False conclusions due to misleading p-values?

Do researchers and students understand p-values and confidence intervals correctly?

Proposal

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

Abstract

In natural sciences, topics around statistics are currently highly discussed and are undergoing a possible change. Concepts like "statistical significance" are being questioned more and more. Several publications criticize the use of p-values by reason of misinterpretation, p-hacking and biased science due to ignoring nonsignificant results.

We assume that a p-value below 0.05 triggers positive emotions in a researcher and leads to misinterpretation. Therefore, we want to examine if people get mislead by p-values. To pursue this question, we will conduct an observational study asking students and scientists either with or without statistical knowledge. We will present statistical figures (with and without misleading p-values) that have to be interpreted.

Testing emotional effects of p-values will have implications on the described discussion about sense and nonsense of statistical significance.

Supervisors:

Prof. Dr. Valentin Amrhein[1]
Dr. Tobias Roth[1]

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

Introduction

Nowadays, statistics plays a very important role in good scientific practice. Concepts like e.g. p-value, confidence interval, sample size, null hypothesis, statistical significance are used by every researcher. However, good scientific conclusions need more than proper statistical models. Well planned study design, conscientious measurements, and careful interpretation is important to represent and reduce uncertainty (Amrhein, 2017; Wasserstein & Lazar, 2016).

The most common concept to represent uncertainty in science is "statistical significance", normally valued by an index called p-value. It indicates the probability of our data, or data more extreme on condition

that the null hypothesis is true (Greenland et al., 2016). Hence, the p-value shows how reliable the observed effect is compared to the ineffectiveness (Amrhein, 2017). The p-value goes back to Ronald Fisher, who advocated "p-value ≤ 0.05" as "statistical significant" (Fisher, 1925). According to Wasserstein and Lazar, 2016, researchers misinterpret the p-value quite often and that is why there are critics since around one century (Amrhein, Korner-Nievergelt, & Roth, 2017; Berkson, 1938; Cohen, 1994). Some recurrent points of criticism on p-value are the following:

- "P-values do not measure the probability that the studied hypothesis is true" (Wasserstein & Lazar, 2016)
- "Scientific conclusion […] should not be based only on whether a p-value passes a specific threshold" (Wasserstein & Lazar, 2016)
- "A p-value, or statistical signicance, does not measure the size of an effect or the importance of a result" (Wasserstein & Lazar, 2016)
- The p-value is hardly reproducable (Amrhein, 2017)

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).

In recent years, the p-value has become very important for some researchers. According to Amrhein, 2017, p-values can be decisive whether a study is published or whether you receive funding. As a consequence of this, "*p-hacking*: trying multiple things until you get the desired result" is becoming more likely (Nuzzo, 2014).

Therefore, we assume that a p-value below 0.05 triggers positive emotions in a researcher and leads to misinterpretation of the issue to be investigated. In this study, we want to examine if people get mislead by p-values. More precisely, if they misinterpret a statistical figure by reason of shown p-values. This leads us to our research question.

Study question: Do p-values have an effect on our perception and interpretation of statistical figures showing confidence intervals or boxplots?

Based on this study question as well as previous observation and knowledge, we formulate the following two hypotheses:

Hypothesis I: Students with no statistical knowledge interpret statistical figures showing confidence intervals and boxplots independently of p-values.

Hypothesis II: Students or researcher with statistical knowledge get mislead by p-values and misinterpret statistical figures showing confidence intervals and boxplots.

Methods

To test the introduced hypothesis, we will conduct an observational study asking students and scientists either with or without statistical knowledge (test subjects). We will present questions concerning interpretation of statistical figures that have to be answered in a predefined amount of time. For that purpose, we will prepare three different questionnaires with identical graphics but different additional informations:

	A	*B*	*C*
Statistical graphics	- ten figures with two mean values and concerning confidence intervals - ten figures with two median values and concerning boxplots		
Additional information	no	p-values	zero-line

In presentation A, graphs are shown without additional information. In presentation B, the p-value will be added and in presentation C, the zero-line will be added. Each presentation will consist of alternate pictures of either two confidence intervals and their mean value or two boxplots and their median value. In total, there will be ten pictures of each type, resulting in twenty pictures over all. After an introduction about the principles of the study and the following issue, every picture is shown for five seconds. We will present the study as a "speed intelligence test" to distract the test subjects from the real content of the study. Furthermore, the short time period per picture should prevent thinking for too long. The reason is that essentially, we want to test the immediate "emotional" reaction of the test subjects (see Introduction: *…we assume that a p-value below 0.05 triggers positive emotions […] and leads to misinterpretation…*).

In each of the twenty pictures, test subjects have to decide which graph of the two shown is more reliable. We as study conductor record their answers in a printed version of the questions. The test subjects will be shown either the presentation A, B, or C according to a predefined alternate manner. The distinguishing parameter of the test subjects is their previous knowledge of statistics. They either have had a class in statistics or not. To obtain objective results, all subjects will be informed in the same way and with the same words.

