P-values seem to mislead about reliability of point estimates

Raphael von Büren[1] raphael.vonbueren@stud.unibas.ch

Julia Fanderl[1] julia.fanderl@unibas.ch

[1] Block course "Zoology and Evolutionary Biology 2018" at the Zoological Institute, Evolutionary Biology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland

Supervisors:

Prof. Dr. Valentin Amrhein[1]

Dr. Tobias Roth[1]

[1] Zoological Institute, Evolutionary Biology, University of Basel, Vesalgasse 1, CH-4051 Basel, Switzerland

Abstract

P-values tell little about reliability of point estimates. They mix information on the effect size and the precision of the measurement. A better statistic to estimate reliability of point estimates are 95% confidence interval or standard error. We suppose that persons with statistical knowledge get mislead by p-values and misinterpret reliability of point estimates. To test this hypothesis, we conduct an observational study asking 172 students and researcher either with or without statistical knowledge. They get presented point estimates (with concerning confidence intervals) either with or without p-values and they have to grade reliability of the point estimates. The questionnaires' analysis indicates that persons without statistical knowledge interpret reliability of point estimates independently of p-values. However, persons with statistical knowledge misinterpret the reliability of point estimates more often when p-values are presented. Therefore, we conclude that p-values seem to mislead about reliability of point estimates.

Introduction

"Reliable information about reliability of results cannot be obtained from p-values nor from any other statistic calculated in individual studies. [] But [p-values] clearly give less information on uncertainty, reliability or replicability of the point estimate than is evident from a 95% confidence interval."

Amrhein, Korner-Nievergelt, and Roth (2017)

The p-value is the probability that a test statistic (such as a t-statistic) would be at least as large as its observed value if every model assumption were correct, including the null hypothesis (Greenland et al., 2016). P-values can thus indicate how incompatible the data are with a statistical model (Wasserstein & Lazar, 2016). We usually hope that the size of p-values is correlated with the strength of the evidence for our obtained results, so that smaller p-values mean that results are more reliable (Amrhein et al., 2017). However, the p-value does not directly measure reliability of the effect that we found in our sample. It mixes information on the size of the effect and how precisely it was measured (Amrhein et al., 2017). According to Cohen (1988), statistical precision of a sample statistic is defined as *"the closeness with which [a point estimate] can be expected to approximate the relevant population value"*. Precision is taken as being synonymous of 'reliability' of a point estimate (Cohen, 1988) and is usually estimated using a standard error or a confidence interval (Cohen, 1988; Greenland et al., 2016). Two confidence intervals with the same width indicate that the point estimates have the same precision and are thus equally reliable.

According to the American Statistical Association, the p-value is commonly misused and misinterpreted (Wasserstein & Lazar, 2016). Nevertheless, the use of p-values in research is increasing, which can be shown in an investigation counting p-values in biomedical publications (Chavalarias, Wallach, Li, & Ioannidis, 2016). According to Amrhein (2017), p-values can be decisive whether a study is published or whether you receive funding. This can lead to p-hacking: trying multiple things until you get a significant result (Bishop & Thompson, 2016). In particular, this is likely in today's environment of studies that chase small effects hidden in noisy data (Nuzzo, 2014).

In our study, we want to examine if p-values mislead about reliability of point estimates. We expect that students and researcher with no statistical knowledge interpret reliability of point estimates (with concerning confidence intervals) independently of p-values. On the other side, we expect that students and researcher with statistical knowledge get mislead by p-values and misinterpret reliability of point estimates (with concerning confidence intervals).

Methods

To test the introduced hypothesis, we conduct an observational study asking students and researcher either with or without statistical knowledge. We define statistical knowledge as a continuous variable indicating number of years' experience working with statistics. To our test subjects, we present questions concerning interpretation about reliability of point estimates. They have to be answered in a predefined amount of time. The constraint is chosen to test the persons immediate reaction on the presented question. We do not want to let them think too long. Like this, we simulate a potential situation where a researcher or student rapidly scans a figure out of a scientific paper.

