Random Numbers Demonstrate the Frequency of Type I Errors: Three Spreadsheets for Class Instruction

Sean Duffy
Rutgers University - Camden

Journal of Statistics Education Volume 18, Number 2 (2010),
www.amstat.org/publications/jse/v18n2/duffy.pdf

Copyright © 2010 by Sean Duffy all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the author and advance notification of the editor.

Key Words: Power, Simulation, Correlation, P-value, Hypothesis testing.

Abstract

This paper describes three spreadsheet exercises demonstrating the nature and frequency of type I errors using random number generation. The exercises are designed specifically to address issues related to testing multiple relations using correlation (Demonstration I), t tests varying in sample size (Demonstration II) and multiple comparisons using analysis of variance (Demonstration III). These demonstrations highlight the purpose and application of hypothesis testing and teach students the dangers of data dredging and a posteriori hypothesis generation.

1. Introduction

Undergraduates and graduate students across many fields of the social, biological and physical sciences are taught null hypothesis significance testing (NHST) as part of most introductory statistics courses. A hybrid of procedures suggested by Fisher (1926; 1956) and Neyman and E. Pearson (1928/1967), researchers identify a null hypothesis ($H_{Ø}$) of no relation or difference between one variable (i.e., the independent variable that a researcher manipulates) and another variable (i.e., the dependent variable that the researcher measures as a function of a change in the independent variable). This null hypothesis is tested against an alternative hypothesis (H_A) that a statistically significant relation or difference is observed between the dependent and independent variables using an inferential statistical test, such as a t test or ANOVA. A relation or difference between variables is considered statistically significant if there is strong evidence that the observed relation or difference is unlikely to be due to chance. In the NHST most commonly practiced in applied fields, researchers reject the null hypothesis when the probability of incorrectly rejecting a true null hypothesis falls beneath an established criterion level alpha (α).

Alpha represents the maximum level that the researcher will accept for incorrectly rejecting the null hypothesis when the null is true, which by convention (in some fields, and by some researchers) is set at five percent.[1] Once the researcher finds evidence for a significant relation or difference between an independent and dependent variable, the null hypothesis can be rejected, and the alternative hypothesis is assumed to be true.

NHST is not without its critics (e.g., Oakes, 1986; Falk & Greenbaum, 1995; Gigerenzer, 2004; Loftus, 1991; Killeen, 2005) and controversies over the practice and application of inferential statistics exist in many disciplines. Rather than engage in the debate over the utility or futility of NHST, this paper describes three exercises and associated spreadsheets that I have found valuable in teaching NHST, and specifically about the problem of rejecting the null hypothesis when the null is true. Through the use of the three simple spreadsheet demonstrations described below, students can learn to avoid common mistakes that arise in the design of experiments and the analysis of data.

2. Inferential statistics and the two types of errors

Around the time students learn inferential statistics, many teachers and books elaborate upon four possible outcomes in a hypothesis testing situation. Figure 1 depicts my own version of a common figure that can be found in various forms in different textbooks (e.g., Davis & Smith, 2004, p. 208; Thorn & Giesen, 2002, p. 194; Steinberg, 2008, p. 149; Vernoy & Kyle, 2001, p. 270; Wilson, 2004, p. 109).

Figure 1: My figure demonstrating the consequences of type I and II errors.

Hypothesis		Reality	
		Null Hypothesis is True	Alternative Hypothesis is True
Findings	Null Hypothesis is True	😐 Accurate: But unfortunate! (Back to the drawing board!) 1 - α	☹ Type II Error: Bad! (Serious, but not deadly!) β
	Alternative Hypothesis is True	☠ Type I Error: Very Bad! (Abandon all hope, ye who enter here!) α	☺ Accurate: Publish and prosper! (Have a great career!) 1 - β

[1] It is important to emphasize that alpha is a criterion level set by researchers, and should only be equated with the probability of making a type I error when the null hypothesis is true. When the null hypothesis is not true, there is no chance of a Type I error. Of course, the reason that we conduct a hypothesis test is because we do not know the truth of the null hypothesis."

The top axis represents the actual state of the world (i.e., the "correct" decision if we could accurately measure all members of a population), and the side axis represents the conclusion based on observed data (the sample randomly selected from the population). Assuming that the alternative hypothesis is more theoretically interesting than the null hypothesis, the best outcome is to find evidence to reject the null hypothesis when the alternative hypothesis is actually true. This results in an interesting and publishable result. Slightly unfortunate is finding no evidence to reject the null hypothesis when the null is actually true. This outcome is not bad either because it is accurate, yet it may be unfortunate, as it implies returning to the drawing board and designing a new study.

Then there are the two errors. One error is failing to find evidence to reject the null hypothesis when the alternative hypothesis is true. Known as a type II error, such errors are bad for science in that an actual result never sees the light of day. The other type of error is generally considered more detrimental: finding evidence to reject the null hypothesis when the null is true, known as a type I error. Such errors are bad for science because they potentially allow a false result to enter the literature. Other researchers may unsuccessfully try to replicate the false finding, wasting crucial resources and time. Once a scholar develops a reputation for non-replication, a career is tarnished, and the once-earnest scientist may end up unemployed and destitute in the poorhouse alongside characters from a Charles Dickens novel, themselves having made similar type I errors. This is why my version of the error chart (Figure 1) uses a poison symbol to represent type I errors.

