Spreadsheets in Education (eJSiE)

Volume 7 | Issue 1 Article 1

1-14-2014

A Pedagogic Exploration of Researcher Degrees of Freedom

Christopher R. Fisher
Miami University, fisherc2@miamioh.edu

A Pedagogic Exploration of Researcher Degrees of Freedom

Abstract
In this article, I present a spreadsheet that demonstrates how researcher degrees of freedom (RDoF) increase type 1 errors in scientific research. RDoF refer to the flexibility in analyzing and reporting data. The overarching goal is to instill good research practices in students through awareness of the problems associated with RDoF and the mechanisms through which specific types of RDoF increase type 1 error rates. To accomplish this goal, the spreadsheet is organized into four modules. The first three modules use Monte Carlo simulations to demonstrate common examples of RDoF—dichotomization, optional stopping and multiple testing. The modules allow students to manipulate factors that control the type 1 error rate. The fourth module demonstrates how multiple types of RDoF combine in practice to produce high type 1 error rates. The article concludes with a set of pedagogic questions that instructors may use to teach core concepts associated with RDoF.

Keywords
Type 1 errors, Statistics, Researcher Degrees of Freedom, Null Hypothesis Significance Testing

Cover Page Footnote
I would like to thank Mary E. Frame for her helpful comments on a previous version of the manuscript.

Introduction

Due to the interdependent nature of scientific knowledge, a single false finding may have many downstream consequences, causing cumulative errors, confusion and misallocation of resources. For this reason, scientific progress requires accurate and complete dissemination of information. False findings come in two forms—non-descriptively termed type 1 and type 2 errors. A type 1 error occurs when a researcher concludes there is an effect when one does not exist. A type 2 error occurs when a researcher fails to find an effect that does exist. Type 1 errors are arguably more common, problematic, and easier to calculate. For these reasons, type 1 errors will be the primary focus of the present article. The infiltration of type 1 errors into the scientific literature is inevitable in disciplines that rely on the use of inferential statistics to generalize from samples to their corresponding populations. Biology, economics, medicine, political science and psychology are among the disciplines that use inferential statistics. In an ideal world, replication and complete reporting would minimize the number of type 1 errors in the long run. However, in practice, replication and complete reporting are not adequately incentivized at an institutional level, resulting in systemic distortions in scientific knowledge [1] [2]. For example, publication bias—the selective publication of statistically significant effects—greatly increases the proportion of type 1 errors in the literature [3]. Under a wide range of assumptions about pre-study odds, statistical power and publication bias, false findings in the literature may be higher than 50% [1].

Recently, there has been increased interest in the decisions made by individual researchers that produce type 1 errors. Simmons and colleagues demonstrated empirically the ease with which an impossible effect can be supported with statistical evidence [3]. In the experiment, participants reported their age before listening to one of two songs: "When I'm Sixty-Four" by The Beatles or "Kalimba" by Mr. Scruff. The results indicated that listening to "When I'm Sixty-Four" caused a reduction in age—an obviously impossible result. The type 1 error was due to what Simmons and colleagues termed researcher degrees of freedom (RDoF)—the flexibility with which data are analyzed and

selectively reported. For example, there are several reasonable procedures for excluding outliers from reaction time data, including nonlinear transformations, the elimination of the top 1% or 2% of reaction times, an absolute cut-off criterion or a combination of these procedures. Other decisions include optional stopping in data collection, the separate treatment or aggregation of similar dependent variables, the inclusion of covariates and the dichotomization of data to list only a few examples. Simmons and colleagues showed that RDoF quickly exert cumulative effects on the rate of type 1 errors. It is important to note that researchers are not necessarily acting fraudulently in these situations. Several reasonable methods of analysis may exist in a given situation and self-serving bias may unwittingly influence the researcher's justification to report a statistically significant result over one that is not. In some cases, researchers may be unaware of the detrimental effects of some types of RDoF, such as optional stopping and dichotomization. In other cases, journal space limitations preclude the full reporting of results.

