Data peeking: a quantitative and qualitative exploration of the use of interim analysis

Mandy Woelk
Utrecht University
M.Woelk@students.uu.nl

Esther Klinkenberg
Utrecht University
E.Klinkenberg@students.uu.nl

ABSTRACT
Data peeking, quitting data collection early or adding more participants at the end, offers the advantage of saving time and money. However, performing an interim analysis without correction leads to a Type-I error inflation. Using alpha spending function could be used to solve this problem. In this paper, we simulated the effects of interim analysis with and without an alpha spending function on type-I error, power and expected sample size. We also offer a Bayesian perspective to interim analysis. In the last part, we discuss the use of interim analysis in psychological research using a qualitative approach.

Keywords
Data peeking, interim analysis, alpha spending Bayesian statistics.

INTRODUCTION

A researcher might have looked at his data while the data collecting was still going, and might have been tempted to stop the study early because the results were already significant. Alternatively, a researcher might have analyzed his complete dataset, only to find a result that was just not significant, and decided to collect some additional data. This so called *data peeking* is not an uncommon scene within research. There are some good reasons for a researcher for wanting to stop early or add more data since conducting research is a time and money consuming practice. Why continue to collect expensive data when you already found an effect? Or why throw away your whole research when you can just add some extra data in order to find an effect? The answer to these questions can be found in the consequences of data peeking on the type-I error.

In hypothesis testing within the social sciences it is common to use an alpha level of .05 [1]. Thus, the null-hypothesis will be rejected when the probability of the data under this null-hypothesis is smaller than .05. Choosing this alpha level means that the chance of finding an effect when actually there is not is at most 5%. This is what is called a Type-I error rate, or a false-positive rate.

This type of error plays an important role in explaining the harm of data peeking, as this error rate increases when peeking at data [2].

As data peeking offers several advantages in terms of time and money savings, it is useful to explore how to employ it without encountering the Type-I error increase. In the next section, we will have a look at a solution called *alpha spending.*

ALPHA SPENDING FUNCTIONS

Alpha spending implies that the total allowable Type-I error is spread out over the number of interim analyses. The functions depend on t^*, the information fraction. This fraction indicates how much of the data has been collected in terms of the accumulated information, and thus indicates how much of the total allowable Type-I error rate should be allocated. There are several alpha spending functions, but in this paper we will only focus on the uniform spending function (UNI) and the O'Brien-Fleming spending function (OF). The functions are as follows:

Uniform: $\alpha(t^*) = \alpha \times t^*$

O'Brien-Fleming: $\alpha(t^*) = 2 - 2\,\varphi(Z\alpha/2/\sqrt{(t^*)})$

Where φ denotes the standard normal cumulative distribution function [3].

SIMULATIONS

Simulations were performed in R [4] using the gsDesign package [5] to calculate the p-value boundaries for every sequential analysis. That is, the p-value needed to reject the null-hypothesis for each analysis. These boundaries were calculated for three situations; no correction, the OF function and the UNI function. For each number of planned analysis, 100.000 data sets were created. Data were generated from a standard normal distribution with N1 = N2 = 64 and a pre-specified mean difference *d*.

Type-I error

The Type-I error plays an important role in the risks of data peeking. Since the Type-I error is the probability of rejecting the null hypothesis when the null hypothesis is true, we set the effect size to $d = 0$ in this simulation, in order to show the actual effects of uncorrected interim analysis on the Type-I error compared to performing an interim analysis with an alpha spending function (UNI or OF). The results concerning the Type-I error are presented in Figure 1.

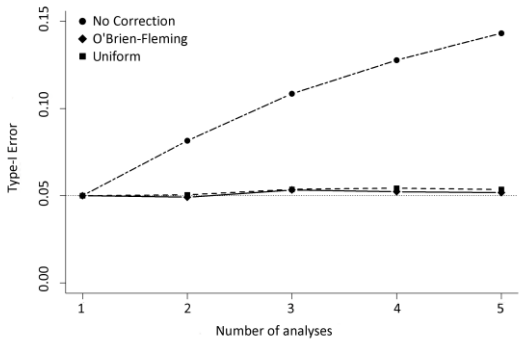

Figure 1. Type I error rate for every interim analysis when using no correction (NC), the OF function or a UNI function.