To test, whether the test subjects get mislead by p-values, we design two different questionnaires in Microsoft PowerPoint (Figure S1): Questionnaires 'type P' and 'type X'. Both start with an introduction, where concepts like confidence interval, boxplot and point estimate are shortly explained. After the question and further explanation about the figures which the test subjects have to interpret, ten slides are presented. Each slide consists of either two confidence intervals together with the point estimate 'mean' or two boxplots with the point estimate 'median' (Figure 1). In total, there are five slides with two confidence intervals each and five slides with two boxplots each, resulting in ten slides over all. The slides with confidence intervals and the slides with boxplots alternate. Every slide is shown for six seconds and between the slides, there is a three second break. We focus mainly on the reliability of point estimates regarding confidence intervals. However, if we just present confidence intervals, the test subjects could understand the aim of our survey. To distract them from the real content, we present the study on the first slide as an investigation to find out whether confidence intervals or boxplots are better to interpret findings (Figure S1).

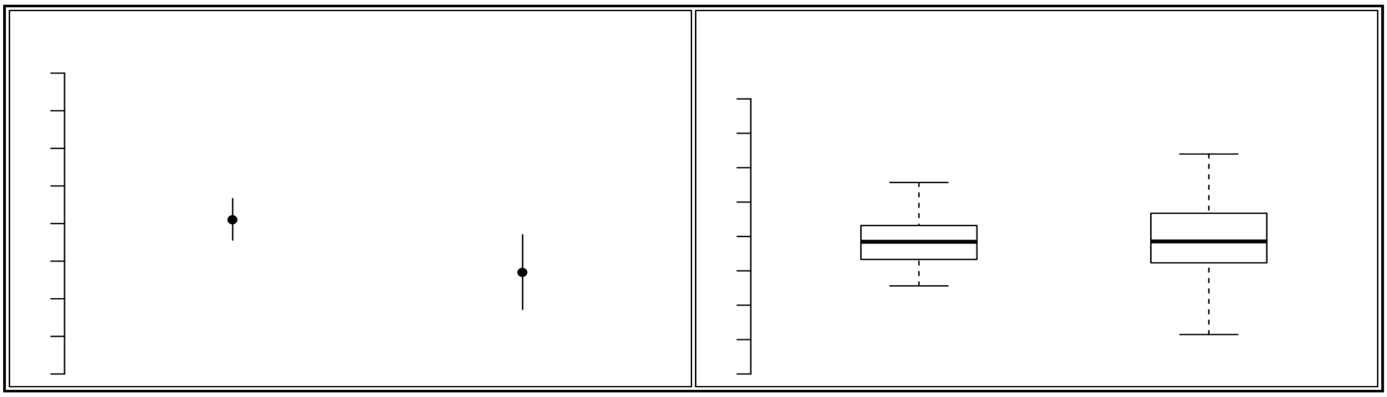


Figure 1: Figures out of questionnaire 'type X'. On the left slide, there are two point estimates (mean) with concerning 95% confidence intervals. The test subjects have to answer which of the two point estimates is more reliable. On the right slide, there is the same situation, but with boxplots. 'Type X' questionnaires are designed without presented p-values.

