Basic Statistical Issues for Reproducibility: Models, Variability, Extensions

Werner Stahel

Seminar für Statistik, ETH Zürich

Cortona, Sep 6, 2015

Extended Slide Version

0. Thoughts on the Role of Reproducibility

0.1 Paradigms

ETH produces knowledge about facts. Facts are reproducible.

... as opposed to belief, which is “irrational” for some of us.

Science is the collection of knowledge that is “true”.

Reproducibility defines knowledge: “the scientific method”

Well, not qite: Big Bang is not reproducible, but is

a theory, nevertheless is called scientific knowledge.

In fact, empirical science needs theories as its foundation.

“Critical thinking” is needed to purify and advance science.

$\longrightarrow$ Critical thinking initiative started at ETH.

Reproducibility of facts defines science

– physics, chemistry, biology, life science = “Exact” Sciences

Some of you come from

– economy, sociology, psychology, philosophy, theology = Humanities

– literature, painture and sculpture, music = Arts

What is the role of reproducibility in Humanities and Arts?

- Humanities try to become “exact sciences” by adopting “the scientific method”.
- Arts: A composition is a reproducible piece of music.
  Reproducibility achieved by fixing notes.
  Intonation only “reproducible” with recordings.
- Improvization in music ; mandalla in “sculpture”:
  Intention to make something unique, irreproducible.

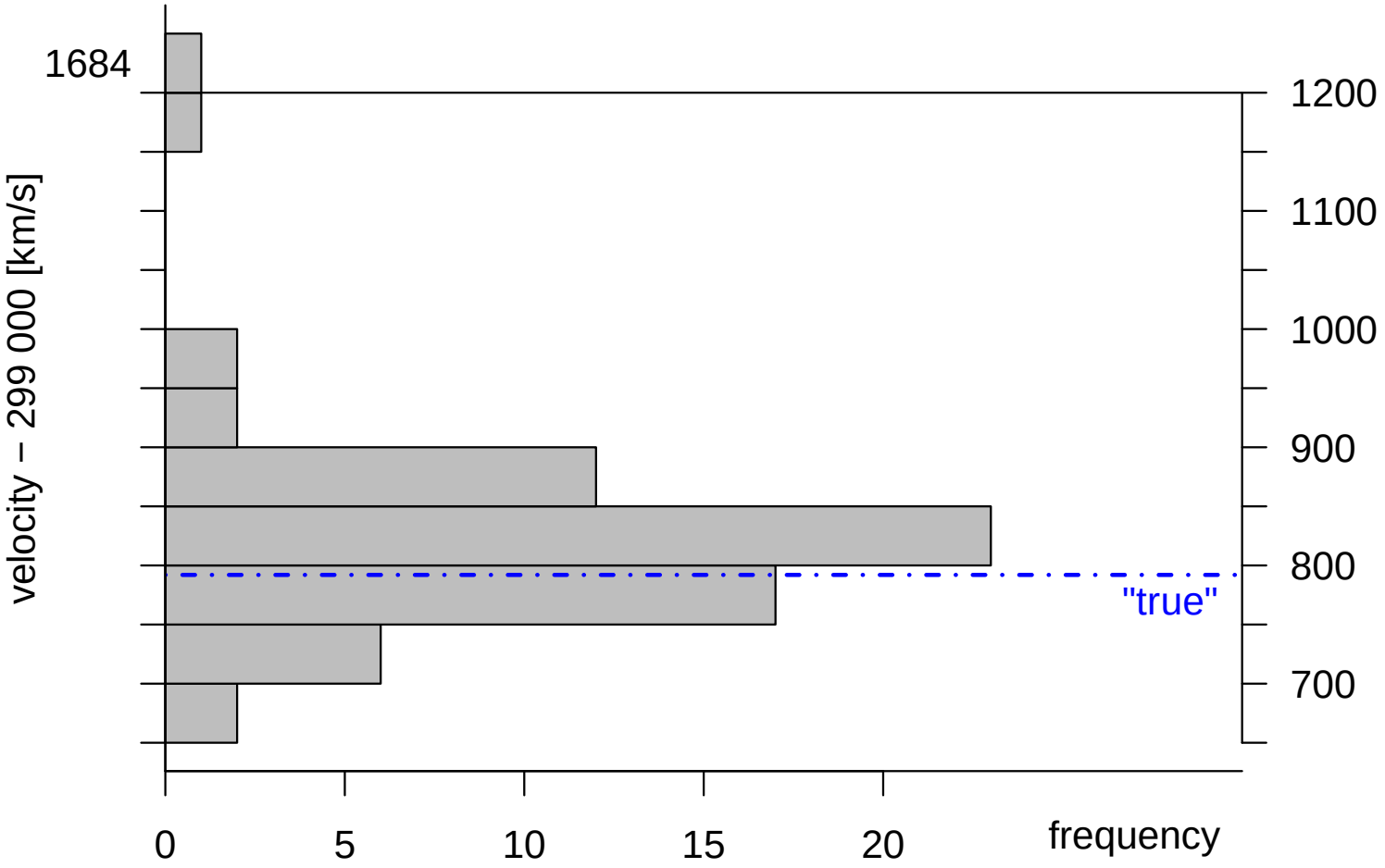
Back to “exact” sciences!

0.2 The Crisis

Reproducibility is a myth in most fields of science!

- Ioannidis, 2005, PLOS Med. 2:
  Why most published research findings are false.
  $\longrightarrow$ many papers, newspaper articles, round tables,
  editorials of journals, ...,
  Topic of Collegium Helveticum $\longrightarrow$ Handbook
- Tagesanzeiger of Aug 28, 2015:
  "Psychologie-Studien sind wenig glaubwürdig"
  $\longrightarrow$ Science (journal)
  We come back to this publication.

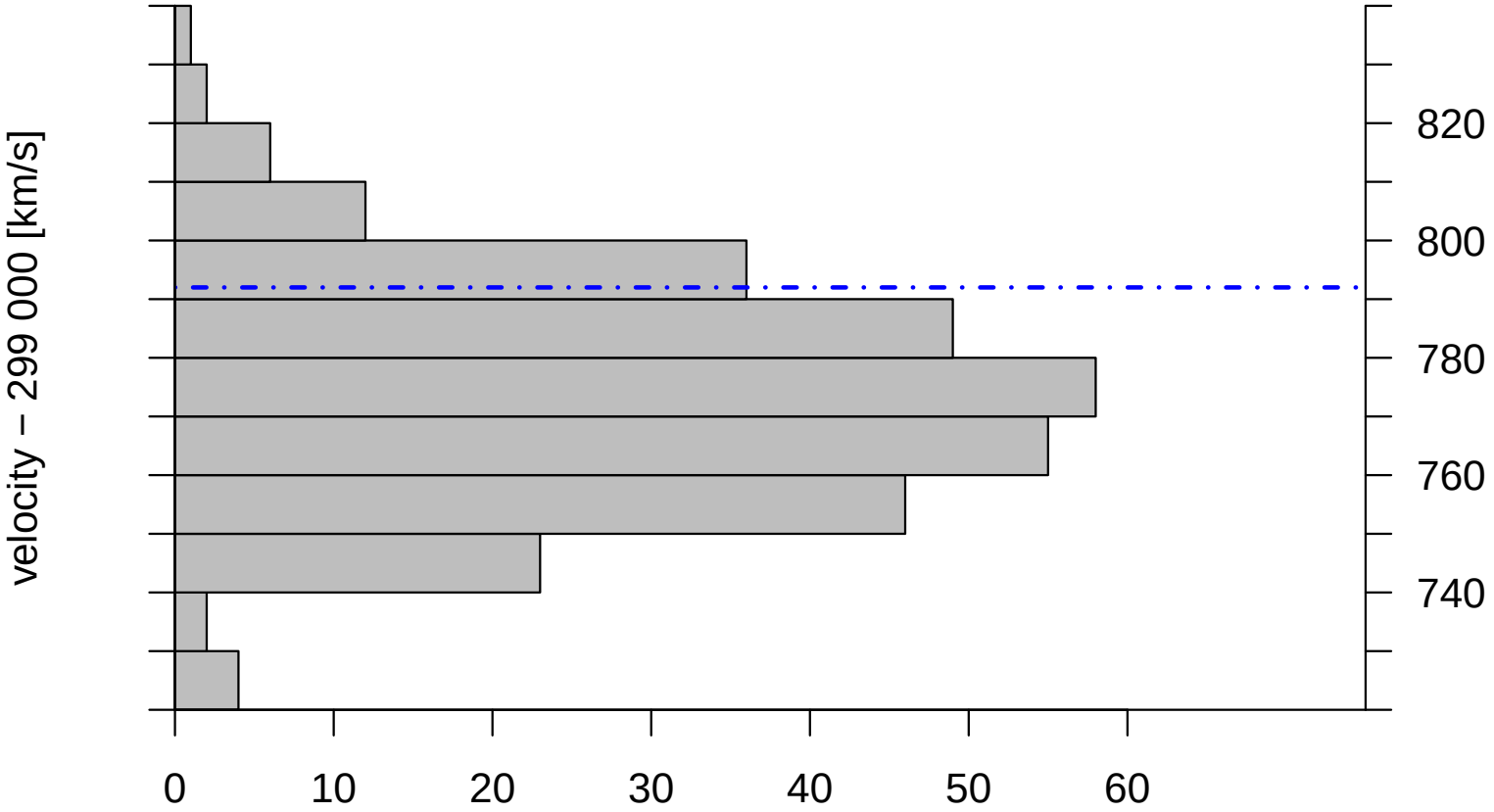
An Example



66 Measurements of the velocity of light by Newcomb, 1882.

Reproduction?

294 measurements by Michelson



Note: smaller scale, narrower range, see later!

0.3 Outline

1. A random sample: Quick rehearsal of basic statistical concepts
2. The significance testing controversy
3. Structures of variation, Correlation, Regression
4. Model development
5. Conclusions: Is reproducibility a useful concept?

1. A Random Sample

Most simple situation. (Velocity of light)

Measurements = random variable X.

Distribution given by "cumulative distribution function" (cdf)

$$F_{\underline{\theta}}(x) = P(X \leq x)$$

Normal distribution $\mathcal{N}(\mu, \sigma^2)$.

