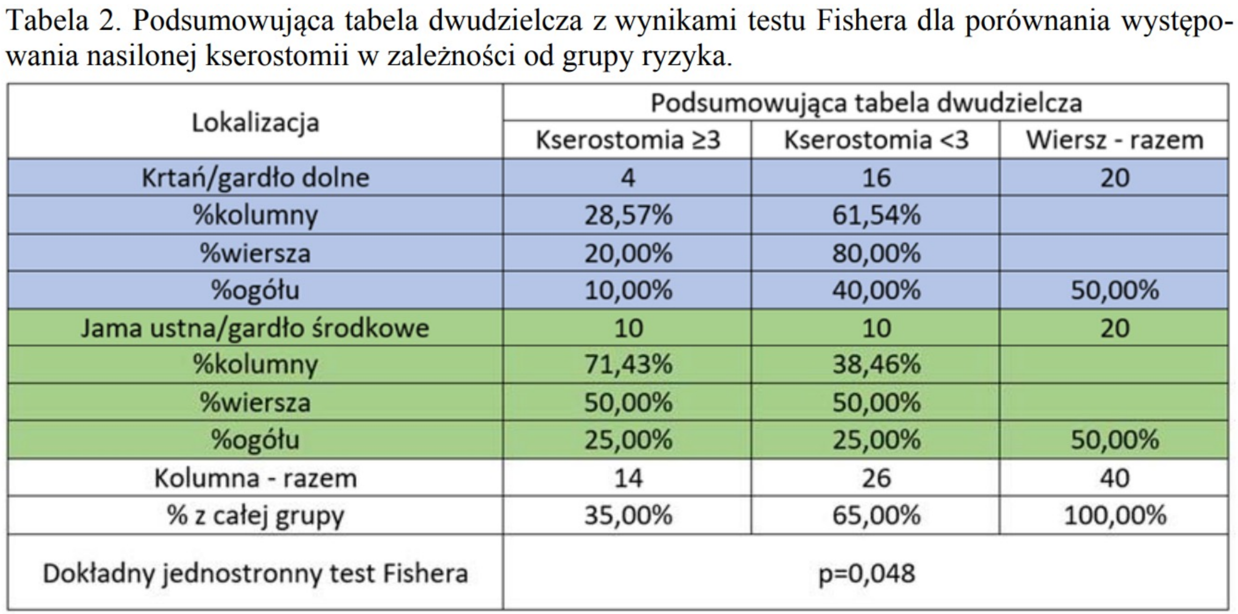


Tabela 2. Podsumowująca tabela dwudzielcza z wynikami testu Fishera dla porównania występowania nasilonej kserostomii w zależności od grupy ryzyka.

Lokalizacja	Podsumowująca tabela dwudzielcza		
	Kserostomia ≥3	Kserostomia <3	Wiersz - razem
Krtań/gardło dolne	4	16	20
%kolumny	28,57%	61,54%	
%wiersza	20,00%	80,00%	
%ogółu	10,00%	40,00%	50,00%
Jama ustna/gardło środkowe	10	10	20
%kolumny	71,43%	38,46%	
%wiersza	50,00%	50,00%	
%ogółu	25,00%	25,00%	50,00%
Kolumna - razem	14	26	40
% z całej grupy	35,00%	65,00%	100,00%
Dokładny jednostronny test Fishera	p=0,048		

Data torturing

Some of the questionnaires were repeated after several years

Q Cochran’s test
- p=0.083
- observable trend (**!?**)

- initial lack of statistical power
- decreasing number of subjects in analysis

- one-sided test (**!**)
  (trying to find an argument *post factum*)

