# The Power and Limits of Statistics

DPRRGSP 2018-11-29

@ReinhardFurrer

Applied Statistics

Department of Mathematics

Department of Computational Science

# Contents

– Preamble

– Good statistical practice
– P-values and their proper use

– Epilogue

# Preamble

This set of slides

- is available at www.math.uzh.ch/furrer/slides/181129FurrerDPRRGSP.pdf
- is a subset of the slides to be shown during the lecture

The full set of slides will be posted after the lecture at www.math.uzh.ch/furrer/download/181129FurrerDPRRGSP.pdf

# Preamble

About me:

- Chair of Applied Statistics
- Minor Applied Probability and Statistics, MSc Biostatistics (STA470 Good Statistical Practice, … )
- Consulting Service MNF
- Commitment to Research Transparency and Open Science

About the lecture:

- Interactive
- Something for everyone

University of Zurich[UZH]

Applied Statistics, IMath

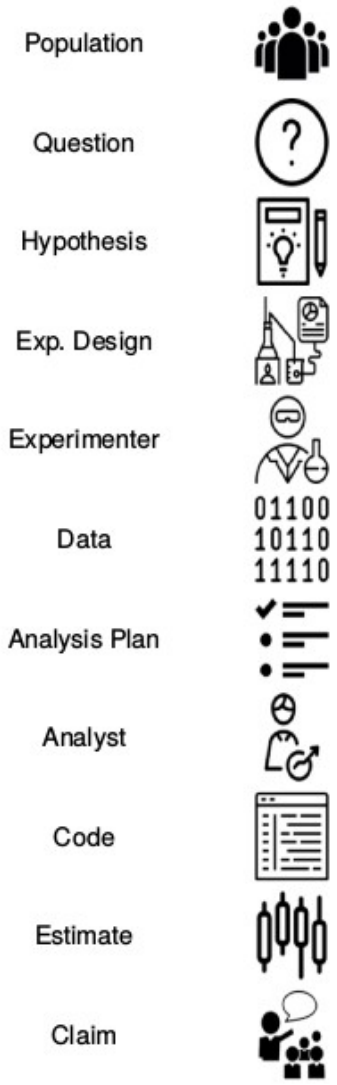
# Good Statistical Practice

# "Scientific Study" Protocol

- General approach:

- Estimate consists of:
  - Model choice
  - Model fitting
  - Model validation

scifigure::sci_figure(scifigure::init_experiments(1,""))