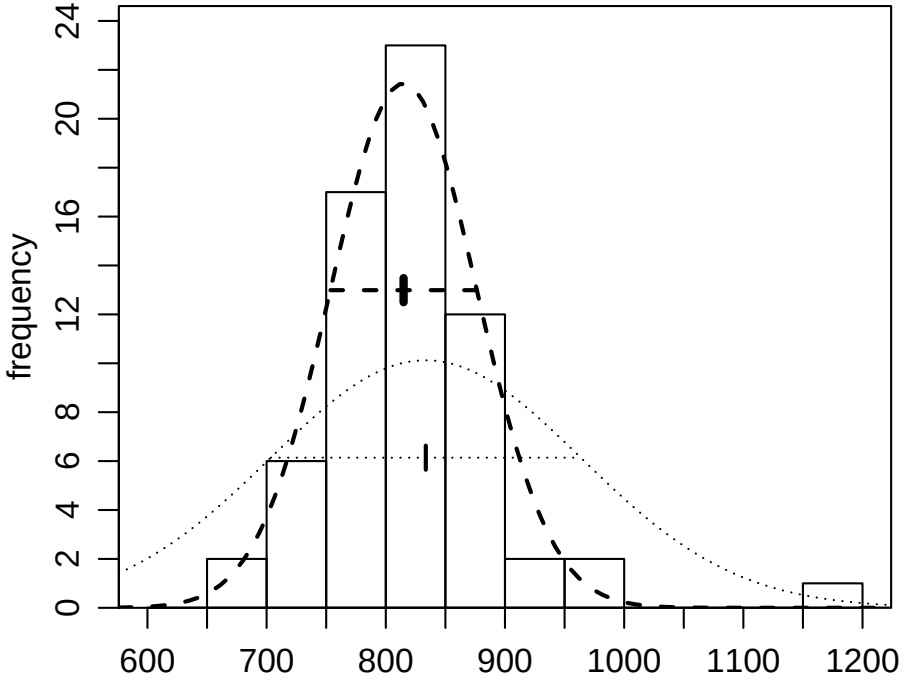
Sample ("simple random sample"):

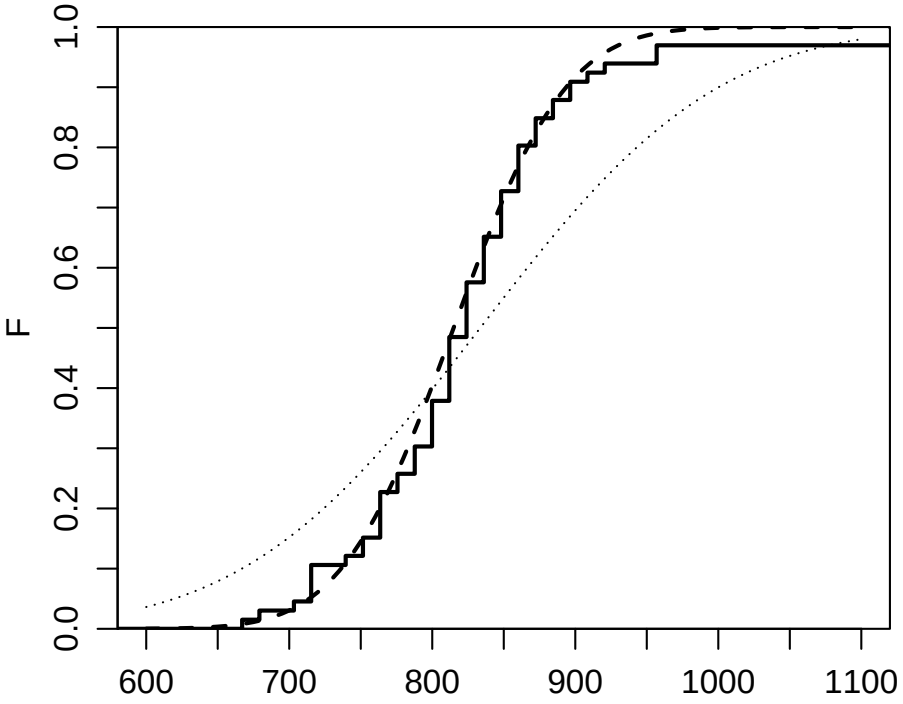
n observations $X_1, X_2, ..., X_n$, $X_i \sim \mathcal{N}(\mu, \sigma^2)$

statistically independent.

Empirical distribution histogram cdf $\widehat{F}(x) = \#\{i|X_i \leq x\}/n$

Theoretical distribution density cdf $F_{\underline{\theta}}(x) = P(X \leq x)$





"Good model": Histogram $\approx$ density and $\widehat{F}(x) \approx F_{\underline{\theta}}(x)$

$\approx$ means: For $n \to \infty$, $\widehat{F}(x) \to F_{\underline{\theta}}(x)$.

Probability theory tells us how fast this happens.

1.1 Statistical Inference

The basic scheme of parametric statistics

A. Postulate a Parametric Model for the Data

B. Find methods for the 3 basic questions of statistical inference:

1. Which value of the parameter(s) is most plausible in the light of the data? $\longrightarrow$ Estimation
2. Is a certain, predetermined value plausible? $\longrightarrow$ Test

3. Which values are plausible (in the sense of the test)?

   $\longrightarrow$ Confidence Interval

Inference for a random sample

A. Model: "Simple Random Sample" $X_i \sim \mathcal{N}(\mu, \sigma^2)$, indep.

B.1 Estimation of μ: mean $\overline{X} = (1/n)\sum_i X_i$.

B.2 Test for null hypothesis $H_0 : \mu = \mu_0$:

Use estim. as a test statistic!: If $|\overline{X} - \mu_0|$ is large, "reject" H_0.

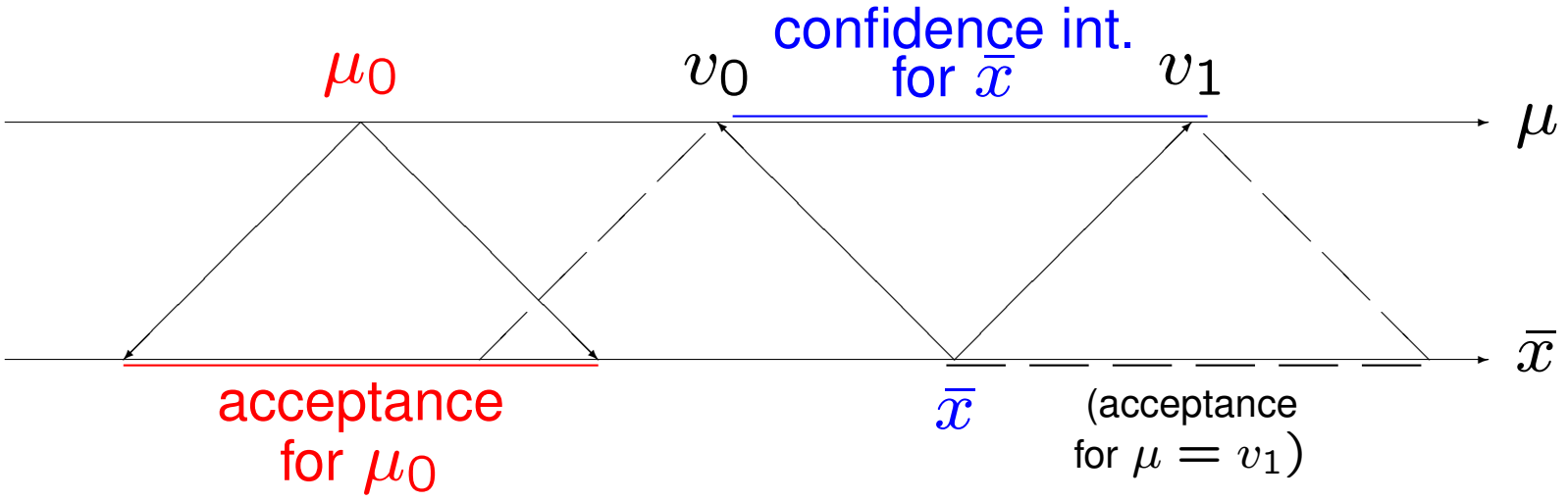
What is large? Need distribution of the test statistic under H_0.

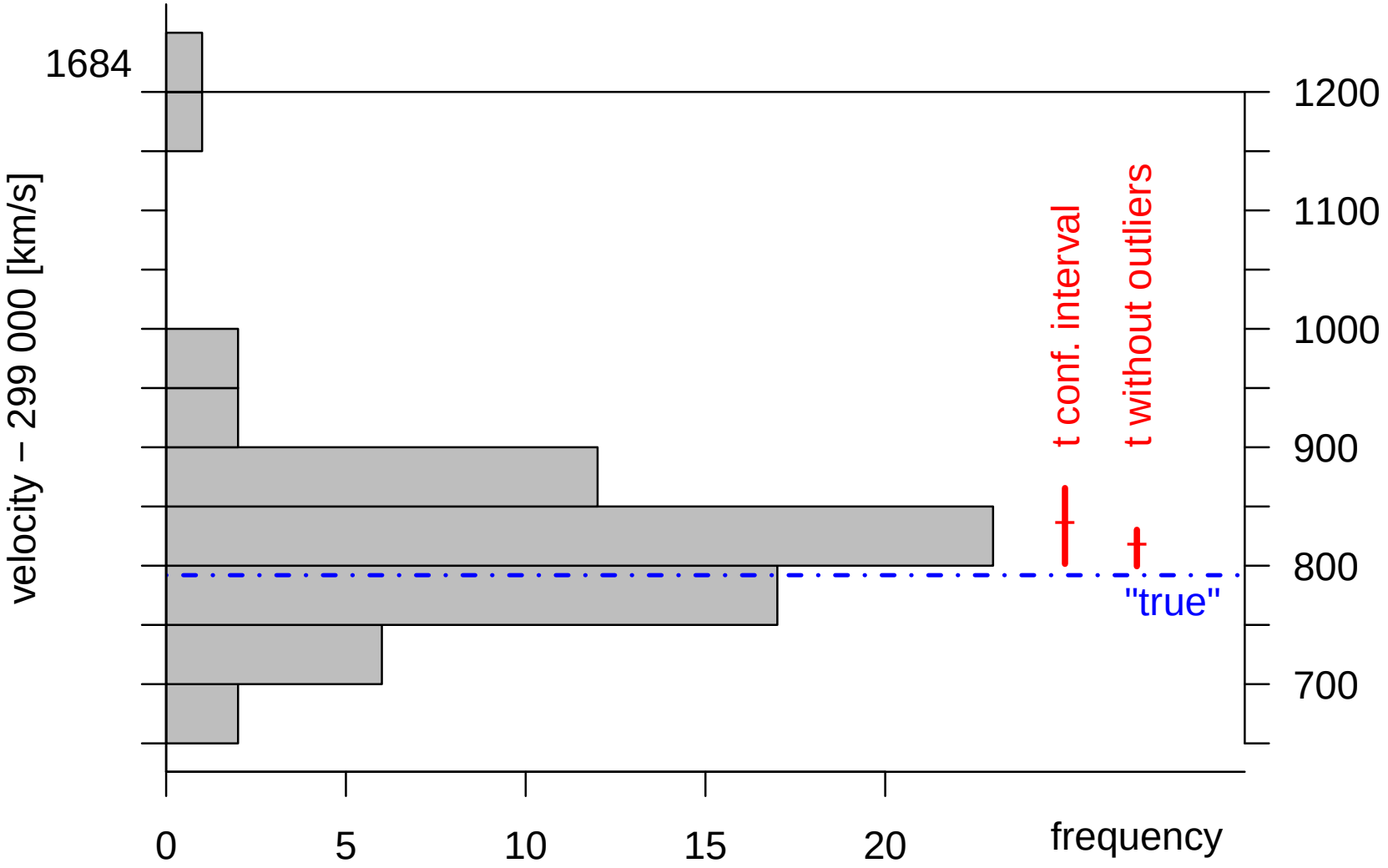
Trick: Standardize t.st. $\longrightarrow$ distr. indep. of parameters (μ_0, σ).

$\longrightarrow$ t-test

B.3 Plausible values of μ? $\longrightarrow$ confidence interval:

$\longrightarrow$ $\bar{x} \pm q\,\mathrm{se}_{\overline{X}}$, $\mathrm{se}_{\overline{X}} = \widehat{\sigma}/\sqrt{n}$, $q \approx 2$.