Type I errors arise because chance perturbations in the sampling process sometimes accumulate in a single sample to provide false evidence for a relation or difference where none actually exists in the population. However, psychologists have long understood that people experience difficulty understanding chance or random processes such as those involved in inferential statistics (e.g., Falk & Konold, 1997; Nickerson, 2002; Mlodinow, 2008). The inability to understand chance affects various psychological processes, producing cognitive biases such as in the perception of illusory correlation in social and non-social cognitive processes (Chapman, 1967; Hamilton & Gifford, 1976; Tversky & Kahneman, 1973). The mind seems to crave certainty, and the fact that random processes are part of hypothesis testing may be one of the reasons students experience difficulty comprehending the subject.

3. Researchers (Mis)understanding the frequency of type I errors

Students learn that the probability of making a type I error is set by the researcher or by an established convention set by the field. Psychologists, for instance, generally set their acceptable level of risk for making a type I error at five percent (.05). If researchers set their level of comfort over making a type I error (alpha) to the conventional level of .05, they will have a 5 percent chance of rejecting the null hypothesis when the null is true every time they conduct an analysis testing a true null hypothesis. In my own courses, I teach students about different types of error rates (family-wise, experiment-wise), and that in a study testing k independent comparisons or relations, the experiment-wise error rate can be calculated as $1 - (1 - \alpha_{\text{per comparison}})^k$. And my students typically nod in agreement, hand in homework in which they correctly calculate the different types of error rates since the formula are so simple, and I assume they understood the issue and move on from there.

In a recent graduate statistics course, I also assigned a hands-on project in which students designed and carried out their own research study. Prior literature demonstrates that carrying out one's own unique investigation is often an effective strategy for learning in statistics courses (e.g., Martinez-Dawson, 2003; Vaughan, 2003) and students develop important skills about carrying out an entire project from start to finish. To my consternation and chagrin, a considerable proportion of the students handed in project proposals similar to the following: "I propose to use a questionnaire to collect data on 7 personality factors, 7 demographic variables, and 6 Likert scale survey questions." In the proposed analysis section, the student wrote, "I plan to use SPSS to perform bivariate correlation coefficients on the variables and report which ones are significant." And at that point, I started pulling out my hair, because apparently a considerable number of students failed to understand the fundamental point about the frequency of type 1 errors across multiple comparisons, such as those involved in a 20 × 20 correlation matrix.

Frustrated, I returned to my class and revisited the issue of hypothesis testing. I explained once again the issue of type I errors, multiple comparisons, and after the fact (*a posteriori*) hypothesis generation. Because Gigerenzer and Hoffrage (1995) demonstrate that probabilities expressed as natural frequencies (1 in 20) are more comprehendible than those expressed as relative frequencies (.05), I altered the message a bit. I explained that setting alpha at .05 means that even if the null hypothesis is true and there is no correlation between any of the variables that 1 in 20 tests are likely to return type I errors. In conducting a 20 X 20 correlation matrix, there are 195 separate hypotheses tested (discounting the symmetry along the diagonal), and 195/20 ≈ 10, so that they could expect at least 10 type I errors in such a study design if there is no correlation between any of the variables. Their reactions…

> Student 1: *"Well, that may be true, but I will only interpret those that are significant and that also make sense. I'll just ignore the ones for which I can't come up with a good explanation."*
> Student 2: *"But if I do things the way you suggest, I would never get anything published, because how do I know before the fact what relations are supposed to be significant?"*
> Student 3: *"Isn't this what everyone does?"*
> Student 4: *"This is what Professor X (tenured member of the department) told me to do."*

After pulling out what was left of my hair, I decided that the students needed an actual demonstration to learn the nature and frequency of *real* type I errors if one were to follow such dubious research practices.

3.1. Demonstration 1: Data dredging and correlations using (secret) random data

For the next class, I developed an Excel spreadsheet consisting of a "dataset" of 30 "participants." I included 20 "variables" such as income (a demographic variable), the often-used "Big Five" personality traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism), data from a hypothetical "Life Satisfaction Scale" (Satisfaction with Marriage, Job, Health, Family, Friends, Sexuality), and a measure of "Emotional

Experience" (Calm, Anger, Happiness, Sadness, Fear, Disgust, Loneliness, and Annoyance).[2] I explained, "In this dataset, income is coded in ten thousands, the satisfaction and personality ratings are on a Likert scale ranging from 0 (does not describe me at all) to 10 (describes me very well), and the emotional experience scale is on a Likert scale ranging from 0 (feel this emotion infrequently) to 10 (feel this emotion frequently). (This spreadsheet is available as an Excel file titled "TypeIErrorExercise.xls" which may be obtained by following the link here or at the end of this article).

What I did not tell them was that the participants and data were merely the =rand() function in excel, with the decimal points neatly cleaned away by formatting the cells to show only integers, and the orders of magnitude altered by multiplying the random decimals produced by the rand function by 10 or 100. So although the data appeared to be actual data, the dataset consisted only of a set of 20 lists of random numbers. On a second worksheet, (correlation) I provided the resulting 20 × 20 correlation matrix, as well as the resulting p values associated with the correlations (based on the t test formula for determining the significance of a correlation coefficient). I set up the conditional formatting of the spreadsheet so that p values significant at the .01 level would show up red, and those significant at the .05 level would show up blue.

