Running head: APRIOT

Is intermediately inspecting statistical data necessarily a bad research practice?

Computing cumulated error probabilities in ANOVA with the help of APriot

Albert-Georg Lang

(ORCID iD: 0000-0002-6449-3209)

Heinrich-Heine-Universität Düsseldorf

Düsseldorf, Germany

Please send correspondence to:

Dr. Albert-Georg Lang

Institut für Experimentelle Psychologie

Heinrich-Heine-Universität Düsseldorf

D-40204 Düsseldorf, Germany

Email: albert.lang@hhu.de

Phone +49 (0) 211 811 4155

Abstract

Intermediately inspecting the statistical data of a running experiment is justifiably referred to as a bad research practice. With only a few intermediate inspections, Type I error rates inflate to a multiple of the previously defined critical alpha. On the other hand, there are research areas where intermediately inspecting data is extremely desirable if not even necessary. For this reason, in medical research, mathematical methods are known as "group-sequential testing" which compensate Type I error cumulation by adjusting critical alpha. In the field of psychological research, these methods are widely unknown or at least used very rarely. One reason may be that group-sequential tests focus on test statistics based on the normal distribution, mainly the *t*-test, while in psychological research often more complex experimental designs are used. The computer program APriot has been developed to enable the user to conduct Monte-Carlo simulations of what happens when intermediately inspecting the data of an ANOVA. The simulations show clearly how bad a research practice intermediately inspecting data (without adjusting alpha) is. Further, it is shown that in many cases adjusted values of alpha can be found by simulations such that the ANOVA can be used together with group-sequential testing similarly as the *t*-test. A last set of demonstrations shows how the power and the required number of participants of a group-sequential test can be estimated and that group-sequential testing can be favorable from an economic point of view.

Introduction

In 2011, Simmons and his collegues published an article in which they described an "undisclosed flexibility in data collection and analysis" that "allows presenting anything as significant" (Simmons, Nelson, & Simonsohn, 2011, p. 1359). Simmons et al. (2011) describe frequent mistakes made by researchers in the field of behavioral sciences that lead to an unacceptably high rate of false positive results. One of these errors—which Simmons et al. (2011) refer to as "flexibility in sample size"—is intermediately inspecting the data of a running experiment and then deciding to either add further participants or to end the experiment depending on whether the intermediately conducted statistical test shows statistical significance or not. John, Loewenstein, and Prelec (2012) conducted a survey based on questionnaires sent to scientists to investigate the prevalence of "questionable research practices"; special questions were used to estimate the true prevalence beyond the tendency to cheat. According to this study, the estimated percentage of scientists who (at least once) intermediately inspected their data is about 70%. From a mathematical point of view, the problem with intermediately inspecting data is obvious. When intermediately inspecting data, not *one* statistical test is conducted but *multiple*. This means, that if critical alpha is set to .05 for each inspection, the cumulated Type I error for *all tests* will be greater than .05. This is the main reason why intermediately inspecting data is often referred to as a bad research practice that must be avoided (there are other concerns, for more details, see the discussion at the end of this article).

There are however research areas where intermediately inspecting data is extremely desirable or even essential. This is most obvious for clinical research. If a new drug is tested longer than necessary to prove its superiority over an older drug, patients may be treated in a suboptimal

way. On the other hand, if the new medicament is not tested long enough, possible adverse effects or other negative consequences could be overseen (Proschan, Lan, & Wittes, 2006). What is needed, is a way to inspect the data of an experiment in sufficiently small intervals such that evidence for both, superiority over an older drug and negative consequences are detected as early as possible without decreasing the validity of the statistical test used. For this reason, so called "group-sequential trials" are a common practice in medical research. With group-sequential tests, data are inspected intermediately and the acquisition of new data is aborted as soon as the statistical test shows statistical significance. In order to avoid a cumulation of the Type I error, critical alpha of each intermediate statistical test is adjusted such that the overall alpha error probability does not exceed a predefined value, e.g., .05. In principle, the problem is similar to the problem occurring if multiple *t*-tests are conducted. The more *t*-tests are conducted, the larger is the probability that at least one test reaches statistical significance by chance. With the methods proposed by Bonferroni (Abdi, 2007) and Holm (1979), critical alpha of each *t*-test can be adjusted so that the Type I error rate does not inflate. Analogously, the Type I error rate can be kept from inflating if data are inspected intermediately by mathematical methods developed for group-sequential testing (for an overview, see Proschan et al., 2006).

While sequential testing is a common practice in medical research, it is virtually not used in many other scientific areas like psychology. This is true despite the fact that the huge percentage of scientists involved in the "bad research practice" of intermediately inspecting data (John et al., 2012) suggests that there is a need for applying sequential designs. One example where a sequential design would be useful in psychology is trying to detect an effect of unknown size. The classical procedure of planning a psychological experiment is to specify an effect size of interest with respect to the object of investigation and conduct an a-priori sample size analysis. Then, the

required number of participants is tested “in one block”. The problem with a-priori sample size analyses lies in determining a well-founded value for the effect size of interest. This may be feasible in application-oriented research such as when testing the ergonomics of a new computer system. In this case, the minimum effect size may be specified with respect to the costs of development of the new system. However, in basic research fields like perception or memory it is often difficult or even impossible to specify a reasonable effect size a priori. The starting point in a new research area often is the question whether an effect—of whatever size—does exist or not. Thus, one would have to specify an effect size as small as possible or, in other words, choose a sample size as large as affordable. This approach may end up in a huge dissipation of time and money if with the completion of the experiment, the effect turns out to be large enough to be shown with a smaller sample size. With intermediate inspections of the data, an effect of large size would be detected early. Then, as more participants are added, the power of the experiment would raise such that effects of smaller and smaller size could be shown.

So why is sequential testing widely unknown in research areas like psychology while it is common practice in medical research? There are exact mathematical procedures for group-sequential testing (see Proschan et al., 2006) and there are multiple computer programs with which group-sequential tests can be calculated (e.g., Pahl, 2015; Reboussin, DeMets, Kim, & Lan, 2014).

One reason why group-sequential tests are rarely used in psychology could be that the typical experimental designs used in psychological research differ from those used in medical research. One of the most common questions of medical research is whether a new drug is more effective than an older drug. There is a treatment group of patients receiving the new drug and a control group receiving the older drug or a placebo. A *t*-test would be applied for this design. In psychology, experimental designs often are more complex; there are multiple factors expected to

have an influence on a dependent variable. Thus, main effects and interactions have to be computed. One of the most common statistical tests for such designs is the analysis of variance (ANOVA) since it offers a great flexibility making it possible to be applied to a large number of different questions. Most mathematical approaches to group-sequential testing, however, only deal with test statistics based on the normal distribution. Thus, they can be applied to designs based on the *t*-test if the number of participants is large enough since with a growing number of participants, the *t*-distribution approximates the normal distribution. However, these approaches cannot be used with designs based on the ANOVA since the ANOVA is based on the *F*-distribution. This must not be confused with the fact that normality is a prerequisite for the ANOVA; the point is that the *F*-statistic used with the ANOVA is not normally distributed.