In light of these issues, I developed a pedagogic spreadsheet that demonstrates how RDoF increase the rate of type 1 errors. The spreadsheet is organized into four modules. The first three modules provide a detailed treatment of three common examples of RDoF: dichotomization, optional stopping, and multiple testing. In each of these three modules, students can systematically explore the factors that increase type 1 errors. The fourth module allows students to explore how RDoF combine in typical research situations to produce cumulative effects. The remainder of the present article is organized as follows. First, a brief overview of null hypothesis significance testing is provided to acquaint the reader with concepts necessary to understand RDoF. Readers who are already familiar with null hypothesis significance testing can skip that section without loss of understanding. In the sections that follow, the modules for dichotomization, optional stopping and multiple testing are described separately in detail. The section for the final module describes how multiple sources of RDoF combine to exert cumulative effects on type 1 errors. Next, the implementation of the spreadsheet is described in a

separate section for clarity of presentation. Lastly, the article concludes with several pedagogic questions instructors may use in the classroom.

Null Hypothesis Significance Testing

Although its use is controversial, null hypothesis significance testing (NHST) is currently the prevailing method of statistical analysis is many fields of science [4,5]. In NHST, a researcher formulates two complementary hypotheses. The null hypothesis states there is no effect between two groups whereas the alternative hypothesis states there is an effect. Within the NHST framework, the alternative hypothesis is supported indirectly via contradiction. The researcher assumes the null hypothesis is true and if the results are sufficiently at odds with this assumption, the null hypothesis is rejected. When the null hypothesis is rejected, there is some chance that the result is due to sampling error. In other words, there is some chance that researcher will conclude there is an effect when one does not exist—a type 1 error. The rate of type 1 errors is controlled by a decision criterion called alpha. By convention, alpha is set to .05. More specifically, an observed test statistic is compared to a theoretical sampling distribution for the null hypothesis. A p-value is computed as the probability of obtaining a test statistic at least as extreme as the one observed (conditional on the null hypothesis being true). When the p-value is less than or equal to alpha, the null hypothesis is rejected, indicating either an extreme result was observed (i.e. a type 1 error) or the null hypothesis is false. Under ideal conditions, alpha sets an upper limit on the probability of a type 1 error. As will be discussed below, RDoF can cause this limit to be exceeded, sometimes by a large margin.

Dichotomization

One example of RDoF occurs when a researcher decides whether or not to dichotomize a continuous variable. A continuous variable, such as age, might be dichotomized as young and old, according to a median split. Although dichotomizing is rarely advisable, it remains in practice to some degree [6]. Some researchers reason that dichotomizing averages out noise inherent in the continuous scale. In actuality, dichotomizing decreases the effect size on average and introduces considerable variability. Although

the dichotomizing decreases the effect size on average, it will produce a larger effect size than the continuous variable in some cases. As a result, the chance of a type 1 error will increase when a researcher chooses between both analyses. The tab titled "Dichotomization" demonstrates how dichotomizing increases the chance of a type 1 error. In the following example, the True Correlation in cell B1 was set to .50 and the macro-enabled button was clicked to initialize a Monte Carlo simulation. On each iteration, 25 pairs of X and Y values were selected from a normal distribution. Each simulated dataset had a fixed correlation of .50. By fixing the sample correlation, it is possible to separate variability due to dichotomizing from variability due to sampling error in estimating the correlation. Next, the X values were dichotomized using a median split and the resulting correlation was recorded. This process was repeated 10,000 times to produce a smooth histogram for demonstration. However, 1000-2000 iterations are sufficient for most purposes.

As shown in Figure 1, the resulting distribution is highly variable and shifted to the left, even though the continuous data were correlated at exactly .50. Although the mean is .38, approximately 26% of the distribution is higher than the fixed correlation of .50. This additional source of variability is responsible for the increase in type 1 errors. The effects of RDoF can be observed by setting the correlation to zero and comparing the percentage of p-values ≤ .05 for continuous only versus continuous or dichotomous (cells B5 and B6). As expected, the rate is .05 when only the continuous variable is tested. However, the type 1 error rate increases to .08 when there is a choice between continuous or dichotomized data. It is worth noting that the percentage of p-values ≤ .05 in each module represents statistical power when the correlation does not equal zero. RDoF increase power but at the expense of increasing type 1 errors.

