Name: Lennart Stipulkowski
Supervisor: Prof. Dr. Ullrich Köthe
Date: December 27, 2019

Seminar: „How do I lie with statistics?“

Flexible Data Collection

How do I lie with flexible data collection?

Lennart Stipulkowski

Contents

"A reader quick, keen, and leery
Did wonder, ponder, and query
When results clean and tight
Fit predictions just right
If the data preceded the theory."
- Anonymous

1 Introduction

In all disciplines data is used for the justification of the hypotheses of the researchers. Studies require the data to be collected and analyzed and therefore underlie the effects of the process of collection and analysis. Scientific evidence is often measured and described by the commonly used „p-value“. The p-value is a measure for the evidence but simply describes the likeliness of the null hypothesis H_0 given a certain measured value. In many disciplines, a p-value of $p = .05$ is considered to be significant but this is just a convention. The popularity of the p-value began with Ronald Fisher in the 1920s as he introduced the theory of hypothesis testing with p-values [1]. The popularity of the p-value as a measure for significance led to consequences as p-hacking and HARKING. Scientists discovered the many possibilities to manipulate the data collection and data analysis to improve the significance of their results and even accidentally manipulate their results. In data collection and analysis there are a lot of commonly accepted “researchers degrees of freedom” that improve the p-value but in reality have no effect or a small effect on the significance of the hypothesis.

In the paper of Simmons et al. [9] an experimental study was created to estimate the effects of degrees of freedoms researchers commonly use. The effects of different degrees of freedoms were tested in computer simulations and showed alarming results as it is easy to reach significant results for datasets with no underlying effect by simply combining well-accepted methods of data collection and data analysis.

Head et al. [3] tried to estimate the extent of p-hacking which is the use of the manipulations in order to minimize the p-value for the publications. Their approach was to analyze the p-curve gathered from sets of publications to estimate to which extent p-hacking is happening in different fields of study. They used text-mining to collect the p-values from publications in order to analyze the distribution of p-values.

Shun-Shin et al. [8] had a different approach to find evidence for p-hacking. They looked at the distribution of medical data and found a method to detect the effects of different manipulations on the distribution. This method promises a simple but sensitive test for p-hacking.

In the end, there is the general question of how to deal with the hypothesis in the scientific process. The textbook way of formulating the hypothesis before the collection and analysis is commonly replaced by hypotheses that are postulated after the results of the study are known. This process has been labeled as “HARKing” which stands for

"Hypothesising After Results are Known". Kerr et al. [5] describe how scientists use HARKing for their publications.

2 Statistical Hypothesis Testing

Statistical hypothesis testing is used to test for evidence of the given hypothesis. Basically, statistical hypothesis testing is done by comparing a sample of data to the population considering the theory and model of the population. The result of the hypothesis test is given by the likeliness to obtain the sample or a more extreme sample, given the null hypothesis H_0 being true [2]. This is done by several statistical tests that are used for different constellations. The null hypothesis is the contradiction of a hypothesis for an effect which means that if the null hypothesis H_0 is true there is no effect at all for the tested hypothesis. Alternative hypotheses are labeled as H_1.

3 The Problem of Flexible Data Collection

The main problem for studies is a false positive rate in favor of the studies hypothesis (Type 1 errors). This could lead to false evidence for effects that do not exist. There are various problems connected to the false-positive effect findings. If false-positive findings are published in literature they persist in literature. A revocation is unlikely due to the risk of the author to lose the reputation if he revokes his results. The same applies to the Journal that shares this interest. [9] Against this there is the risk of losing credibility if other scientists expose these errors or find evidence against the hypothesis. This can lead to a loss of credibility of the author, journal or in fact the field of study. A different effect is the practice to not publish null findings as they are not considered to be valuable. There is no incentive to publish null findings as journals often reject null findings for publication and null findings mostly do not contribute to the reputation of the scientist and thus are not published. They can even have a negative influence on their reputation as they lower the fraction of "significant findings". As the p-value gained such importance it led to a simplified view on "significance". Reducing the p-value became a common practice in all disciplines and is often misunderstood as a gain of significance. But many methods that are commonly used reduce the p-value but the significance for the hypothesis remains the same. This leads us to the problems of the researchers' degrees of freedom in data acquisition and analysis. The scientist faces many decisions in data collection and analysis that are connected to the researchers' degrees of freedom. This raises the question of when to do these decisions. The common practice is to specify the hypothesis when interpreting the data and not beforehand