Because the excel sheet is programmed so that typing anything into any cell re-randomizes the data, for the following demonstration I avoided changing the value of any cells.

In class, I showed the students the Excel sheet on a projector. That randomization contained 5 correlations significant at the .01 level and 10 significant at the .05 level. I tell the students that we would conduct a class exercise in interpreting the results of a correlation matrix. After explaining the variables, I asked them to interpret the significant relations based on the results. I pointed to student 1.

> Student 1: *"It looks like people who are extraverted are likely to have higher incomes, as the correlation is .54. That makes sense, because outward focused people probably are better at getting better jobs than introverted people, who probably lack people skills necessary to secure a position."*
> Student 2: *"Conscientious people are more likely to be satisfied by their marriage, at .63. That's logical, because conscientious people are probably more likely to be sensitive to the needs of their spouses than non-conscientious people, resulting in higher marital satisfaction."*
> Student 3: *"People experiencing the emotion of annoyance right now are positively correlated at .34 with being neurotic, that makes sense, because neurotic people are probably worried, and in worrying all the time are annoyed all the time as well."*
> Student 4: *"Experiencing fear is negatively related with experiencing calm, which you'd predict, since if you're afraid, you're not exactly calm."*

[2] The courses in which I have used this demonstration were psychology classes. In a statistics course for other disciplines, the spreadsheet could be easily altered to address specific issues addressed by that discipline. For instance, for economists, the numbers could represent economic indicators of a sample of stocks, or for biologists, various biometric variables about a sample of animals. Also, the number of fake participants can be increased or decreased at the instructor's discretion.

We went through all the significant correlations. Of course, *a posteriori*, it is possible to come up with a hypothesis that explains any relation, as twisted as the prediction might be. For instance, one of the students 'hypothesized' that the strong positive relation between the personality trait of conscientiousness and the emotional experience of anger (r = .78) was readily explained away as resulting from the fact that conscientious people might often feel anger at having to be nice all the time, so that outwardly they might be conscientious, but inwardly they are furious at having to be nice all the time. Of course that makes sense!

To bring the demonstration to a close, after we explained away all the significant correlations, I go back to the original data and tell them, "I have a confession to make. Although I told you this was actual data, I wasn't completely honest. Each of these data points is simply a random number." I show them that the original data sheet consists of only =rand() functions. I further demonstrate this by hitting the delete key in an empty cell, which changes the data and results in a new set of spurious correlations. I have them try and interpret the new correlations again, which are just as "meaningful" as the ones they had just finished interpreting. Their reactions were classic!

> Student 1: *"No! You're fooling us."*
> Student 2: *"Impossible! But why are all these so significant if it is just random???"*
> Student 3: *"No way! There is a correlation of .93 in there! That can't be random!"*
> Student 4: *"Then why are all those correlations significant? It's not random, something else must be going on."*

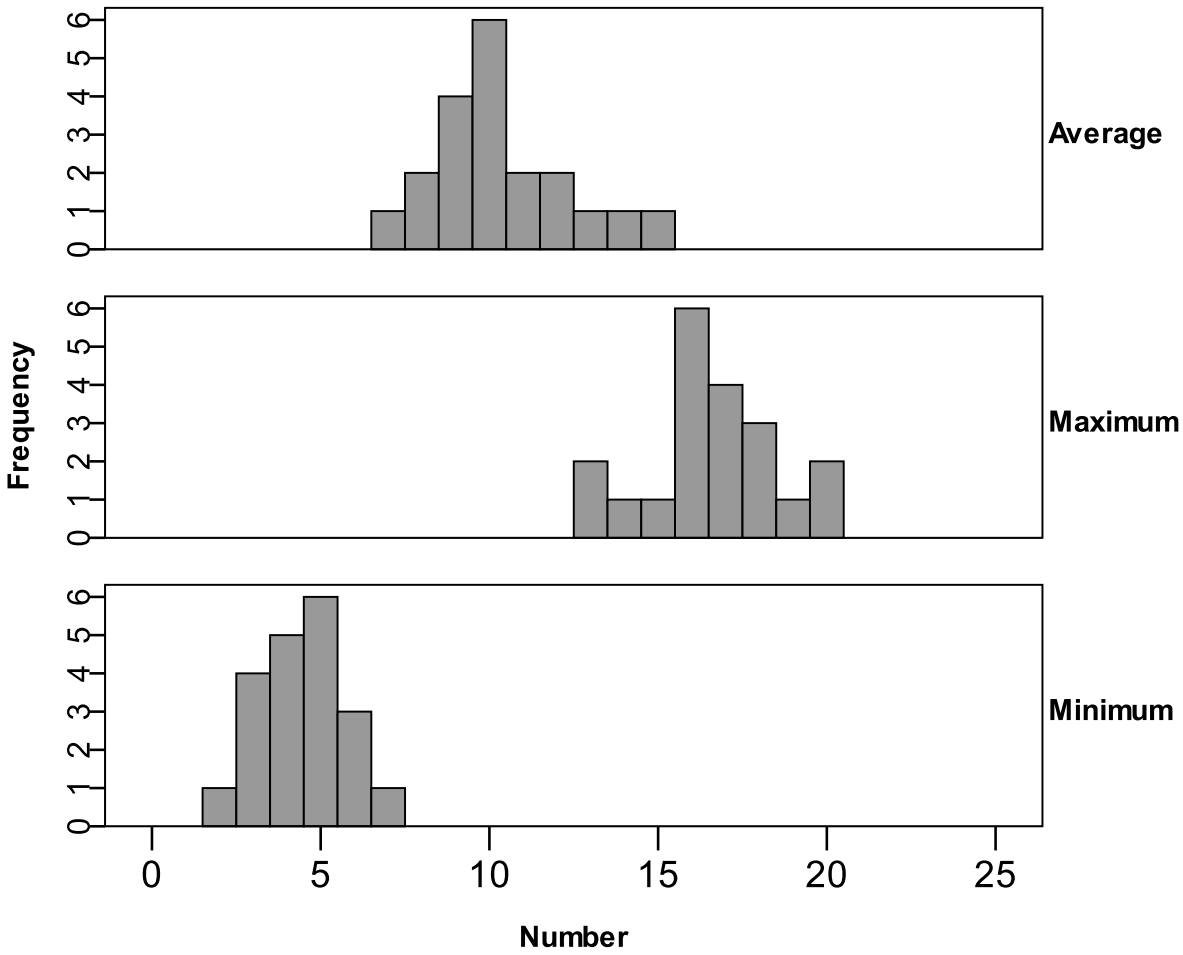
"No!" I exclaimed. "What you have witnessed is simply the power of chance and the alarming frequency of type I errors if you are comfortable with testing 195 separate hypotheses while setting alpha at .05 and willing to generate hypotheses for significant relations after the fact. Don't do this or you're guaranteed to make type I errors and publish bad research!"