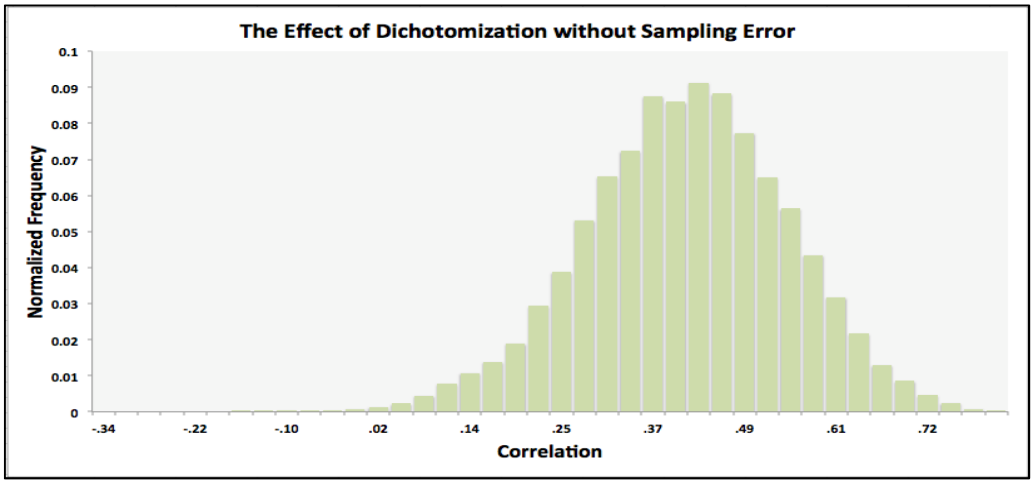


Figure 1: Variability produced by dichotomizing X values when the fixed sample correlations r = .50.

Optional Stopping

Another decision researchers must make is when to stop collecting data. Collecting data can be costly and time consuming. However, collecting more data will increase the power of the experiment—the ability to detect a true effect. Researchers must balance these competing goals. In NHST, theoretical sampling distributions assume a fixed sample size [4]. For this reason, it is necessary to specify the sample size a priori and continue the experiment until the predetermined sample size is achieved. However, some researchers may engage in a different sampling plan called optional stopping. As an example, consider a researcher who plans to collect a maximum sample of 30 participants but peaks at the data midway through. If an effect is detected with only 15 participants, the researcher terminates data collection. However, if no effect is detected, data remaining data are collected. Researchers may reason that collecting additional data once an effect is detected is wasteful. As appealing as this intuition may be, the reality is that optional stopping increases the chance of a committing a type 1 error. When the null hypothesis is true, the p-value will converge on .50 in the limit because it is uniformly distributed [4]. However, the p-value is volatile initially, fluctuating up and down until eventually converging on .50. Thus, optional stopping allows the researcher to capitalize on chance fluctuations.

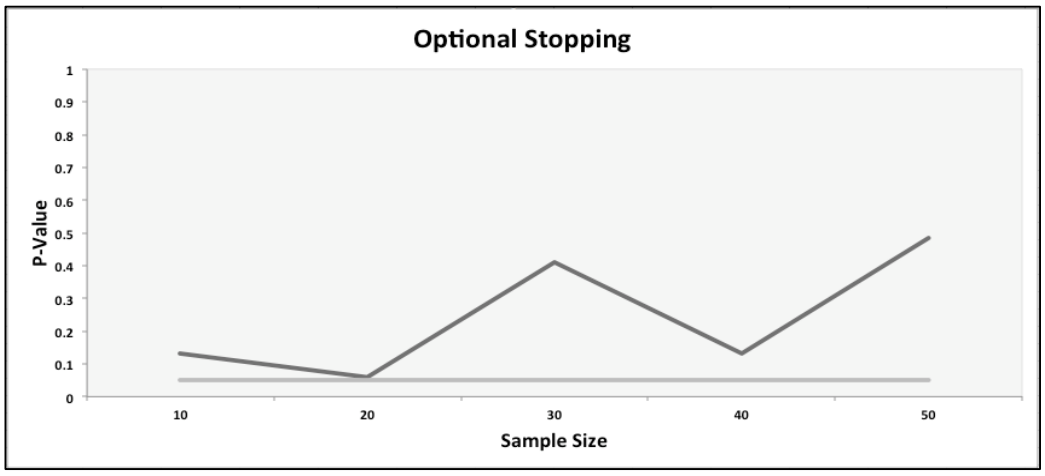


Figure 2. An illustration of optional stopping. A test is performed after every 10 participants and stops once p-value ≤ .05 or a maximum of 50 participants have been collected. All testing points are shown to illustrate how increased opportunities increase the type 1 error rate.