# “Scientific Study” Protocol: Data

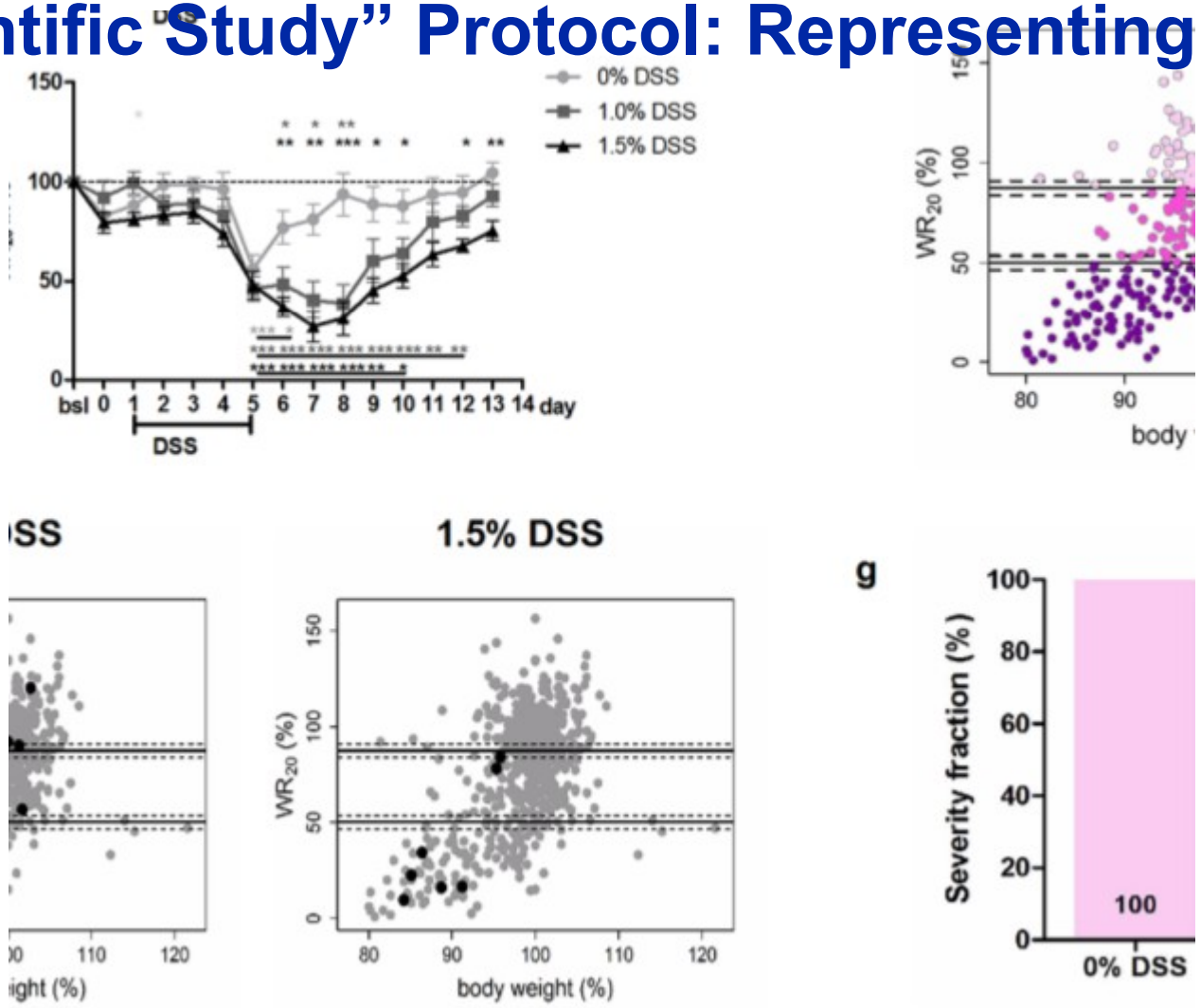
- Text file
- Long or wide format
- Simple but meaningful column names
- Numerics are numerics (not `>` etc),
  missing values are 'NA' (not empty, 9999, -9999, ...)
- Dates: 2018-11-29
- Separate CodeBook with basic information for all variables
  units, possible range, factors and encoding
- No colors, formating or calculations allowed
  [10.1080/00031305.2017.1375989][10.1080/00031305.2017.1375987]

# "Scientific Study" Protocol: Representing Data

Exploratory data analysis (EDA)

- Carefully consider type of data (nominal, ordinal, interval, ratio) and adapt plotting (barplot histogram, boxplot)
- Add: ***n*** and standard errors, uncertainties, ranges
- Think four times before using a pie chart
- No fancy thrills!

# "Scientific Study" Protocol: Representing Data



# “Scientific Study” Protocol: Code

- Scripting, R or better with Markdown
- Accessible data, code and documentation
- Reproducible images and figures
- Ideally version control [10.1080/00031305.2017.1399928]
- Sharing using a 'Research Compendium':
  - files according convention of the community
  - separation of data, method, output
  - specifying the computational environment

[10.1080/00031305.2017.1375986]

# "Scientific Study" Protocol: Estimate/Claim

Estimate:

- Model choice:

  Typically a parametric description
  Statistical model that is defendable

- Model fitting:

  Estimation, fitting, prediction

- Model validation:

  Assessing appropriateness, adjustments

Claim: Discussed in the second part

# Summary

- Proper data storage
- Accessable data, code and documentation
- Fair, accessible figures
- Scripting, with Markdown
- Ideally version controlled compendium
- Statistical modeling as craftmanship and art

University of Zurich[UZH]