The x-axis shows the planned number of analyses and the y-axis shows the Type-I error probability (α). The upper line indicates that the error increases with every performed interim analysis when using no correction at all. When planning 5 analyses, the type-I error is as high as .14 instead of the planned .05. The simulation shows that, when using an alpha spending function, the type-I error rate is controlled by being spread out over the different interim moments. Thus, concerning the type-I error rate, Figure 1 shows that it would be better to use an alpha spending function than no correction if a researcher wants to keep this rate as low as planned.

Power

Another statistical parameter that a researcher should take into account is the statistical power of his research. We simulated the effect of doing interim analyses, with or without an alpha spending function, on the statistical power. The results of the power simulation (N1 = N2 = 64, α = .05) are presented in Figure 2. Since the power is the chance of finding an effect when there actually is one, we used *d* = .5. The power increases when peeking at the data without using an alpha spending function. Since the p-value boundaries of the OF and UNI functions are lower than the ones in the NC situation, a possibly present effect is found earlier in the latter situation than in the alpha spending situations, resulting in a higher power level.

When comparing the power of both alpha spending functions, the figure shows that the values of the OF and the UNI function differ somewhat. Relative to a power of .80 for only one planned analysis, the power of the study decreases to a value of approximately .79 for the OF function and the UNI function results in a power of .75 when planning 5 interim analyses. The difference between the spending functions arises from the amount of alpha spent per analysis. The UNI function spends this value equally over each analysis. Thus, when planning 5 analyses with α = .05, the alpha spent at each analysis is .01. When using the OF function, however, the amount of alpha spent at the beginning is very low, but increases to .0122 at the last analysis, where the largest sample size is attained. This means that the chance of finding an effect gets higher at the end of data analysis when using the OF function compared to the UNI function, resulting in a higher power for the first function.

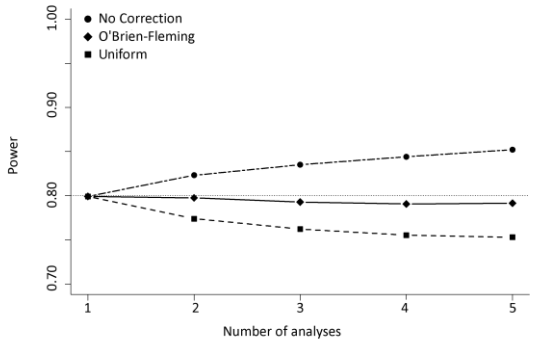

Figure 2. Power for every interim analysis when using no correction (NC), the OF function or a UNI function.

BAYESIAN APPROACH

The simulations presented in Figures 1 and 2 are based on the frequentist approach. There is, however, another approach called the Bayesian approach. From a Bayesian perspective it is not necessary to control the Type-I error in order to make valid inferences [6]. Therefore, we examined what the implications are for doing a Bayesian interim analysis.

In Bayesian statistics, the hypothesis testing procedure does not use a p-value. Instead, a Bayes factor is used to test hypotheses. This factor is a measure of the likelihood of one hypothesis against the other based on the observed data. So, when testing a hypothesis there is not necessarily a dichotomous decision to be made as with using a p-value to decide which hypothesis is 'true'.

Edwards, Lindman and Savage [7] stated that for Bayesian methods, the stopping rules that govern when data collection stops are irrelevant to the interpretation of the data. In order to illustrate the actual effects of sequentially adding more data on (the interpretation of) the Bayes factor, we conducted a Bayesian t-test on a simulated dataset in JASP [8]. We took a random sample from a normally distributed simulation with N1=N2=64 and an effect size of d = .5.

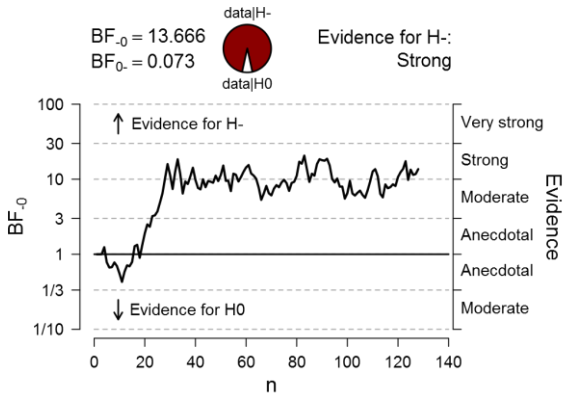

Figure 3. Bayes factor per data point in sequential analysis.