For the reasons mentioned above, the question arises whether and to what extent the methods for sequential designs used in medical research are applicable in psychology. While it remains a challenge for mathematics to continue developing further methods that cover a growing variety of statistical tests, Monte-Carlo simulations may be helpful to see what happens if an ANOVA is used with sequential testing and whether alpha can be adjusted similarly as with the *t*-test. The computer program APriot has been developed to enable the user to conduct Monte-Carlo simulations of the ANOVA with the possibility to simulate the effects of intermediate inspections on Type I error rate and power. The aim of developing APriot was to answer two questions: The first question concerns intermediately inspecting data as a bad research practice. How large is the cumulative Type I error probability if data are inspected intermediately? The second question tries to figure out whether and to what extend the ANOVA can be used with sequential testing. If data are inspected intermediately, is it possible to find a critical alpha value such that the cumulative Type I error probability does not exceed a predefined value?

APriot has been developed to make Monte-Carlo simulations accessible to a large number of users. A graphical user interface is provided helping the user to enter data as in other statistical software. There is no need to enter matrices of mean values, standard deviations, or a variance-covariance matrix of residuals. The user can enter the raw data of a pilot study; APriot will use the raw data to compute the matrices necessary for the simulation. APriot has been designed for simulating all effects of the ANOVA, including interactions of an arbitrary order and both, the univariate and the multivariate approach to repeated-measures analyses.

After an introduction into APriot, simulations with varying numbers of simulation runs will be used to answer the question of how precisely simulations estimate probabilities. In the subsequent section, it will be demonstrated that intermediately inspecting data leads to highly-inflated Type I errors even with relatively few inspections. In the subsequent section it is tried to find adjusted values for alpha such that the overall Type I error rate does not exceed a predefined value. The last section shows how the power of a sequential test can be estimated by simulations. Further, it is tried to give a first answer to the question whether more or less intermediate inspections are preferable from an economic point of view.

Description of APriot

The computations conducted during a simulation

APriot simulates the effects of repeatedly inspecting statistical data on the alpha error probability and the power. For the simulations, random numbers are generated that have the distributional properties—assuming a normal distribution—estimated by the raw data the user enters into APriot.

For this reason, APriot computes the matrices of means, standard deviations, and—for within-subject designs—correlations.

Let us consider a simple example. If the user enters the raw data for a between-subjects ANOVA with 1 factor of 3 levels, APriot computes the matrices of means (e.g., 3, 4, 5 for the 3 factor levels) and standard deviations (e.g., 1.0, 1.8, and 1.5). If "simulate alpha" is selected by the user, the procedure continues as follows: Alpha is the probability of detecting an effect while this effect does not exist in the underlying population. Thus, in the "virtual population" from which the random numbers for the subsequent simulation are drawn the effect does not exist. In the case of our 1 factor – 3 levels example, 'the effect does not exist' means that the means of all three levels are equal. Thus, APriot computes the mean of the three means (in our example: 4) and uses this mean as the basis for generating the random numbers for all three factor levels. Further, the user can choose whether homogeneity of variances should be assumed. Depending on the user's selection, APriot computes the variance (standard deviation) for each of the 3 levels independently or computes one variance (standard deviation) for all 3 levels together.

When the simulation is started, APriot generates random numbers drawn from a normal distribution with mean = 0 and sd = 1. These random numbers are converted into numbers with the means and standard deviations given by the previously computed matrices. If a repeated measures variable is contained in the effect to be simulated, the matrix of random numbers is additionally multiplied by the result of a Cholesky decomposition of the correlation matrix computed from the raw data entered by the user to obtain correlated data.

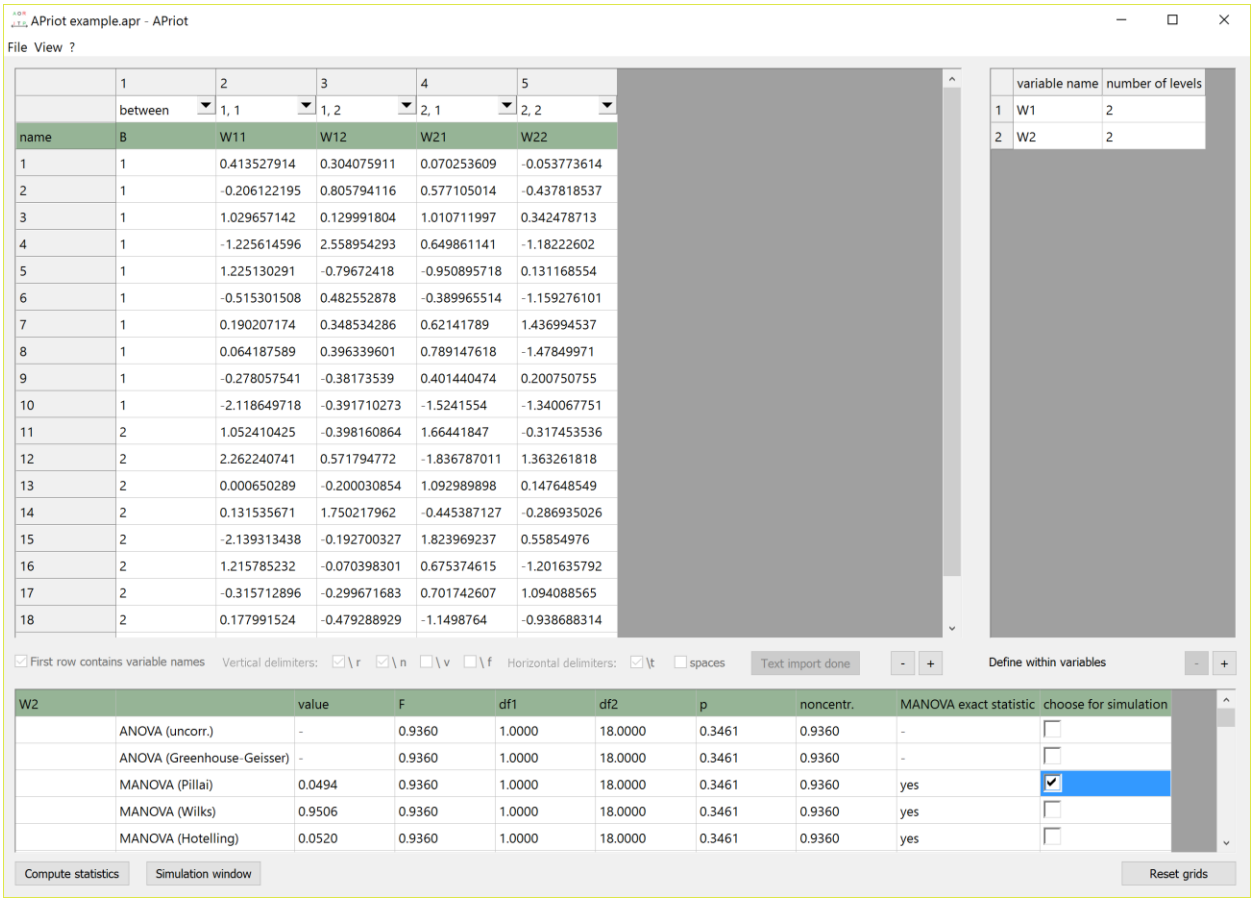
To continue our example, let us assume that the user chose to conduct one ANOVA with 10 participants and to add a maximum of 4 x 10 participants if statistical significance is not reached. Further, the user selected to "Replace the original matrix". Thus, APriot generates 10

(participants) x 3 (levels) random numbers according to the distributional properties described above. With these random numbers, a 1 factor – 3 levels ANOVA is computed. If the result is "significant" the simulation pass is marked as a 'hit' and the next simulation pass is started. If the result is not significant, additional 10 x 3 random numbers are generated and a new ANOVA is computed. This ANOVA is computed with all participants available so far, that is, 20 x 3 random numbers in our example. If the result now is significant, the pass is marked as a "hit" and the next pass is started. If the result is not significant, further 10 x 3 random numbers are generated. This procedure is repeated to a maximum of 4 times. If the last ANOVA does not show significance, this simulation pass is marked as a 'miss'; then, the next simulation pass is started. At the end, the number of hits is divided by the number of simulation passes, and the result is displayed.