Applied Statistics, IMath

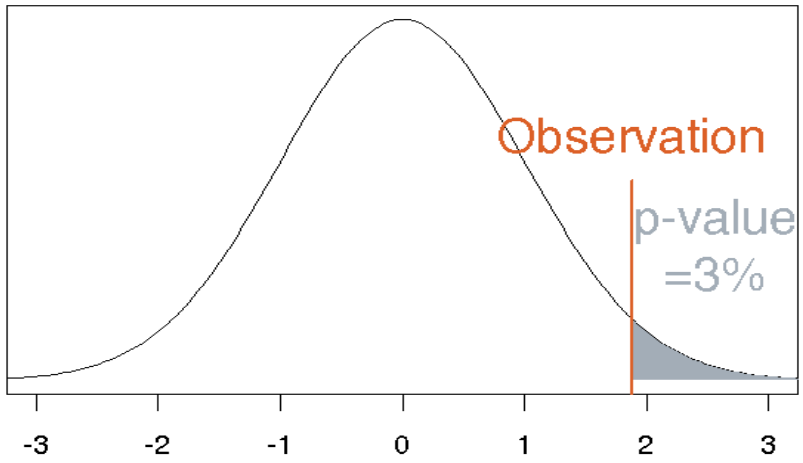
# P-values and Their Proper Use

# Concept of a Statistical Test

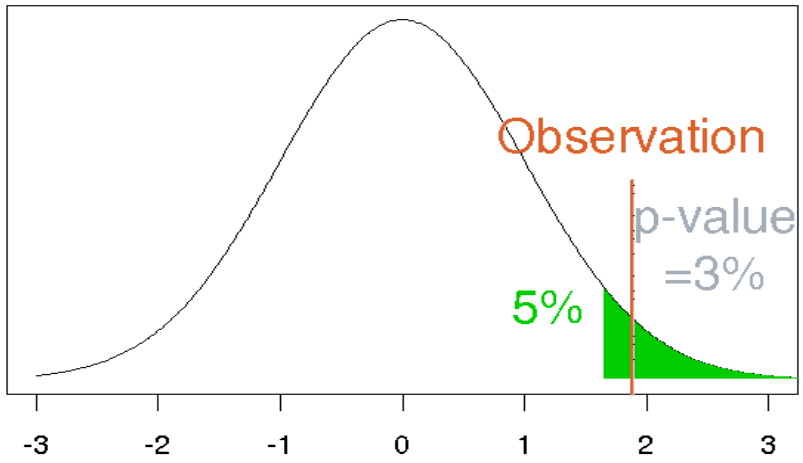
- There is never a proof for a hypothesis
- Data can only provide evidence against
- Based on hypothesis, how does the data compare

**Definition:** The p-value is the probability, under the distribution of the null hypothesis, of obtaining a result equal to or more extreme than the observed result.

# P-value



# “Sufficiently” small P-value



# Hypothesis Tests vs Significance Test

Disimilarities:

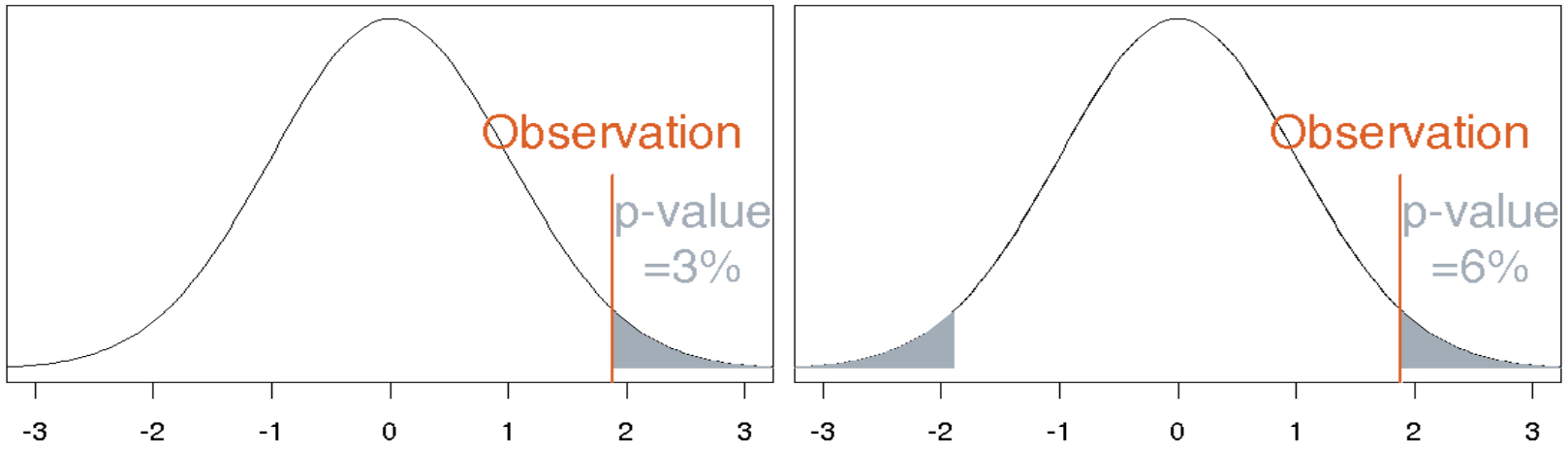
– Continuous evidence against (Hypothesis Tests)
versus
zero/one coding (Significance Tests)