In the module "Optional Stopping", the p-values from a single simulation are plotted at each testing point to illustrate how optional stopping increases type 1 errors. In the following example, the maximum sample size was set to 50 (cell B2) and a test was performed after every 10 participants (cell B3). It is important to note that this number (cell B3) must be a factor of the maximum sample size in order for the macro to work properly. In Figure 2, the p-value satisfies the decision criterion of p-value ≤ .05 once 20 data points were collected. According to the optional stopping rule described above, the researcher would terminate data collection at this point. The remaining testing points are plotted to illustrate an important point: had the criterion not been satisfied, there would have been additional opportunities to commit a type 1 error. The type 1 error rate is .15 for this particular example (see cell B7), well above the nominal type 1 error rate set by alpha.

Multiple Tests

Another source of RDoF is determining which comparisons should be made or whether to combine similar dependent variables. Suppose a researcher conducts an experiment with four groups. Which groups

should be compared? There are six pairwise comparisons in this example. Even more comparisons are possible if some groups can be combined. The type 1 error rate for this "family" of related tests is called family-wise type 1 error [7]. Assuming independence, family-wise type 1 error is defined as the probability of at least one type 1 error within a family of tests:

$$P(familywise\ error) = 1 - (1 - \alpha)^n \quad (1)$$

where n is the number of tests within a family. When independence does not apply, the formula is much more complex. For this reason, it is approximated through simulation in the spreadsheet. Consider a similar example in which a researcher must decide whether to treat two similar variables separately or aggregate them. Unlike the previous example, this provides three possible ways to test the *same* hypothesis rather than a family of related hypotheses. Although the type 1 error rate will increase by testing the hypothesis three ways, the inter-correlations between the dependent variables will mitigate the type 1 error to some degree.

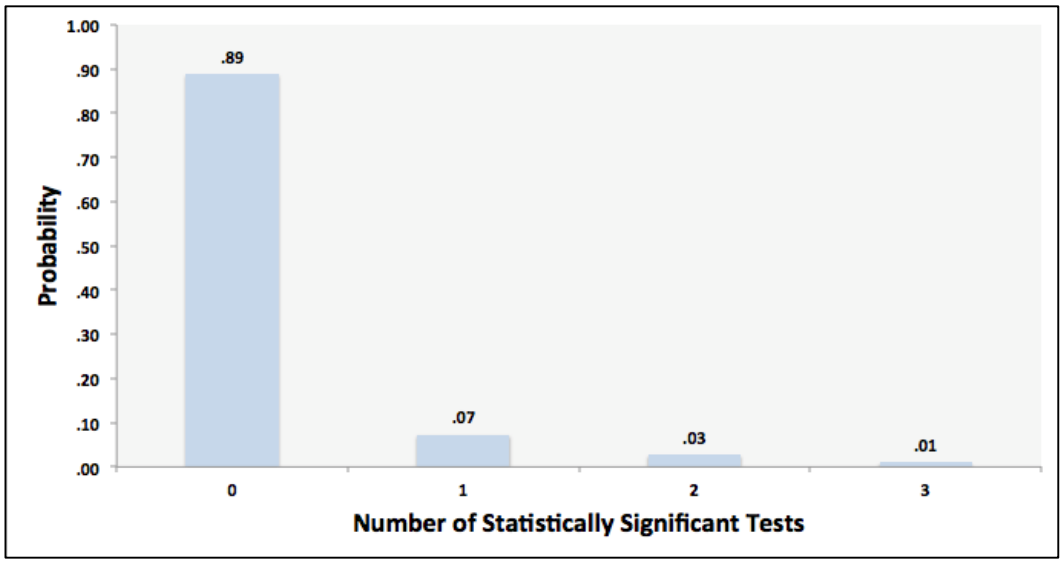
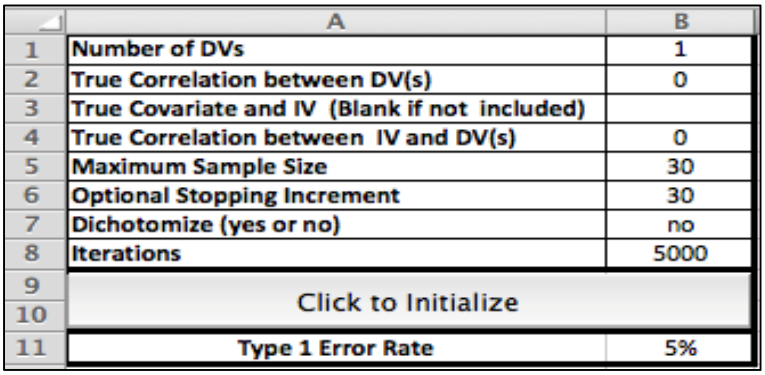