When examining Figure 3 there seem to be no limitations in Bayesian sequential analyses in terms of performing an interim analysis with a stopping rule. For example, if we used a Bayes factor of 10 as a stopping

rule in our simulated Bayesian t-test, the 'data collection' could have been stopped around N = 25 without having to correct for this early stopping. This is in agreement with research of Edwards, Lindman and Savage [7], who stated that for Bayesian methods, the stopping rules that govern when data collection stops are irrelevant to the interpretation of the data. Thus, it seems entirely appropriate to collect data until a certain Bayes factor is reached or until a certain sample size is reached.

QUALITATIVE INTERVIEWS

As described above, the use of interim analysis offers a couple of advantages, especially in time and cost savings. The problems that arise when not using a correction when performing such an analysis are discussed and we examined a frequentist and a Bayesian way to perform an interim analysis in a correct way. However, most articles using an interim analysis are on a medical subject as far as we could find. We wanted to know whether this method could also be useful in the psychological field of research and if so, why it is hardly used or published in psychology research articles. In order to explore these questions, we did several interviews with researchers in the field of psychology at Dutch universities.

Methods

We interviewed two psychology researchers working at the University of Utrecht and one at the University of Groningen. They were all working in the field of clinical psychology, although in different subfields. One researcher did mostly experimental research, the others conducted research with groups of patients that was mostly focused on the development and working of psychological treatments. Since we only interviewed three psychologists, the information we gained is not exhaustive and these questions need to be studied more extensively in possible future research.

We started by asking the researchers to tell something about their work and experience in their research field(s). Then we started talking about data peeking, and explaining the alpha spending functions and showing the simulations we performed concerning the Type-I error and the power. We then turned to the main questions for the qualitative part of this paper: do the researchers think (Bayesian) interim analysis could be useful in the field of psychology and if so, why are there so few publications using these methods in this research field. The results will be split up into three parts. First, we will elaborate on the relevance of interim analysis according to the researchers we interviewed. Secondly, we will explain why the researchers think interim analysis is not used or not published in their field and lastly, we will explain why they think Bayesian statistics are not applied that often.

Relevance

According to the researchers interviewed, the utility of interim analysis depends on the research field within psychology. For example, as reported by the researchers in the field of treatment evaluation, research on this topic could take ten or even twenty years until the data collection has finished. An interim analysis could be a solution here, as the treatment could be applied much earlier than planned. On the other hand, the experimental researcher said that it is not the data collection that takes most of the time. According to this person, "the preparation is more time-consuming in this research field, as the materials for the experiment need to be created, the study needs to be pretested, everything needs to be carefully thought out before the real experiment starts, but the experiment itself does not take much time." For this kind of research, an interim analysis would not save as much time as it could do in studies on treatments. As researchers always need to pay for an interim analysis in terms of statistical power or Type-I error, the advantage of time does not weigh up against the costs in statistical parameters here. For this reason, the experimental researcher thought it would be better not to use an interim analysis in this kind of research.

Use of Interim analysis

Thus, a (Bayesian) interim analysis could be useful at least in the field of research on psychological treatments. In the interviews, we explored the reasons for not using interim analysis. According to two of the researchers interviewed, one of the factors that could play a role in the poor use of interim analysis in psychology could be the fraud that has been committed in the field. One researcher mentioned that "the reputation of the psychological research field has been damaged due to fraud, which means that every study now needs to formulate very clear hypotheses and methods." Another researcher added: "and the power calculation, the number of participants, needs to be clearly mentioned and explained" to prevent researchers from committing more fraud. "The study needs to be registered on a website as well and when publishing an article, the researcher has to prove that the study has been registered." According to one researcher, "it is not possible to register a study with planned interim analyses." Therefore, this researcher thought that an interim analysis may not be the best way to conduct research, taking into account the mistakes commonly made with this research practice, and that it would be better to keep the analyses on a classical level, without any special formulas like interim analysis.

Use of Bayesian statistics

Regarding the use of Bayesian (interim) analysis, a possible reason for the low number of published articles using this method could be that, as they told us themselves, most researchers only have experience with the frequentist way of statistics. Even nowadays, this is the philosophy being taught at universities and people possibly have no idea about the existence of another way of doing statistics. It takes an extra effort to learn Bayesian statistics and people need to be able and willing to make this effort in order to conduct research using this approach. Additionally, as one researcher mentioned, there is a lot of pressure on researchers, as "they need to publish as much as possible, as quick as possible." Thus, there is not much time to learn another way of doing statistics.