Similarities:

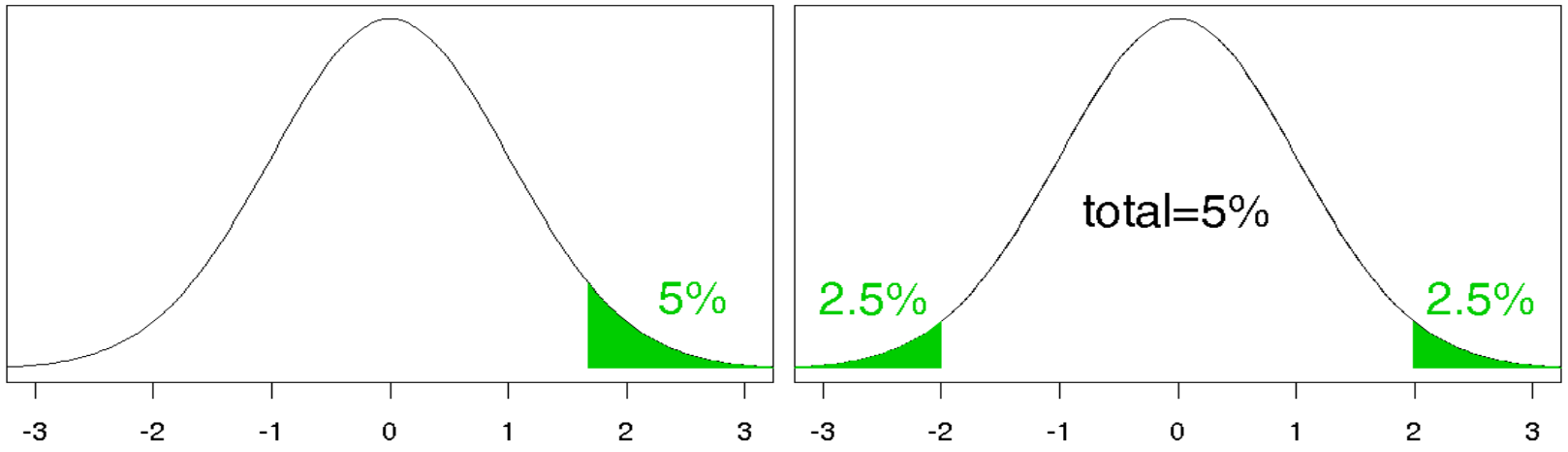
– Null hypothesis $H_0$ and “hidden” alternative hypthesis

– Data only provides evidence against $H_0$

# P-value


Observation
p-value
=3%
-3
-2
-1
0
1
2
3
Observation
p-value
=6%
-3
-2
-1
0
1
2
3

# Rejection Region (Significance Tests)



# Procedure for a Statistical Test

1. Formulation of the scientific question or scientific hypothesis
2. Formulation of the statistical model (assumptions)
3. Formulation of the statistical test hypothesis and selection of significance level
4. Selection of the appropriate test
5. Calculation of the p-value, comparison and decision
6. Interpretation