If the user chooses not to replace the original matrix, the matrix of the raw data the user entered is used as the initial block of participants and only the additional blocks of participants are simulated with random numbers. This may be useful if the user has already started the experiment and wants to simulate the cumulated Type one error when adding participants to the existing dataset.

If "Simulate power" is selected instead of "Simulate alpha error probability", there is no need to compute the distributional properties, as they were if the effect did not exist in the underlying population, since the power is the probability to detect an effect if it really exists in the underlying population. Thus, the means matrix computed from the raw data is left unchanged. All other aspects of the simulation are the same as described before.

The handling of APriot



	1	2	3	4	5
	between	1, 1	1, 2	2, 1	2, 2
name	B	W11	W12	W21	W22
1	1	0.413527914	0.304075911	0.070253609	-0.053773614
2	1	-0.206122195	0.805794116	0.577105014	-0.437818537
3	1	1.029657142	0.129991804	1.010711997	0.342478713
4	1	-1.225614596	2.558954293	0.649861141	-1.18222602
5	1	1.225130291	-0.79672418	-0.950895718	0.131168554
6	1	-0.515301508	0.482552878	-0.389965514	-1.159276101
7	1	0.190207174	0.348534286	0.62141789	1.436994537
8	1	0.064187589	0.396339601	0.789147618	-1.47849971
9	1	-0.278057541	-0.38173539	0.401440474	0.200750755
10	1	-2.118649718	-0.391710273	-1.5241554	-1.340067751
11	2	1.052410425	-0.398160864	1.66441847	-0.317453536
12	2	2.262240741	0.571794772	-1.836787011	1.363261818
13	2	0.000650289	-0.200030854	1.092989898	0.147648549
14	2	0.131535671	1.750217962	-0.445387127	-0.286935026
15	2	-2.139313438	-0.192700327	1.823969237	0.55854976
16	2	1.215785232	-0.070398301	0.675374615	-1.201635792
17	2	-0.315712896	-0.299671683	0.701742607	1.094088565
18	2	0.177991524	-0.479288929	-1.1498764	-0.938688314

W2		value	F	df1	df2	p	noncentr.	MANOVA exact statistic	choose for simulation
	ANOVA (uncorr.)	-	0.9360	1.0000	18.0000	0.3461	0.9360	-	☐
	ANOVA (Greenhouse-Geisser)	-	0.9360	1.0000	18.0000	0.3461	0.9360	-	☐
	MANOVA (Pillai)	0.0494	0.9360	1.0000	18.0000	0.3461	0.9360	yes	☑
	MANOVA (Wilks)	0.9506	0.9360	1.0000	18.0000	0.3461	0.9360	yes	☐
	MANOVA (Hotelling)	0.0520	0.9360	1.0000	18.0000	0.3461	0.9360	yes	☐

Figure 1: The main window of APriot. On the left side, there is the data area where the variables and variable combinations are entered; on the right side, the names and numbers of levels of the within-subject variables are specified. The results of the ANOVA are displayed on the bottom of the window. The example shows a design with one between-subjects variable ("B") with 2 levels and 2 within-subjects variables ("W1" and "W2") with 2 levels each.

Figure 1 shows the main window of APriot. The example used in this introduction is part of the software installation of APriot; after the installation, it can be found on the user's desktop ("APriot example.apr"). The first step for conducting a Monte-Carlo simulation with APriot is entering raw data, for example, from a pilot study. Data are entered as would be done with other statistical software. In the data area (left side of the window), each between-subjects variable and

each combination of within-subject variables is entered into a separate column. The levels of between-subjects variables are marked by distinct arbitrary numbers (in the example, ‘1’ and ‘2’). The second step is to declare the within-subject variables in the right area of the main window. For each within-subject variable, a name and the number of levels are specified. Depending on how many within-subject variables and levels have been declared by the user, the pull-down menus in the first row of the data area will contain the entries necessary to declare each column as containing a between-subjects variable or a particular combination of the within-subject variables. In the example, there are two within-subject variables named ‘W1’ and ‘W2’ with two levels each. In the data area, columns 2 to 5 are denoted to each factorial combination of the within-subject variables (‘1, 1’, ‘1, 2’, ‘2, 1’, and ‘2, 2’). The first column is denoted to the only between-subjects variable of the data set, hence “between” is chosen in the pull-down menu of the first column. Row 2 (green) of the data area is used to specify names for the variables and variable combinations. If a researcher wants to add further participants, clicking on the ‘+’ button below the data area will add further rows at the end of the data set, where new data can be entered.

A click on the button “Compute statistics” shows the ANOVA’s results on the bottom of the main window. Analyses of all variables are listed with F-values, degrees of freedom, empirical alpha error probabilities, and noncentrality parameters. For within-subject variables (and interactions with within-subject variables), the result of the uncorrected univariate approach to the repeated measures ANOVA is available as well as the result with Greenhouse-Geisser corrected degrees of freedom. Further, the multivariate test criteria Pillai-*V*, Wilks Lambda, and Hoteling’s T^2 are shown.

