Extended Abstract for the
Symposium on Integrating AI and Data Science into School Education Across Disciplines 2025

SOFTWARE-SUPPORTED AND SIMULATION-BASED INTRODUCTION TO SIGNIFICANCE TESTS

Karin Binder[1] & Michael Rößner[2]
[1]Paderborn University, Germany, Karin.Binder@uni-paderborn.de
[2]LMU Munich, Germany

Focus Topics: Learning materials, tools

Introduction and theoretical background

Significance tests and p-values play an important role in science and their results are reported in many articles. In data science workflows these procedures are, for example, a crucial part in data cleaning or for the detection of outliers.

If interpreted in a probabilistic sense, p-values mean the probability of obtaining the found or an even more extreme test statistic value, under the assumption the null hypothesis is true in reality. However, significance tests and p-values for judging research hypotheses are associated with many misconceptions and are sometimes the subject of critical debate in research (Oakes, 1986, Haller & Krauss, 2002). Some journals even prohibit the mere reporting of p-values (e.g. the Journal „Basic and Applied Social Psychology," Trafimow & Marks, 2005).

One prominent misconception is, for example, to think that the p-value is the probability of the null hypothesis being true (Gigerenzer, 2004). Furthermore, some people commit the replication fallacy: If you repeat the experiment many times, you would obtain a significant result in (1 – p) % of occasions, (Gigerenzer, Krauss & Vitouch, 2004). Another misconception is to interpret significant test results as large effects or relevant effects (Herrera-Bennett, Heene, Lakens, & Ufer, preprint). But even if traditional significance tests are controversial, they are still used intensively in research. And when used correctly, they can provide valuable insights. On the other hand, effect sizes or confidence intervals, for example, are sometimes considered to be useful alternatives or supplements (Cumming, 2012).

Research question and method

This contribution presents a study in which 66 gifted or interested students (grades 8-12) volunteered to attend an intensive workshop on significance tests for one week (see also Rößner, Binder & Ufer, in press). Afterwards, a post-test was carried out with the 41 students who attended the whole workshop to investigate the extent to which the students are nevertheless subject to typical misconceptions and which misconceptions occur only rarely. In this test, 31 closed items were used, addressing seven known classes of misconceptions (e.g., the inverse probability fallacy, or the replication fallacy).

The workshop builds on the following points to develop a better understanding of significance tests:

- A strong focus on the concept of p-values (see e.g., Biehler, Engel & Frischemeier, 2023)
- A simulation-based introduction of p-values (compare Figure 1; see e.g., Chandrakantha, 2020, Podworny, 2019)
- A software-supported introduction of significance tests (instead of manual calculation using spreadsheets, see e.g., Chandrakantha, 2020)
- An explicit discussion of typical misconceptions (see, e.g., Oser, Hascher & Spychinger, 1999, Krauss & Wassner, 2002).

From a content perspective, the workshop centered on the Chi²-test for independence, the two-sample t-test and tests on correlation and regression coefficients. The introduction to significance testing was based on simulations and in connection with p-values (Chandrakantha, 2020). This approach was chosen in order to convey an appropriate concept of significance tests and to prevent misconceptions (Oser, Hascher & Spychinger, 1999). The approach was software-supported with the help of R, for which short, predefined functions were provided in order to avoid overemphasizing the computational procedures. Data on students' performance in mathematics as part of a PISA study served as an introductory example. However, procedural knowledge for the concrete calculation of significance tests

In S. Podworny & S. Schönbrodt (Eds.), *Towards Fostering AI and Data Science Literacy in Schools Across Disciplines. Proceedings of the 1st Symposium on Integrating AI and Data Science into School Education Across Disciplines (AIDEA1 2025)*, Salzburg, Austria.

by hand or with the aid of software–which are strongly focused on in many school and university course concepts–remained in the background.

