Ruprecht-Karls-University Heidelberg

Faculty of Mathematics and Computer Science

The shortcomings of p-values and the multiple testing bias

Seminar Report

Submitted to Ullrich Köthe, Jens Kleesiek, Stefan Radev and Christoph Wiesinger

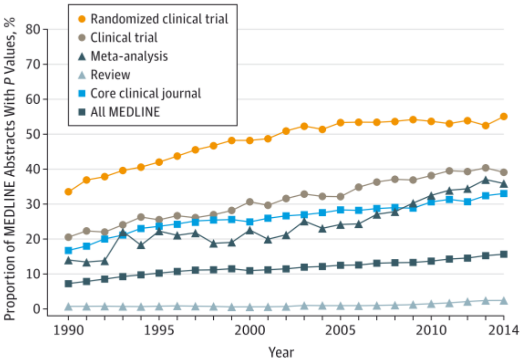
Seminar: How do I lie with statistics?
Submission: 02.01.2020

Author: Frederick Stegmüller
Matriculation number: 3219145
M. Sc. Economics, 3^{rd} semester of study

Contents

1 Introduction

The usage of null hypothesis significance testing (NHST) has become more and more prominent in scientific research. According to a text mining analysis by David Chabalarias et. al. (2016), the amount of MEDLINE articles from 1990 to 2014 has a relative increase of 4.5% per year, while the amount of articles reporting p-values in their abstract had a relative increase of 8.2% per year. As can be seen in Figure 1, the amount of p-values reported in the abstract is even higher for certain categories, such as clinical trials or meta-analyses.



Source: Chabalarias et. al. (2016)

Figure 1: Proportion of MEDLINE Abstracts reporting at least one p-value, 1990-2014

The prominence of NHST in scientific research also gave rise to criticism. Due to concerns about the misuse and misinterpretation of p-values, Ronald Wasserstein and Nicole Lazar (2016) issued a statement in The American Statistician, informing about the correct context and interpretation of the p-value.
The goal of this report is to inform about certain problems and shortcomings of the contemporary NHST process, as well as showing ways on how to improve.

2 Origins of NHST

2.1 Fisher (1925)

As outlined by Denes Szucs and John Ioannidis (2017), the original use of p-values was propagated by Ronald Fisher in his book "Statistical Methods for Research Workers" in 1925. This method calculated the p-value of a previously defined hypethesis. The conventional threshold of the p-value was proposed as $p \leq 0.05$, although the final decision was up to the judgement of the experimenter. The goal of this method was to try to disprove the null hypothesis and H_0 was only deemed demonstrable when multiple experiments rarely gave statistically significant results. It is also important to notice that this method relies on the null hypothesis beeing tested many times, thus it would not be possible to dismiss H_0 simply by performing a single experiment.

2.2 Neyman and Pearson

Szucs and Ioannidis (2017) also summarize the process by Neyman and Pearson, which introduced the concept of the alternative Hypothesis (H_1). This approach was introduced as a formal decision procedure that was motivated by industrial quality control problems. The goal of this process is to minimize the false negative rate β, or equivalently maximize the true positive rate (also known as power) $1 - \beta$ subject to an arbitrary bound α on false positive errors. This approach does not use the specific p-value as a measure of evidence. The p-value is only calculated in order to reject H_0 if p exceeds α.

For this approach, specific assumptions on H_1 have to be made in order to minimize β, as well as an appropriate threshold α has to be chosen. This is possible for a factory context, where the effect size is largely known and the trade off between Type I error rate and Type II error rate can be calculated. In the context of research work, this is not necessarily the case, as effect sizes are usually unknown and the costs or implications of Type I and Type II error rates are not easily measurable.

Lastly, the Neyman-Pearson process was designed for repeated testing in the long-run and does not work efficiently when applied to single experiments.

3 Contemporary use of NHST

The current approach to NHST, as described by Jacob Cohen (1994) and Szucs and Ioannidis (2017), takes some aspects of both Fisher's method and the process by Neyman and Pearsons, while omitting other parts that are crucial to those methods.

In this approach, the concepts of null hypothesis and alternative hypothesis from the Neyman-Pearson process are used, but the null hypothesis is usually set to predict nothing, while the alternative hypothesis is often not defined quantitatively. Defining (or rather not defining) H_1 this way makes it impossible to calculate the power $1 - \beta$ before the experiment, which is neccessary for the Neyman-Pearson process. Then the specific p-value is calculated. If p<0.05, H_0 is automatically rejected, mirroring the mechanical rejection process of Neyman and Pearson, but arbitrarily setting $\alpha = 0.05$ as was the proposed convention by Fisher. After H_0 is rejected, the unspecified alternative hypothesis is accepted as scientific fact.

Lastly, the exact p-value is interpreted to serve as a relative measure against H_0, arguing a small p-value would provide stronger evidence than a bigger one.

