Filedrawer and Publication Bias in Academic Research

The peer review process is the main mechanism through which scientific communities decide whether a research paper should be published in academic journals This exercise is based on:

Franco, A., N. Malhotra, and G. Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345(6203): 1502–5.

and

Franco, A., N. Malhotra, and G. Simonovits. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23(2): 306–12.

By having other scientists evaluate research findings, academic journals hope to maintain the quality of their published articles. However, some have warned that the peer review process may yield undesirable consequences. In particular, the process may result in *publication bias* wherein research papers with statistically significant results are more likely to be published. To make matters worse, being aware of such a bias in the publication process, researchers may be more likely to report findings that are statistically significant and ignore others. This is called *filedrawer bias*.

In this exercise, we will explore these potential problems using data on a subset of experimental studies that were funded by the Time-sharing Experiments in the Social Sciences (TESS) program. This program is sponsored by the National Science Foundation (NSF). The data set necessary for this exercise can be found in the csv files `filedrawer.csv` and `published.csv`. The `filedrawer.csv` file contains information about 221 research projects funded by the TESS program. However, not all of those projects produced a published article. The `published.csv` file contains information about 53 published journal articles based on TESS projects. This data set records the number of experimental conditions and outcomes and how many of them are actually reported in the published article. The tables below present the names and descriptions of the variables from these data sets.

Name	Description (filedrawer.csv)
`DV`	Publication status
`IV`	Statistical significance of the main findings
`max.h`	H-index (highest among authors)
`journal`	Discipline of the journal for published articles
`teasown`	Amount of tea sown in county
`sex`	Proportion of males in birth cohort
`post`	Indicator variable for introduction of price reforms

Name D	escription (published.csv)
`id.p`	Paper identifier
`cond.s`	Number of conditions in the study
`cond.p`	Number of conditions presented in the paper
`out.s`	Number of outcome variables in the study
`out.p`	Number of outcome variables used in the paper

```
published <- read.csv("/Users/clayparham/Documents/Bush631-/Assignment 12/published.csv")
filedrawer <- read.csv("/Users/clayparham/Documents/Bush631-/Assignment 12/filedrawer.csv")
```

Question 1

We begin by analyzing the data contained in the `filedrawer.csv` file. Create a contingency table for the publication status of papers and the statistical significance of their main findings. Do we observe any distinguishable pattern towards the publication of strong results? Provide a substantive discussion.

We see in the results that null levels of significance are almost never even written. Between unpublished and published papers, we see similar levels of weak statistical significance. However, if you want to get it published in a top paper, you'll want to have a strong statistical significance.

```
table(DV = filedrawer$DV, IV = filedrawer$IV)
```

```
##                       IV
## DV                     Null Strong Weak
##   Published, non top    5     35   31
##   Published, top        5     21    9
##   Unpublished           7     31   32
##   Unwritten            31      4   10
```

Question 2

We next examine if there exists any difference in the publication rate of projects with strong vs. weak results as well as with strong vs. null results. To do so, first, create a variable that takes the value of 1 if a paper was published and 0 if it was not published. Then, perform two-tailed tests of difference of the publication rates for the aforementioned comparisons of groups, using 95% as the significance level. Briefly comment on your findings.

We find that published papers are not significantly different between strong and weak, whereas there is some significance between not-published and published papers.

```
filedrawer$published <- ifelse(filedrawer$DV == 'Published, non top' | filedrawer$DV == 'Published, top
filedrawer$strong <- ifelse(filedrawer$IV == 'Strong', 1, 0)
filedrawer$weak <- ifelse(filedrawer$IV == 'Weak', 1, 0)
filedrawer$null <- ifelse(filedrawer$IV == 'Null', 1, 0)

t.test(filedrawer$published, filedrawer$null)
```

```
##
##  Welch Two Sample t-test
##
## data:  filedrawer$published and filedrawer$null
## t = 6.0093, df = 424.73, p-value = 4.011e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1766020 0.3482849
## sample estimates:
## mean of x mean of y
## 0.4796380 0.2171946
```

```
t.test(filedrawer$published, filedrawer$strong)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  filedrawer$published and filedrawer$strong
## t = 1.4355, df = 439.9, p-value = 0.1518
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02505061  0.16079722
## sample estimates:
## mean of x mean of y
## 0.4796380 0.4117647
```

```
t.test(filedrawer$published, filedrawer$weak)
```

```
##
##  Welch Two Sample t-test
##
## data:  filedrawer$published and filedrawer$weak
## t = 2.3178, df = 439.5, p-value = 0.02092
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01651255 0.20068202
## sample estimates:
## mean of x mean of y
## 0.4796380 0.3710407
```

Question 3

Using Monte Carlo simulations, derive the distribution of the test statistic under the null hypothesis of no difference for each of the two comparisons you made in the previous question. Do you attain similar p-values (for a two-tailed test) to those obtained in the previous question?