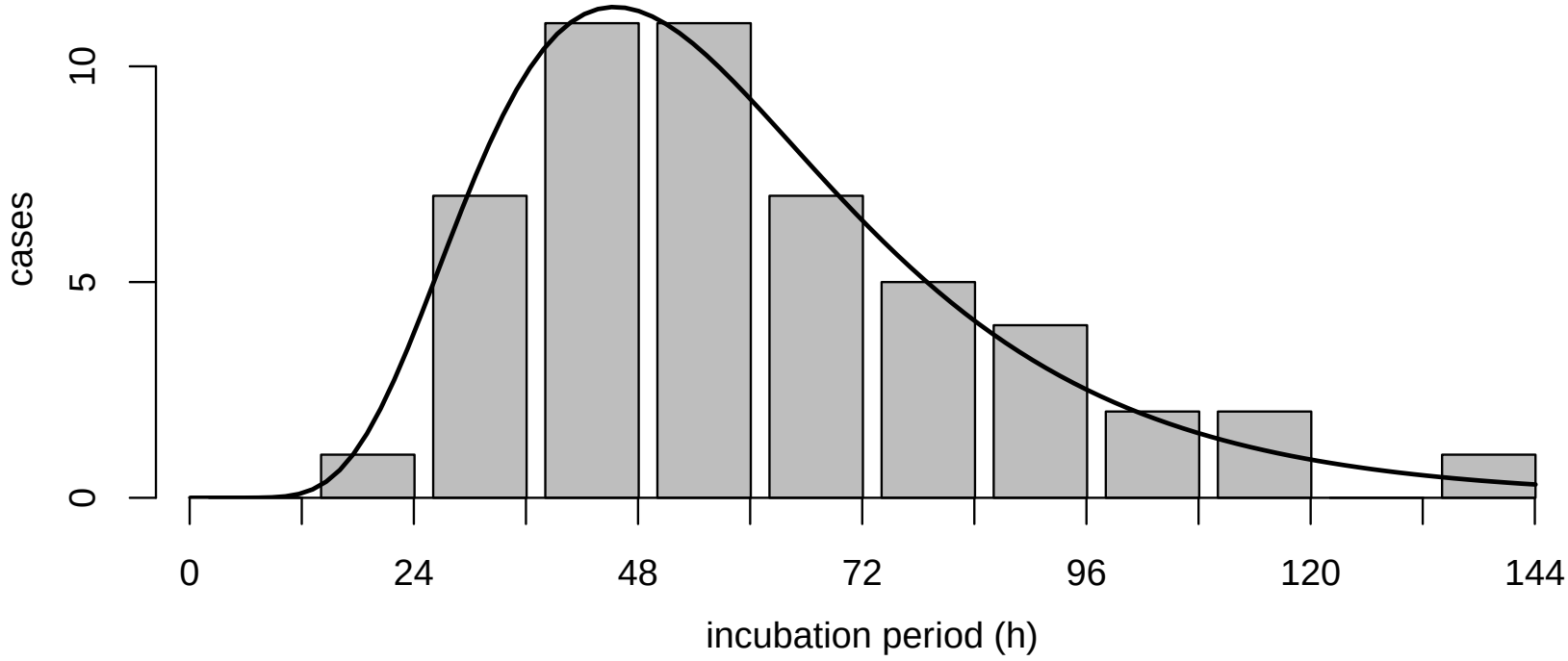


Confidence interval does not cover the true velocity of light.

Too short, for statistical-technical reasons? – Maybe!

Alternative models.

- Observed values from variables that are > 0 usually have a skewed distribution, often a log-normal distribution.
  (Mulitplicative laws of nature lead to the log-normal d.)



- Choose any other model with a good justification.
  $\longrightarrow$ Adjust the methods to the assumed model.

General Parameter

Parametric model F_θ

Estimator $\widehat{\theta}$ obtained by Maximum Likelihood

Distribution of $\widehat{\theta}$ under F_θ: approx: $\widehat{\theta} \approx\sim \mathcal{N}(\theta, V/n)$,

V: "asymptotic variance"

$\longrightarrow$ confidence interval $\widehat{\theta} \pm 2 \cdot \sqrt{V/n}$

1.2 Role of Assumptions

Determination of the distribution requires large dataset.

What if the model for the data is not correct?

(What does “correct” mean? Can a model be correct?)

- “robust statistics”
- Better: choose “nonparametric” methods:

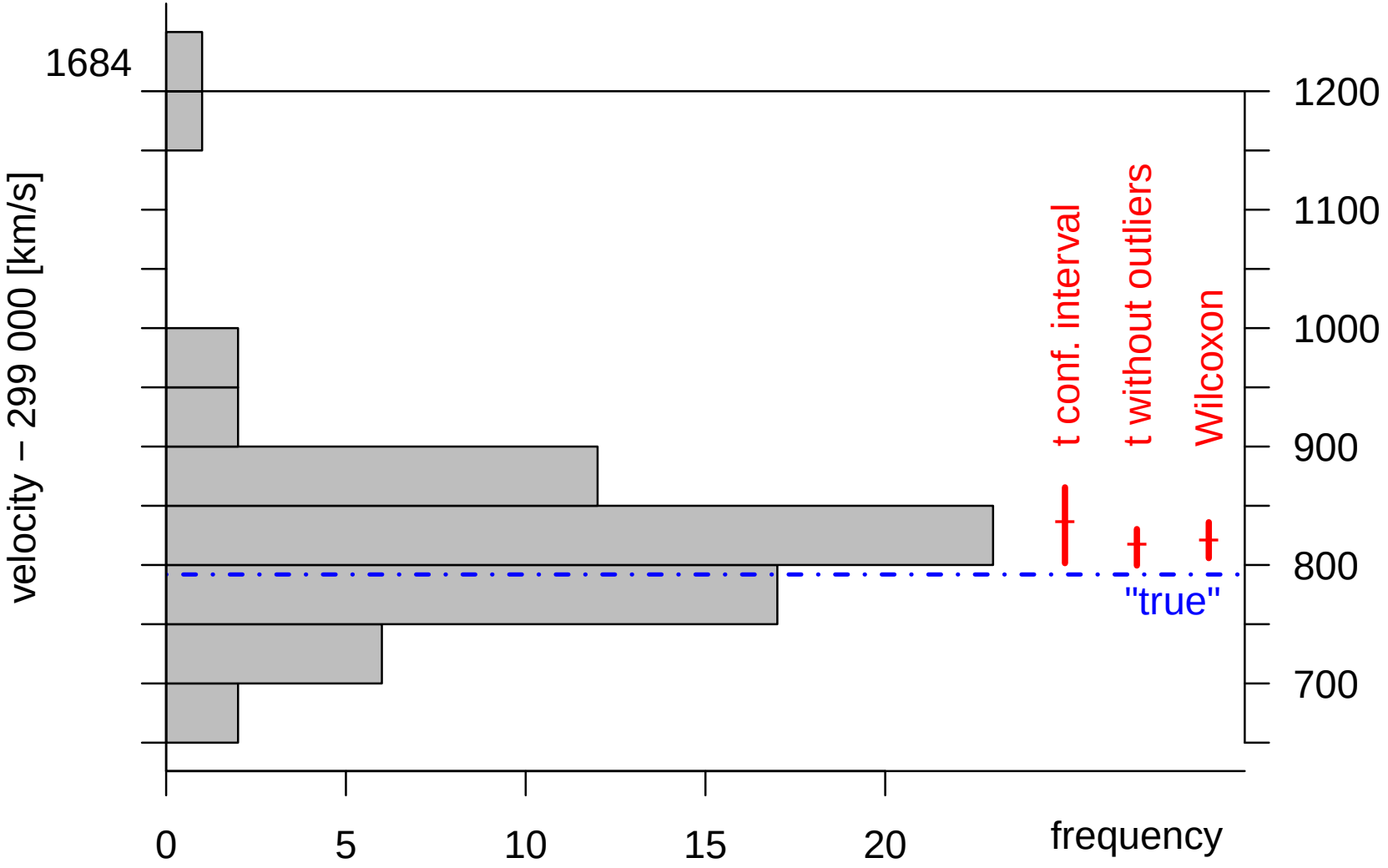
  distribution of test statistic does not depend on model $F_{\underline{\theta}}$

  ... well, as long as it is symmetric...

  $\longrightarrow$ Rank methods, Wilcoxon signed rank test

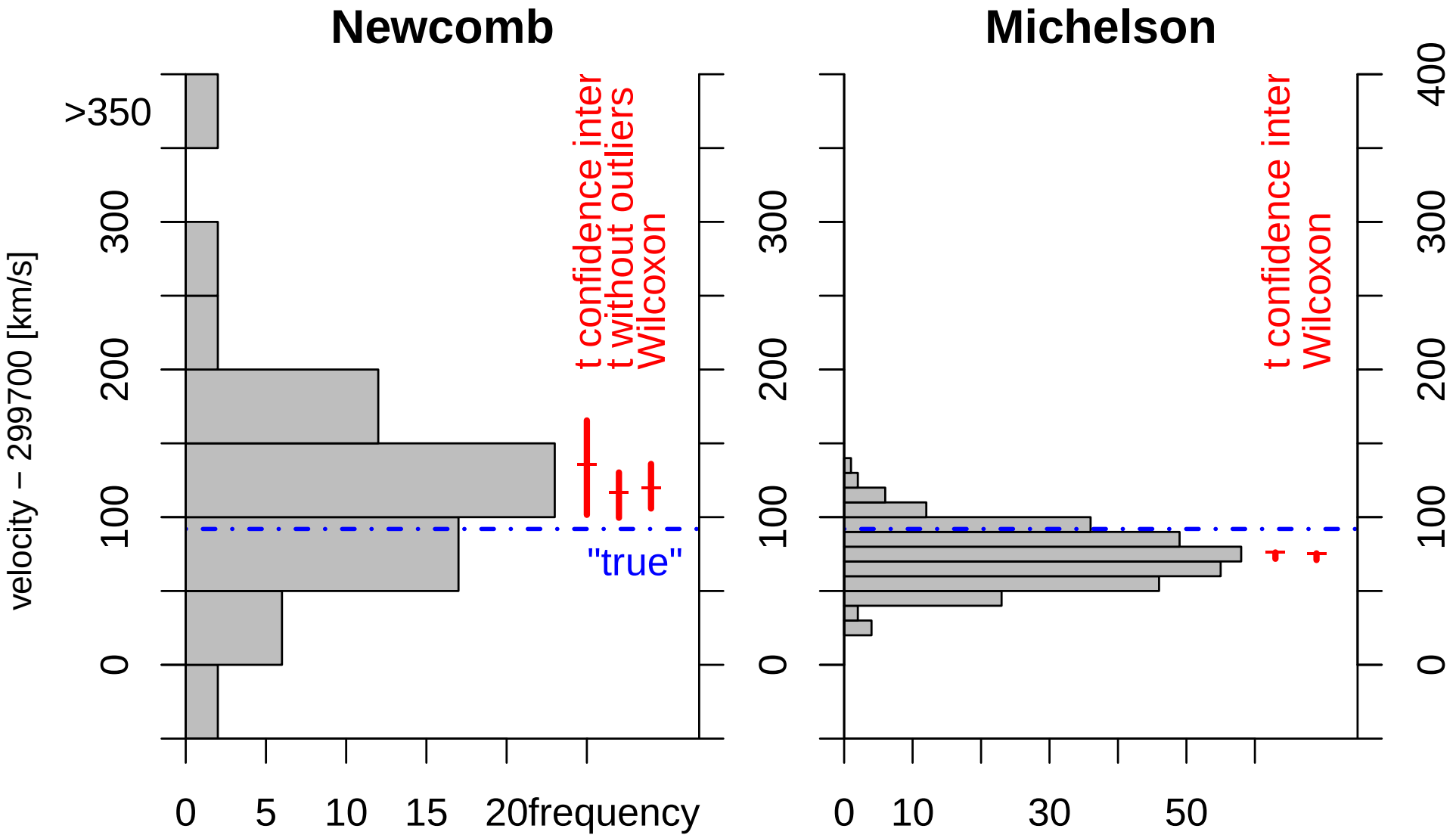
  and respective confidence interval!