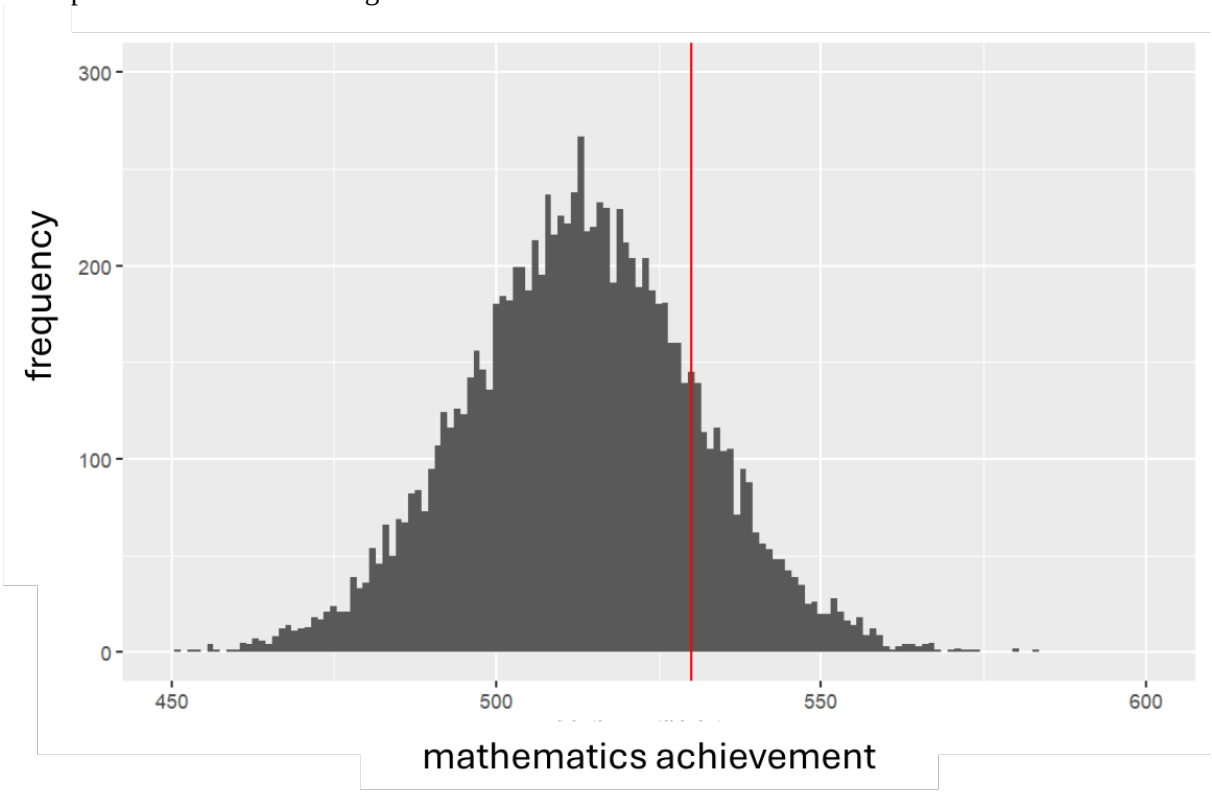

Figure 1: Simulation of the reference distribution with R to elaborate the meaning of the p-value. The red line represents the average mathematics achievement in an (imaginary) sample

Results

Some of the typical fallacies occurred only rarely in the post-test. Only a few students rated single items as correct, interpreting a significant result as clear evidence for the research hypothesis. Even the famous "inverse probability fallacy", which has been witnessed in many earlier studies, even with experts, rarely occurred after training. However, other false conclusions that were not explicitly addressed in the workshop (such as the relevance fallacy) appeared frequently. Many students believed that a statistically significant result is always a relevant result. In order to overcome these errors an explicit discussion seems to be necessary (see e.g., Krauss & Wassner, 2002).

Outlook

The software R played a major role in the course concept described. This led to problems in the exercises, as the focus frequently shifted from talking about statistics to finding errors in the code. For this reason, the use of an educational statistic software could be worthwhile, in particular for younger students. One promising possibility is the software CODAP, that has recently introduced the new plug-in testimate, which enables the significance tests described here (t-tests, Chi^2-test, testing against null-correlation, testing regression coefficients). The CODAP-tool testimate will be presented in Binder & Erickson (in preparation).