Figure 3. A histogram of the number of type 1 errors in a set of three inter-correlated dependent variables.

Continuing with the last example, a total of three tests entered into cell B3. In cell B1, the correlation was set to 0 and the correlation between the dependent variables is set to .70 in cell B3. A high correlation of .70 is reasonable in this example because similar DVs will be partially redundant. Because the covariance matrix increases quickly as more dependent variables are added, the correlations are assumed to be equal for simplicity. Thus, the correlation in B3 can be treated as an average inter-correlation. Clicking the macro-enabled button will initialize the simulation and record the results in a column graph. Figure 3 plots the probability of obtaining a specific number of statistically significant tests (p-value ≤ .05). The probability of committing a type 1 error is displayed in cell B8 as .11. This value can be inferred from the graph two ways. First, the probabilities for 1 through 3 can be summed: $P(x \geq 1) = .07 + .03 + .01 = .11$. Alternatively, it can be calculated as $1-P(x = 0) = 1-.89 = .11$. The column graph makes it clear that increasing the number of tests increases the chance of a type 1 error in most cases because there are more opportunities for a type 1 error to arise.

Combined Effects

The preceding modules focused on specific types of RDoF. The modules were designed to provide a conceptual understanding of the mechanisms through which type 1 errors are produced. In the fourth module, RDoF are combined so that their cumulative effects can be observed under various conditions. By combining the effects of multiple RDoF, it is possible to simulate typical research situations. As shown in Figure 4, cell B3 controls the number of tested dependent variables in the simulation. B2 specifies the inter-correlation among the dependent variables. As before, the inter-correlations can be thought of as an average correlation between the dependent variables.



	A	B
1	Number of DVs	1
2	True Correlation between DV(s)	0
3	True Covariate and IV (Blank if not included)	
4	True Correlation between IV and DV(s)	0
5	Maximum Sample Size	30
6	Optional Stopping Increment	30
7	Dichotomize (yes or no)	no
8	Iterations	5000
9		
10	Click to Initialize	
11	Type 1 Error Rate	5%

Figure 4. A screenshot of the Combined Effects module without any RDoF. As expected, the type 1 error rate is 5%.

A covariate may be entered into cell B3. As its name implies, a covariate is a variable that co-varies with the dependent variable but cannot be controlled experimentally for practical reasons. A researcher may use regression or a partial correlation to statistically control for the effects of the covariate. The mechanism through which the covariate can produce type 1 errors is not easy to illustrate in a spreadsheet and thus did not receive a separate module. However, because it is a common RDoF, it is included in the Combined Effects module. To understand how a covariate can increase type 1 errors, it is important to note that a correlation can be conceptualized as the covariance between the independent variable and dependent variable with respect to the product of their standard deviations:

$$r = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \qquad (2)$$

The effect of the covariate is partialled out of X and Y using the following equation [8]:

$$r_{xy.c} = \frac{r_{xy} - r_{xc} r_{yc}}{\sqrt{(1 - r_{xc}^2)(1 - r_{yc}^2)}} \qquad (3)$$

where the subscript c denotes the covariate. When r_{xc} or r_{yc} is negative, the numerator will increase and the denominator will decrease, thereby amplifying the correlation. Under these conditions, the covariate can produce a type 1 error. When B3 is blank, the covariate will not be used in the simulation. When a value is entered into B3, the effects of the covariate are removed from each of the dependent variables. For simplicity, the simulations assume the covariate is not correlated with the independent variable (i.e. the X variable). It is important to note that the macro requires a positive definite matrix in order to sample from a multivariate normal distribution. Certain configurations involving negative correlations may not satisfy this requirement and will result in a runtime 5 error. However, the correlations between dependent measures are generally positive. Thus, the restriction of a positive definite matrix will have minimal impact on the generalizability of the simulations to practical situations.