  This is a general recommendation!



Examples: similar to t interval (without Newcomb's outliers!)

1.3 Reproducibility?



1.3 Reproducibility?

Overlap of confidence intervals is not quite the correct criterion!

Original study: $\widehat{\theta}_0 \sim \mathcal{N}(\theta_0, \mathrm{se}_0^2)$

Replication: $\widehat{\theta}_1 \sim \mathcal{N}(\theta_1, \mathrm{se}_1^2)$

($\longrightarrow$ Different precision allowed.)

Test for $H_0 : \theta_1 - \theta_0 = 0$? $\widehat{\theta}_1 - \widehat{\theta}_0 \sim \mathcal{N}(0, \mathrm{se}_0^2 + \mathrm{se}_1^2)$

$\longrightarrow$ confidence interval $\widehat{\theta}_1 - \widehat{\theta}_0 \pm 2\sqrt{\mathrm{se}_0^2 + \mathrm{se}_1^2}$.

Does it include 0?

Experience tells that the test usually rejects.

Why?

- Original or replication study not properly done or analyzed
- Improved experimental methods have reduced systematic error
- Statistical model needs improvement!
- ... (see later!)

“Stay with us! We will be back soon!”

2. The significance testing controversy

Rule in most of the sciences:

An effect must not be discussed if it is statistically insignificant.

Filter against publications with spurious effects.

Has been perverted into an

industry producing statistically significant effects!

2.1 The testing paradoxon

- There is “always” a tiny effect – even if clearly irrelevant
- If n increases, the power of any sensible test $\longrightarrow 1$

  $\longrightarrow$ The test does not answer the question if there is an effect (there is “always” one), but whether the sample was large enough to make it significant.

$\longrightarrow$ Only look for relevant effects!

Test $H_0 : \mu \leq c$, where c is the threshold for “relevant”.

How to choose c? – Not needed: use confidence interval for communication!

2.2 Reproducibility of test results

Cases for "truth" and results of original test: 4 cases.

Probability P of obtaining the same result in the replication

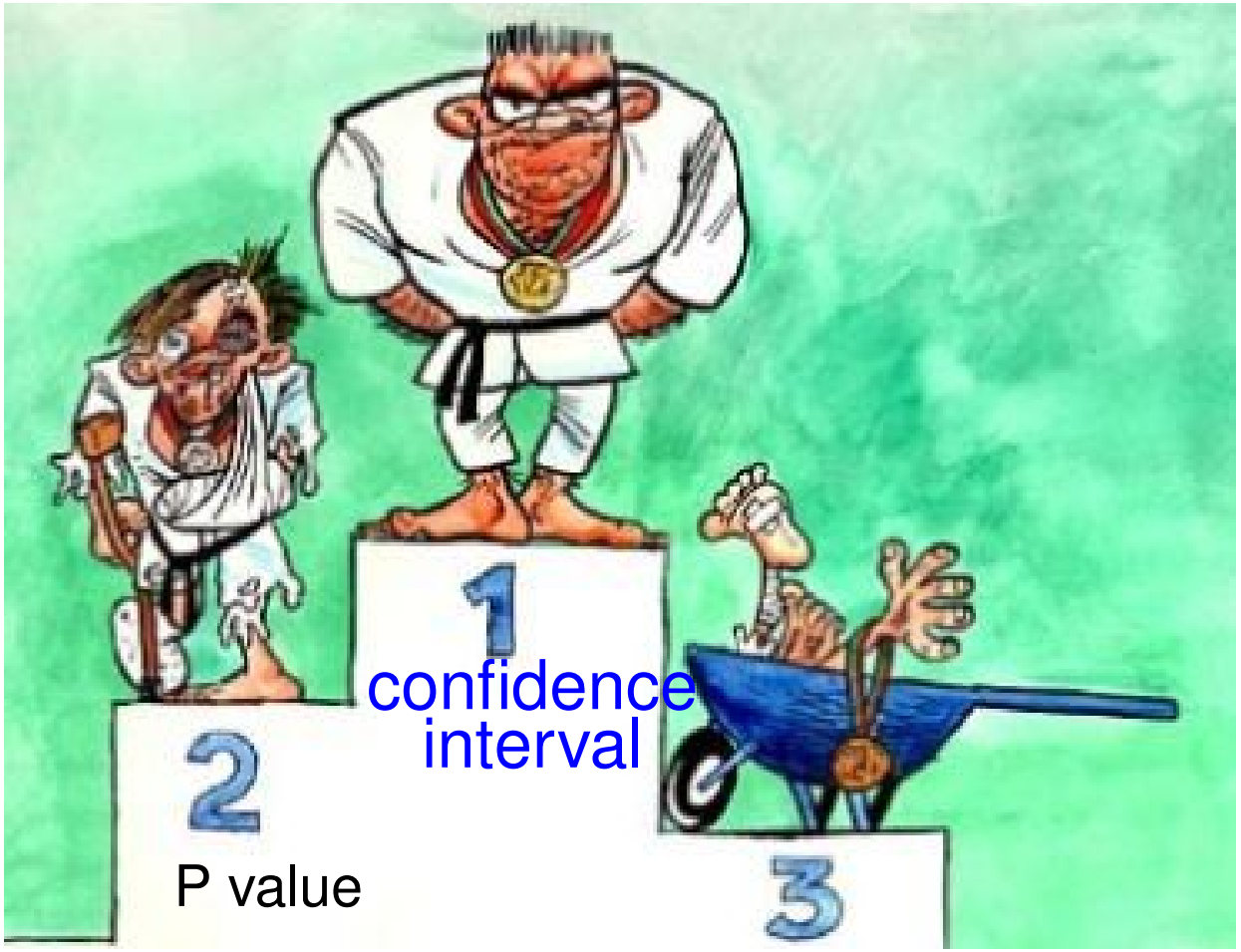
	test result	
	non-significant	significant
H_0	$P = 95\%$	$P = 5\%$ (*)
H_A	$P =$ 1−power (*)	$P =$ power

(*) we do not want to replicate these wrong results!

$\longrightarrow$ The probability of wanting and getting the same result is only high for clear effects and sufficient sample sizes to make the power large in both studies.

In 1999, a committee of psychologists came close to a

ban of the statistical test! $\longrightarrow$ Use confidence intervals!



3. Reproducibility: Empirical results

3.1 The topic of Reproducibility is hot!

Tagesanzeiger of Aug 28:

Psychologie-Studien sind wenig glaubwürdig

(Studies in psychology are little trustworthy)

“Open Science Collaboration”, Science 349, 943-952, Aug 28, 2015:

“Estimating the reproducibility of psychological science”

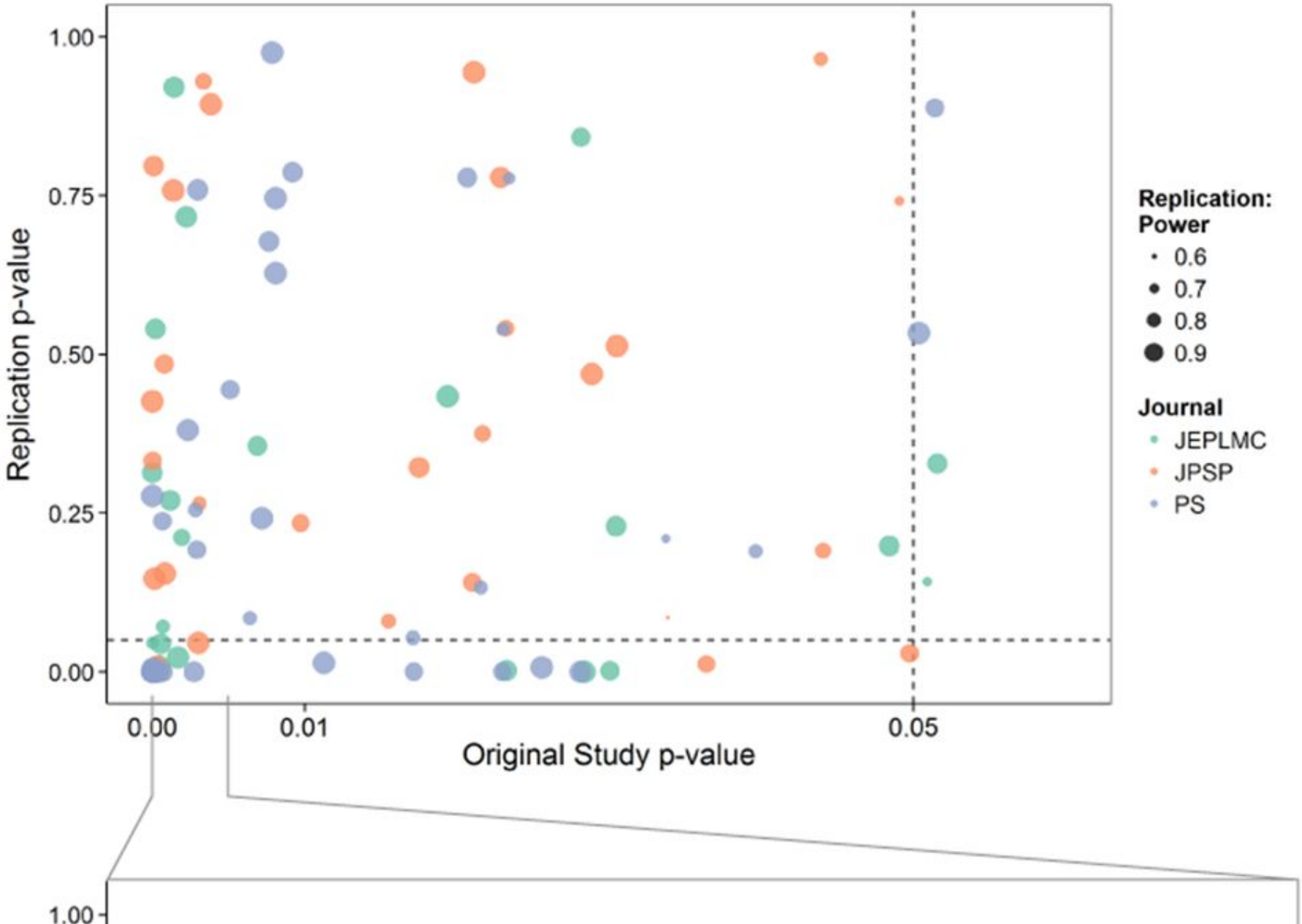
100 research articles from high-ranking psychological journals.