**And this shall not be repeated...**
**… next week ...**

# Errors (Significance Tests)

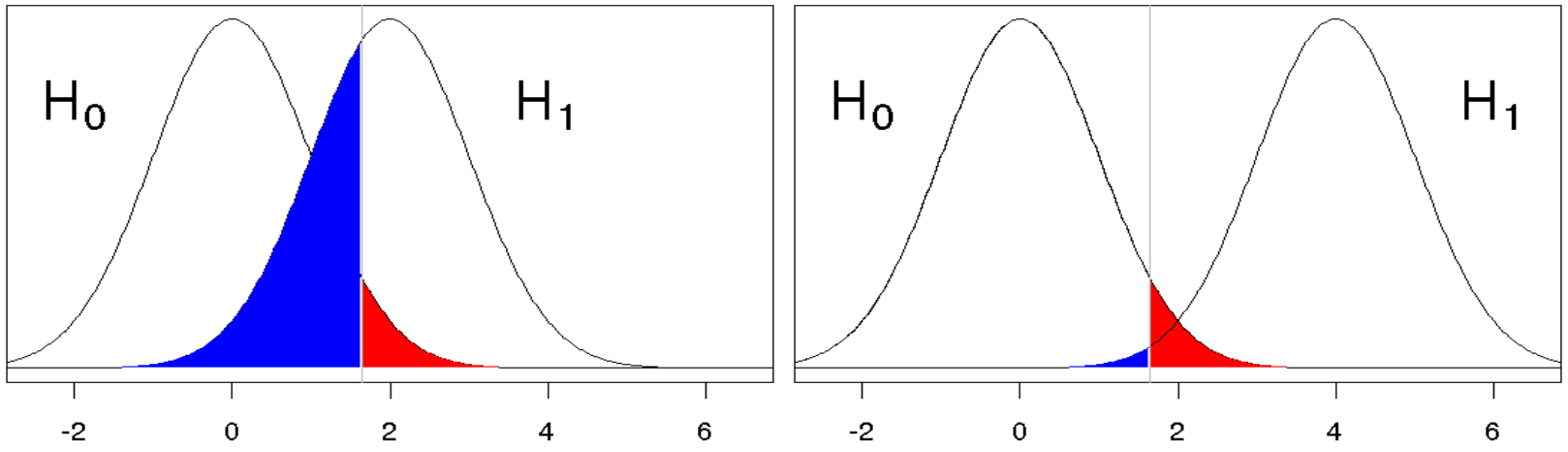
| Test results | True state: $H_0$ true | True state: $H_1$ true |
|---|---|---|
| do not reject $H_0$ | $1-\alpha$ | $\beta$ |
| reject $H_0$ | $\alpha$ | $1-\beta$ |

# Errors (Significance Tests)

| | | True condition | | | |
|---|---|---|---|---|---|
| | Total population | Condition positive | Condition negative | Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$ | Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$ |
| **Predicted condition** | Predicted condition positive | **True positive**, Power | **False positive**, Type I error | Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$ | False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$ |
| | Predicted condition negative | **False negative**, Type II error | **True negative** | False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$ | Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$ |
| | | True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$ | False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$ | Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$ | Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$ |
| | | False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$ | Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$ | Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$ | $F_1$ score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$ |

[wikipedia.org/wiki/True_positive_rate]

# Errors (Significance Tests)



# Effect Size and Power

Type I error, $\alpha$:

– Fixed (for a single statistical test)

Type II error, $\beta$:

– Depends on significance ($\alpha$)

– Depends on sample size ($\boldsymbol{n}$)

– Depends on alternative (which is not observable)

– Depends on the inherent uncertainty

# Effect Size and Power

Type I error, α:

– Fixed (for a single statistical test)

Type II error, β:

– Depends on significance (α)