Figure 2 shows the implementation of a t-test using the tool testimate. It was examined whether the average mathematics achievement differs between two groups of students (for example from different countries).

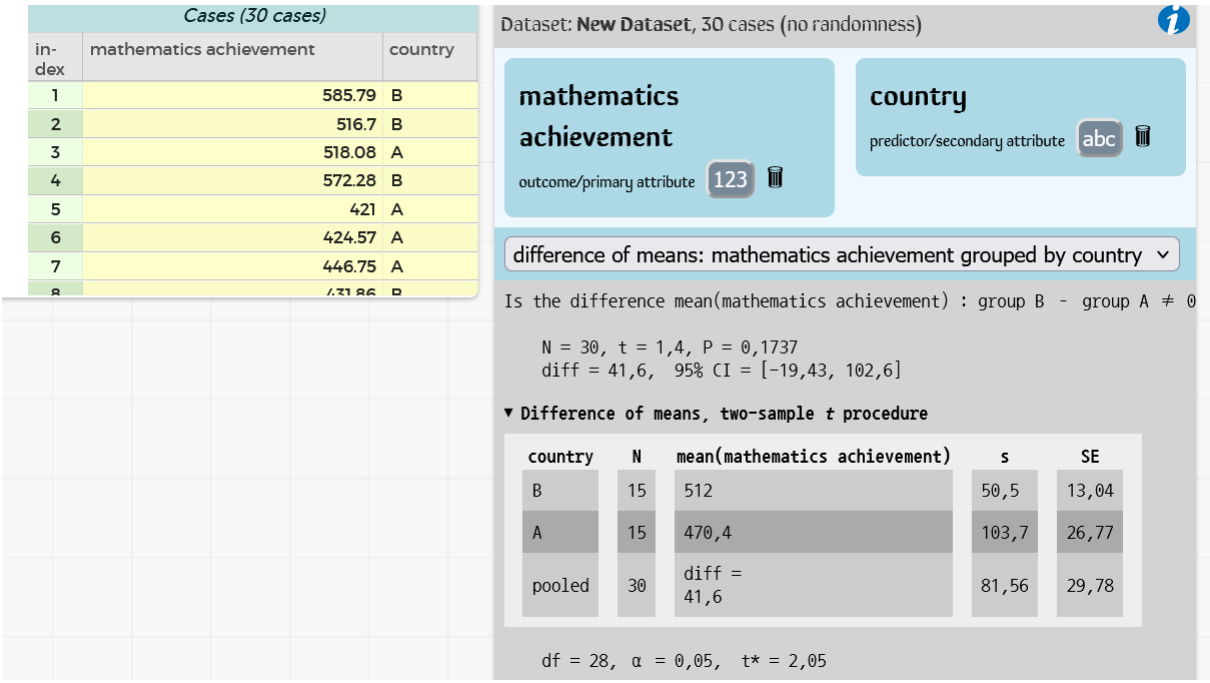

Figure 2: CODAP-tool testimate for significance testing (two-sample t-test)

The results show that there is no significant difference in mathematics achievement between country A and country B. However, what "no significant difference" means exactly must first be worked out–preferably using simulations–in order to understand the results obtained with testimate. Descriptively we see that the mean in mathematics achievement in country A is 470.4, whereas the mean in country B is 512 points, a difference of 41.6 points.

CODAP also offers a possibility for the simulation-based approach we used in our study with the help of R. For example, Figure 3 shows the simulated reference distribution for the PISA example described above.