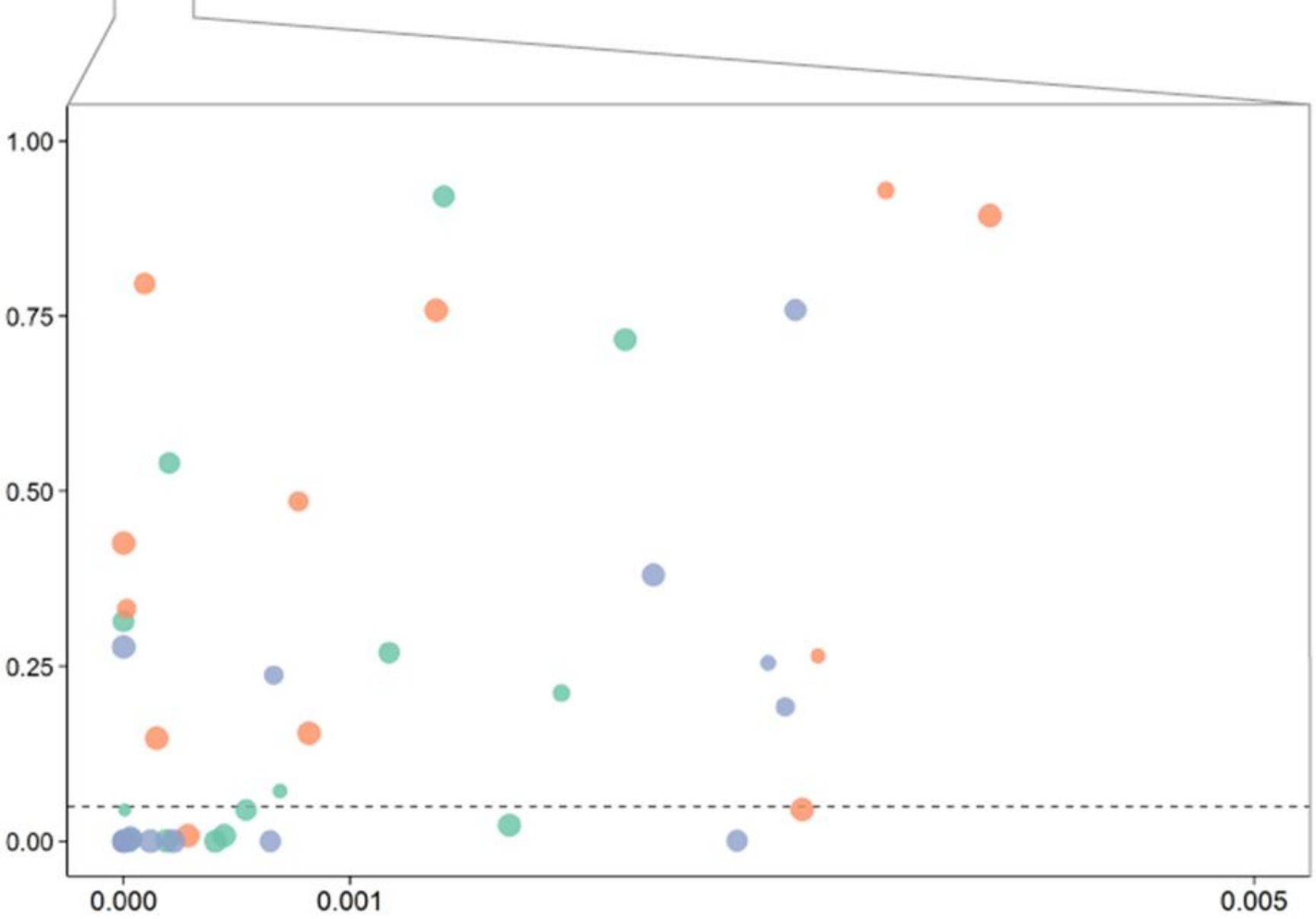
260 collaboraters attempt to reproduce 1 result for each.

Effect size could be expressed as a correlation

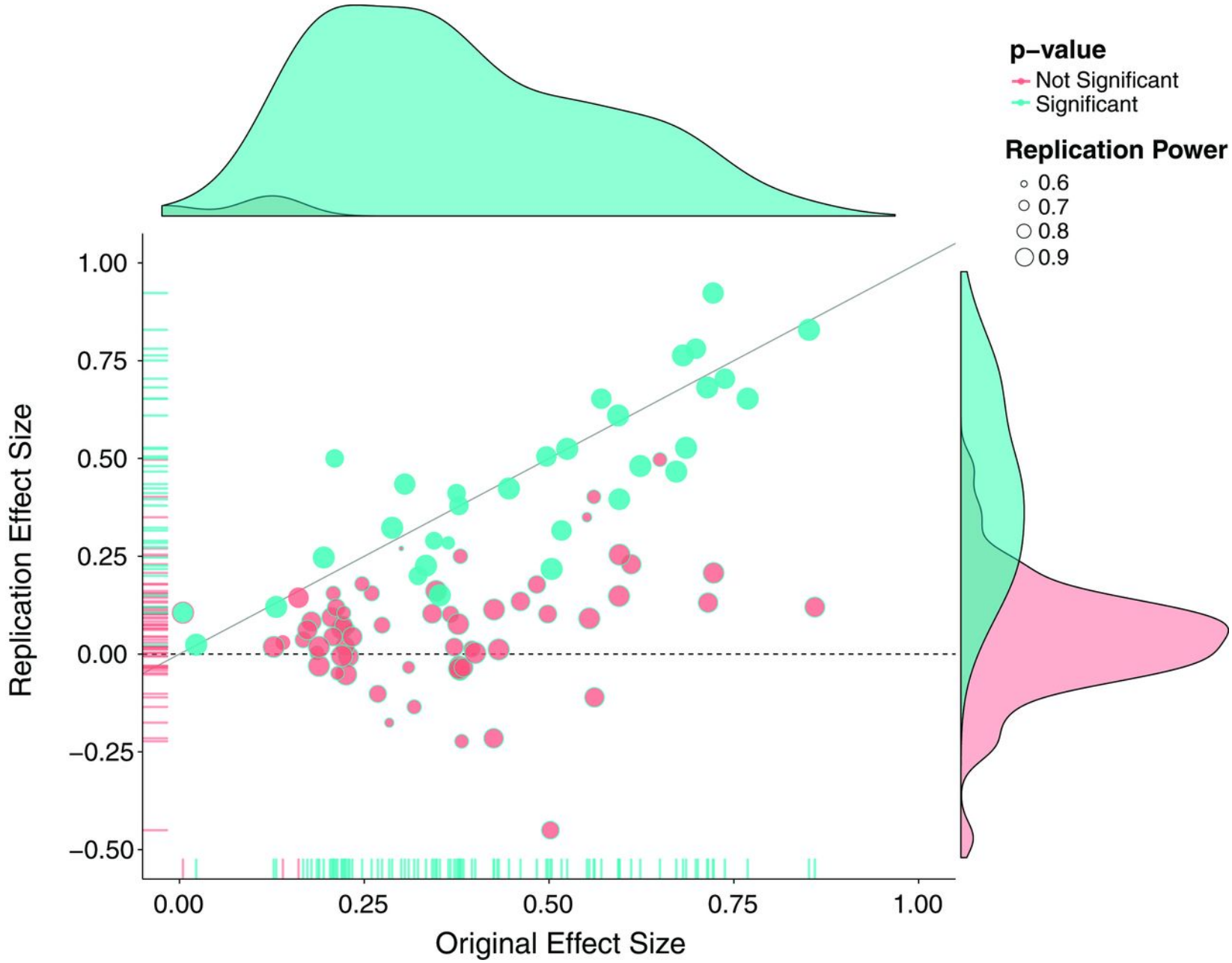
$\longrightarrow$ P-values, confidence intervals.

P-values


1.00
0.75
0.50
0.25
0.00
Replication p-value
0.00
0.01
0.05
Original Study p-value
Replication: Power
0.6
0.7
0.8
0.9
Journal
JEPLMC
JPSP
PS
1.00


1.00
0.75
0.50
0.25
0.00
0.000
0.001
0.005

Effect Size


p–value
Not Significant
Significant
Replication Power
0.6
0.7
0.8
0.9
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Replication Effect Size
0.00
0.25
0.50
0.75
1.00
Original Effect Size

$\longrightarrow$ Effect sizes are lower, as a rule, in the replication.

Significant difference in effect size?

was not studied!!!

Instead: only 47% of the confidence intervals of the repr.study

covered the original estimated effect!

Similar results for pharmaceutical trials, Genetic effects, ...

Note:

What is a success/failure of a reproduction? $\longrightarrow$ **not well defined!**

... not even in the case of assessing just a single effect!

Why does replication fail?

Data manipulation? Biased experiment?

3.2 Multiple comparisons and multiple testing

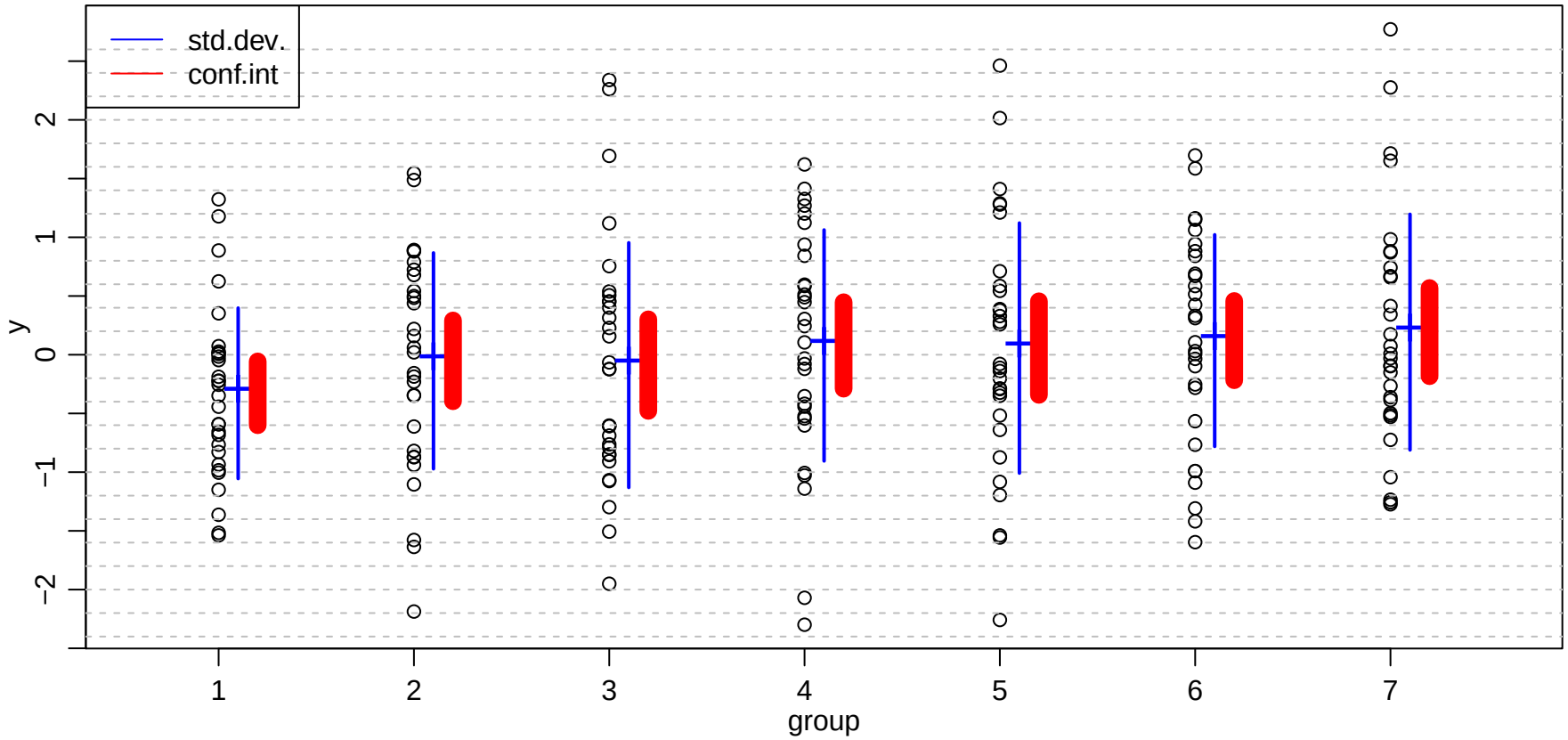
Here is a common way of learning from empirical studies:

- visualize data,
- see patterns (unexpected, but with sensible interpretation),
- test if statistically significant,
- if yes, publish.