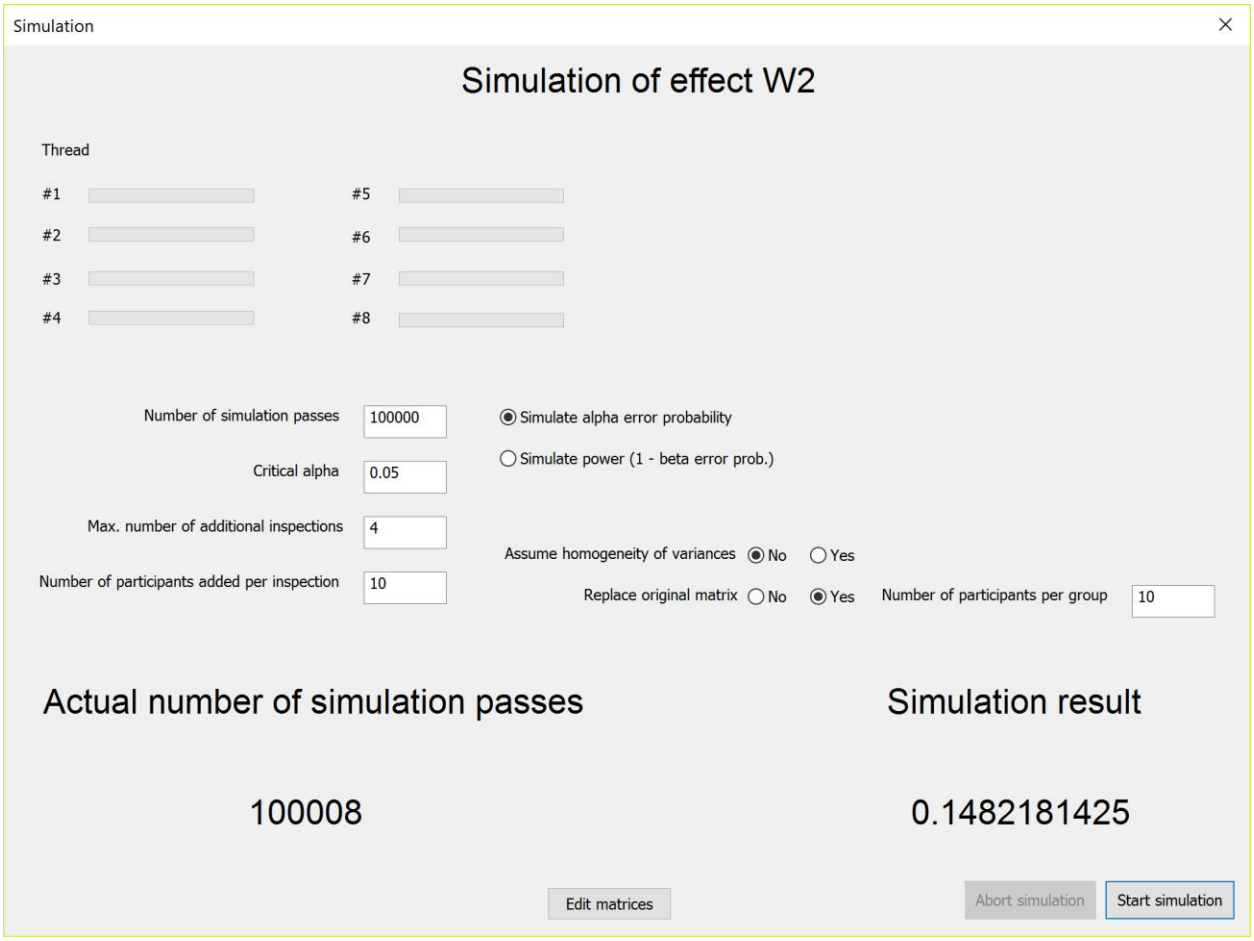


Figure 2: The simulation window of APriot. In this example, the user has chosen to simulate the alpha error probability when a maximum of 4 additional inspections is performed.

The third step is mandatory before simulations can begin: The rightmost column of the results area contains check boxes allowing the user to choose the analysis which will be the basis for the subsequent simulation. Only one analysis may be chosen at one time. With a click onto the "Simulation window" button, a new window appears that permits to specify all parameters of interest for the subsequent Monte-Carlo simulation (Figure 2). The following parameters can be specified:

- Number of simulation passes. This is the number of "virtual experiments" done during the simulation.
- Critical alpha. This is the alpha value needed to count a single simulation run as "significant".
- Simulate alpha error probability/power. The user can choose whether the alpha error probability or the power (1 – beta) will be subject to the simulation.
- Max. number of additional inspections. Here the user can specify how often the sample size will be simulated to be enlarged if a test is not significant.
- Number of participants added per inspection. The number of participants added after each inspection is specified here.
- Assume homogeneity of variances. The variance of each variable combination is estimated from the raw data entered in the main window. The user can specify whether APriot should assume that all variances are equal[1] in the underlying population.
- Replace original matrix. If "No" is chosen, the raw data entered in the main window are kept and only the data of the participants added are simulated with random numbers. If "Yes" is chosen, the raw data are replaced by random numbers with identical distributional properties. In this case, the user can choose how many participants per group the randomly generated initial block of samples will have.

After the simulation parameters have been specified, the Monte-Carlo simulation is started by clicking on the button "Start simulation". On the top of the window, one progress indicator is displayed for each thread of APriot[2]. When the simulation is finished, the "Actual number of

simulation passes" is displayed which is in most cases slightly higher than the number specified by the user due to rounding up when the simulation passes are allocated to different threads. In the field "Simulation result", the result of the simulation (alpha error probability or power) is shown.

Precision of estimating probabilities with APriot

Before performing the simulations, the question should be answered of how precise estimating probabilities using Monte-Carlo simulations is, or in other words, how many simulation runs are necessary to obtain a reliable result. To answer this question, two fictional examples were used. In the first example, a between-subjects effect with three levels was tested; in the second example, the multivariate approach (Pillai-*V*) to a within-subject effect with three levels (r = .5) was tested. Critical alpha was set to .05. Both examples started with 10 participants (per group). The cumulative Type I error probability was tested with four further intermediate inspections, each with 10 additional participants (per group). Monte-Carlo simulations were conducted with 1,000, 10,000, 100,000, and 1,000,000 simulation passes. Each of the conditions was tested 30 times in order to estimate the variability of the simulation results. Figure 3 shows the results. With 100,000 simulation runs, the standard deviations in both cases were 0.001 which seems to be a good value when probabilities are to be estimated with a precision of two decimal places. On the test system (Intel Core i7-4930K CPU with a core frequency of 3.9 GHz), 100,000 simulation runs took about 13 seconds for the between-subjects example and 4 seconds for the within-subject example. Thus, at least 100,000 simulation should be used in order to get precise estimates by simulations.

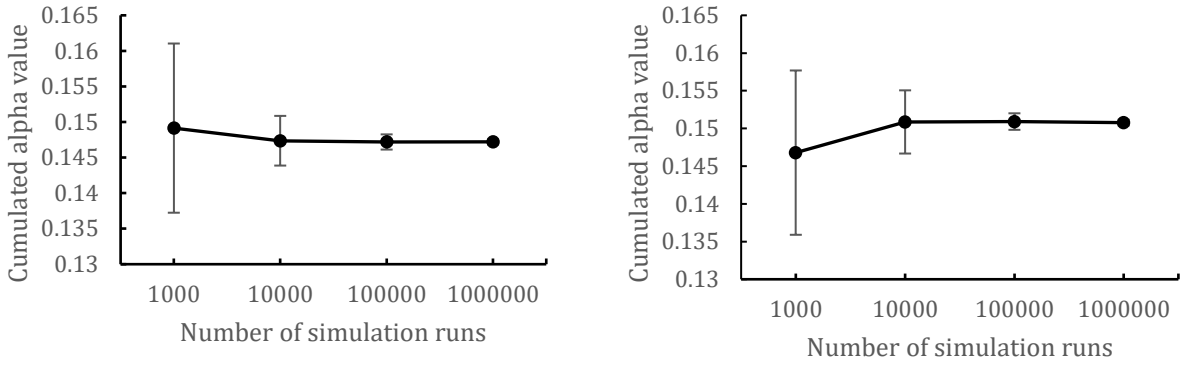


Figure 3: Cumulated alpha error probabilities of two fictional examples with different numbers of simulation runs (left panel: between-subjects, right panel: within-subject – Pillai's *V* test). Error bars denote standard deviations.

Cumulated alpha values if critical alpha is not adjusted

In this section, several examples will be discussed in which intermediate inspections of data are simulated without adjusting critical alpha.

t-test

Since a two-sided *t*-test reveals the same alpha error probability as a one-factorial (two levels) ANOVA, the simulation of a *t*-test offers the chance to verify the results of APriot by comparing them with the results of programs doing Monte-Carlo simulations for the *t*-test.

1 Method

The first example is taken from Lakens' blog (Lakens, 2014a) who discusses the use of the R-function phack by Sherman (2014) which does Monte-Carlo simulations of cumulated error probabilities for the *t*-test. The example is about a fictional experiment in which two groups are compared. A *t*-test is conducted with 50 participants. If this *t*-test does not reveal statistical significance, further 50 participants are tested. This is repeated once more if needed such that there is a maximum of 150 participants. In this fictional experiment, the mean values of both groups were 0.0 and the standard deviations were 1.0. Critical alpha for each test was set to .05; 100,000 simulation runs were used.

As a next step, APriot was used to test different numbers of inspections with the (maximum) total number of participants always being 150 (1 x 150, 2 x 75, …, 30 x 5). With every number of inspections, 100,000 simulation runs were used.