CONCLUSION

In this paper, we discussed that data peeking is wrong, but that there are also correct ways of performing an interim analysis. We offered several solutions from two different statistical approaches; the frequentist and the Bayesian approach. The main problem with data peeking is the increase of the Type-I error rate. In order to control this error an alpha spending function could be used, spreading out the total allowable Type-I error over the total number of interim analyses.

The frequentist simulations showed that before planning an interim analysis, a researcher should make up a balance concerning the desired Type-I and Type-II error. If it is more important to have a high power level and subsequently a low Type-II error rate, it would still be better to use an alpha spending function than not using any correction. In this case the researcher could set the alpha level at a higher rate, as this allows him to control both the Type-I error and the Type-II error rate. On the other hand, when power is less important than the Type-I error rate, this value could be set at the usual .05 level or even lower, still with the use of an alpha spending function. The O'Brien-Fleming function would be the better option up to this point, as the Type-I error rate for both alpha spending functions is approximately the same, but the power is somewhat higher when using the OF function.

The Bayesian analysis showed that there seem to be no limitations in doing sequential analysis using the Bayesian method. This is because the Bayes factor can be interpreted the same at any moment of analysis and the use of a stopping rule does not change the interpretation of the results. Therefore, there seems to be nothing wrong with collecting data until a certain Bayes factor is reached or until a certain sample size is reached. So, for Bayesian hypothesis testing, it seems to be completely appropriate to examine data before the data collection is complete and stopping the data collection early.

Performing an interim analysis offers several advantages. However, this method is barely used, as far as we could conclude from found publications in psychological research. In the qualitative part of this paper, we attempted to explore why this is the case by interviewing researchers in the field of clinical psychology. According to these researchers, the relevance of interim analysis depends on the research field within psychology. In experimental research, data collection does not take much time so an interim analysis will not yield many advantages within this field. However, treatment studies usually take years to complete and an interim analysis could offer advantages to the researcher in terms of costs and time.

Nonetheless, the questionable research practices in psychological research have resulted in stricter rules with regard to conducting research, which possibly makes it harder to perform an interim analysis. Researchers need to register their study before starting data collection, and according to one researcher interviewed, it is not possible to register a study with planned interim analyses.

A Bayesian analysis could be a solution if a researcher wants to be able to add participants and inspect the results after every participant without encountering the costs of the frequentist method of an interim analysis. However, the difficulty here is probably the lack of training and experience researchers have with this statistical approach.

ROLE OF THE STUDENT

Mandy Woelk and Esther Klinkenberg were both undergraduate students working under the supervision of Irene Klugkist. The subject has been proposed by Irene, the students have developed the structure and have written the paper. The introduction and theoretical background were written by both students. After those parts, the tasks were more divided. Esther has written the solutions part and the Bayesian approach, Mandy has written the simulations part and worked out the qualitative interviews. However, the simulations in R and the interviews themselves were conducted by both Esther and Mandy, as well as the general conclusion at the end.

ACKNOWLEDGMENTS

We would like to thank Erik-Jan van Kesteren for providing and explaining the R-code and providing the needed information for the analysis in JASP. Also, we would like to thank the researchers who were willing to give an interview for the qualitative part of this paper. Finally, we are very grateful for the opportunity we got to learn so many new things under the supervision of Irene Klugkist.

REFERENCES

1. Lewis-Beck, M., Bryman, A. E., & Liao, T. F. (2003). *The Sage encyclopedia of social science research methods*. Sage Publications.
2. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*.
3. DeMets, D. L., & Lan, G. (1995). The alpha spending function approach to interim data analyses. In *Recent Advances in Clinical Trial Design and Analysis* (pp. 1-27). Springer US.
4. Chambers, J. (2008). *Software for data analysis: programming with R*. Springer Science & Business Media.
5. Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials.* Boca Raton: Chapman-Hall/CRC.
6. Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, *54*(4), 343-349.
7. Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review*, *70*(3), 193.
8. JASP Team (2016). JASP (Version 0.7.5.5)[Computer software]