(cf. “industry producing statistically significant effects”)

The problem, formalized

7 groups, generated by random numbers $\sim \mathcal{N}(0, 1)$. $\longrightarrow H_0$ true!



Test each pair of groups for a difference in expected values.

$\longrightarrow 7 \cdot 6/2 = 21$ tests. $P(\text{rejection}) = 0.05$ for each test.

$\longrightarrow$ Expected number of significant test results = 1.05!

significant differences for 1 vs. 6 and 1 vs. 7 Publish the significant result!

You will certainly find an explanation why it makes sense...

$\longrightarrow$ Selection bias.

Solution: for multiple ("all pairs") comparisons:

- Make a single test for the hypothesis that all μ_g are equal!
  $\longrightarrow$ F-test for factors.
- Lower the level α for each of the 21 tests such that
  $P(\geq 1 \text{significant test result}) \leq \alpha = 0.05$!
  Bonferroni correction: divide α by number of tests.
  $\longrightarrow$ conservative testing procedure $\longrightarrow$ You will get no significant results $\longrightarrow$ nothing published

(Are we back to testing? –
Considerations also apply to confidence intervals!)

In reality, it is even worse!

When exploring data, nobody knows how many hypotheses are "informally tested" by visual inspection of informative graphs.

Exploratory data analysis – curse or benediction?

Solution?

One dataset – one test!

(or: only a small number of planned tests/confidence intervals)

3.3 Stepping procedure of advancing science:

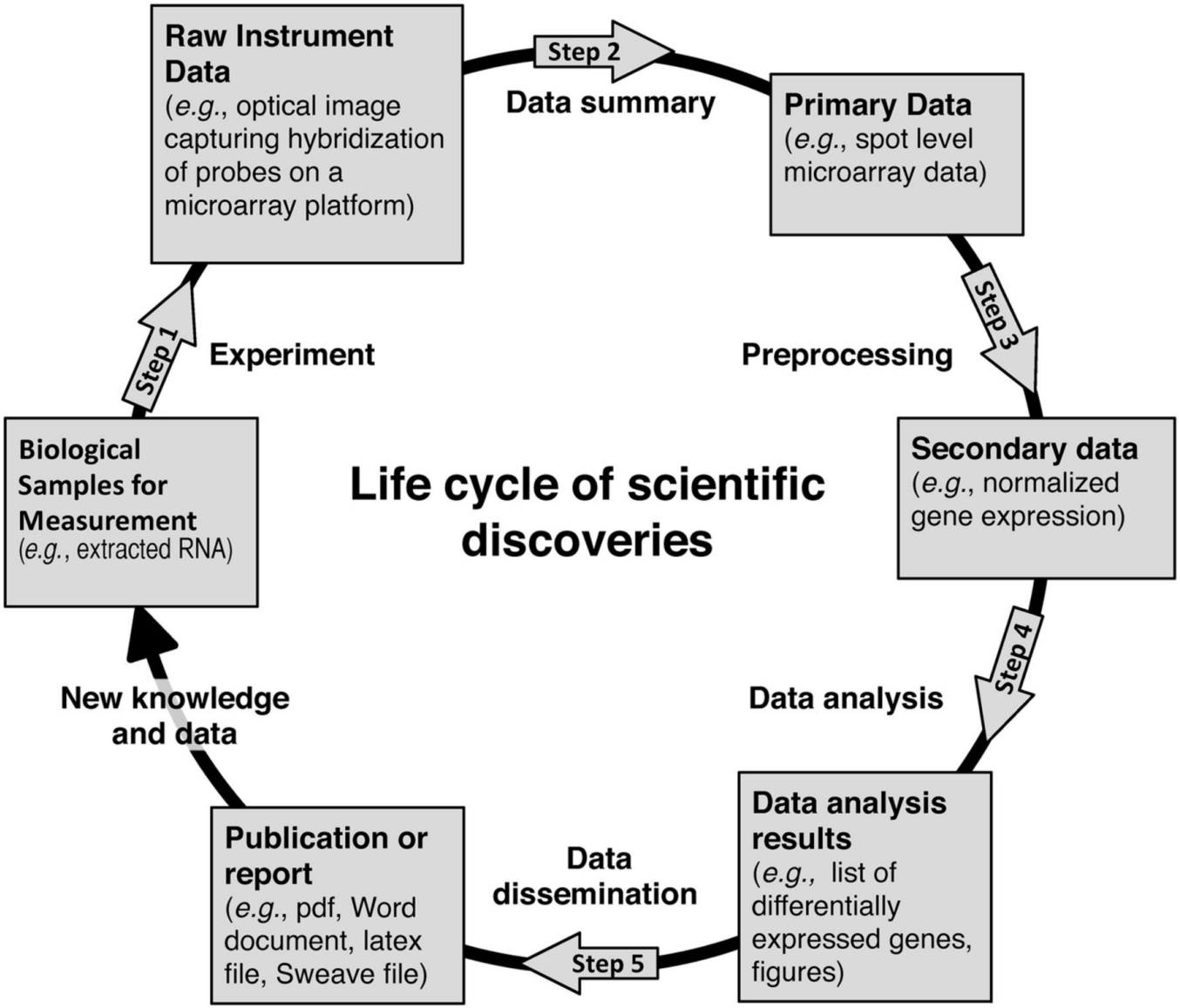
1. Explore data freely, allowing all creativity
   Create Hypotheses about relevant effects

2. Conduct a new study to confirm the hypotheses (not H_0!)
   "Believe" effects that are successfully confirmed
   (with a sufficient magnitude to be relevant!)

1.* Use dataset in an exploratory attitude to generate new hypotheses.

it. Iterate until retirement.

Note that step 2 is a phony replication!



Huang & Gottardo, Briefings in Bioinformatics 14 (2012), 391-401

4. Structures of variation, correlation, regression

Experience: Measurements of the same quantity made

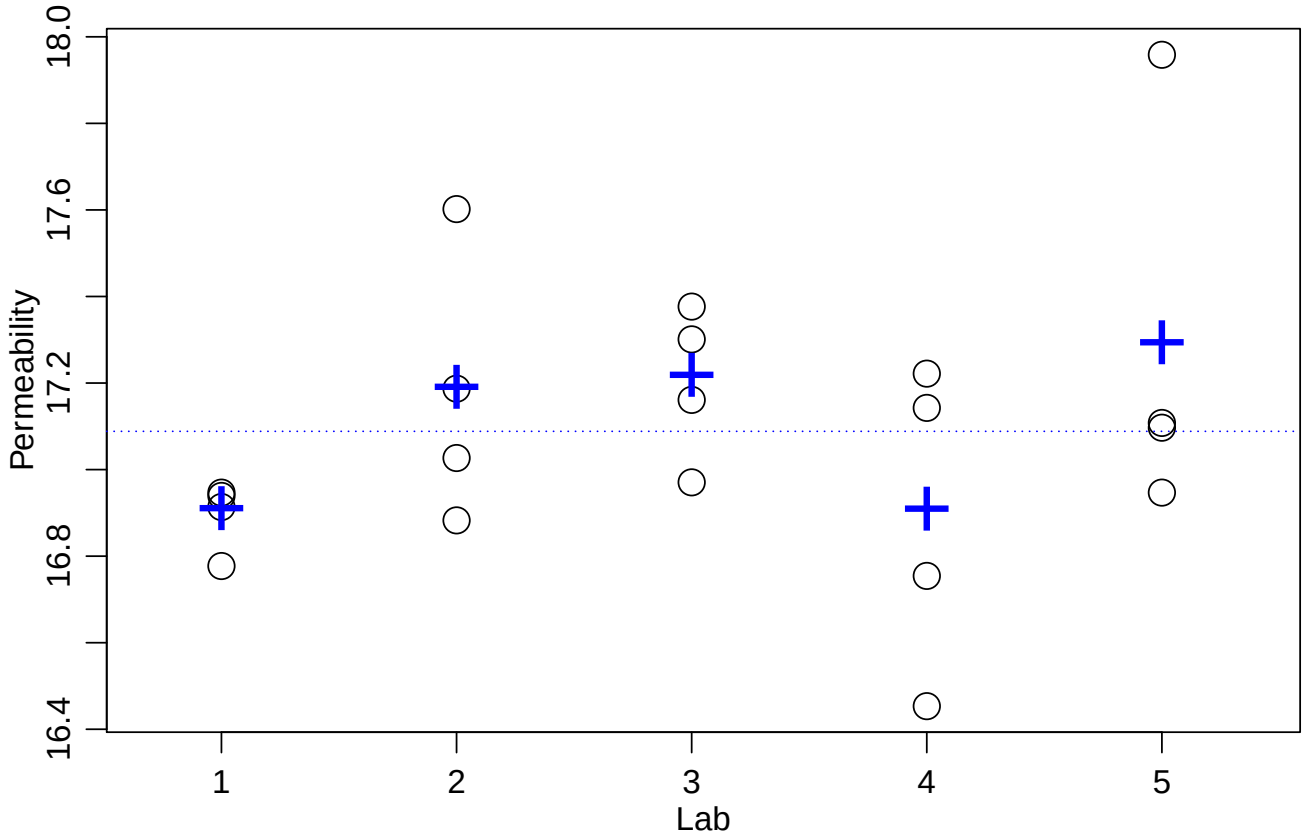
- on the same day
- by the same device / person / ...
- on the same field, genotype, subject, ...
- in the same study

are more similar than if made on different days, devices, ...

4.1 Interlaboratory studies

Send $I = 4$ samples of the same material

to each of $G = 5$ laboratories g.



permeability of concrete

Is there a group (lab) effect? $\longrightarrow$ Model!

$$Y_{gi} = \mu + A_g + E_{gi}\ , \quad E_{gi} \sim \mathcal{N}(0, \sigma^2)\ .$$

A_g: Effect of the laboratory, modelled as random, $A_g \sim \mathcal{N}(0, \sigma_A^2)$.

Think of an analogy between labs and studies.