with an estimated prevalence of 50-90% [3, 5]. This promotes the appearance of p-hacking as this simplifies the whole process to choosing the methods that result in a low p-value. All results gathered in this process are often self-justified to be the significant results.

4 Researchers Degrees of Freedom

The researchers' degrees of freedom are several techniques, methods, and manipulations in data collection and analysis that are often well accepted and commonly used in studies. Simmons et al. is mentioning the following degrees of freedom [9]:

- Amount of data to be collected
- Exclusion of observations
- Selection of combined conditions and which one to compare
- Which control variables to use
- Combining measures
- Transforming measures

4.1 Amount of data to be collected

When collecting data for a study there are many degrees of freedom of the amount of data to be collected. For example, the size of a sample must be chosen or the rule of stopping the data collection for flexible data collection can be chosen.

4.2 Exclusion of observations

In many studies, certain rules of data exclusion are applied. This is often used to filter extreme values that might not fit the model. In many studies, probands under the age of 18 are excluded and thus underlie an effect of exclusion. Another example is a test of reaction times of students in a certain task. If one of the student answers too fast one can suppose that this is an error. But the problem is that it is not possible to find an accurate threshold that is differentiating all errors from all true values.

4.3 Selection of combined conditions and which one to compare

When combining conditions the likeliness to find a combination considered to be significant rises as one increases the number of checks for significance. The same applies to the selection of which one to compare. Even if the null hypothesis is true one increases the likeliness to find a combination with significance.

4.4 Which control variables to use

If there is the need to use a control variable for the study the result is another degree of freedom to choose from. This is the simple result due to the fact that it increases the amount of tests that can become true and then be selected.

4.5 Transforming measures

In all disciplines several methods of transformation exist that can influence the result to become a Type 1 error in favor of the studies hypothesis. The estimated p-value might not be effected and fools the studies results.

5 HARKing - Hypothesizing after Results are Known

HARKing is an acronym of "Hypothesizing after Results are Known" and refers to hypotheses that are specified while and after the data collection and analysis. According to Kerr et al. and Head et al. [3, 5] HARKing occurs with a prevalence of estimated 50-90% and thus has a great impact on most published studies. The classical approach of specifying the hypothesis in the beginning, is often replaced by several forms of HARKing where findings of the data collection and data interpretation are changing the hypothesis of the study. This practice encourages the techniques of p-hacking as the hypothesis is fitted to the p-hacked findings. Kerr et al. suggests the following categorization to categorize the forms of hypothesis specification:

	After Results Are Known	
Before the Study	**Plausible**	**Implausible**
Anticipated & Plausible	a	b
Anticipated & Implausible	c	d
Unanticipated	e	f

Table 1: Cross-Classification of Hypotheses by A Priori and Post Hoc Status [5]

Hypotheses in publications are often combinations of hypotheses from different categories. The classic hypothetico-deductive approach is a combination of hypotheses of category a or b. Whereas HARKing is based on hypotheses of the categories a, c and e.

6 How bad can it be?

Simmons et al. [9] conducted different simulations on sampled data to estimate the influence of the different researchers' degrees of freedom of data collection and data analysis. They considered four different degrees of freedom in their simulations. All four degrees of freedom are commonly used in different fields of studies. The four simulated degrees of freedom are:

(a) choosing among depending variables
(b) choosing sample size
(c) using covariates
(d) reporting subsets of experimental conditions

For example, the researchers' degree of choosing the sample size is used by roughly 70% according to a survey where behavioral scientists were asked. Often there is the belief that it has a trivial influence on the false-positive rate of the significance test.