I then sent the students home with the spreadsheet and the following assignment:
You now have the same spreadsheet we used in class. You now know that the data is simply random numbers. I would like you to simply count the number of significant correlations at the .05 or higher level for 20 re-randomizations of the data. You can re-randomize the data simply by clicking on an empty cell in the spreadsheet and typing the delete key. Repeat this 20 times, each time recording the number of significant correlations you find in the resulting simulation. Calculate the average number of significant correlations, and also provide the minimum and maximum number of significant correlations that you found across the 20 randomizations. Please write a paragraph reflecting upon how this exercise helped you understand the difference between hypothesis testing and data dredging. Also, reflect upon why it is dangerous to generate hypotheses after running a study and analyzing data. Also, reflect upon what assumptions are violated by the data as described on the spreadsheet, for instance, is it correct to assume that personality traits are normally distributed? How will you deal with these problems in the future?

Figure 2 presents histograms of the average number of significant correlations (type 1 errors) across the class of 20 students, as well as the minimum and maximum numbers of significant correlations. Recall that the formula for the experiment-wise error rate predicts one would expect

about ten type I errors of the 195 hypotheses being tested. The average for the class was 10.35 (SE = 0.45), which is spot on.

Figure 2: Histogram depicting the average, minimum and maximum numbers of type I errors (with alpha set to .05) from the random number correlation matrix of 195 biserial comparisons from Demonstration I.



The student essays suggested that they now understood the difference between hypothesis testing and data dredging, and most indicated how they would revise their research projects in order to avoid spurious correlations by developing specific hypotheses based upon the published literature rather than testing hundreds of hypotheses and seeing which are statistically significant. Several suggest that they might at first conduct an exploratory analysis by conducting a preliminary study, then conduct a confirmatory replication in order to determine whether the significant correlations hold across two sets of data from independent samples.

3.2. Demonstration II: Larger Ns do not protect you from Type I errors!

After the correlation demonstration, my course moved onto inferential tests of differences between means, beginning with the t test. In a class discussion over the research projects, one