2 Results

For the first simulation with a maximum of three inspections, Phack revealed a cumulative Type I error probability of .107. With the result being rounded to three decimal places, APriot revealed the same result. Figure 4 shows the results for different numbers of inspections. With only one inspection (no intermediate inspections) the Type I error is nearly .05, as is to be expected; with 30 inspections, the cumulated Type I error is about .29. These cumulated error probabilities are in good agreement with the values obtained by Proschan et al. (2006) who used numerical integration for a number of inspections <= 20 and simulations for more than 20 inspections.

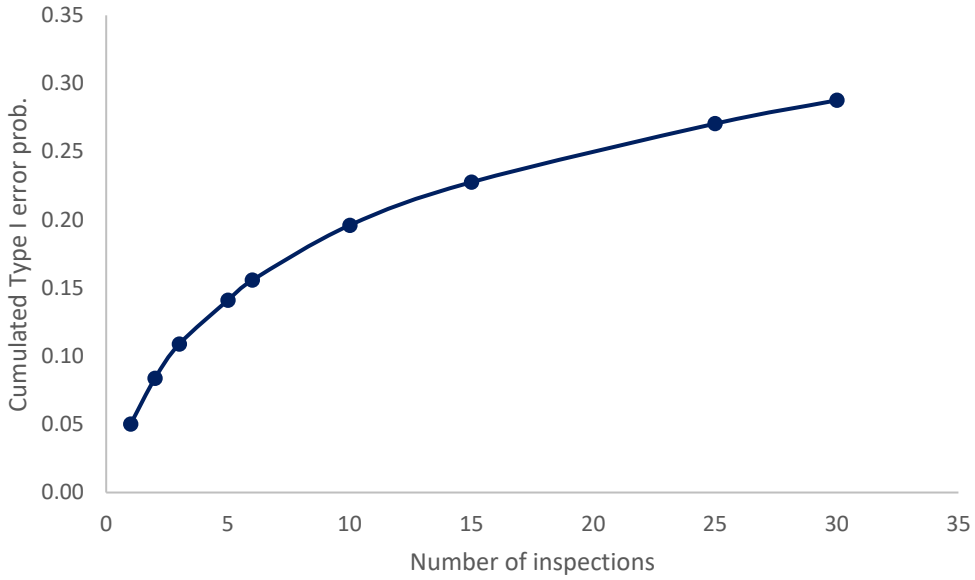


Figure 4: Cumulated Type I error probabilities as a function of the number of inspections.

The results show clearly that intermediately inspecting data (without adjusting critical alpha) is indeed a bad research practice. With only three inspections, the Type I error probability already raises to more than 10%. Maybe many researchers are not fully aware of the large error probability intermediately inspecting data is accompanied with since they only see that the error probability of the statistical test they have conducted is .05. However, they oversee that they would have accepted H1 not only with this test going significant (e.g., significance after having tested 50 participants) but with multiple other tests, as well (significance after having tested 25 participants, 50 participants, 75 participants, etc.). The probability that *any* of these tests reaches statistical significance is much larger than the probability for a *single* test. One can say that "presenting anything as significant" (Simmons et al., 2011) is possible with this kind of research practice, "p-hacking" (Sherman, 2014) is another way to express the same.

ANOVA

The simulations of the *t*-test in the above paragraph were conducted as a demonstration of how bad intermediately inspecting data (without adjusting alpha) is as a research practice and to test whether the results of APriot are in accordance with Phack. In practice, a mathematical approach would be applied to the *t*-test. However, for the ANOVA, computing cumulated Type I error probabilities is not as straightforward since the test statistic of the ANOVA is based on the *F*-distribution rather than the normal distribution.

1 Method

The next example of a Monte-Carlo simulation is the test of a fictional experiment using a 2 x 2 ANOVA with one between-subjects and one within-subject variable. All three effects were tested, that is the effect of the between-subjects and the within-subject variable as well as the interaction; for the effects containing the repeated-measures variable, Pillai-*V* was used as a test statistic. The same numbers of inspections have been simulated as for the *t*-test in the preceding paragraph (1 x 150, 2 x 75, …, 30 x 5 participants). The last simulation of this series is the test of a repeated measures ANOVA (multivariate approach, Pillai-*V*) with four levels. Correlations of -.9, 0.0, and .9 were used.

2 Results

Figure 5 shows the results of both ANOVAs. In order to facilitate comparisons, the result of the *t*-test has been replotted. Obviously—at least for the examples used here—cumulated error probabilities for the ANOVAs are similar to the probabilities of the *t*-test. However, as the number of inspections increases, some of the ANOVA probabilities get larger than the probabilities of the *t*-test.

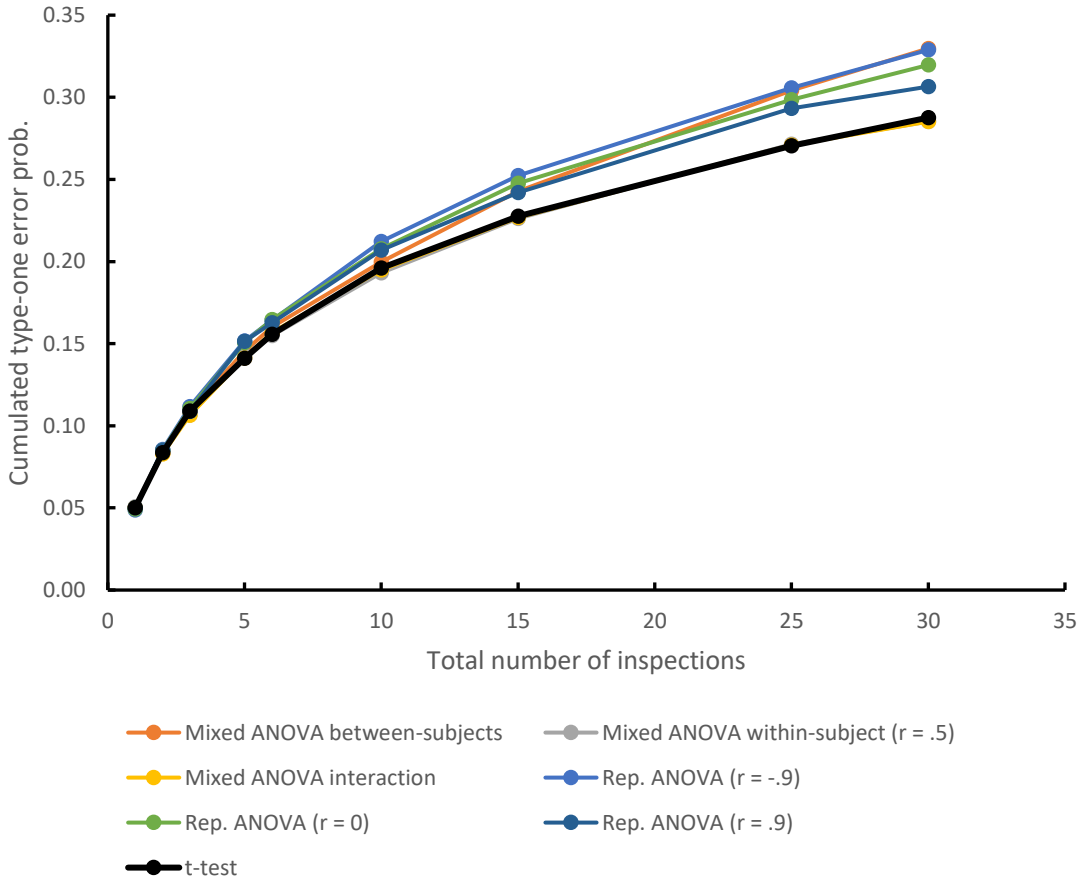


Figure 5: Cumulated Type I errors in the simulations as a function of the total number of inspections. The curve of the *t*-test is plotted a bit thicker as black line. The grey line for the "Mixed ANOVA within subject (r = .5) is covered by the yellow line "Mixed ANOVA interaction" because these two lines are nearly identical.