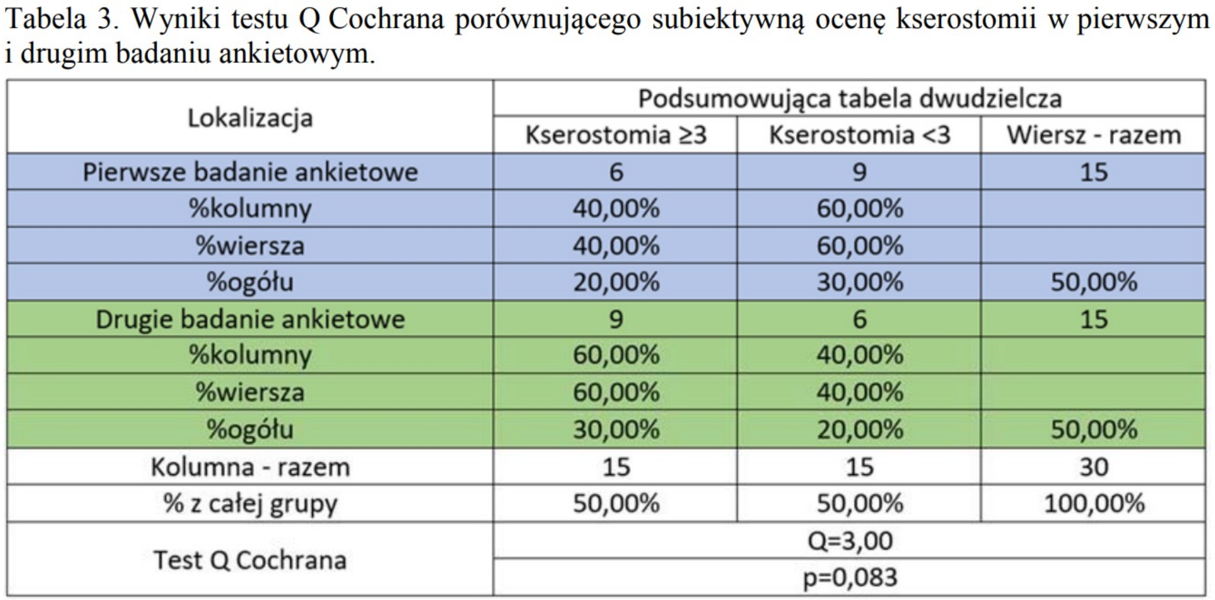


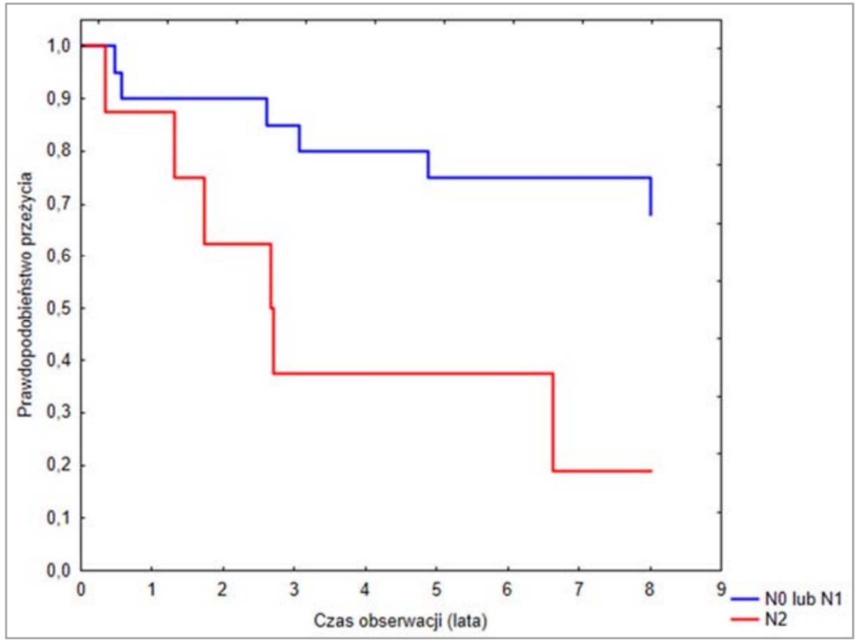
Tabela 3. Wyniki testu Q Cochrana porównującego subiektywną ocenę kserostomii w pierwszym i drugim badaniu ankietowym.

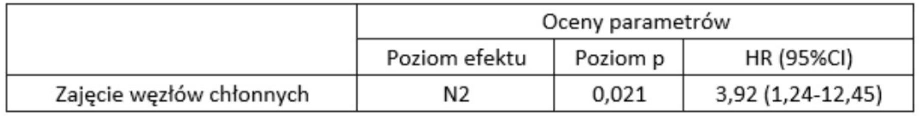
Lokalizacja	Podsumowująca tabela dwudzielcza		
	Kserostomia ≥3	Kserostomia <3	Wiersz - razem
Pierwsze badanie ankietowe	6	9	15
%kolumny	40,00%	60,00%	
%wiersza	40,00%	60,00%	
%ogółu	20,00%	30,00%	50,00%
Drugie badanie ankietowe	9	6	15
%kolumny	60,00%	40,00%	
%wiersza	60,00%	40,00%	
%ogółu	30,00%	20,00%	50,00%
Kolumna - razem	15	15	30
% z całej grupy	50,00%	50,00%	100,00%
Test Q Cochrana	Q=3,00		
	p=0,083		

Data torturing

A survival analysis was planned in addition

- the log-rank test
- after several trials and divisions of the whole study groups into subgroups, a statistically significant result was obtained (p=0.021)
- researchers were satisfied ☺
- have they planned the survival analysis?
- have they checked whether the study provides enough power for survival analysis?