Most importantly, while both the methods by Fisher or by Neyman and Pearson rely on testing the hypothesis multiple times using different data, the contemporary approach to NHST mostly relies on single experiments to reach conclusions.

4 Shortcomings of the contemporary NHST

4.1 Flaws in logical reasoning of contemporary NHST

The logical reasoning of rejecting the null hypothesis and accepting the alternative hypothesis is represented by P. Pollard and J. Richardson (1987) as:

1 If H_0 is correct, the data (D) are highly unlikely

2 D occured

$\Longrightarrow$ $P(D|H_0)$ is highly unlikely, thus we can reject H_0 and accept H_1.

An example of this logic they provided would be:

1 If a person is an American (H_0), he is unlikely to be a member of congress (D)

2 A member of congress is found

$\Longrightarrow$ That member of congress is probably not an American.

As this example should make clear, such a conclusion can not be drawn from these premises. The p-value states the probability of the data given the null hypothesis ($P(D|H_0)$), while rejecting H_0 would need a small probability of the null hypothesis given the data ($P(H_0|D)$). While the probability of a person being a member of congress given only the information that that person is an american is indeed quite low, the information needed to reject H_0 would be the probability of a person beeing an American given that they are a member of congress. According to the U.S. Constitution Art. I, Sec. 1-3 the congress consists of the Senate, for which one has to be a U.S. citizen for at least nine years, and the House of Representatives, for which one has to be a U.S. citizen for at least seven years to be a member. Thus the probability of a person being an American given that they are a member of congress is actually 100%.

As was shown, $P(D|H_0)$ does not inform on $P(H_0|D)$, the probability actually needed to argue against the null hypothesis. Secondly, only the probability concerning H_0 is calculated, while no information on the probability of H_1 is needed. Using the example above, the alternative hypothesis would be defined as $H_1 = \neg H_0$, meaning the person is not an American. Even though $P(D|H_0)$ might be very small, $P(D|H_1) = P(H_1|D) = 0\%$, meaning the alternative hypothesis in this case is even more unlikely than the null hypothesis. This demonstrates that only using p-values of H_0 is logically not enough to accept the alternative without any further investigation of H_1 itself.

4.2 NHST is not suitable for Big Data

P-values are determined by calculating the probability of a test statistic of the dataset, specifically if the test statistic increases, the probability of the test statistic decreases. When defining a threshold α at which the null hypothesis can be rejected, it thus suffices to check whether the test statistic exceeds the constant c at which point the probability of the test statistic given H_0 is smaller than α (Wooldridge, 2013: 774ff).

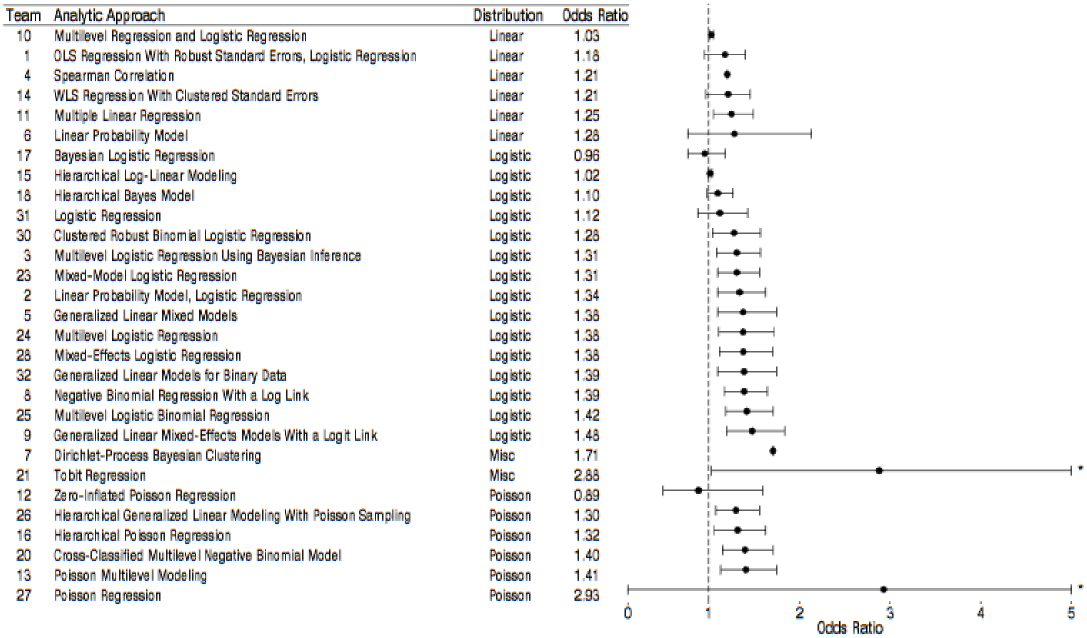
The way test statistics are typically constructed they are dependant on the number of observations of the dataset in such a way that including more observations leads to a higher value of the test statistic. When defining H_0 to predict zero effects it can suffice to increase the size of the dataset to be large enough for the test statistic to exceed c. As follows, any miniscule effect that does not equal exactly zero can be determined to be statistically significant at the α-level, given enough data (Szucs and Iannidis, 2017).