The statistics of our investigation will be done in R Studio (Version R Studio: 1.1.423, Version R: 3.4.3) and will probably be an ANOVA. Figure 1 shows how our data possibly will look like.

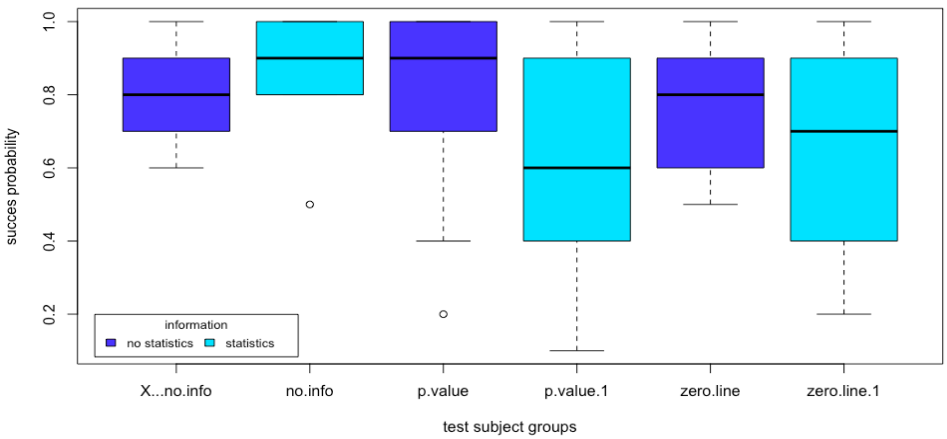

Figure 1: *Possible data outcome (boxplot) of our investigation with imaginary data*

Research schedule

Table 1: *Timetable with dates and workflow*

	Date	*Workflow*
Week 1	Mon, 12.03.18	Elaborate **questionnaire**; data modeling (confidence intervals and boxplots); reading up on the concerning scientific literature (evening)
	Tue, 13.03.18	Elaborate and pretest **questionnaire**; define data management (Microsoft Excel or Microsoft Access); reading up on the concerning scientific literature (evening)
	Wed, 14.03.18	**Data collection** (Botanical Institute, Zoological Institute, Kollegienhaus, Alumneum); reading up on the concerning scientific literature (evening)
	Thu, 15.03.18	**Does it work? – first analyses**; **Data collection** (Kollegienhaus, Faculty of Theology, Mensa); reading up on the concerning scientific literature (evening)
	Fri, 16.03.18	**Data collection** (Biozentrum); reading up on the concerning scientific literature (evening)
Week 2	Mon, 19.03.18	**Data collection** (Swiss TPH); reading up on the concerning scientific literature (evening)
	Tue, 20.03.18	**Data collection** (Place: Depending on which study group is underrepresented); reading up on the concerning scientific literature (evening)
	Wed, 21.03.18	Evaluation and statistical **analysis** (R Studio)
	Thu, 22.03.18	Evaluation and statistical **analysis** (R Studio)
	Fri, 23.03.18	Evaluation and statistical **analysis** (R Studio)
Week 3	Mon, 26.03.18	**Report** (Methods, Results)
	Tue, 27.03.18	**Report** (Discussion)
	Wed, 28.03.18	Excursion Petite Camargue Alsacienne
	Thu, 29.03.18	No class
	Fri, 30.03.18	No class
Week 4	Mon, 02.04.18	No class
	Tue, 03.04.18	**Report** (Introduction, Discussion)
	Wed, 04.04.18	**Presentation** (Prep)
	Thu, 05.04.18	**Presentation** (Prep)
	Fri, 06.04.18	**Symposium:** Presentation of group projects

Outlook

Some scientist advertise a rethinking and recommend either to redefine statistical significance from 0.05 to 0.005 (Benjamin D. J. et al., 2017) or to publish results "regardless of statistical significance" (Amrhein & Greenland, 2017). If our hypothesis *Students or researcher with statistical knowledge get mislead by p-values and misinterpret statistical figures showing confidence intervals and boxplots* can be supported by our study, more investigation would be required. Furthermore, there would be an urgent need to rethink the current gold standard "p-value".

References

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

Amrhein, V., & Greenland, S. (2017). Remove, rather than redefine, statistical significance. *Nature Human Behaviour, 2*, 4-4. doi:10.1038/s41562-017-0224-0

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

Benjamin D. J., J. O. B., Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers,, Richard Berk, K. A. B., Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles E erson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster,, Edward I. George, R. G., Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Had eld, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell,, Michael McCarthy, D. M., Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa,, Brendan Nyhan, T. H. P., Luis Pericchi, Marco Perugini, Je Rouder, Judith Rousseau,, & Victoria Savalei, F. D. S. n., Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*, 6-10.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536. doi:Doi 10.2307/2279690

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. (1994). The Earth Is Round (p < 0.05). *American Psychologist, 49*, 997-1003. doi:Doi 10.1037/0003-066x.50.12.1103

Fisher, A. R. (1925). *Statistical Methods for Research Workers*.

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