Variance of a deviation between measurement and wanted value:

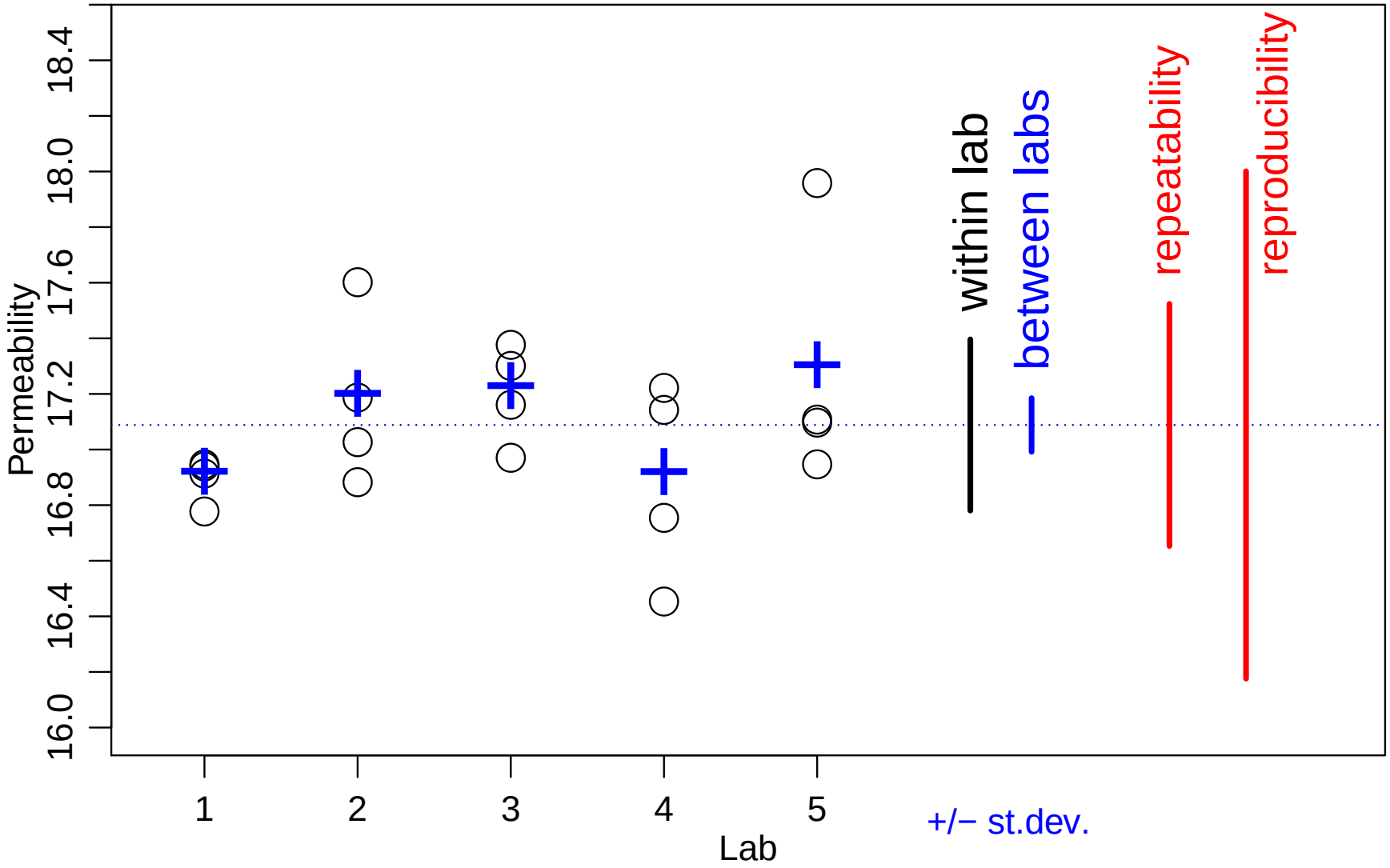
$$\mathrm{var}(Y_{gi} - \mu) = \mathrm{var}(A_g) + \mathrm{var}(E_{gi}) = \sigma_A^2 + \sigma^2$$

σ_A^2, σ^2: "variance components". (There may be > 2 of them.)

σ : standard deviation within lab (study)

σ_A: standard deviation between labs (study effects)

Estimation needs a version of Maximum Likelihood.


Permeability
16.0
16.4
16.8
17.2
17.6
18.0
18.4
1
2
3
4
5
Lab
within lab
between labs
repeatability
reproducibility
+/− st.dev.

Consequences:

- Difference of Y's within a lab (study):

  $$Y_{gi} - Y_{gi'} = E_{gi} - E_{gi'}$$
  $$\longrightarrow \mathrm{var}(Y_{gi} - Y_{gi'}) = 2\sigma^2$$
  $\longrightarrow$ Interval of length $\ell_{repeat} = 2\sqrt{2}\,\sigma$ covers difference between 2 measurements in the same lab (study).

  $\longrightarrow$ ℓ_{repeat} called repeatability.

- Difference of Y's from 2 different labs (studies):

  $$Y_{gi} - Y_{g'i'} = A_g + E_{gi} - A_{g'} - E_{g'i'}$$
  $$\longrightarrow \mathrm{var}(Y_{gi} - Y_{g'i'}) = 2(\sigma_A^2 + \sigma^2).$$
  $\longrightarrow$ Interval of length $\ell_{reprod} = 2\sqrt{2}\,\sqrt{\sigma_A^2 + \sigma^2}$ covers diff. between 2 measurements in different labs (studies).

  $\longrightarrow$ ℓ_{reprod} called reproducibility.

Useful for (replication) studies?

Each study should estimate the same effect.

$\longrightarrow$ 2 variance components, "within study" and "between studies"!

Difficulty: Need many studies (!) to estimate σ^2_A

$\longrightarrow$ or instead, need additional, possibly informal,
information on study-to-study variability.

$\longrightarrow$ in any case, these considerations provide a (valid)
excuse for missing the reproducibility goal!

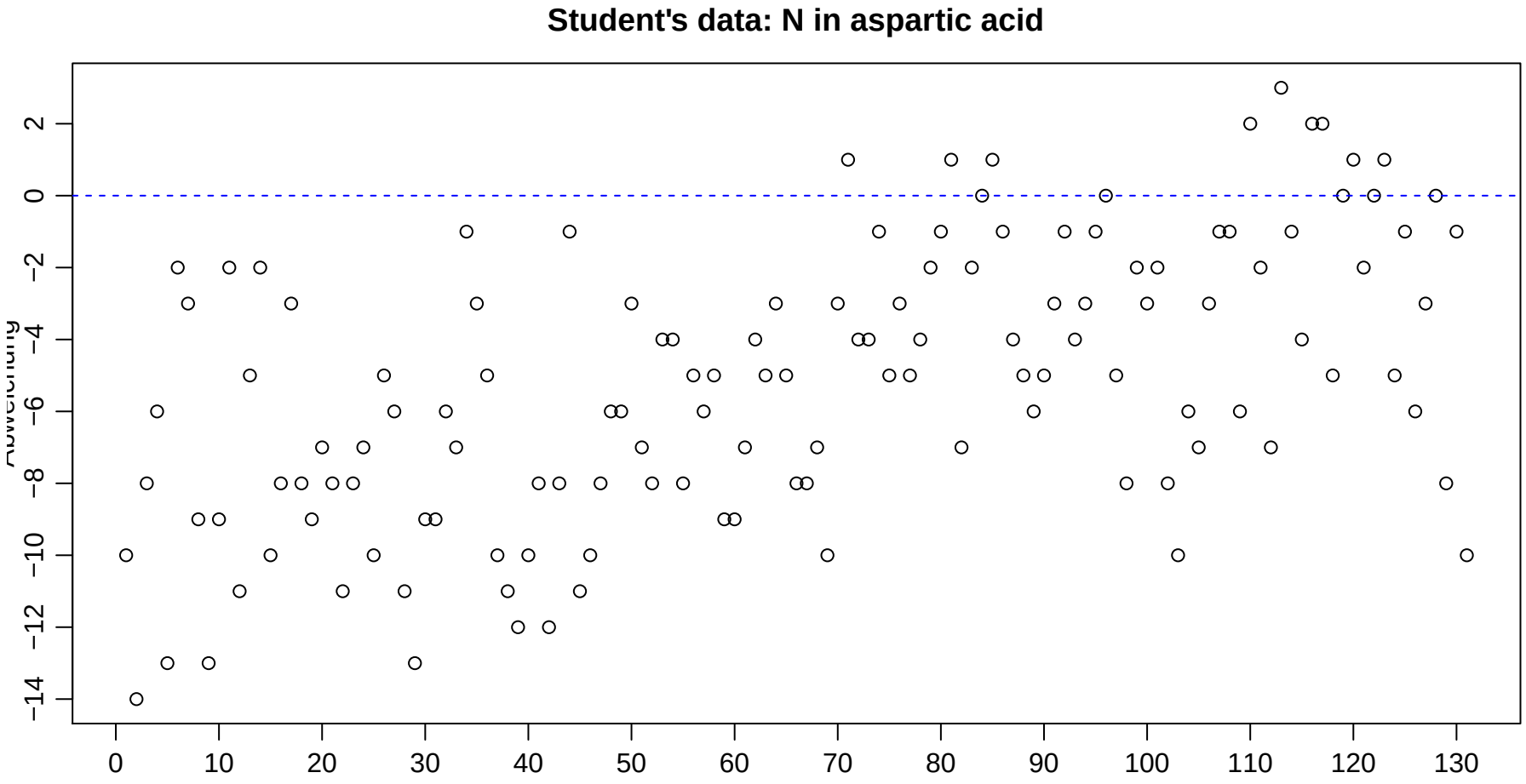
4.2 Correlation

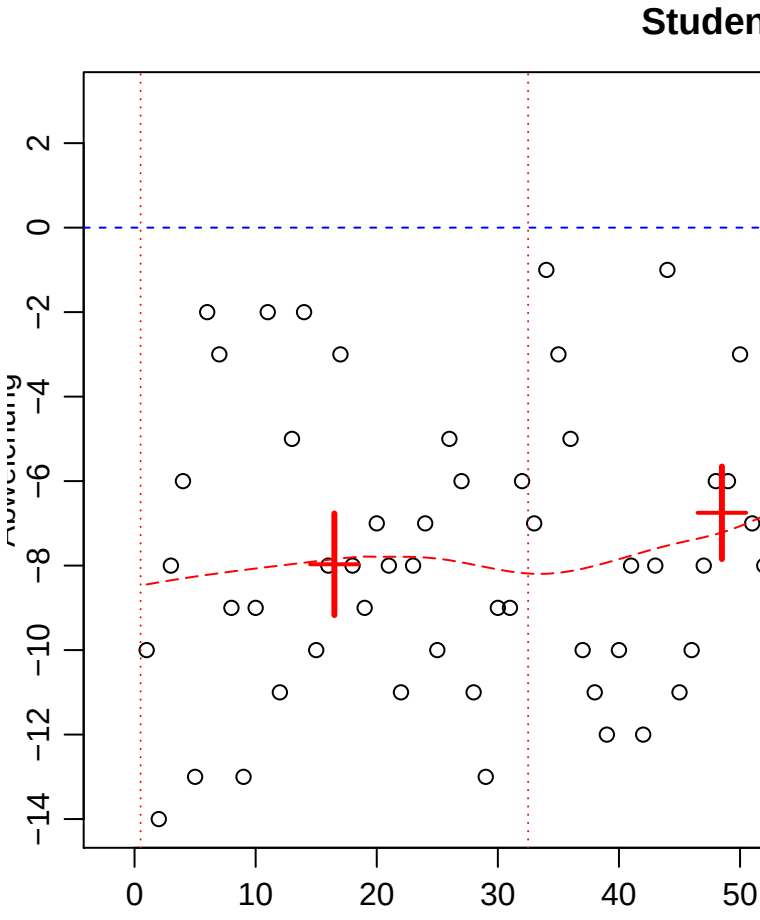
Historical example.