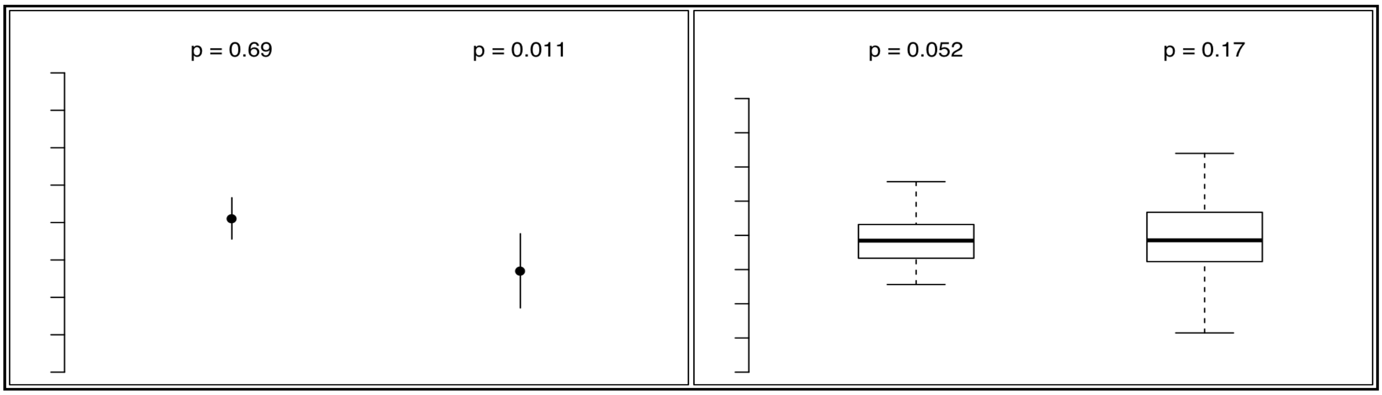


Figure 2: Figures out of questionnaire 'type P'. On the left slide, there are two point estimates (mean) with concerning 95% confidence intervals. The test subjects have to answer which of the two point estimates is more reliable. On the right slide, there is the same situation, but with boxplots. 'Type P' questionnaires are designed with presented p-values.

The questionnaires 'type P' and 'type X' exists in English ('P_e' and 'X_e') and in German ('P_d' and 'X_d'). This results in four questionnaires with the same confidence intervals and boxplots respectively ('questionnaire group'). The only difference besides the language is the fact, that 'type P'-questionnaires additionally presents p-values above the confidence intervals and boxplots (Figure 2). Furthermore, there are the following words written on one slide of the 'type P'- questionnaires: *The two p-values refer to one-sample t-test against the same null hypothesis; sample sizes are equal.*

Thus, we create 'questionnaire groups' consisting of four questionnaires. The different 'questionnaire groups' differ from each other due to modified point estimates, their confidence intervals and boxplots respectively. We generate these statistical figures using RStudio (Version RStudio: 1.1.423, Version R: 3.4.3) and the following simulation parameters (Figure S2):