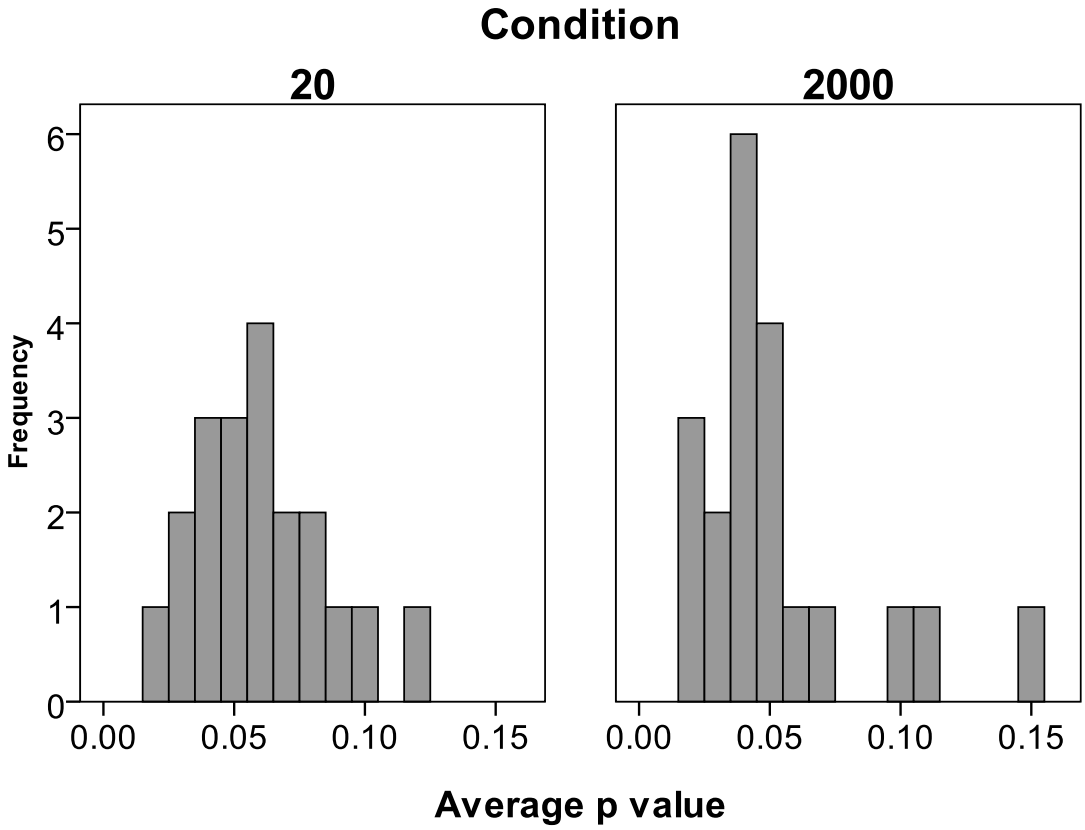
student suggested he would not have to worry about type I errors because he was working with a nationally representative dataset consisting of 2000 participants. "With a sample so large, any observed effect couldn't be due to chance because the sample is so large." I explained to him after class that the sample size was relatively independent from the probability of making type I errors given a reasonable number of observations. Type 1 errors are a function of alpha, which the researcher sets, and that sample size only affects the statistical power (β) of the test. So although a larger sample might reduce the probability of making a type II error, it would have relatively little impact upon the type I error rates, because smaller spurious random perturbations in the sampled data would be detected (and deemed statistically significant) due to the increased power of the test. He did not seem persuaded by the end of the discussion, so I felt another demonstration would be in order.

For the following class, I created a spreadsheet called SampleSizeTypeIError.xls. (This spreadsheet is available as an Excel file titled "SampleSizeTypeIError.xls" which may be obtained by following the link here or at the end of this article). In it, I produced two worksheets consisting of random data using the =rand() function in Excel, "Twenty" and "Two Thousand." The twenty sheet consisted of two columns of twenty numbers each, and the two thousand sheet consisted of two columns of twenty thousand numbers each. In class, I explained that these were random numbers (since they knew what was up my sleeve from the previous assignment). Because I was teaching t tests, I set up the file to demonstrate testing the difference between means using non-directional (2-tailed) independent sample t tests. Using a projector, I followed the instructions provided in the spreadsheet and clicked on the suggested cells in order to generate new random lists of numbers to demonstrate that across many re-randomizations of numbers, there were many "significant" results (type I errors) with the list of twenty and the list of two thousand numbers. We did an informal count in class, finding 2 type I errors out of 50 for the sheet with 20 data points, and 3 type I errors out of 50 for the sheet with 2000 random numbers. The expected number, of course, is 2.5 per 50 (.05) in both cases.

I sent the students home with the following assignment: *Please download the spreadsheet we used in class. I would like you to re-randomize the data set 100 times for both the worksheet of 20 cases and 2000 cases and count the number of "significant" differences you observe between the two sets of "data." Please count a) the number of significant cases you found, and b) keep track of the largest and smallest p value you find within this dataset. Please hand these values in and write an essay describing what you learned by doing this assignment about a) the frequency of type I errors b) sample size and type I and II errors and c) the value of replication.*

Figure 3 presents histograms from the 20 students' simulations. Unsurprisingly, they found an average of 6 (SE= 0.6) significant differences out of the 100 simulations for the dataset of 20 random numbers and 5.7 (SE = .01) significant differences out of the 100 simulations for the dataset of 2000 random numbers. The proportion of "significant" results in both sample sizes is very close to alpha regardless of the sample size.

Figure 3: Histograms depicting the average number of type I errors out of 100 simulations with 20 and 2000 cases of data that are random numbers from Demonstration II.



In their essay responses, several students mentioned their first attempt at re-randomizing the data resulted in a significant difference, which suggested to them that it might be possible that their first attempt at an experiment might result in a type I error. Several noted that replication would help build confidence in their findings, because it would be unlikely to obtain three sequential type I errors independently. Most reflected on their surprise that sample size did not protect them from the probability of making type I errors, and suggested a deeper understanding of the relation between power and alpha. By this point, hair began growing back on my head.

3.3 Demonstration III: How ANOVA controls for multiple comparisons, but type I errors still occur!

A third demonstration I produced using random numbers later in the course aimed to instruct ANOVA procedures, and specifically why ANOVA is used in making multiple comparisons over multiple t tests. I developed this spreadsheet in response to several of the student project proposals that planned to conduct multiple comparisons using several t tests. Since I had not at that point discussed analysis of variance (ANOVA) in detail, I used this mistake as an opportunity to develop a random number spreadsheet to help demonstrate how ANOVA controls for multiple comparisons. However, one misconception I found in teaching ANOVA is that

students erroneously believe that ANOVA eliminates the probability of type I errors, which of course is not the case.