The correlation between the independent and dependent variables can be specified in cell B4. Again, when B4 is zero the simulation computes the type 1 error rate. For optional stopping, the maximum sample size is specified in cell B5 and the test increment (which must be a factor of B5) is specified in cell B6. To exclude optional stopping, simply set B5 = B6. Finally, the independent variable can be dichotomized when computing the correlation between the independent variable and each of the dependent variables. To dichotomize, simply type "yes" into cell B7. As expected, the type 1 error rate in Figure 4 is 5% because only one test was performed. What happens in a typical research situation in which there are several RDoF? A typical situation is exemplified in Figure 5. For this example, assume the researcher has two dependent measures of the same variable. These variables can be treated separately or aggregated, resulting in a total of 3 dependent variables. Due to their similarity, the inter-correlations will be high. In Figure 5, the inter-correlations are set to .75. A covariate is included and assumed not to correlate with the dependent variables. Performing analyses can be costly in terms of time for some research. Bearing this in mind, the optional stopping rule allows the researcher to peak at the data once midway through the experiment, with the option to proceed to a maximum sample size of 30. Thus, cell B5 is set

to 30 and B6 is set to 15. At each of the two possible testing points, the researcher will perform the test with the independent as a continuous and dichotomous variable. Collectively, there are 18 possible tests. There are three dependent variables. For each dependent variable, the independent variable is treated as continuous or dichotomous. For each dependent variable, separate tests are performed with the inclusion of a covariate. Each of these 3+3+3 = 9 tests are performed once with 15 subjects and again with 30 subjects, for a total of 18 tests. The cumulative effect of RDoF for this example produces a type 1 error rate of 33%.

	A	B
1	Number of DVs	3
2	True Correlation between DV(s)	0.75
3	True Covariate and IV (Blank if not included)	0
4	True Correlation between IV and DV(s)	0
5	Maximum Sample Size	30
6	Optional Stopping Increment	15
7	Dichotomize (yes or no)	yes
8	Iterations	5000
9	Click to Initialize	
10		
11	Type 1 Error Rate	33%

Figure 5. The cumulative effects of RDoF in a typical research situation. The type 1 error rate is much higher than .05 set by the criterion alpha.

Implementation

Dichotomization

In this section, the macros for each module are described in detail for the interested reader. The simulation in the tab "Dichotomization" produces three outputs: (1) the type 1 error rate for continuous correlations in cell B5, (2), the type 1 error rate for continuous and dichotomous correlations in cell B6 and (3) the fixed dichotomized correlations displayed in the histogram. Each of the correlations and respective p-values are computed within a for loop that repeats until it reaches a maximum iteration specified in cell B3. Upon each iteration, X and Y values are generated from a bivariate normal distribution using code for Cholesky decomposition described in [9]. First, WorksheetFunction.NormSInv(Rnd)