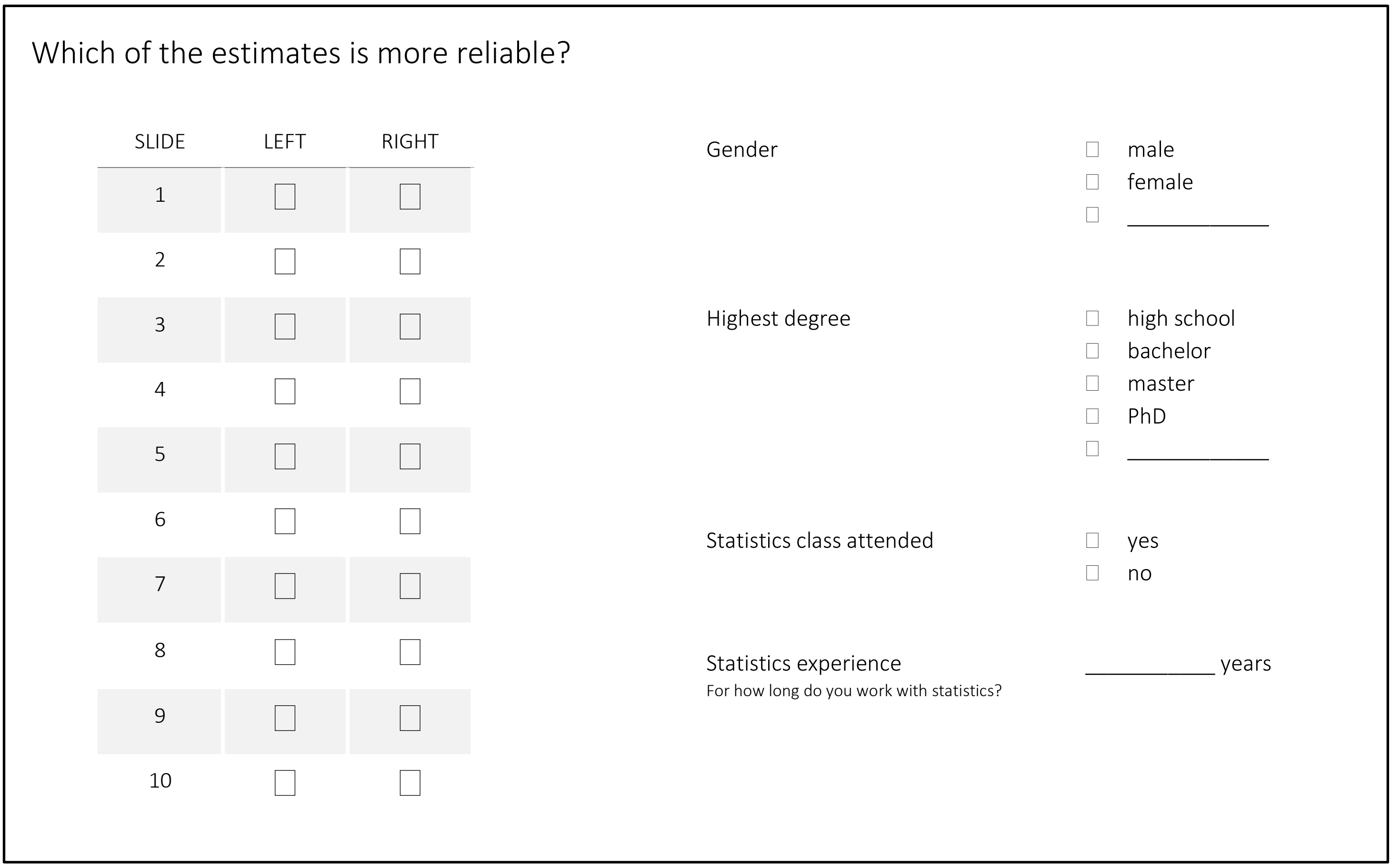
Sample size:	n = 100
Effect size:	randomized
Variance:	randomized; confidence interval lengths have to be easily distinguishable by eye, meaning a difference in the standard deviations of '> factor 1.5' or '< factor 0.5'

We simulate the data so that the p-value is in 50.93% of all our 1720 generated figures smaller when the confidence interval length is smaller as well.

In each of the ten slides, test subjects have to decide which point estimate of the two shown on each slide is more reliable. They record their answers on a printed form (Figure 3). On the first page, the form consists of two columns of ten boxes, each box corresponding to a point estimate. The second page consists of four questions for the additional parameters: sex, highest degree, statistics class attended (yes/no) and number of years with statistical experience. The test subjects have to answer the questions on page one. If they have no idea, they have to guess. The questions on the first page have to be answered before seeing the second page. To obtain objective results, all test subjects are approached in the same way and with the same words.

The results are digitally recorded in a Microsoft Excel 'Masterfile'. The statistical analysis of our investigation is done in RStudio (Version RStudio: 1.1.423, Version R: 3.4.3). As main analysis, we run a generalized linear mixed-effects model based on a logistic regression (outcome 0 or 1) with the following setting (Figure S3):

Experimental treatment:	figures with or without p-value
Covariate:	statistical experience of person
Interaction:	treatment x experience



Which of the estimates is more reliable?

SLIDE	LEFT	RIGHT
1	□	□
2	□	□
3	□	□
4	□	□
5	□	□
6	□	□
7	□	□
8	□	□
9	□	□
10	□	□

Gender
□ male
□ female
□ ____________

Highest degree
□ high school
□ bachelor
□ master
□ PhD
□ ____________

Statistics class attended
□ yes
□ no

Statistics experience
For how long do you work with statistics?
___________ years

Figure 3: *Printed form. Left part is presented during the questionnaire, right part in the end of the questionnaire.*