131 measurements of a known quantity

(nitrogen content of aspartic acid, by Student 1927).

Prototype experiment for replication! Simple random sample!


Student's data: N in aspartic acid
Asparticacid
2
0
−2
−4
−6
−8
−10
−12
−14
0
10
20
30
40
50
60
70
80
90
100
110
120
130



Cut into 4 parts. (“Simulation of replication studies”)

Failure to reproduce the result

within the statistically allowed margins as obtained

under the assumption of independence.

Clear time series type dependence, autocorrelation > 0.

$\longrightarrow$ Model correlation with a time series model!

Probability theory then yields longer confidence intervals!

Note correspondence with the model of variance components!

Is statistics hopeless?

- Generate contrasts!

  Compare 5 treatments $\longrightarrow$ ask 5 measurem. from each lab.

  $\longrightarrow$ Differences between treatments

  will not be affected by the lab effect.

$\longrightarrow$ Experimental design!

- Use blocks of experimental units that are homogeneous

  (location, time, conditions)

- Use blocks as different as possible for generalizability of results.

- Randomize the treatments (or use special designs like latin sq.)!

- Include all potential nuisance effects into the model.

4.3 Regression

Simple regression: Response variable Y "depends on"

"input variable" X

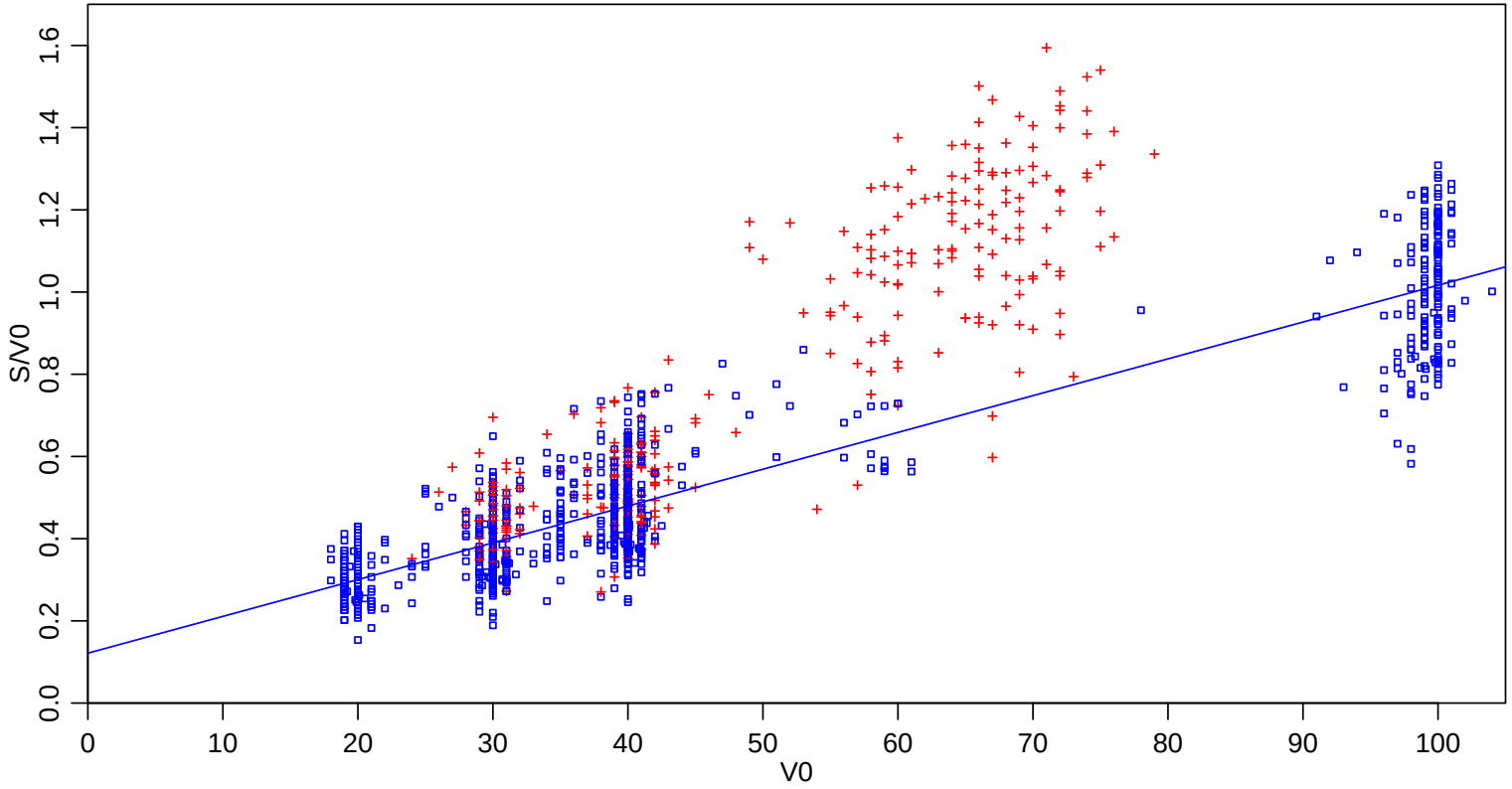
$$Y_i = \beta_0 + \beta_1 x_i + E_i \ , \quad E_i \sim \mathcal{N}\left(0, \sigma^2\right) \ , \text{ independent}$$

Example: Distances needed for stopping freight trains.

Distance S, velocity V0

$$\mathtt{S}_i = \beta_0 \mathtt{V0}_i + \beta_1 \mathtt{V0}_i^2 + \widetilde{E}_i \qquad \text{quadratic in V0}$$

$$(\mathtt{S/V0})_i = \beta_0 + \beta_1 \mathtt{V0}_i + E_i \qquad \text{linear in V0}$$



Multiple regression: Response variable Y "depends on"
several to many "input variables" $X^{(j)}$

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... + E_i$$

Example: `Inclination` as another input variable,
and many more, see later.

No assumptions on $x^{(j)}$. This makes the model very flexible:

- binary variable $\longrightarrow$ model for 2 groups
- factors, grouping variables
- nonlinear relationships (transformed original variables and Y !)
- functions (nonlinear) of other X's: $X^{(j)} = X^{(k)2}$
- interactions

4.4 Reproducibility and Regression

- Variables that should be kept constant but cannot:

  $\longrightarrow$ Include in regression model!

- Fit a joint model for the data of the original and
  the replication study (if applicable)
  with a grouping variable "`Study`"
  and all interactions of it with the interesting variables.
  (Possibly with a model for correlation of errors E_i)

  This allows for a differential interpretation of
  the parts where reproducibility has and has not been achieved.

5. Model development

... consists of adapting the (structure of the) model to the data.

Select:

a. the explanatory and nuisance variables ("full model")

b. functional form (transformations, polynomials, splines)

c. interaction terms

d. possibly a correlation structure of the random errors $\longrightarrow$ systematically select the best fitting terms! $\longrightarrow$ overfitting!

Tradeoff between flexibility and parsimony.

Why should models be parsimonious?

Here: Intuition says that simple models reproduce better.

Example: Distances needed for stopping freight trains $\longrightarrow$ Result:

S/V0 $\sim$ Inclin + Lambda + Length + Type + Lambda^2 + V0 +
V0: (Inclin + Lambda + Length + Type) +
V0: (Inclin:Lambda + Inclin^2 + Lambda^2) +
V0^2 + V0^2):Length

$\longrightarrow$ The resulting model is certainly not the correct one!

What is “the correct model”, anyway?

Reproduciblily: Model selection is a non-reproducible process

(except for formalized procedures)

Should it be banned?

$\longrightarrow$ Yet another version of the dilemma of advancing science!

Summarizing:

Model development leads to severe reproducibility problems

because of “Researcher degrees of freedom”

Adequate statistical procedures can solve the more formalized types

of such problems. $\longrightarrow$ Model Selection Procedures.

6. Conclusions

Where and when is reproducibility a useful concept?

6.1 "Exact" sciences ...

(well: "quantitative, empirical part of sciences")

... Physics, Chemistry, Biology, Medicine, Life sciences.

- Reproducibility is an important principle to keep in mind.
  Feasible? Sometimes. Needs motivation, skill & luck. Recognition?

- **Data Challenge, Confirmation**
  Science is not only about collecting facts
  that stand the criterion of reproducibility, but about
  generating theories (in a wide sense) that connect the facts.

Types of confirmation:

+ Reproduction: Same values of input variables

  $\longrightarrow$ should produce response values within variability of error distr.

+ Generalization = extrapolation: Extend the range of input var's

  Check if regression function is still appropriate.

  Data Challenge

+ Extension: Vary additional input variables

  to find adequate extension of the model.

Recommendation: Perform combined study for reproducibility and generalization and/or extension.

6.2 Psychology: Reproducibility of Concept

Quantify “concepts” such as intelligence.

Questionnaires or “tests” $\longrightarrow$ quantified concept.

Study relationships between concepts (response) and

e.g., socio-economic variables, or between concepts.

Confirm concepts by using different questionaires / tests

hopefully getting “the same” concepts and their relations.

	same second-study features	different second-study features	level of validation
repetition, repeatability	all settings, experimenters	—	all data features, compatible estimated effects
replication, reproducibility	all settings, procedures	experimenters, institution	compatible estimated effects
data challenge	model	settings of explanatory and nuisance variables	model fits both studies, conclusions
replication of concepts	concepts (constructs) and relations between them	methods (instruments)	stable concepts and relations, cnclusions

6.3 Social sciences, ...

– Macro-Economics: Economy only exists once, no reproduction.

– Society, History: same

– Psychology: Circumstances (therapist, institution, culture)

are difficult to reproduce.

These sciences should not be reduced to quantitative parts!

What about philosophy and religion?

Good for discussions over lunch.

Messages

- Avoid significance tests and P-values. Use confidence intervals!
- Precise reproducibility in the sense of compatibility of quantitative results (non-significant difference) is rare (outside student's lab classes in physics)
- It becomes somewhat more realistic if models contain a study-to-study variance conponent and/or a correlation term.
- Dilemma of advancing science: Exploratory and confirmatory steps.

  $\longrightarrow$ Data Challenge

  Instead of mere reproducibility studies, perform confirmation / generalization studies!

- Reproducibility is only applicable to empirical science.

  There are other modes of thinking that should be recognized as "science" in the broad sense ("Wissenschaft").

  What is confirmation in these fields?

- $\longrightarrow$ In what sense / to what degree should reproducibility be a requirement for serious research?

- It is Sunday. My sermon in 2 sentences:

2 dimensions of life

- Dimension of facts $\longrightarrow$ Science,
  including empirical science ; reproducibility
- Dimension of meaning, significance (Bedeutung)
  relevant for conducting my life $\longrightarrow$ religion

Thank you for your endurance