The results of the "How bad can it be" simulations are collected in the following diagram 1. It shows the likeliness of a false-positive finding for different combinations of the researchers' degrees. All results are based on sampled data with no effect size and thus correspond to the null hypothesis H_0. The results are divided for different levels of significance.

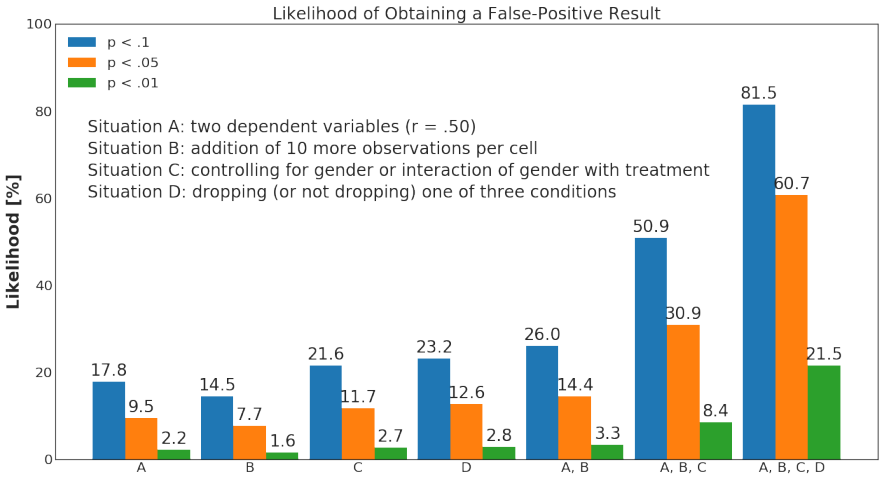

Figure 1: Results of the "How bad can it be" simulations

For the standard significance level of $p = .5$ we find a likeliness around 10% for false-positive results when one of the researchers' degrees is applied. The combination of all four degrees of freedom shows an extreme false-positive rate of 60.7% and thus shows that hypothesis significance testing can easily be manipulated with common practices.

7 Analysing p-Hacking: p-Curve

Head et al. [3] analyzed the distribution of p-values in sets of publications of many disciplines. They used text mining to extract the p-values from the papers. The collected p-values are then plotted to analyze the distribution of the p-values which is called p-curve. P-curves can be assessed to get an estimation of the amount of p-hacking and the effect of the publication bias for the set of publications. The usual distribution of p-values is a right-skewed distribution if the effect size is greater zero. With no effect, the distribution has a uniform value.

There are two main effects that can be measured using the p-value distribution. One is the "selection bias" or "file drawer effect" which shows as a low publication rate with results with a p-value over .05. This is because many authors do not publish results that do not reach significance as it is not improving their reputation. The other effect is the "inflation bias" or "p-hacking". P-hacking results in a distribution with some more results right under the significance level e. g. $p = .05$. This is because authors are trying to optimize their analysis and if the analysis yields significance they are publishing the results. This means that many studies just above significance $p > .05$ are reanalyzed and reach significance after applying methods of the researchers' degrees of freedom.

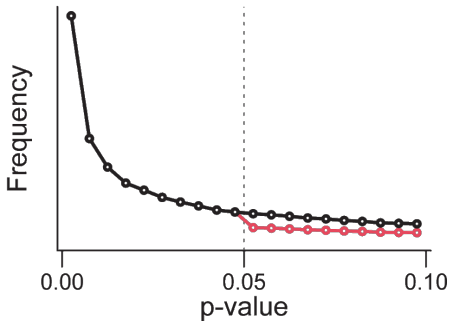

Figure 2: Effect of the publication bias on the p-curve [3]

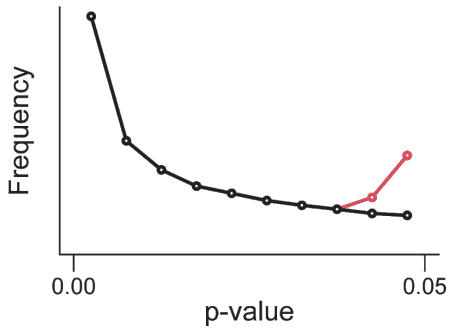

Figure 3: Effect of p-hacking on the p-curve [3]