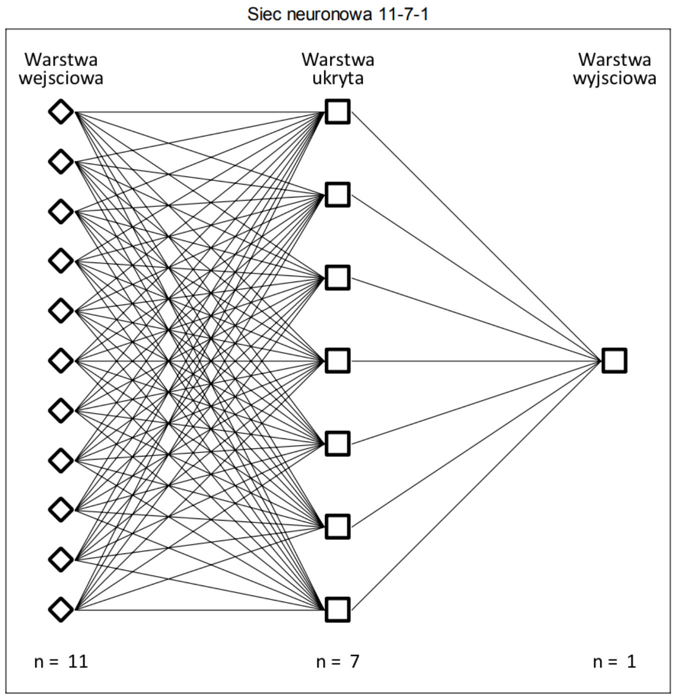
	Oceny parametrów		
	Poziom efektu	Poziom p	HR (95%CI)
Zajęcie węzłów chłonnych	N2	0,021	3,92 (1,24-12,45)

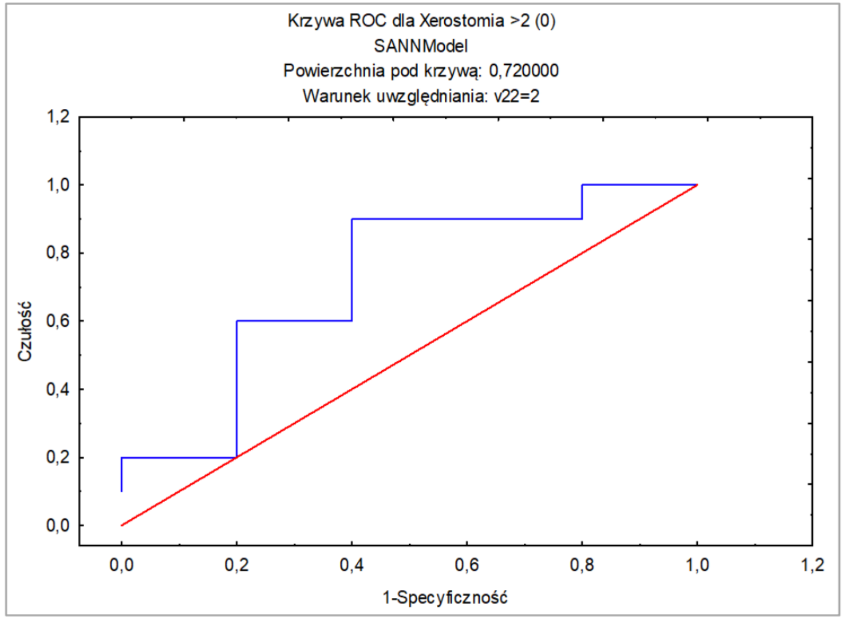
Rys. 6. Krzywe Kaplana-Meiera przedstawiające różnice w przeżyciu w zależności od zajęcia węzłów chłonnych oraz tabela z wynikami analizy proporcjonalnego hazardu Coxa.