4.3 Selective Reporting

In an effort to show how much influence the choice of different approaches on analyzing the same hypothesis, a project supervised by Brian Nosek gave the same dataset to 29 different research teams in order to analyze whether football players whith darker skin tone were more likely to receive red cards by referees than players whith lighter skin tone (Silberzahn et. al. 2016).

As Figure 2 shows, the large variety of analytic approaches that were used resulted in varying results for both the point estimates and the confidence intervals. In this example using $p < 0.05$, 69% of the research teams reported significant results, while 31% reported results that where not significant. The odds ratio varied from a minimum of 0.89 to a maximum of 2.93 and a median of 1.31, although none of the two odds ratios smaller than one where significant. These research teams simply reported their results without any incentive for p-hacking, HARKing or fabricating significant results to get published, so the variability of the results shows how influencial simple analytical choices can be.

Once those incentives are present, such a set of possible results might tempt researchers to pick out a result with a small enough p-value while ignoring the results of other approaches that did not reach significance, possibly leading to bias towards a specific hypothesis that is favored by the researcher.



Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93

Source: Silberzahn et. al. (2016)
*: CI cut off for better interpretability of the plot

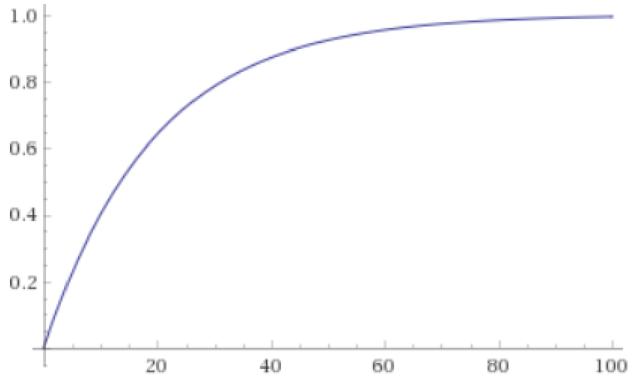
Figure 2: Point estimates and 95% CI for the effect of the skin tone of football players on the number of red cards received by referees

4.4 Publication bias

When researchers do not get statistically significant results, these are less likely to be submitted or published than significant results, producing publication bias (Philippa Easterbrook, Jesse Berlin, Ramana Gopalan and David Matthews: 1991). As Szucs and Ioannidis (2017) argue, setting $\alpha = 0.05$ will still lead to 5% of the tests on a true H_0 to report significant results when performed multiple times. The bias towards reporting only significant results, however, can lead to those 5% false findings to be published while the remaining, insignificant results are being mostly ignored. Such practice can cause any true null hypothesis to be rejected in the long run.

5 Multiple testing bias

The multiple testing bias, as outlined by Szucs and Ioannidis (2017), arises when multiple related hypotheses are tested using NHST. In such a case, the probability of a true H_0 having at least one false positive result (family wise error rate) of k independent tests is $\alpha_{\text{TOTAL}} = 1 - (1 - \alpha)^k$



Source: own depiction

Figure 3: Plot of α_{TOTAL} for p<0.05

As Figure 3 shows for p<0.05, the family wise error rate (FWER) already increases strongly for a few added hypotheses. Simply analyzing a second hypothesis increases FWER to 9.75%, while testing 5 and 10 increases a FWER to 22.62% and 40.12% respectively.

There are various possibilities the multiple testing error can be accounted for, one example would be the Bonferroni correction for k tests correcting the p-value for the amount of tests: $p_B = \frac{\alpha}{k}$. Though such corrections do decrease the Type I error rate, they increase the Type II error rate.

The alternative to these corrections that was proposed by Szucs and Ioannidis (2017) is the False Discovery Rate (FDR). To calculate the FDR, first Q is defined as the ratio of false positive (FP) to both false and true positive (TP) results: $Q = \frac{FP}{FP+TP} = \frac{FP}{R}$. As in scientific research the amount of true positive results is unknown, it can be considered a random variable. Q cannot be controlled directly, so FDR is defined as $FDR = E[Q|R > 0] \cdot P(R > 0)$. FDR can be controlled using α and the power $1 - \beta$. In comparison, the FWER α_{TOTAL} can also be defined as $FWER = P(FP \leq 1) = 1 - P(FP = 0)$. FDR and FWER can then be compared, as if H_0 is true in all tests, FDR=FWER, while FDR<FWER when some H_1 are true.