Results

We ask 172 persons (44% female) in eleven different locations: Theologisches Alumneum Basel, Biocentre Basel, Botanical Institute Basel, BSSE ETH Basel, Geographical Institute Basel, Kollegienhaus Basel, NLU Basel, Pavillonweg Bern, Psychological Institute Basel, Swiss TPH Basel and Zoological Institute Basel. Both observers (Raphael von Büren, Julia Fanderl) ask the same number of persons (86 each). 24% of the test subjects have 0 years statistical experience, 69% have 1-9 years and 7% have more than 9 years (Figure 4). 67% of the 1720 questions are answered correctly. To test, if our generalized linear mixed-effects model fits our data, we exclude all persons with more than 9 years statistical experience because they could have a high leverage. However, the results remain stable. Figure 5 shows the main result: Correct answers increase with statistical experience, but less so when p-values are presented. Within the first year of statistical experience, the treatment (p-value presented or not) has not a considerable influence on the proportion of correct answers.

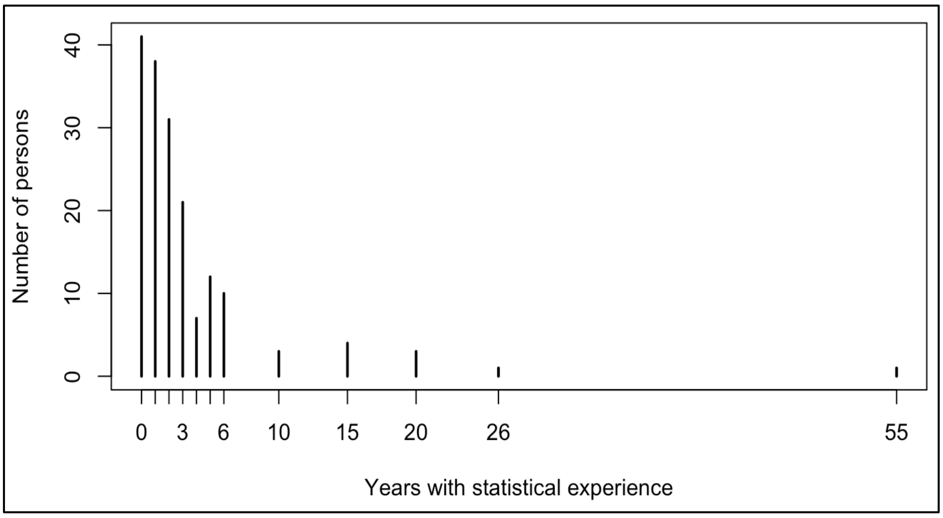


Figure 4: *Survey sample.*

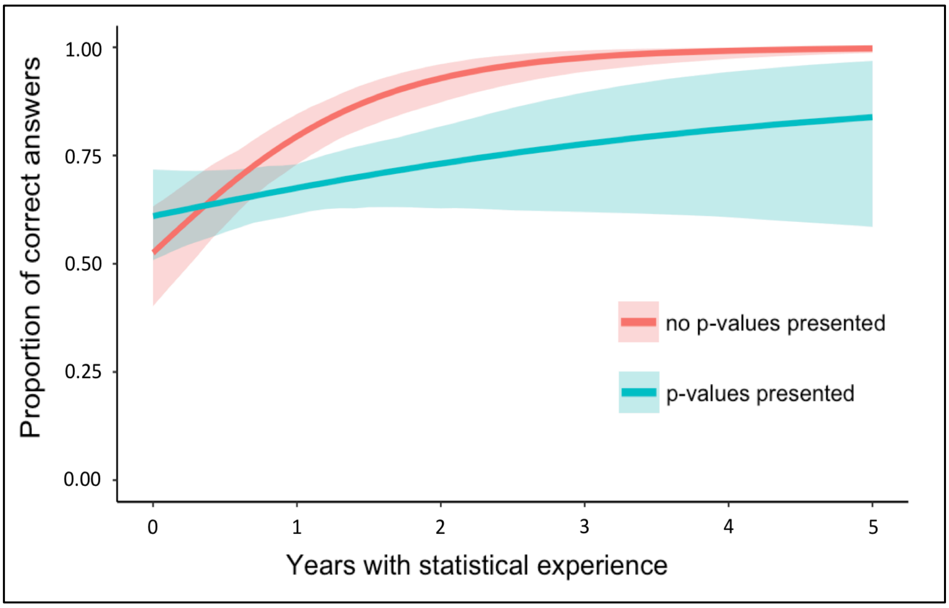


Figure 5: *Correct answers increase with stat. experience, but less so when p-values are presented. Given are model predictions (lines) and 95% credible intervals (shaded areas).*