3 Discussion

While it would be premature to draw general conclusions from the simulations of these examples, it is nevertheless interesting that the results of the ANOVAs do not differ much from the result of the *t*-test, which can be computed with a mathematical approach to sequential testing. Proschan et al. (2006) describe a similar phenomenon. With the mathematical approach, the boundaries for the interim analyses of the data are usually not computed as critical alpha values but as critical values on the normal distribution (*z*-values). As mentioned before, with small sample sizes, the approximation of the normal distribution by the *t*-distribution is inexact and thus, mathematical methods cannot be applied with small sample sizes. Proschan et al. (2006) however found that when using *p*-boundaries instead of *z*-boundaries, the calculations are rather exact even for small sample sizes. Thus, maybe the *p*-boundaries computed for the normal distribution are a good approximation for (some) statistical tests not based on the normal distribution. To answer this question reliably, simulations with many different scenarios or a mathematical proof will be necessary.

Adjusting critical alpha within sequential testing

As discussed before, sequential testing—with adjusting alpha—is something completely different than (illegally) inspecting data within interim analyses. As we have seen in the introduction, there are good reasons for sequential testing, for example, ethical reasons in medical research or the wish to find an effect of unknown size in fundamental research. For statistical tests based on the normal distribution, there are computer programs like Lan-DeMets (Reboussin et al.,

2014) or GroupSeq (Pahl, 2015) helping the researcher to find adjusted boundaries such that the cumulated Type I error probability does not inflate. In the preceding section, simulations were used to see how the cumulated Type I error probability increases with intermediate inspections of the data. In this section, examples are shown in which the attempt is undertaken to use simulations to find an adjusted value of critical alpha such that the cumulated Type I error does not exceed a previously specified value.

t-test

We start with the same example by Lakens (2014a) as used in the previous section. In this example, a *t*-test with a maximum of 3 x 50 participants was conducted. The simulations showed that, with the value of critical alpha set to .05 for each interim analysis, the cumulated Type I error probability for all three inspections is .107. Now, GroupSeq (Pahl, 2015) was used to find an adjusted value of alpha. The analysis revealed that with critical alpha set to .022 for each test, the cumulated Type I error probability for all three inspections is .05. The result of a simulation with APriot is in good agreement with this. Critical alpha was set to .022 within APriot, all other settings of the simulation were the same as in the previous section. With 100,000 simulation runs, the cumulative Type I error probability was .049.

ANOVA

More interesting than the *t*-test is the question whether in a similar way, a corrected value of alpha can be used to control the overall Type I error in a sequentially conducted ANOVA. Thus,

both ANOVA examples of the previous section were computed with the adjusted alpha value just found for the *t*-test (.022). For all effects of both ANOVAS, APriot revealed values of cumulated alpha between .050 and .052. With slight modifications of critical alpha (.021 and .020), the cumulated alpha value was <= .050 for all tests.

With the examples presented here, the adjusted alpha values for the ANOVA found by simulations are similar to the adjusted alpha value for the *t*-test. While further evidence is needed to decide whether the methods applicable to the *t*-test are also valid for the ANOVA, the question should be answered under which circumstances simulations can be used to adjust alpha in ANOVA-designs. The obvious problem lies in the fact that in order to conduct a simulation, the matrices of means, standard deviations and—in the case of repeated-measures designs—correlations of residuals are necessary. Thus, the question whether simulations can be used to find critical alpha values for sequential testing with the ANOVA depends largely on the question whether there are good estimates for the matrices mentioned above. If there is a pilot study, the matrices can be estimated based on the corresponding matrices of the pilot study. If there is no pilot study, one could use a so called *internal pilot study* (Wittes & Brittain, 1990). With an internal pilot study, a first sample of participants taken from an intermediate inspection serves to estimate parameters needed for the running experiment. For example, internal pilot studies are often used to compute the "conditional power", that is estimating the effect size from an intermediate inspection in order to compute the sample size needed for the complete experiment (Lakens, 2014b). Analogously, one could use the parameters found in an intermediate inspection to estimate the parameters needed to run a simulation[3]. It is not completely clear until now how large the influence of the matrices is on the value of cumulated alpha resulting from the simulation. From the tests conducted until now, one could presume that cumulated alpha values are rather similar

with different experimental designs and that the simulation result is quite robust against variations of the matrices. Thus, while having a real pilot study is ideal, internal pilot studies can be a good means of finding the parameters needed to conduct a Monte-Carlo simulation. As with real pilot studies, it is important to keep the sample size of an internal pilot study not too small.

Power and sample size

In the previous paragraphs, the question has been discussed of how the Type I error probability cumulates with intermediate inspections of statistical data and how this cumulation can be corrected. In the remainder of this article, we will take a closer look at how the power of a sequential test is compared to a classical test.

Mathematically computing and simulating the power of a group-sequential test

In the case of a test statistic based on the normal distribution, the computer programs Lan De-Mets (Reboussin et al., 2014) and GroupSeq (Pahl, 2015) can be used to compute the number of participants required to conduct a group-sequential analysis with a given power (Proschan et al., 2006). More precisely, a multiplier is computed with which the number of participants needed for a conventional test has to be multiplied to obtain the number of participants needed for a group-sequential test. A fictional between-subjects *t*-test served as an example. An a-priori power analysis with G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) revealed that a total of 92

participants (46 per group) was needed to show an effect of size d = .76 with a power of .95. A simulation with APriot paralleled this result. With the help of Lan-DeMets (Reboussin et al., 2014), a multiplier was computed with which the number of participants needed for the conventional test had to be multiplied for a sequential test. With a total of three inspections and Pocock bounds, this computation revealed that the number of participants needed with the conventional test must be multiplied by 1.069 for the group-sequential test. Thus, 46 x 1.069 = 50 (rounded up to the next integer value) participants are needed per group with a group-sequential design with three inspections. A simulation with APriot is in good agreement with this result. With critical alpha set to .022 (see above) the cumulative power was .936. Thus, with sequential testing the maximum number of participants needed is larger than with a classical test.

Is there an economical advantage of group-sequential testing?

As seen in the previous example, the maximum sample size of a group-sequential test has to be increased as compared to an experiment with a fixed sample size in order to achieve a comparable power. On the other hand, a sequential test may reach statistical significance *before* the maximum number of participants is reached. Thus, the question arises whether—from an economical point of view—it is better to do a classical a-priori sample size analysis and then conduct the experiment with a fixed number of participants or to conduct a sequential test with a larger maximum sample size but with the chance to finish with less participants. What is needed, is an “expected” number of participants needed when conducting a sequential test.