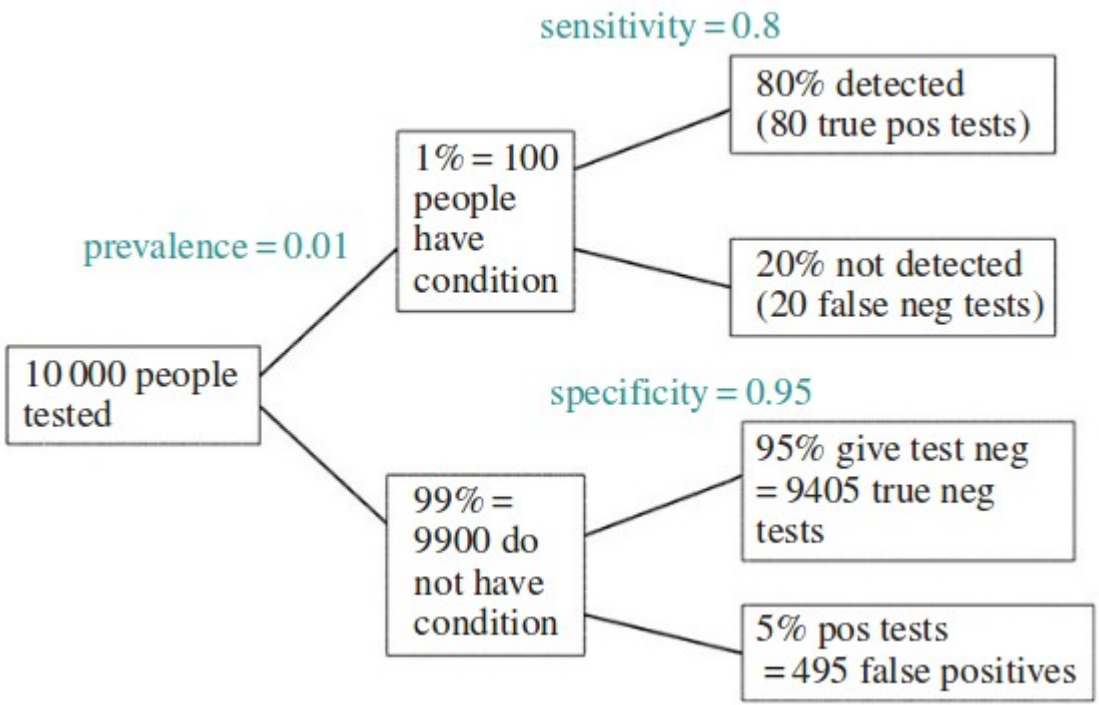
– Depends on sample size (***n***)

– Depends on effect size (normalized difference of hypotheses)

Cohen's ***d***

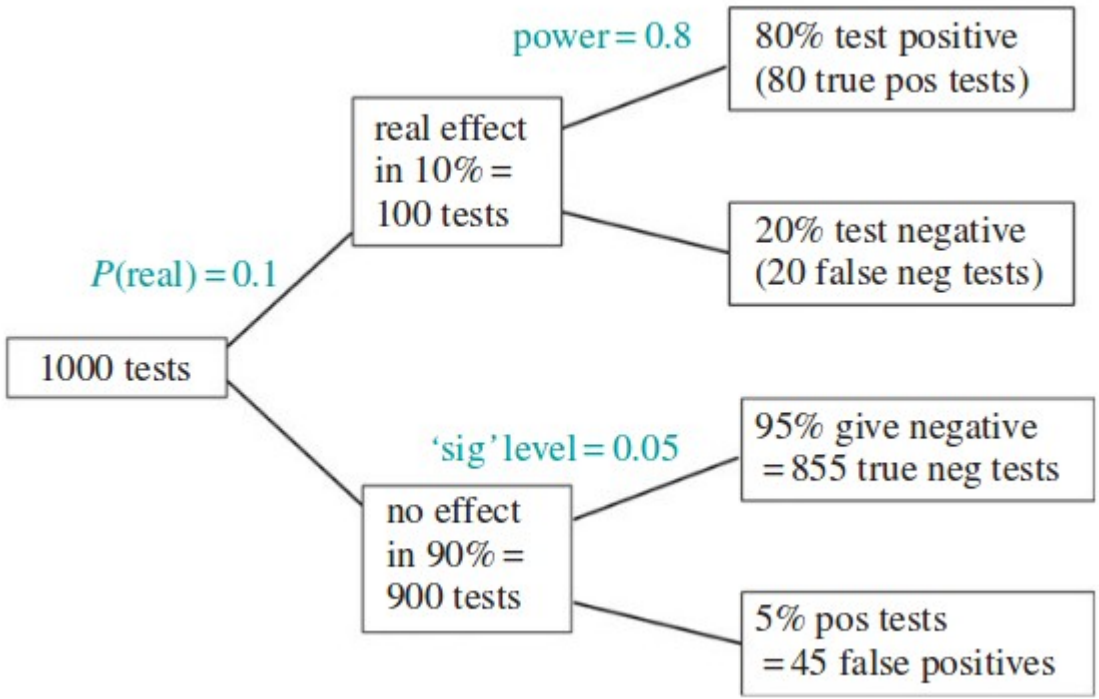
Easy: https://rpsychologist.com/d3/NHST/
Advanced: https://lakens.shinyapps.io/p-curves/