is used to generate independent X and Y values. Second, a new variable Y′ is created using a special bivariate case of Cholesky decomposition. In particular Y′ is computed as $X*r + Y(1-r^2)^{.50}$, where r is the desired correlation from cell B1. The resulting X and Y′ values represent data sample from a bivariate normal distribution with a correlation specified in cell B1. The correlation between continuous variables X and Y′ is computed using WorksheetFunction.Correl. The corresponding p-value is computed in two steps. The first step takes advantage of the fact that a t test (for which excel can compute a p-value) is a special case of the correlation, r. r is transformed to a t-statistic according to the following formula: $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$, with df = N-2 [7]. The p-value is computed with WorksheetFunction.TDist and finally recorded to column AH. The correlation in which X is dichotomized is computed in a similar manner, except for the inclusion of an additional step. X is dichotomized according to a median split with WorksheetFunction.Median. X values greater then or equal to the median are assigned a value of 1 whereas X values less than the median are assigned a value of 0. The correlation and p-value is computed as previously described. To find the type 1 error rate when the researcher chooses between continuous and dichotomous X values, the minimum p-value is selected with WorksheetFunction.Min and recorded to column AI. Fixed correlations are generated with one additional step. To remove any incidental correlation between X and Y, Y is regressed on X, using WorksheetFunction.Slope and WorksheetFunction.Intercept. Second, the residual of Y is computed as the difference between the predicted and observed Y values, denoted $Y_{residual} = Y_{predicted} - Y$ The resulting $Y_{residual}$ is uncorrelated with X. Third, a new variable Y_{fixed}' is computed using the simple bivariate case of Cholesky decomposition to produce an exact correlation between X and $Y_{residual}$ (free of sampling error): $Y_{fixed}' = X*r + Y_{residual}(1-r^2)^{.50}$. Finally, the X variable is dichotomized according to a median split and the correlation between the dichotomized X variable and Y_{fixed}' is recorded in column AJ. The Frequency function is then used to produce a histogram to show the effect of dichotomization without sampling error in r.

Optional Stopping

The macro for Optional Stopping follows the same basic procedure described in the subsection Dichotomization. Two outputs are generated: the estimated type 1 error rate based on multiple iterations and an illustrative line graph based on a single iteration. Both are based on a very similar simulation process. Beginning with the estimation of the type 1 error rate, data of maximum size N_{max} (cell B2) are sampled from a bivariate normal distribution using Cholesky decomposition. The data are checked in increments, I, specified in cell B3. For example, if the maximum sample size, $N_{max} = 50$, and the increment is I = 10, the correlation and p-value will be calculated when N = 10, N = 20 and so on until all N_{max} data points are included. For each test, a counter variable counts whether the current p-value ≤ .05 or not. A type 1 error is recorded if the counter variable is greater than 0. To accomplish this in the macro, a for loop repeats N_{max} /I iterations. For example, on the first iteration, the data points 1-10 are referenced in a nested for loop and the correlation and p-value are computed. On the second iteration, data points 1-20 are referenced in a nested for loop as before. A type 1 error is recorded. if at least one of the optional stopping points (N = 10, N = 20 etc.) yields a p-value ≤ .05 This process continues until all 50 (N_{max}) data points are included in the calculation of the correlation and p-value. This process is repeated as specified in cell B4. The results are recorded in column AJ as 1's (type 1 errors) and 0's and averaged in cell B7 to approximate the type 1 error rate associated with optional stopping. This process is repeated once to produce the illustrative line graph. The p-values and corresponding sample sizes are recorded to column AG and AH, respectively. A line graph is displayed based on these data.

Multiple Tests

As before, the macro for Multiple Tests follows a similar procedure. A multivariate normal distribution is generated with one independent variable and D dependent variables defined in B3. The correlation between the independent variable and dependent variable(s) is defined in B1 and the inter-correlations between the dependent variables are defined in B4. A covariance matrix based on these inputs is submitted to Cholesky decomposition to sample from the multivariate normal distribution, with

a sample size specified in cell B2. Next, the correlation and p-values are computed for each independent and dependent variable. The number of type 1 errors is recorded during each iteration. The proportions of 0-D type 1 errors are recorded in columns AG and AH and displayed in a bar graph. The type 1 error rate is recorded in C8 as 1 – the probability that no type 1 errors occur.

Combined Effects

The Combined Effects macro combines the other modules in a straightforward manner. Because the modules are duplicated in the Combined Effects module, some of the details are omitted and the reader is deferred to the subsections above. The basic structure of the macro is organized as follows. Upon each iteration, multivariate normally distributed data are generated and processed by up to 3 nested for loops, depending on user inputs: (1) a for loop for computing correlations between continuous variables, (2) a for loop for computing correlations using a dichotomized independent variable and (3) a for loop for computing a partial correlation, which removes the effect of the covariate. As described in the previous modules, each of the three for loops is capable of computing correlations for multiple dependent variables and implementing optional stopping. This structure allows the macro to enumerate the appropriate number of tests. The number of tests performed can be formulated as follows:

$$N_T = TO(1 + C + D) \tag{4}$$

where T is the number of dependent variables; O is the number of optional stopping points, C is an indicator that equals 1 if a covariate is included and 0 otherwise; and D is an indicator variable that equals 1 if the independent variable is dichotomized and 0 otherwise.