1 Method

A fictional experiment inspired by current research of the author and his colleagues (Ignaz, Lang, & Buchner, 2013) was used as a pilot study for the simulation with APriot. The variable of interest was a within-subject variable with 7 levels (r = .24), the sample size was 80. The Greenhouse-Geisser-corrected effect of the within-subject variable reached statistical significance, $F(5.586, 441.309) = 3.381$, $\varepsilon = .931$, $p = .004$, $\eta_{part}^2 = .041$. Based on this pilot study, the design of a follow-up study was constructed. The "classical" variant of the follow-up study was planned conducting an a-priori power analysis using G*Power (Faul et al., 2007). The power analysis revealed that, given alpha = .05 and the effect size and sphericity correction of the pilot study, 88 participants were needed to reach a power of .95. A simulation with APriot showed a similar result (G*Power: 1 – ß = .951; APriot: 1 – ß = .963).

A sequential test was planned to be conducted with intermediate inspections after each testing of 10 participants. Simulations were used to find an adjusted value of critical alpha and the maximum sample size required to have a cumulated Type I error probability of not more than .05 and a power of at least .95, as was the case with the classical test. Thus, what was to be found, was a pair of two parameters, a value of adjusted alpha and a maximum number of participants, such that the total Type I error probability and the power met the requirements. By successive approximation, the following pair was found: adjusted alpha: .014, maximum number of participants: 10 x 10 = 100. For reasons of verification, the simulation was repeated with the identified pair of critical alpha and number of participants with 1 million simulation runs: cumulated alpha = .050; power = .950.

With the classical variant of the experiment, a fixed sample size of 88 participants is needed to reach a power of .95. With group-sequential testing, a maximum of 100 participants is needed, with the chance to detect the effect with less participants. So the question is, which variant is more economical – testing 80 participants "in one block" or testing up to 100 participants with up to 10 blocks. Since the probability of a stochastic event can be calculated as the sum of multiple events, provided that the events are disjoint, the following procedure was applied. The probability to find an existing effect with a maximum of 10 blocks is the power of the group-sequential test, which has been found to be .950 by simulation. The probabilities to find the effect with less blocks are disjoint since the experiment is ended as soon as the criterion for critical alpha is met (*either* the effect is found after the first block *or* after the second block *or* after the 3rd … 10th block). Thus, the total power of .950 is the sum of the probabilities of finding the effect with 1, 2, …, and 10 blocks. Finding the probability of the effect to occur with the first block is straightforward; a simulation was run with 1 block of 10 participants which resulted to a power of .019. The simulation was repeated with 2 blocks (result: .107). .107 is the probability to find the effect with 1 block *or* 2 blocks. Thus, the probability to find the effect with *exactly* 2 blocks is the difference between .107 (1 block *or* 2 blocks) and .019 (exactly 1 block). This procedure was applied to all numbers of blocks up to 10. Each number of blocks is not only associated with a probability but also with a "value". The value is the saving of participants as compared to the classical variant of the experiment (88). Positive values indicate that with a certain number of blocks, participants are saved as compared to the classical experiment, negative savings indicate that more participants are needed. As a last step, for each number of blocks the product of probability and saving was computed and the products were summed up. This way, an expectation value of the saving is available.

2 Results

In the example, the expectation value of the saving with sequential testing as compared to a classical design is about 34 participants. This means, if this experiment would be repeated for an infinite number of times, on average 34 participants would be saved. Table 1 shows the results of the simulations and the computation of the savings in detail.

Table 1

Cumulative power depending on the number of blocks.

Number of blocks	1	2	3	4	5	6	7	8	9	10	sum
cumulative power	0.019	0.107	0.249	0.408	0.561	0.694	0.792	0.868	0.916	0.950	
difference	0.019	0.087	0.143	0.159	0.153	0.132	0.099	0.075	0.049	0.034	0.950
saving	78	68	58	48	38	28	18	8	-2	-12	
product	1.50	5.94	8.29	7.63	5.82	3.70	1.78	0.60	-0.10	-0.41	34.75

Of course, this test is only one example and has to be repeated with different experimental designs. However, from the test, it becomes plausible that group-sequential testing can be a more economical way of detecting an effect.

How many intermediate inspections are most economical?

In the preceding paragraph, it has been shown that sequential testing can help to save a considerable number of participants. The question unanswered until now is whether, under

economical aspects, it is better to frequently inspect the data or whether it is preferable to use large "blocks" of samples.

Method

The fictional experiment described in the preceding paragraph has been repeated with different sample block sizes. Together with the preceding experiment, 4 different block sizes were compared: 5, 10, 20, and 40 participants. The values of critical alpha and the maximum number of additional inspections were adjusted so that the accumulated Type I error was nearly identical in all four variants of the experiment (.047 - .049) and the power was at least .95. The savings of participants were computed the same way as described in the preceding paragraph.

Of course, there are multiple parameters that could affect the result, for example, the effect size, variances, and correlations. In order to test whether the effect size—or the (maximum) number of participants—influences the result, the simulation was repeated with the effect size being manipulated. Thus, there were simulations with three effect sizes, a smaller effect ($\eta_{part}^2 = 0.023$, number of participants required for classical experiment: 158), a medium effect (the initial study with $\eta_{part}^2 = 0.041$, 88 participants), and a larger effect ($\eta_{part}^2 = 0.101$, 35 participants).

Results

Figure 6 shows the results. The left panel shows how critical alpha had to be adjusted in order to ensure that the total Type I error probability did not exceed .05. Although for smaller block sizes, critical alpha had to be adjusted to lower values as compared to larger block sizes,

smaller block sizes tended to be accompanied by larger savings of participants (right panel). Thus—at least for the current series of simulations—medium to large numbers of intermediate inspections of the data are preferable from an economic point of view.

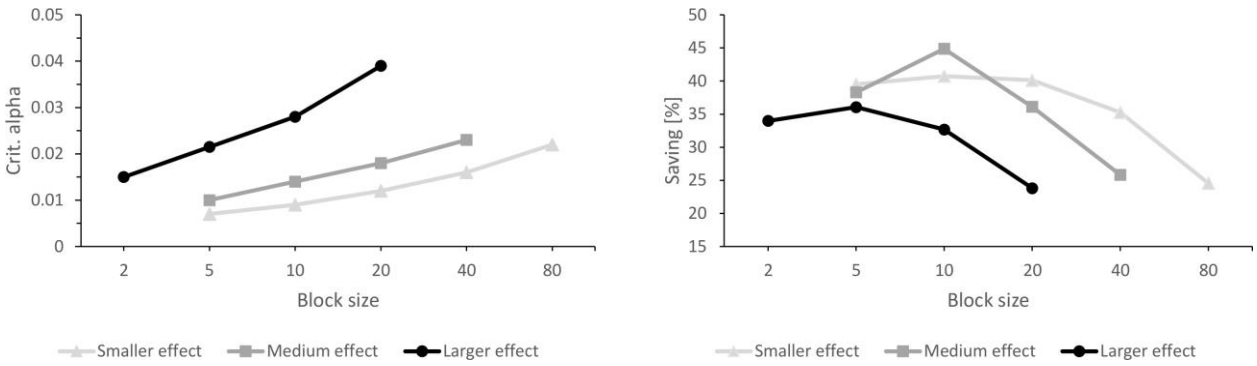


Figure 6: Critical alpha (left panel) and saving of participants relative to classical experiment (right panel) as a function of the sample block size. Larger effects sizes go along with smaller total sample sizes; thus, not all block sizes are used with all three effects.