It is important to note, though, that this approach of utilizing FDR does relie on the NHST method. Thus it might help against the multiple testing bias, but the above mentioned problems are still equally valid. Szucs and Ioannidis (2017) thus argue the possibility of using Bayesian methods to analyze multiple comparisons, which have been shown to report more conservative results than NHST.

6 Ways of improvement

This report has shown so far, that NHST is not as strong of an analysis tool as often assumed, as well as not necessarily the best application depending on the context. As only informing about problems is not constructive, this section focuses on how the current approach to NHST can be improved. Ronald Wasserstein, Allen Schirm and Nicole Lazar (2019) compiled many recommendations on improving the use of NHST, some of which will be summarized below.

To decrease p-hacking or publication bias, the incentives need to be changed so that all findings get reported and not only significant ones. One such incentive could be pre-registration of the hypothesis and methodology before conducting the research. Also, researchers should not simply use arbitrary thresholds to classify findings into significant or non-significant. Such practice can, for example, lead to simply rejecting a null hypothesis whith p=0.049, while there might not be much of a difference to a non-significant result whith p=0.051 that would not reject the null hypothesis, even though such a difference between significant or not significant results could very well just be the result of a difference in analytical choices. Rather the p-values should be reported as continuous values in order to be interpreted as descriptive statistics besides effect sizes and confidence intervals.

More importance should be put on reproducing findings and meta-analyses than on the significance of single experiments. It has been suggested, to include a dynamic display of "Reproduced by", to see whether findings have already been reproduced or not. Other recommendations include large scale replication projects and registered replication reports.

To increase transparency, Szucs and Ioannidis (2017) argue that researchers should publish both raw data and analysis scripts, as these can provide deeper understanding of the analysis, considering simple decisions during the research process can influence the final outcome. Providing the analysis script might also help researchers trying to reproduce the findings.

Lastly, Wasserstein and Lazar 2016 reported these questions and answeres from an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach p=0.05?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use p=0.05?
> A: Because that's what they were taught in college or grad school.

Such circular reasoning should not be the foundation of a curriculum, thus one of the most important ways of improvement should be to teach alternative methods and approaches and not only NHST.

7 Conclusion

The use of NHST has become prevalent in contemporary statistics, even though there are flaws in its usage and it is not always suitable in the context it is used in. Furthermore there are often flaws in the way the p-values are interpreted, going so far that the ASA issued a statement about its correct use and interpretation. Systemic problems such as only significant results being reported cause further problems in drawing conclusions from the available research as they cause incentives for practices such as p-hacking or HARKing.

It is clear that more focus has to be set to use NHST correctly and that scientific reasoning takes more work than simply determining the success of ones analysis using a single number without context.

There is also a need to reform the publication requirements in order to get rid of those misconceptions and harmful incentives, as well as setting incentives of getting findings reproduced.

Finally, besides teaching the correct usage and conditions for NHST, reasonable alternatives should be taught as well.

8 References

- **Chavalarias, David, et. al.** "Evolution of Reporting P Values in the Biomedical Literature, 1990-2015." *JAMA*, 2016, Vol. 315(11), pp. 1141-1148

- **Cohen, Jacob.** "The earth is round (p<.05)." *What if there were no significance tests?* Routledge, 2016, pp. 69-82

- **Easterbrook, Philippa J., Berlin, Jesse A. Gopalan, Ramana and Matthews, David R.** "Publication bias in clinical research." *The Lancet*, 1991, Vol. 337, pp. 867-872

- **Ioannidis, John P. A.** "What Have We (Not) Learnt from Millions of Scientific Papers with P-Values?" *The American Statistician*, 2019, Vol. 73, pp. 20-25

- **Pollard, P. and Richardson, J. T. E.** "On the probabilty of making Type I Errors" *Psychological Bulletin*, 1987, Vol. 102, pp. 159-163

- **Silberzahn, R., et. al.** "Many Analysts, One Data Set: Making Transparent How Variations in Analytical Choices Affect Results." *Advances in Methods and Practices in Psychological Science*, 2018, Vol. 1(3), pp. 337-356

- **Szucs, Denes and Ioannidis, John.** "When null hypothesis significance testing is unsuitable for research: a reassessment." *Frontiers in human neuroscience*, 2017, Vol. 11, Article 390

- **U.S. Constitution** Art. I, Sect. 1-3

- **Wasserstein, Ronald L., Lazar, Nicole A.** " The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician*, 2016, Vol. 70, pp. 129-133

- **Wasserstein, Ronald L., Schirm, Allen L., and Lazar, Nicole A.** "Moving to a World Beyond "p<0.05"." *The American Statistician*, 2019, Vol. 73, pp. 1-19

- **Wooldridge, Jeffrey M.** "Introductory Econometrics - A Modern Approach." *Delhi: Cengage Learning*, 2013, 5th edition