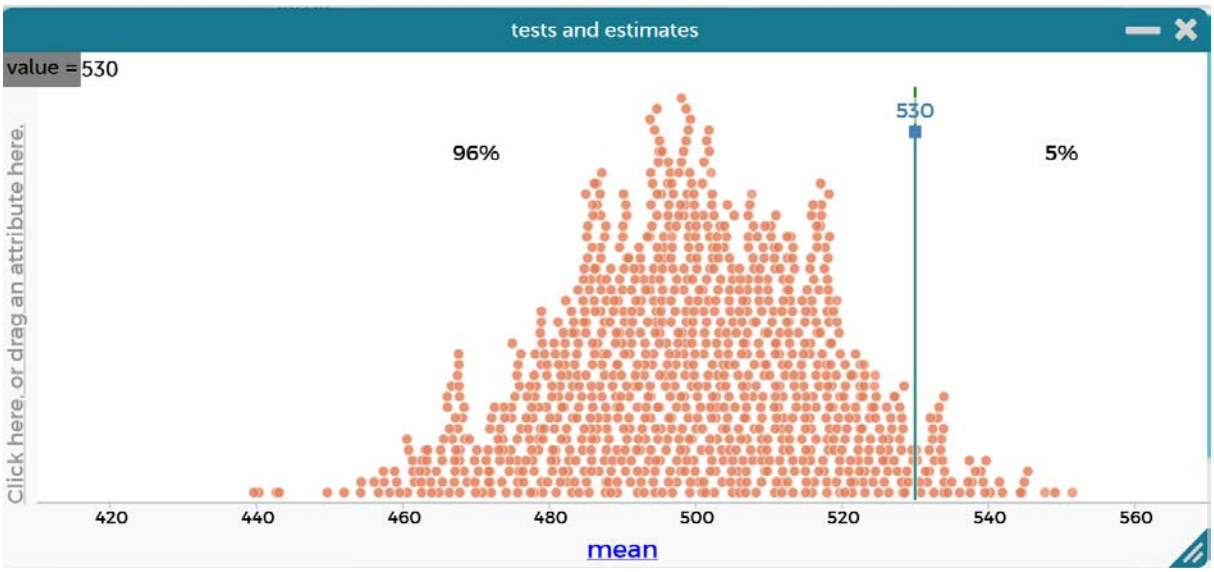

Figure 3: Simulation of a theoretical distribution of the PISA results with a mean of 500 points and a standard deviation of 100 points with CODAP to elaborate the meaning of the p-value (compare Figure 1)

References

Binder, K. & Erickson, T. (in preparation). Traditional tests through a randomization lens: treating p-values as data with CODAP.

Biehler, R., Engel, J., & Frischemeier, D. (2023). Stochastik: Leitidee Daten und Zufall. In *Handbuch der Mathematikdidaktik* (pp. 243-278). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chandrakantha, L. (2020). Visualizing the p-value and understanding hypothesis testing concepts using simulation in R. *Electronic Journal of Mathematics & Technology*, *14*(3).

Cumming, G. (2012). Understanding The New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis, Routledge, New York.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of socio-economics*, *33*(5), 587-606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. *The Sage handbook of quantitative methodology for the social sciences*, 391-408.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of psychological research*, *7*(1), 1-20.

Herrera-Bennett, A. C., Heene, M., Lakens, D., & Ufer, S. (Preprint). Improving statistical inferences: Can a MOOC reduce statistical misconceptions about p-values, confidence intervals, and Bayes factors? https://doi.org/10.31234/osf.io/zt3g9

Krauss, S., & Wassner, C. (2002, July). How significance tests should be presented to avoid the typical misinterpretations. In *Proceedings of the Sixth International Conference on Teaching Statistics. Cape Town, South Africa: International Association for Statistics Education. Online: www. stat. auckland. ac. nz/~ iase/publications.*

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences, Chichester: Wiley.

Oser, F., Hascher, T., & Spychiger, M. (1999). Lernen aus Fehlern Zur Psychologie des „negativen "Wissens. In *Fehlerwelten: Vom Fehlermachen und Lernen aus Fehlern.* Beiträge und Nachträge zu einem interdisziplinären Symposium aus Anlaß des 60. Geburtstags von Fritz Oser (pp. 11-41). VS Verlag für Sozialwissenschaften.

Podworny, S. (2019). *Simulationen und Randomisierungstests mit der Software TinkerPlots: Theoretische Werkzeuganalyse und explorative Fallstudie.* Springer-Verlag.

Rößner, M., Binder, K., & Ufer, S. (in press). Simulationsbasiert Signifikanztests verstehen. *Mathematica Didactica*.

Trafimow, D., & Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology*, *37*, 1–2.