***Table 1:** Multivariate logistic regression output: Effect of observer, gender and interaction treatment x experience.*

	Estimate	**Std. Error**	**p-value**
Intercept	0.268	0.321	0.403
Statistical experience	1.247	0.238	0.000
Treatment	0.341	0.381	0.372
Observer	-0.113	0.243	0.643
Gender	-0.203	0.246	0.409
Treatment x Experience	-0.971	0.314	0.002

We do find high evidence neither for observer bias nor for gender as an influence on the proportion of correct answers (Table 1). We analyze if the boxplots or the confidence intervals are correctly interpreted more often. We only use the answers from questionnaire 'type X' (no p-values presented), as the p-value could be a confounding variable. The proportion of correct answers is slightly higher for figures with confidence intervals (85%) than for figures with boxplots (83%)(Table S1). Furthermore, we analyze if correct answers increase with the number of already answered questions (learning curve). Correct answers increase marginally over time and could be caused by a few individuals (Table S1).

Discussion

The survey's results support our hypothesis that students and researcher with statistical knowledge get mislead by p-values and misinterpret reliability of point estimates. A reason for that could be the following: According to Amrhein et al. (2017), the p-value reflects our observed evidence against a null hypothesis. For describing the evidence against a null hypothesis, the size of the effect matters. If we want to describe the precision and thus the reliability of a point estimate, we should look at the confidence interval. Because p-values mix information on the effect size and the reliability of the point estimate and thus do not directly measure the reliability of the effect, p-values often mislead about reliability of point estimates. Apparently, some students and researcher do not recognize the fact that p-values mix information. They probably look at p-values as a measure for reliability of the effect that they found in their sample.

We wonder whether we got the same evidence for misleading by p-values giving the test subject more time to think (e.g. 20 seconds instead of 6 seconds). Or giving more information about p-values, reliability and effect size. Furthermore, a third type besides 'type P' and 'type

X' with presented zero line (null hypothesis) can be presented or 'type P' can be presented additionally with the zero line.

Our survey is a first explorative study with the finding that p-values seem to mislead about reliability of point estimates. However, further investigations are needed to explore not only correlation but also a possible causality.

"It seems that the only way to know how replicable our results are, is to actually replicate our results. Science will proceed by combining cumulative knowledge from several studies on a particular topic. [] A single replication can neither validate nor invalidate the original study. It simply adds a second data point to the larger picture."

Amrhein et al. (2017)

References

Amrhein, V. (2017). Das magische P. *Süddeutsche Zeitung 23.09.2017,* p. 37.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ, 5*. doi:10.7717/peerj.3544

Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ, 4*, e1715. doi:10.7717/peerj.1715

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting p-values in the biomedical literature, 1990-2015. *Jama-Journal of the American Medical Association, 315*, 1141-1148. doi:10.1001/jama.2016.1952

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol, 31*, 337-350. doi:10.1007/s10654-016-0149-3

Nuzzo, R. (2014). Statistical Errors. *Nature, 506*, 150-152. doi:DOI 10.1038/506150a

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *American Statistician, 70*, 129-131. doi:10.1080/00031305.2016.1154108

Supplementary Material

Figure S1 – Entire questionnaire (digital available – raphael.vonbueren@stud.unibas.ch)

Figure S2 – R Script Data Simulation (digital available – raphael.vonbueren@stud.unibas.ch)

Figure S3 – R Script Analysis (online available: https://github.com/TobiasRoth/BK-pvalues/blob/master/Praesentation/index.Rmd)

Table S1 – Results (online available: https://tobiasroth.github.io/BK-pvalues/#12)