I created a spreadsheet called "ANOVAExercise.xls." (This spreadsheet is available as an Excel file titled "ANOVAExercise.xls" which may be obtained by following the link here or at the end of this article). The file contains two worksheets, one consisting of a colorful series of boxes and charts (ANOVA), and another sheet that can be used to save the simulations (TALLY). On the ANOVA sheet, on the left there are five "groups" of data ranging from 0 – 10 in different colors. The summary statistics appear below the data, and a bar graph appears below that, summarizing the means and standard errors. The pink top panel contains the formula for a one-way ANOVA comparing the means of the five "groups." All the necessary information to compute analysis of variance is included in the box, which proved useful for demonstrating the mathematical formula used in calculating the F ratio. In the blue box beneath the pink box, there are pairwise t tests comparing the means of all the group differences. The spreadsheet contained the following instructions:

> *To begin this simulation, start in the green box below. Click the "delete" key. This will produce a new set of random numbers. The spreadsheet will calculate whether there is a significant difference between the conditions using an ANOVA versus using ten independent t tests for multiple comparisons of the group means. Please look in the red boxes for the results of both tests (ANOVA and T TESTS). Please record on the tally sheet A) the F ratio, B) if the ANOVA yielded a significant result by typing a 1 for "yes" and a 0 for "no." C) For the t test, count the number of significant "results" you obtained, and indicate this number on the tally sheet as well. Repeat this 50 times. Once you do 50 simulations, print out the tally sheet and hand it in with your homework. Also, email your professor the completed file. As always, have fun with statistics and LEARN. Oh, and don't forget...these data are simply random numbers!*

The purpose of this exercise was to determine the number of pairwise comparisons occurring by chance that are type I errors, relative to the proportion of ANOVA F tests that are type I errors. With alpha set at .05, the *proportion* of type I errors would be around .05 for both the ANOVA as well as the multiple t tests. However, the error in the students' reasoning is that the *absolute number* of type I errors would be much greater for the multiple t tests, simply because the absolute number of t tests performed is ten times the absolute number of ANOVAs performed.

I provided students with the following assignment to accompany the demonstration: *On the tally sheet, please examine the proportion of type I errors made for both the ANOVA and the t tests. Please also examine the absolute number of type I errors for both the ANOVA and t tests. Why are the proportions so equal yet the absolute numbers so different? How does this inform you about why it is advantageous to use ANOVA over multiple comparisons using t tests? Reflect on your own research projects. Would you revise your proposed analysis based on this simulation?"*

The students found an average of 2.85 (SE = 0.25) type I errors with the ANOVA simulation, and an average of 25.2 (SE = 0.74) type I errors with the t-tests. Taking into account the proportion of tests actually run (50 ANOVAs and 500 t tests) the average proportion of type I

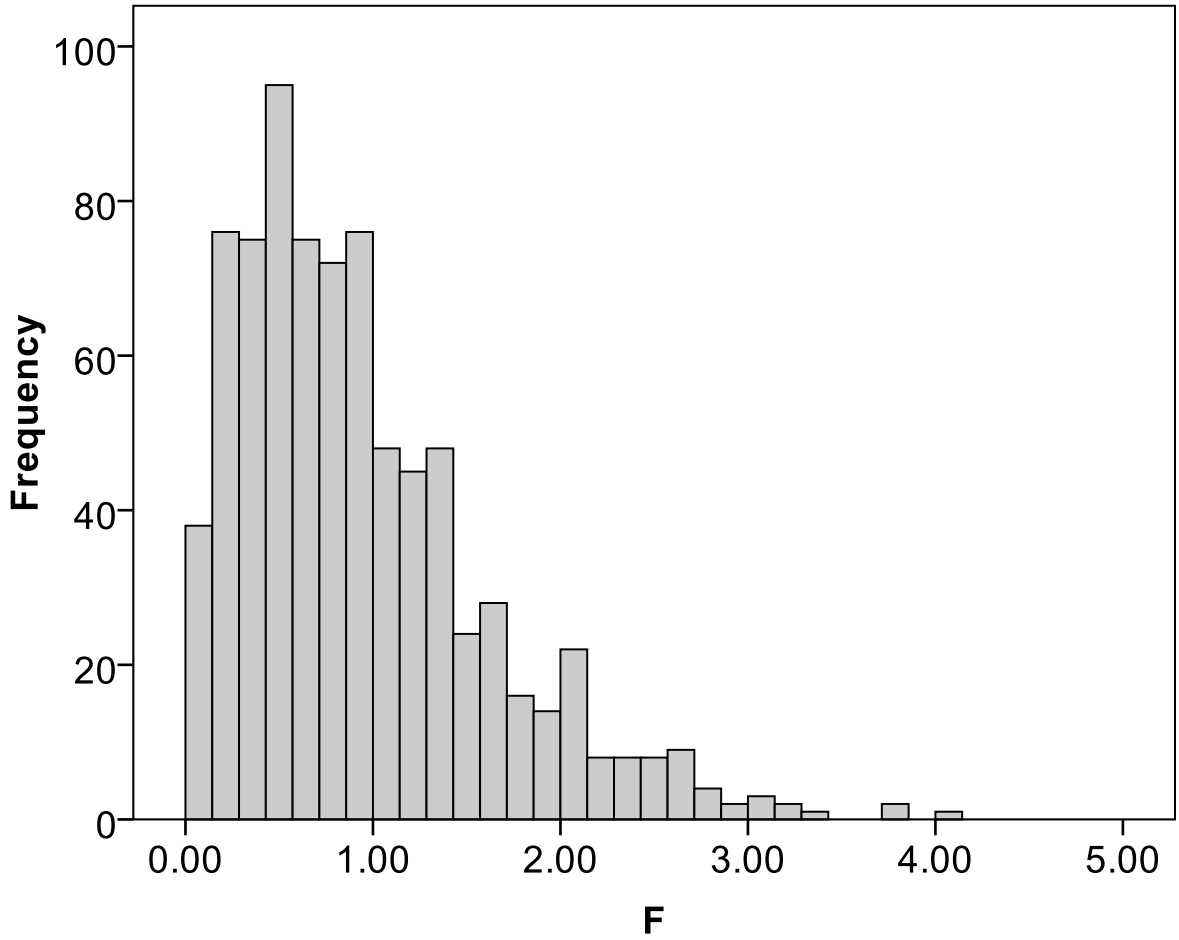
error per test performed was 0.057 (SE = .005) for the ANOVAs and 0.050 (SE = .005) for the t tests. I used these data to start a class discussion on how ANOVA and t tests have the same alpha, but in making multiple comparisons, the fact that one is conducting so many independent tests of hypotheses means that the absolute number of type I errors is greater than that found with ANOVA. One student pointed out that ANOVA is testing only the hypothesis that at least one of the groups exhibits performance that is different from at least one of the other groups in such a way that is unlikely explained by chance variation across all groups. The class discussed how in many cases (when the number of levels of an independent variable is greater than 2) a significant omnibus ANOVA requires further analysis through the use of contrasts (for planned comparisons) or post-hoc tests (for unplanned comparisons).