To determine the dimension of p-hacking Head et al. [3] is comparing the amount of p-values for two different bins. The lower bin is in the range of $0.04 < p < 0.045$ the higher in the range of $0.045 < p < 0.05$. In theory the amount for both bins should be nearly equal with a few more p-values in the lower bin as one can see in figure 3 (black curve). But the effect of p-hacking is increasing the amount of p-values in the higher bin in comparison to the lower bin (red curve). Using a binomial test one can estimate the probability for p-hacking for a set of studies.

7.1 Results

The study of the p-curves in different fields yields evidence for p-hacking being a widespread practice. Another finding is that most studies are in fact studying effects with existing effect size as the p-curve is right skewed. The results for p-hacking in "Medical and health sciences" and "Multidisciplinary" yield a significant result of $p < 0.001$ that p-hacking occurs. Many other fields have similar results with a high probability for p-hacking. In the study they distinguished p-values in the results section from p-values in the abstract. They found that in abstracts the amount of believed p-hacking is lower as the authors often write the most significant results in the abstract and results with lower significance in the results section.

8 Analysing p-Hacking: Tadpole-Test

Another way to find evidence for p-hacking is the tadpole-test developed by Shun-Shin and Francis [8]. They analyzed medical datasets and found a way to detect p-hacking by looking at the distribution of the data itself. The distribution of a certain attribute mostly corresponds to a normal distribution or constrained normal distribution. When plotting the distribution one would see an evenly distributed pattern or if constrained a pattern that looks like a tadpole. If two groups are compared the distribution would normally look like two tadpoles swimming away from each other. But remeasurement, removal, and reclassification of the data would result in a reshaped distribution that looks like two kissing tadpoles. The change of the distribution can be quantified using the D'Agostino Z-Score, which is a measure of the skewness of the distribution. The following figure 4 shows the natural normal distribution and normal distribution with a constrained range in comparison with the unnatural distribution looking like kissing tadpoles. The figure 5 shows the effects of remeasurement, removal, and reclassification on normal distributed data. Remeasuring data is empathizing the heads of the tadpole as remeasurements often happen when measurements are showing values outside the expected range. Removing data points is removing the second tail from the head which results in the tadpole shape. The effect of reclassification is empathizing the remaining tails.

In three different scenarios Shun-Shin and Francis [8] show that the manipulations of remeasurement, removal, and reclassification are very likely in clinical environments as the result of missing knowledge of the effect of these manipulations. One can assume that similar scenarios take place in other fields. As one example they gave the following scenario:

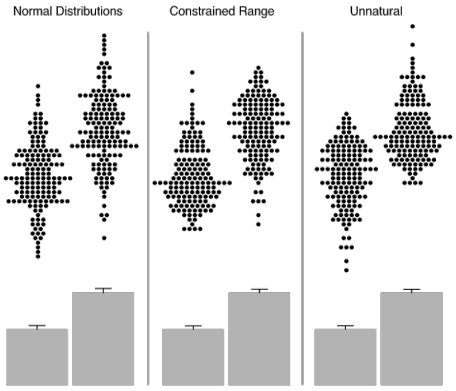

Figure 4: Natural/Unnatural distributions (tadpoles) [8]

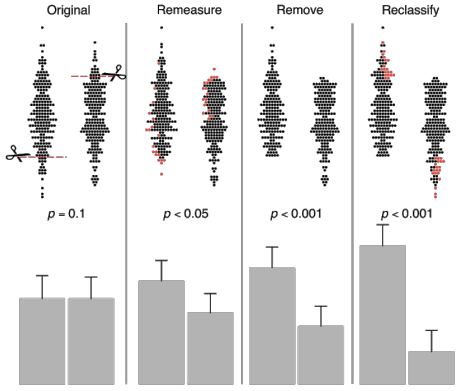

Figure 5: Effects of remeasurement, removal, reclassification [8]

Imagine being a medical doctor in a clinic. Being on the round a student nurse is measuring the oxygen saturation of a patient. The expected value for healthy people is in the range of $spO_2 = 95 - 100\%$. All previous measurements have been $> 97\%$ The measuring device now measures a value of 85%. Now the question is: What do you do? 1. Do you confine to bed and initiate to give the patient oxygen? 2. Do you document a value of 85% and request a test for a pulmonary embolism? 3. Or do you remeasure the oxygen saturation yourself and document the value of the remeasurement?