Discussion

The simulations demonstrate that intermediately inspecting data without adjusting alpha is indeed bad. With only a few intermediate inspections, the cumulated Type I error probability raises to a multiple of the value of critical alpha for each single inspection. By applying this "technique", there is an enlarged risk of effects reaching statistical significance that do not exist. On the other hand, intermediately inspecting data is desirable in many scientific contexts. In psychological research, there is often the problem of not knowing the to-be-expected effect size a priori. This makes computing the necessary number of participants difficult since an a-priori power analysis can only be conducted if there is a well-founded estimate of the effect size. Intermediately inspecting data solves this problem in an elegant way since with each test of additional participants,

the power of the experiment increases. This way, it is possible to successively adept the sample size to the sample size necessary to show an effect of unknown size.

From medical research, a mathematical approach is known as group-sequential testing. With this approach, the shortcoming of an inflated Type I error rate is overcome by adjusting critical alpha of the intermediate inspections (Proschan et al., 2006). However, most of the mathematical methods for group-sequential testing only deal with test statistics based on the normal distribution. Another possibility to overcome alpha error cumulation is applying Bayesian statistics. For a detailed description on how conducting sequential tests with Bayesian statistics and a discussion of the advantages and disadvantages of Bayesian sequential testing, see Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2015).

APriot has been developed to estimate cumulated error probabilities of the ANOVA by Monte-Carlo simulations. In this article, it has been demonstrated that in many cases, Monte-Carlo simulations can be used for planning sequential tests with the ANOVA. With the help of APriot, the method of adjusting critical alpha can be applied to the ANOVA similarly as to the *t*-test. It is interesting that—at least with the examples discussed in this article—the mathematical approach for group-sequential tests based on the normal distribution seems to be a rather good approximation for the ANOVA. It remains to be seen whether and to what extent these results can be generalized. If this would be the case, sequential tests with the ANOVA would be especially easy to plan using the well-known mathematical methods.

Simulations sometimes seem “more concrete” than mathematical computations. So, Monte-Carlo simulations may help within an educational context to “demonstrate” what happens if data are inspected intermediately. This can help researchers and students to understand both that intermediately inspecting data without adjusting alpha makes it impossible to interpret the results

of a statistical analysis in a meaningful manner and that with a suitable correction value the overall Type I error does not inflate to an unknown value but can be kept within predefined bounds.

To make simulations as available as possible, APriot has been designed to compute the matrices necessary for simulations (means, standard deviations, and correlations of residuals) from raw data (e.g., of a pilot study). Data are entered as is done with other statistical software; this helps users not so experienced with the mathematical basis of the ANOVA to conduct Monte-Carlo simulations and "see" what (would have) happened with an actual experiment.

To conclude, whether a *t*-test is to be conducted or an ANOVA: Intermediately inspecting data without a correction of alpha leads to a highly inflated Type I error probability. In many cases, there are mathematical ways to adjust alpha and/or Monte-Carlo simulations can help to find one. With APriot, the researcher is provided with a tool allowing Monte-Carlo simulations of relatively complex experimental designs without much effort. The concern that adjusting alpha may make intermediately inspecting data uneconomical is unfounded as could be demonstrated in the previous section. Thus, sequential testing—with adjusting alpha—allows for intermediately inspecting data without the shortcoming of an inflated Type I error probability and in many cases is a more economical way to detect an effect.

Limitations of APriot and group-sequential testing

While APriot and mathematical methods of group-sequential testing help the researcher to adjust alpha so that the Type I error probability does not inflate, there are some concerns with group-sequential testing which must be taken into account. The most important problem is that too many effects might be reported as statistically significant. As discussed in the introduction, if a

researcher adds more and more participants to an experiment, effects of very small size can be detected. While this is desirable when an effect of unknown size is to be found, there is a risk of finding effects which are so small that they lack relevance. The misleading expectation could arise that any effect could show statistical significance just by adding further blocks of participants. Thus, it is important for an experimenter to keep track of the effect size. At this point, one could argue that it would be better to select a minimum effect size and do an a-priori sample size analysis. The advantage of this method is that a clear criterion is available when to stop the experiment and report the result as "not significant". However, with group-sequential testing, the effect will be found with less participants if the effect size is actually larger than the previously defined minimum effect size. So, group-sequential testing with strictly watching the effect size and stopping the experiment if it gets to small can be a reasonable compromise.

At last, it must be mentioned that group-sequential testing and adjusting alpha is not the only possibility to detect an effect of unknown size. As mentioned in the introduction, defining a small minimum effect size and then doing an a-priori sample size analysis still is an attractive way. Further, replication studies are an important instrument to find out whether an effect really exists or has been found by chance. With repeated-measures designs, it is important to have multiple measures per subject in order to reduce within-subject variability. Many of these and similar techniques to enlarge the reliability of an experiment can be used as an alternative to or together with group-sequential testing.

Taken together, adjusting alpha with group-sequential testing does not solve the problem of false positive results as a whole, but it addresses one important reason for false positive results. Adjusting alpha with group sequential tests is an important instrument to make results reported in literature more reliable.

References

Abdi, H. (2007). Bonferroni and Sidak corrections for multiple comparisons. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage.

Diehl, J., & Arbinger, R. (1990). *Einführung in die Inferenzstatistik*. Eschborn near Frankfurt on the Main: Klotz.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191. doi:10.3758/BF03193146

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70. doi:10.2307/4615733

Ignaz, A., Lang, A.-G., & Buchner, A. (2013). The impact of practice on the adjustment of interaural cues in a lateralization task. *The Journal of the Acoustical Society of America, 134*(2), 901-904. doi:10.1121/1.4812861

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science, 23*(5), 524-532. doi:10.1177/0956797611430953

Lakens, D. (2014a). Data peeking without p-hacking. Retrieved from http://daniellakens.blogspot.nl/2014/06/data-peeking-without-p-hacking.html

Lakens, D. (2014b). Performing High-Powered Studies Efficiently with Sequential Analyses. *Social Science Research Network*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2333729 doi:10.2139/ssrn.2333729

Pahl, R. (2015). GroupSeq: A GUI-Based Program to Compute Probabilities Regarding Group Sequential Designs. Retrieved from https://cran.r-project.org/web/packages/GroupSeq/index.html

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials*. New York: Springer.

Reboussin, D. M., DeMets, D. L., Kim, K., & Lan, K. K. G. (2014). Lan-DeMets Method - Statistical Programs for Clinical Trials. Retrieved from https://www.biostat.wisc.edu/content/lan-demets-method-statistical-programs-clinical-trials

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological methods*. doi:10.1037/met0000061

Sherman, R. (2014). phack: An R Function for Examining the Effects of p-hacking. Retrieved from http://rynesherman.com/blog/phack-an-r-function-for-examining-the-effects-of-p-hacking/

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*, *22*(11), 1359-1366. doi:10.1177/0956797611417632

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine, 9*, 65-72. doi:10.1002/sim.4780090113

[1] Homogeneity of variances is a precondition for conducting an ANOVA. However, the ANOVA is considered to be relatively “robust” against a violation of this precondition (e.g., see Diehl & Arbinger, 1990, p. 214).

[2] By default, APriot starts as many threads as processor cores are available. However, in the “Preferences” menu, the user can specify the number of threads manually.

[3] Please note that the internal pilot study is not used in a classical way here. It is not used to compute the conditional power and adopt the overall sample size; rather, it is used to obtain the parameters needed for a simulation.

Appendix A

Program Availability

APriot has been developed to run with Windows 7, 8, and 10. There are 32-bits and 64-bits versions of the program. APriot is a noncommercial program. It can be downloaded for free at http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/apriot.html.

The program has been developed carefully and extensively tested. However, no warranty of any kind is given. Please report bugs to albert.lang@hhu.de.