In their essay responses, most of the students demonstrated that they understood that multiple comparisons increase the absolute number of type I errors because so many more hypotheses are being tested, relative to the ANOVA, which tests only the hypothesis that at least one of the groups exhibits data that seems unlikely due to chance. In class, I used this revelation to introduce the idea of testing for differences between groups after finding a significant omnibus ANOVA through a priori contrasts for predicted hypotheses or post-hoc tests such as Tukey's HSD or Scheffe's test for ones that were not predicted.

Additionally, I took advantage of this opportunity to show the students that their simulations using random numbers produced an F distribution. The previous week, I had explained the F distribution and showed them examples of F distributions with different degrees of freedom. One student asked me how statisticians 'produce' such distributions. I told them that they had already produced one, and that I would show this in the following class.

Because the students emailed me their spreadsheets, I copied and pasted the 50 F ratios produced by the simulation from the tally sheet into an applied statistics software package and produced a histogram of the resulting distribution of F ratios (See Figure 4). In class, I asked the students to open their textbook to the page depicting a line graph of the F distribution. I showed them the resulting histogram, and told them that this was a graph of the concatenated F ratios that they had submitted to me.

Figure 4: F ratios for 800 simulations using random data from Demonstration III.



"Where have you seen something that looks like this before?" After a few moments of silence, the student who asked me how statisticians produce the F distribution exclaimed, "Wait! That looks like the graph of the F ratio in the book!" I pushed her, "Why would this make sense from the simulations you produced?" After a moment, she responded, "Because we were using random numbers, it would be rare to find large F ratios by chance, but you'd predict small F ratios by chance which is why there were so many values near 1 or 2, and so few above 3 or 4!"

It was one of those Eureka moments that occur so rarely in an evening graduate statistics class. I then showed them that out of the 800 simulations produced using random numbers, 35 of them were above the critical region for $\alpha = 0.05$ and $F(4,800) = 2.38$. I asked them to open their calculators and do the math: 35/800 = .0425, which I pointed out was very close to their alpha of .05 (These data are included in the spreadsheet as the "800Simulation" worksheet). The collective a-hahs were quite rewarding, as was the sense that my hair was beginning to grow back in.

4. General Discussion

Applied statistics software packages have revolutionized the practice of statistical analysis by saving time and energy that would otherwise go towards long and tedious calculations. Such programs also afford fewer errors in simple arithmetic that have long frustrated students and researchers alike. At the risk of sounding like a Luddite, I have found that the ease and rapidity of testing hundreds of hypotheses in a matter of milliseconds allows students and researchers to conduct sloppy research that violates basic statistical assumptions and best (or at least better) research practices. Rather than forgetting to carry over a remainder of 1, students forget that every hypothesis they test comes at a risk of committing a sin of omission (a type II error), or a more critical sin of commission (a type I error). It is easy to toss a dozen variables into a textbox or run a few dozen t tests and see which hypotheses "stick," then write up a research report as if the resulting analyses were produced from genuine hypothesis testing rather than disingenuous *a posteriori* hypothesis generation.

The use of random numbers in a spreadsheet application helps teach students the dangers of such dubious research practices. By interpreting spurious correlations and significant group differences using actual random numbers, students learn the danger of data dredging, and witness first-hand the actual frequency of type 1 errors. They also begin to understand how easy it is to come up with after-the-fact explanation for any significant relation or difference. Additionally, this exercise helps students better understand the logic of hypothesis testing, and provides a platform for discussing more detailed topics about various issues, such as Dunn's procedure for controlling for error rates in multiple comparisons. Most importantly, students learn how random processes can create false beliefs, which is an important realization for researchers at the start of their careers.