Most people would choose 3 and remeasure the value as they expect an error in the last measurement. This clarifies that remeasurement is likely to happen under certain circumstances.

In simulations Shun-Shin and Francis [8] calculated the number of manipulations needed to attain significance of $p = .05$ for groups of patients. The following figures (6, 7) show the amount of manipulations needed in absolute count of manipulations and relative amount of manipulations.

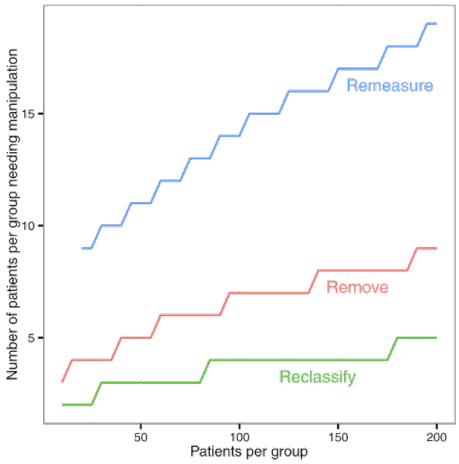

Figure 6: Absolute count of manipulations needed [8]

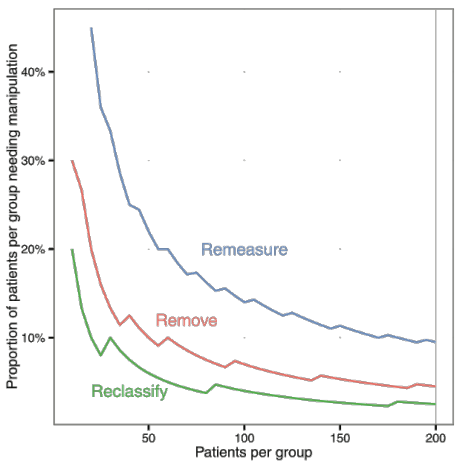

Figure 7: Relative count of manipulations needed [8]

One can see that for large patient groups $N > 150$ the relative proportion of patients needing manipulation is relatively low. In comparison, the effect of reclassification has the greatest impact. For a patient group of $N = 800$ we only need around 1% of the patients to be reclassified. Removal needs 2% and remeasurement needs 4% of the patients. This shows that a relatively small proportion of the patient data needs to be manipulated to get a positive result.

9 Solutions

Many authors are not aware of the problems of a flexible collection of data and the effects on the false-positive rate of the findings in favor of their hypothesis. In order to give some principles as a guideline, the authors Simmons et al. [9] wrote six rules for the authors and four rules for the reviewers of the publications to diminish these problems. An organized way to diminish the problems of flexible data collection is the idea of registered reports that also prevents the practice of HARKing.

9.1 Rules for the authors and reviewers

Simmons et al. [9] wrote the following six rules for the Authors:

- Rule for terminating data collection before collecting
- Enough observations per cell
- List all variables collected
- Report all experimental conditions
- Report the statistical results as if no observations would be excluded
- If analysis includes covariates also reporting the results without covariate

The rules for the reviewers:

- Author should follow the rules above
- Tolerance of imperfections in results
- Require the authors to report their analytic decisions
- If justification of data-collection or analysis is not compelling require authors to conduct exact replication