There were similar means and similar p-values to what we found above for everything but the strong files. That might show that there is some validity to what the data found.

```
sample.strong <- sample(filedrawer$strong, size = 20000, replace = TRUE)
sample.weak <- sample(filedrawer$weak, size = 20000, replace = TRUE)
sample.null <- sample(filedrawer$null, size = 20000, replace = TRUE)
prop.table(table(sample.strong))
```

```
## sample.strong
##       0       1
## 0.58585 0.41415
```

```
prop.table(table(sample.weak))
```

```
## sample.weak
##       0       1
## 0.62605 0.37395
```

```
prop.table(table(sample.null))
```

```
## sample.null
##      0      1
## 0.7829 0.2171
```

```
mean(sample.strong)
```

```
## [1] 0.41415
```

```
mean(sample.weak)
```

```
## [1] 0.37395
```

```
mean(sample.null)
```

```
## [1] 0.2171
```

```
t.test(filedrawer$published, sample.null)
```

```
##
##  Welch Two Sample t-test
##
## data:  filedrawer$published and sample.null
## t = 7.7656, df = 223.31, p-value = 2.913e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1959146 0.3291615
## sample estimates:
## mean of x mean of y
##  0.479638  0.217100
```

```
t.test(filedrawer$published, sample.strong)
```

```
##
##  Welch Two Sample t-test
##
## data:  filedrawer$published and sample.strong
## t = 1.934, df = 224.73, p-value = 0.05437
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001238946  0.132214964
## sample estimates:
## mean of x mean of y
##  0.479638  0.414150
```

```
t.test(filedrawer$published, sample.weak)
```

```
##
##  Welch Two Sample t-test
##
## data:  filedrawer$published and sample.weak
## t = 3.1218, df = 224.56, p-value = 0.002034
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03897318 0.17240284
## sample estimates:
## mean of x mean of y
##  0.479638  0.373950
```

Question 4

Conduct the following power analysis for a one-sided hypothesis test where the null hypothesis is that there is no difference in the publication rate between the studies with strong results and those with weak results. The alternative hypothesis is that the studies with strong results are less likely to be published than those with weak results. Use 95% as the significance level and assume that the publication rate for the studies with weak results is the same as the observed publication rate for those studies in the data. How many studies do we need in order to detect a 5 percentage point difference in the publication rate and for the test to attain a power of 95%? For the number of observations in the data, what is the power of the test of differences in the

publication rates?

Question 5

The H-index is a measure of the productivity and citation impact of each researcher in terms of publications. More capable researchers may produce stronger results. To shed more light on this issue, conduct a one-sided test for the null hypothesis that the mean H-index is lower or equal for projects with strong results than those with null results. What about the comparison between strong versus weak results? Do your findings threaten the ones presented for Question 2? Briefly explain.

Question 6

Next, we examine the possibility of filedrawer bias. To do so, we will use two scatterplots, one that plots the total number of conditions in a study (horizontal axis) against the total number of conditions included in the paper (vertical axis). Make the size of each dot proportional to the number of corresponding studies, via the `cex` argument. The second scatterplot will focus on the number of outcomes in the study (horizontal axis) and the number of outcomes presented in the published paper (vertical axis). As in the previous plot, make sure each circle is weighted by the number of cases in each category. Based on these plots, do you observe problems in terms of underreporting?

Question 7

Create a variable that represents the total number of possible hypotheses to be tested in a paper by multiplying the total number of conditions and outcomes presented in the questionnaires. Suppose that these conditions yield no difference in the outcome. What is the average (per paper) probability that at the 95% significance level we reject at least one null hypothesis? What about the average (per paper) probability that we reject at least two or three null hypotheses? Briefly comment on the results.