One of the benefits of using Excel is that the =rand() function re-samples random numbers every time any cell value is changed (which is why in several of the exercises, students hit the delete key within an empty cell – this causes the program to resample the data, producing a new simulation). Although it is possible to employ macros so Excel re-samples the data without student input (and do so thousands of times in less than a second) I have found that the use of such macros limits the pedagogical value of the exercise. By having to simulate the sampling of new random distributions, students experience the equivalent of a "replication" of an experiment, and in tallying up data themselves, develop a better feel for the process as a whole than if a macro did the work. Students acquire a sense of "ownership" of the numbers because of the time and effort put into resampling over several iterations, more than what they would get out of pressing a single button. Finally, these exercises give students practice with the Excel program, which is an important computer skill more generally. For example, several students did not know that Excel allows for multiple worksheets, or that it is possible to calculate simple statistics (t tests, ANOVA) using Excel alone.

On a precautionary note, one possible limitation of these demonstrations is that students might believe that different rules apply to random and real data. For example, in response to the first correlation demonstration, one student wrote that "Well, the production of type I errors may occur for random numbers, but when I collect my data I will not be using random numbers, but actual data, and so I can be more confident in the results than if I were using just random

numbers." In response to this, I emphasized in class that whether the data is "real" or "random" the same risks of type I errors still apply. I asked my students to imagine a scenario in which they hand out surveys to a class of bored and disinterested undergraduates, all of whom would rather be doing anything on earth than answering a boring survey. I propose that each of the participants might decide to simply answer the survey by randomly circling responses, in which case the data are "real" but just as "random" as those produced by the Excel program. I also emphasized that real data can create type I errors by providing them with a real data set of my height between the years of 1976 – 1996, and U.S. national murder rates during the same period, showing a strong (0.75) and significant ($p < .001$) relation between the two variables. I emphasized both sets of variables were "real," but they were unlikely related in any meaningful way (unless my physical development somehow created a murder epidemic, or if the increase in murder rates somehow spurred my growth – both of which are highly unlikely).

Additionally, the simulation of the F distribution in Demonstration III provided them with a clear example of the reason why ANOVA is used when making multiple comparisons, as well as a better understanding of the F distribution. The students learned that the proportion of type I errors were the same for ANOVA and the multiple t tests, but that for any given analysis, there would be a greater absolute number of type I errors with t tests because so many of them are performed concurrently when considering all pairwise comparisons of the various levels of an independent variable. Further, being able to produce a graph of an F distribution from their own random simulation data allowed the students to better understand how the analysis of variance procedure works. Also, having a spreadsheet consisting of the simple equations used to calculate the F ratio helped demystify the often complex and obscure output common to most statistical packages. The students could trace the equations forward and backwards and understand that the mathematics underlying these tests is simple arithmetic, and not complicated and abstract algebra or calculus, but more importantly, not magic!

Did these demonstrations work? Well, the students passed the final exam with flying colors, and I once again have a full head of hair. But this assessment could just be a type I error, so I will reserve judgment until replication by others.

Links to Excel spreadsheets discussed in article:
Type I Error Exercise - TypeIErrorExercise.xls
Sample Size Type I Error - SampleSizeTypeIError.xls
ANOVA Lesson - ANOVAExercise.xls

References

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior, 6,* 151-155.

Davis, S.F., & Smith, R.A. (2004). *An Introduction to Statistics and Research Methods.* Upper Saddle River, NJ: Pearson.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology,12,* 392-407.

Faulk, R., & Greenbaum, C.W. (1995). Significance tests die hard. *Theory and Psychology, 5,* 75-98.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. Psychological Review, 104, 301-318.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd (originally published in 1926).

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33,* 587-606.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning with instruction: Frequency formats. *Psychological Review, 102,* 684-704.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102-105.

Killeen, P.R. (2005). *An alternative to null-hypothesis significance tests. Psychological Science, 16, 345–53.*

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* New York: Wiley.

Martinez-Dawson, R. (2003). Incorporating laboratory experiments in an introductory statistics course. *Journal of Statistics Education, 11.*

Mlodinow, L. (2008). The Drunkard's Walk: How Randomness Rules our Lives. New York: Random House.

Neyman, J. & Pearson, E.S. (1967). "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, Part I", reprinted at pp.1–66 in Neyman, J. & Pearson, E.S., *Joint Statistical Papers*, Cambridge University Press, (Cambridge), (originally published in 1928).

Neyman, J. (1950). *First Course in Probability and Statistics.* New York: Holt.

Nickerson, R. S. (2002). The production and perception of randomness. Psychological Review, 100, 330-357.

Steinberg, W.J. (2008). *Statistics Alive.* Thousand Oaks, CA: Sage.

Thorne, B.M. & Giesen, J.M. (2002). *Statistics for the Behavioral Sciences* (4th Ed.) Boston: McGraw Hill.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207-232.

Vaughan, T.S. (2003). Teaching statistical concepts using student specific datasets. *Journal of Statistics Education, 11*.

Vernoy, M., & Kyle, D.J. (2001). *Behavioral Statistics in Action.* Boston: McGraw Hill.

Wilson, J.H. (2004). *Essential Statistics.* Upper Saddle River, NJ: Pearson.

Sean Duffy
Department of Psychology
Rutgers University
311 North Fifth Street
Camden, NJ, 08102
seduffy@camden.rutgers.edu

You may edit these Excel files as you wish, and if you have improvements or suggestions feel free to email them to me.