9.2 Registered Reports

Registered Reports is a process of publishing results by the organisation “Center for Open Science” (COS). The principal idea is to hand in the concept of the study with the

studies hypothesis before the process of data collection and data analysis begins. The concept and hypothesis is then reviewed in a first peer-review process. After the data collection and analysis the finished report is again peer-reviewed. This idea should prevent HARKing and p-hacking. With registered reports the opportunity to make hypotheses after collecting the data is diminished and the peer reviewers will see the change of the hypothesis if it changes to the final report. P-hacking is also exacerbated because the different methods of analysis must be selected beforehand. This does not allow the author to test many different methods to choose the method with a positive result. According to COS currently 217 journals are using the practice of registered reports [7]. The publishing of registered reports does not inhibit all exploratory analyses done in advance of the data collection. But findings and analyses done after the collection of data are marked and identificable as findings that are not preregistered. Nosek and Lakens [6] further emphasize that registered reports do not prevent all Type 1 and Type 2 errors and for replication studies a registered report disagreeing the original study may be a problem as it raises more questions and may reduce the trustworthiness.

10 Conclusion

The problem of a flexible collection of data and HARKing is widespread. Head et al. [3] show that many disciplines are likely to practice p-hacking and thus threaten the reputation and credibility of their fields. Also alarming is the fact that this affects the quality of the results of meta-studies relying on these studies. The quality of meta-analyses is strongly depending on the quality of the individual studies. According to anonymous surveys done by John et al. [4] where 2.000 psychologists have been asked a lot of them admit practices of the researchers' degrees of freedom that are disputable.

The simulations done by Simmons et al. show the effect of the different degrees of freedom of data collection and analysis on the Type 1 error rate. Alarming is the fact that the results are alarmingly high which affects the believability of many studies in the past.

This leads us to the question of how to react to the findings. Important is a general understanding of the effects of the researchers' degrees of freedom by the researchers. This would improve the self-convincing practice of the p-value optimization to be questioned. A critical discussion about the practices of HARKing should be the result

of the alarming prevalence. Concepts like registered reports should be more popular and accepted helping to diminish today's practice.

"There are three kinds of lies: lies, damned lies, and statistics."
- Anonymous

References

1. D. J. Biau, B. M. Jolles, and R. Porcher. “P value and the theory of hypothesis testing: An explanation for new researchers”. *Clinical Orthopaedics and Related Research* 468:3, 2010, pp. 885–892. ISSN: 15281132. DOI: 10.1007/s11999-009-1164-4.

2. F. Emmert-Streib and M. Dehmer. “Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference”. *Machine Learning and Knowledge Extraction* 1:3, 2019, pp. 945–961. DOI: 10.3390/make1030054.

3. M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. “The Extent and Consequences of P-Hacking in Science”. *PLoS Biology* 13:3, 2015. ISSN: 15457885. DOI: 10.1371/journal.pbio.1002106.

4. L. K. John, G. Loewenstein, and D. Prelec. “Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling”. *Psychological Science* 23:5, 2012, pp. 524–532. ISSN: 14679280. DOI: 10.1177/0956797611430953.

5. N. L. Kerr, S. Adamopoulos, T. Fuller, S. Greenwald, P. Kiesler, D. Laughlin, and A. Mcglynn. *HARKing: Hypothesizing After the Results are Known*. Technical report 3. 1998, pp. 196–217.

6. B. A. Nosek and D. Lakens. “Registered reports: A method to increase the credibility of published results”. *Social Psychology* 45:3, 2014, pp. 137–141. ISSN: 21512590. DOI: 10.1027/1864-9335/a000192.

7. C. for Open Science. *Registered Reports: Peer review before results are known to align scientific values and practices.* 2019. URL: https://cos.io/rr/.

8. M. J. Shun-Shin and D. P. Francis. “Why Even More Clinical Research Studies May Be False: Effect of Asymmetrical Handling of Clinically Unexpected Values”. *PLoS ONE* 8:6, 2013. ISSN: 19326203. DOI: 10.1371/journal.pone.0065323.

9. J. P. Simmons, L. D. Nelson, and U. Simonsohn. “False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant”. *Psychological Science* 22:11, 2011, pp. 1359–1366. ISSN: 14679280. DOI: 10.1177/0956797611417632.