The new addition to the Combined Effects macro is the computation of a partial correlation. The partial correlation removes the effect of a covariate according to Equation 3. Aside from computing the partial correlation, the

for loop for the covariate proceeds in the same manner as the previous procedures.

Upon each iteration, a counter variable sums the number of type 1 errors across all the relevant for loops. A 1 is recorded in column AG if the counter variable is greater than 0. Otherwise, a 0 is recorded. The type 1 error rate is computed as the average of column AG and recorded in cell B11.

Problems

In this section, 4 problems are provided for instructors to use in the classroom. Suggested answers are included to verify comprehension.

Problem 1

In the tab "Dichotomization", set the True Correlation to 0 and initialize the simulation. In the simulation, data were sampled from a bivariate normal distribution with a correlation of exactly 0. One variable in each data set was dichotomized and the resulting correlations were recorded in the histogram. Explain how the variability due to dichotomization produces increased type 1 errors in cell B8.

Suggested answer: The histogram illustrates that dichotomizing produces additional variability in the correlation. This additional source of variability allows the correlation to sometimes exceed the correlation based on the continuous data. The type 1 error rate increases from 5% to 8% when the results of both tests are selectively reported.

Problem 2

Some researchers reason that optional stopping is an efficient research strategy because a large effect can be detected with fewer data points than a small effect. Thus, if an effect is detected with fewer data points, it is wasteful to continue data collection. Repeat the simulation in the tab "Optional Stopping" to find patterns in the p-value line that explain how optional stopping can increase type 1 errors.

Suggested answer: The p-value fluctuates up and down on each simulation. If it the p-value is greater than .05, it may fall below .05 on subsequent tests. Thus, optional stopping provides more opportunities to capitalize on chance.

Problem 3

In the multiple tests tab, set the True Correlation to 0, the number of DVs to 1 and the correlation between DVs to 0. What is the type 1 error rate and how does it compare to alpha (i.e. .05)? What happens when more DVs are added? What happens when the DVs have a high correlation? Why?

Suggested answer: When the True Correlation is 0 and only one test is performed, the type 1 error rate is .05, which is equal to alpha. As more tests are performed, the type 1 error rate increases because each test provides another opportunity for an error. When the correlation between the DVs is high, the type 1 error is partially mitigated because the outcomes tend to produce similar results. In other words, the DVs are redundant and reduce the opportunities for increased type 1 errors.

Problem 4

The tab titled "Combined" allows you to combine the multiple RDoF. Construct a few scenarios that you think are typical in research by varying the RDoF. How much greater are the type 1 error rates in those situations compared to alpha? Try to construct a scenario in which the type 1 error rate is high, e.g. greater than 50%. Do you think the scenario is realistic or typical?

Suggestion answer: Answers will vary.

Conclusions

The interactive spreadsheet demonstrates how RDoF increase type 1 errors in scientific research. Common research decisions—when combined with flexibility in reporting—increase the chance of a type 1 error. Some of these decisions include the aggregation or separate treatment of similar

dependent variables, multiple testing, optional stopping in data collection, the use of covariates and whether to dichotomize a continuous variable. Under reasonable assumptions, just a few RDoF can increase the type 1 error rate to .15 - .30, well above the nominal rate of .05 set by the alpha criterion. Collectively, the pedagogic simulations give credence to the adage "If you torture your data enough, it will confess to anything". In conjunction with the pedagogic problems, the interactive spreadsheet may be a valuable tool for increasing awareness of RDoF and how they increase type 1 errors. In addition, the spreadsheet may help instill good research habits in aspiring researchers.

References

1. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.
2. Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528-530.
3. 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.
4. Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin & Review*, 14, 779-804.
5. Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301.
6. MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*(1), 19.
7. Howell, D. C. (2011). *Statistical methods for psychology*. Cengage Learning.
8. Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC.

9. Moore, T. (2001). Generating Multivariate Normal Pseudo Random Data. *Teaching Statistics*, *23*(1), 8-10.