# False Discovery Rate (FDR)




[10.1098/rsos.140216]

# FDR, p-values and Discoveries




http://shinyapps.org/apps/PPV/

[10.1098/rsos.140216]

# Properties: what p-values can do

- P-values can indicate how incompatible the data are with a specified statistical model reflecting the null hypothesis
- P-values can indicate if the hypothesis should be further scrutinized
- P-values are part of proper inference which is required for full reporting and transparency

# Properties: what p-values can *not* do

- A p-value does not measure the probability that the studied hypothesis is true
- A p-value does not measure the size of an effect or the importance of a result
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis
- By itself, a p-value should not be the sole factor for scientific conclusions and business or policy decisions

# “Stats” Sports

- 6 principles from the ASA statement
  [http://retractionwatch.com/]
- 12 missconeptions of p-values
  [10.1053/j.seminhematol.2008.04.003]
- 25 missinterpretations of p-values, confidence intervals, and power
  [10.1007/s10654-016-0149-3]
- Ride the wave: “Lies, damned lies and statistics ...”
  [10.1016/j.prrv.2017.02.002]

# Six Principles from the ASA Statement

1. P-values can indicate how incompatible the data are with a specified statistical model
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis

[http://retractionwatch.com/]

# Recommendations, Solutions ...

Only of “temporary” relief:

- Bann p-values
- Lower p-value threshold

Conceptually better:

- Bayesian approaches

**BEST:**

- Statistical literacy and statistical knowledge

University of Zurich[UZH]



# Appendix

# References

Altman, DG: Statistics and ethics in medical research. Misuse of statistics is unethical. *Br Med J,* 1980, 281:6249, 1182–1184 [PMC1714517]

Broman KW, Woo, KH: Data Organization in Spreadsheets, *Am Stat*, 2018: 72:1, 2-10 [10.1080/00031305.2017.1375989]

Bryan J (2018) Excuse Me, Do You Have a Moment to Talk About Version Control?, *Am Stat*, 2018, 72:1, 20-27 [10.1080/00031305.2017.1399928]

Colquhoun D: An investigation of the false discovery rate and the misinterpretation of p-values, *R. Soc. open sci.* 2014; 1: 140216; [10.1098/rsos.140216]

Ellis SE, Leek JT: How to Share Data for Collaboration, *Am Stat*, 2018, 72:1, 53-57 [10.1080/00031305.2017.1375987]

Goodman S: A Dirty Dozen: Twelve P-Value Misconceptions, *Seminars in Hematology*, 2008, 45(3): 135-140 [10.1053/j.seminhematol.2008.04.003]

Greenland S, Senn SJ, Rothman KJ, et al.: Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*, 2016; 31(4):337-50 [10.1007/s10654-016-0149-3]

Marwick B, Carl Boettiger C, Mullen L: Packaging Data Analytical Work Reproducibly Using R (and Friends), *Am Stat*, 2018, 72:1, 80-88, [10.1080/00031305.2017.1375986]

Mellis C: Lies, damned lies and statistics: Clinical importance versus statistical significance in research, *Paediatric Respiratory Reviews,* 2018, 25, 88-93 [10.1016/j.prrv.2017.02.002]