Data torturing

Searching for predictors of xerostomia worsening

- several ROC curve analyses
- no success, therefore the authors used neural networks (11 L_0 neurones; 7 layers)
- obtained network provided accuracy of 80% and 0.72 area u the curve (AUC)
- researchers were fully satisfied ☺
- lack of interpretability?





Data torturing

Problems

- excessive optimism in the planning phase regarding data collection (not taking into account that some patients may not meet the inclusion criteria)
- very low statistical power
- analyses within subgroup that were not previously planned (additional subdivisions of groups)
- conducting previously unplanned analyses (survival analysis, ROC analysis)
- multiple hypotheses testing without appropriate corrections
- abuse of more and more complicated analytical methods in order to prove the assumed thesis („hypothesis driven science"; Fisher's test → Neural Network)
- the blind pursuit of statistically significant results without considering their clinical significance

Data torturing

What should we do?

- it cannot be avoided

But we should be cautious to:

- were the other (new?) hypotheses made before or during the experiment?
- were all the subgroup analyses planned in advance, prior to experiment?
- are all the obtained results related to the main hypothesis of the study?
- is it not necessary, at lease in the case of some statistical tests, that appropriate multiple testing corrections be applied?

GDAŃSKI UNIWERSYTET MEDYCZNY

Thanks for your attention...

Literature

CO MOŻNA WYCISNĄĆ Z TYCH DANYCH?

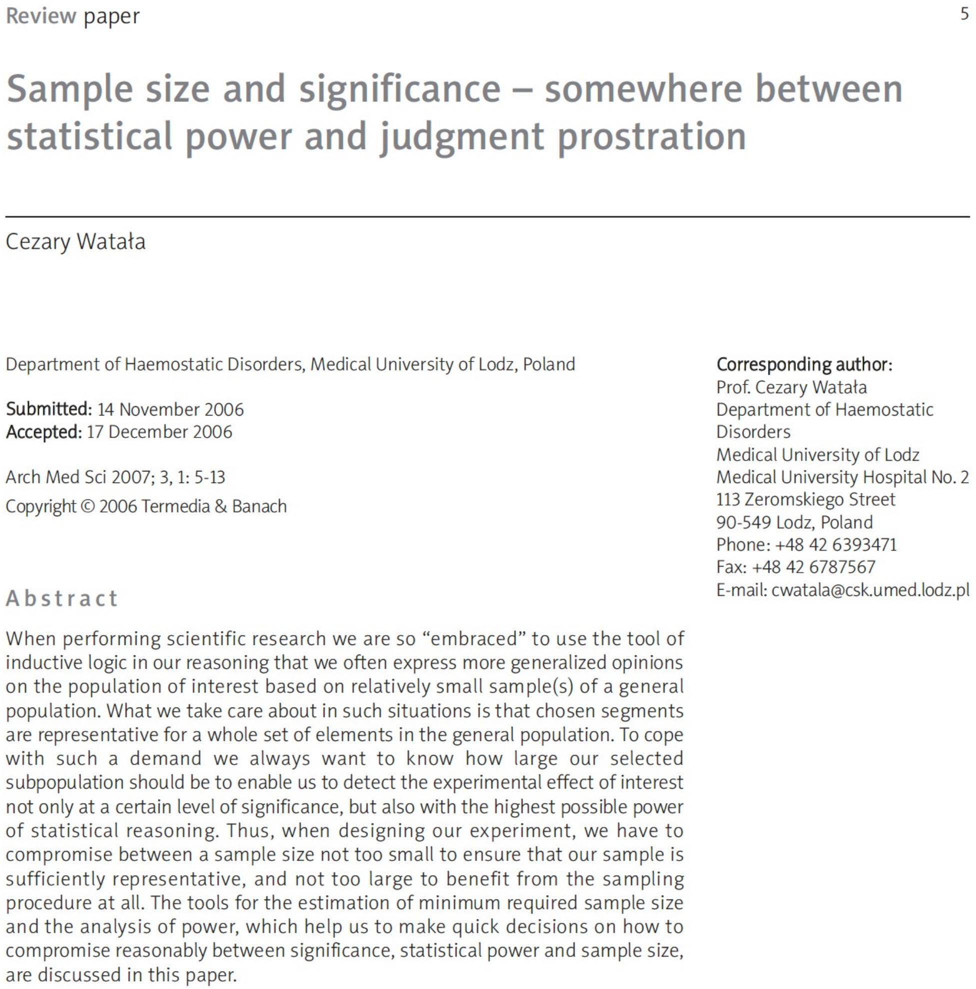
Andrzej Stanisz, Collegium Medicum Uniwersytetu Jagiellońskiego w Krakowie, Zakład Biostatystyki i Informatyki Medycznej

JAK SKUTECZNIE WYKORZYSTYWAĆ METODY STATYSTYCZNE W PLANOWANIU I PRZEPROWADZANIU EKSPERYMENTU NAUKOWEGO?

Cezary Watała, Uniwersytet Medyczny w Łodzi, Zakład Zaburzeń Krzepnięcia Krwi KDL; Uniwersytecki Szpital Kliniczny nr 2 im. WAM

JAK PLANOWAĆ DOŚWIADCZENIA NAUKOWE Z WYKORZYSTANIEM METOD STATYSTYCZNYCH?
TESTOWANIE HIPOTEZ STATYSTYCZNYCH: MIĘDZY MOCĄ STATYSTYCZNĄ A NIEMOCĄ DECYZYJNĄ

Cezary Watała, Uniwersytet Medyczny w Łodzi, Zakład Zaburzeń Krzepnięcia Krwi; Uniwersytecki Szpital Kliniczny nr 2 im. WAM



Review paper

5

Sample size and significance – somewhere between statistical power and judgment prostration

Cezary Watała

Department of Haemostatic Disorders, Medical University of Lodz, Poland

Submitted: 14 November 2006
Accepted: 17 December 2006

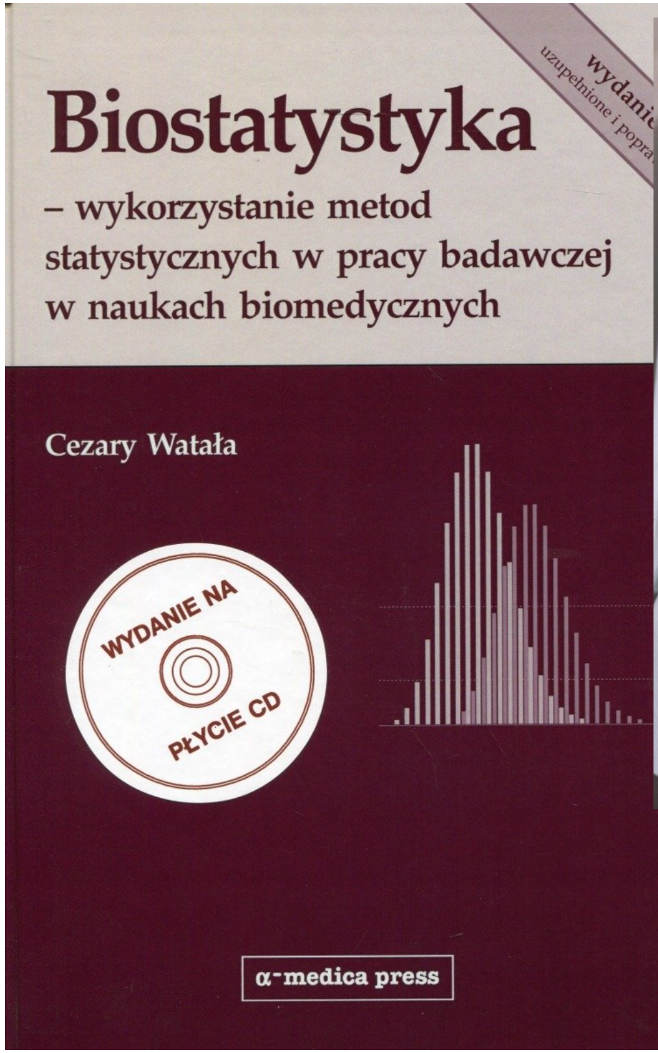
Arch Med Sci 2007; 3, 1: 5-13

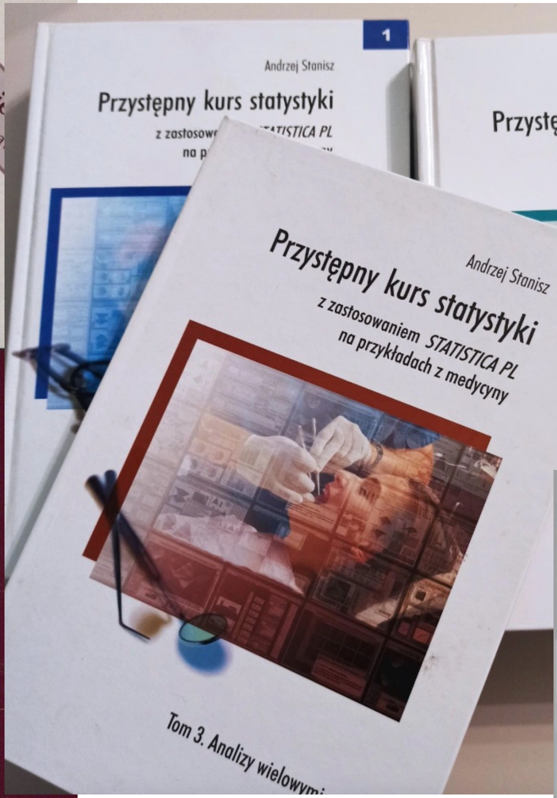
Corresponding author:
Prof. Cezary Watała
Department of Haemostatic Disorders
Medical University of Lodz
Medical University Hospital No. 2
113 Zeromskiego Street
90-549 Lodz, Poland
Phone: +48 42 6393471
Fax: +48 42 6787567
E-mail: cwatala@csk.umed.lodz.pl

Abstract

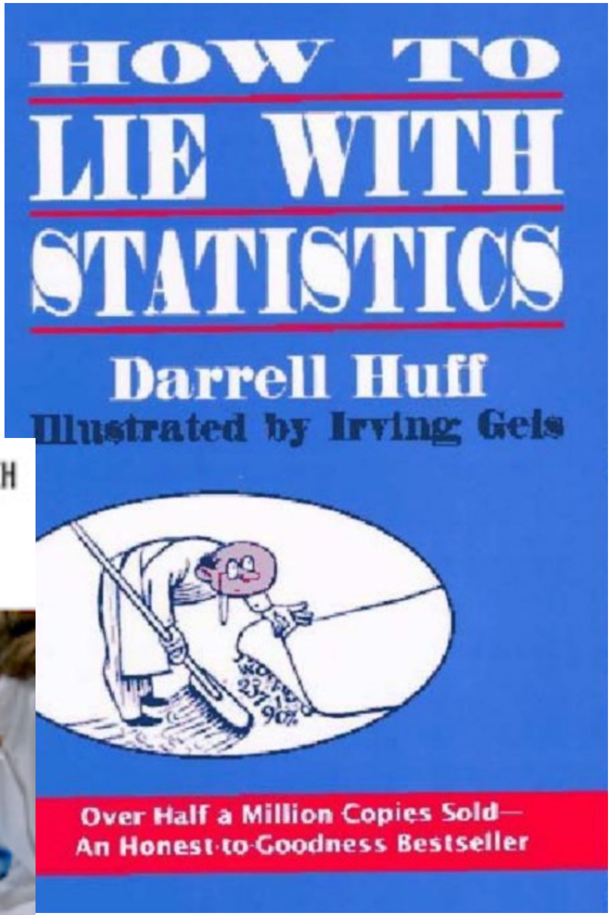
When performing scientific research we are so “embraced” to use the tool of inductive logic in our reasoning that we often express more generalized opinions on the population of interest based on relatively small sample(s) of a general population. What we take care about in such situations is that chosen segments are representative for a whole set of elements in the general population. To cope with such a demand we always want to know how large our selected subpopulation should be to enable us to detect the experimental effect of interest not only at a certain level of significance, but also with the highest possible power of statistical reasoning. Thus, when designing our experiment, we have to compromise between a sample size not too small to ensure that our sample is sufficiently representative, and not too large to benefit from the sampling procedure at all. The tools for the estimation of minimum required sample size and the analysis of power, which help us to make quick decisions on how to compromise reasonably between significance, statistical power and sample size, are discussed in this paper.

Literature


Biostatystyka
– wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych
wydanie uzupełnione i popra
Cezary Watała
WYDANIE NA PŁYCIE CD
α-medica press


Andrzej Stanisz
Przystępny kurs statystyki
z zastosowaniem STATISTICA PL
na przykładach z medycyny
1
2
3
Tom 3. Analizy wielowymi

ZASTOSOWANIA METOD STATYSTYCZNYCH
W BADANIACH NAUKOWYCH IV
Kraków 2012
StatSoft


HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller