Homework 2

Potential Outcomes and Randomization Based Inference

Jaiden Atterbury

Potential Outcomes

1. Suppose that we have a population in which 42% are of type Helped, 6% of type Hurt, 23% of type Always Good and 29% of type Never Good. Let $Y = 1$ indicate a good outcome; $Y = 0$ a bad outcome; $X = \text{B}$ indicates the new customer experience; $X = \text{A}$ indicates the existing customer experience.

Table of Proportions Under Randomization:

	HE	HU	AG	NG
$X = \text{B}$	42%	6%	23%	29%
$X = \text{A}$	42%	6%	23%	29%

(a) Describe the distribution of outcomes that would be observed if everyone experienced A.

If everyone experienced $X = \text{A}$, all 42% of the people of type helped would experience the outcome $Y = 0$, all 6% of the people of type hurt would experience the outcome $Y = 1$, all 23% of the people of type always good would experience the outcome $Y = 1$, and lastly, all 29% of the people of type never good would experience the outcome $Y = 0$. Thus, in total, 29% of the individuals would experience the outcome $Y = 1$, and 71% of the individuals would experience the outcome $Y = 0$.

(b) Describe the distribution of outcomes that would be observed if everyone experienced B.

If everyone experienced $X = \text{B}$, all 42% of the people of type helped would experience the outcome $Y = 1$, all 6% of the people of type hurt would experience the outcome $Y = 0$, all 23% of the people of type always good would experience the outcome $Y = 1$, and lastly, all 29% of the people of type never good would experience the outcome $Y = 0$. Thus, in total, 65% of the individuals would experience the outcome $Y = 1$, and 35% of the individuals would experience the outcome $Y = 0$.

(c) Under random assignment, find $P(Y = 1|X = \text{A})$, the proportion having a good outcome in the group assigned to receive the existing experience A.

Under random assignment, the proportion of individuals having a good outcome in the group assigned to receive the existing experience, $X = \text{A}$, is the same as calculating $P(Y = 1|X = \text{A})$. Given that an individual was assigned to $X = \text{A}$, the only types fo people who observe the outcome $Y = 1$ are those of type always good and hurt. Since units were randomly assigned, proportions of types in treatment and control are the same. Thus, we know the proportion of individuals of type hurt in control is 6%, and the proportion of individuals of type always good in control is 23%. Therefore, $P(Y = 1|X = \text{A}) = 29\%$.

(d) Under random assignment, find $P(Y = 1|X = \text{B})$, the proportion recovering in the group assigned to receive the new experience B.

Under random assignment, the proportion recovering in the group assigned to receive the new experience, $X = \text{B}$, is the same as calculating $P(Y = 1|X = \text{B})$. Given that an individual was assigned to $X = \text{B}$, the only types fo people who observe the outcome $Y = 1$ are those of type always good and helped. Since units were randomly assigned, proportions of types in treatment and control are the same. Thus, we know the proportion of individuals of type helped in treatment is 42%, and the proportion of individuals of type always good in treatment is 23%. Therefore, $P(Y = 1|X = \text{B}) = 65\%$.

(e) Again assuming random assignment, what extra piece of information is required to find $P(Y = 1)$?

Since we already know that $P(Y = 1|X = \text{A}) = 29\%$ and $P(Y = 1|X = \text{B}) = 65\%$, we only need one extra piece of information to find $P(Y = 1)$. In particular, the piece of infromation that we'd need to know in order to calculate $P(Y = 1)$ is the proportion/percentage of individuals assinged to treatment and control. With this information we could solve for $P(Y = 1, X = \text{A})$ and $P(Y = 1, X = \text{B})$ in the equations $P(Y = 1|X = \text{A}) = \frac{P(Y=1,X=\text{A})}{P(X=\text{A})}$, and $P(Y = 1|X = \text{B}) = \frac{P(Y=1,X=\text{B})}{P(X=\text{B})}$ and thus we'd have all the information we'd need to find $P(Y = 1)$.

A/B Testing Simulation

2. Log in to the Wharton Interactive A/B Testing Simulation. Play the simulation.

Your goal is to maximize the conversion rate and maximize profits for Nano. Specifically you are asked to test out strategies for maximizing conversion rates using A/B tests. You will have the opportunity to play through 12 rounds (or weeks) of experiments.

Keep track of your progress and as you form your hypotheses, run your experiments, gather data, track your strategies and consider what is working in terms of strategy and what isn't working.

Note: You'll notice that you'll see a score for your round and your profits. If you don't get anywhere close to 100%, don't worry! It isn't possible to score 100% in this simulation. Think of Practice Mode as practice – you can make mistakes and learn from your mistakes.

After you have completed the practice mode simulation answer the following:

(a) In 1-2 paragraphs, briefly describe strategies that you employed.

After running through 12 weeks of the A/B test simulation, I scored 80% with a total profit of $\$3,856,718$. The main strategy I employed was choosing the order of aspects of the webpage that I thought would be the most important to change first, second, etc. I chose this order before starting the simulation and I stuck to it throughout the entirety of the simulation. I started off by changing the main image as I thought that was the aspect of the webpage that would be most likely capture a users attention when they first enter the webpage. Then, I targeted the tagline as I felt that if the users could get something stuck in their head/something that they could get behind they would be more likely to purchase the phone. The next aspect of the webpage I choose to target was the call to action, I changed this ealier on as, just like the main image, if I could capture the audiences attention they would be more likely to add to the cart, which is one step away from purchasing. The call to action aspect was followed by the features aspect, which is probably one aspect of the webpage that I would change earlier in my next trial as I feel this was something that really bolstered my profits once I got it right. Lastly, once everything else was optimized, I tried to find the target price. I saved price for last as it is the most sensitive topic, and it makes sense to mess with it only when everything else is where you want it to be.

Once I solidified my predetermined order of attack the next strategy I employed was making sure there was always an experiment going on in each subcategory. For example, if I optimized a certain aspect in one subcategory before finishing it in the others, I wouldn't wait to optimize that aspect in all subcategories before moving on to the next aspect in the subcategory that had already been optimized. The next, and most important strategy I employed was figuring out when I would lock in a decision for each aspect in each subcategory. The main rule of thumb I used was if all levels of each aspect were equally allocated in terms of web traffic and one performed significantly better than the rest, I would automatically choose it, and vice versa if one performed significantly worse than the rest I would automatically ditch it. In the situation where none of the levels performed any better, I would usually run the experiment again without changing anything to wait and see if a trend would occur. If all of this failed, I would change the allocation based on what I thought was the optimal level of the aspect and then run the experiment again. On average, it usually took me around 2 weeks to optimize each aspect for each subcategory, with 3 weeks being the longest it took me to optimize a single aspect in a single subcategory.

(b) What worked and what didn't work? What would you advise another student playing this simulation to do? Again write 1-2 paragraphs.

The aspects of my strategy that seemed to work the best were my order of execution, and making sure I always had an experiment running in each subcategory at all times. By saving price as the last aspect to optimize in each subcategory, I noticed my profits skyrocketed near the end. While at the same time, by making sure I always had an experiment running in each category, I was able to keep my profits increasing throughout the entirety of the simulation. In particular, there was no point in which my profits decreased and the rate of increase stayed positively constant pretty much the whole time. On the other hand, my most flawed and least effective/logical strategy was my stopping conditions, in particular knowing when and how much to change traffic allocations when no trend seemed apparent. I was good at noticing when things were significant, but when things weren't significant I had a hard time figuring out what my next move would be. I'm still a little unsure of how I should change traffic allocation when all four levels of an aspect have the same allocation but none seem to be better or worse than the others.

If I were to give advice to another student playing this simulation on what to do, I would first tell them to look at all of the aspects and figure out which ones they think are the most important to change, and create a running order on which ones they would experiment on first. I would also tell any student planning on doing this simulation to keep track of trends they see in each subcategory, as I noticed some subcategories were more responsive to different aspects than others. In my simulation, international mobile users made me "work" the most, as trends seemed to not be as apparent for each aspect in comparison to the other subcategories. By keeping track of these trends, people playing this simulation might be able to come up with strategies for all aspects of the webpage that work best for specific subcategories.

Randomization based inference

For the next two questions review the files neyman-fisher-498.R and hyper-geometric.R.

3. A randomized experiment is performed to investigate the effect of a change in an algorithm on the time required to respond to a user query. There are 20 users with 10 in Treatment and 10 in Control. The ordered response times (in seconds) in each group are as follows:

$$\begin{aligned} \text{Treatment} &: 1.12, 1.81, 1.86, 3.21, 3.55, 5.49, 5.70, 6.24, 14.60, 19.96, \\ \text{Control} &: 0.66, 2.25, 5.76, 6.59, 9.66, 10.38, 10.49, 13.66, 18.05, 20.57. \end{aligned}$$

First consider Neyman's approach:

(a) Estimate the average causal effect of the change.

```
# Setting up the data
x <- c(rep(0,10), rep(1,10))
y <- c(0.66, 2.25, 5.76, 6.59, 9.66, 10.38, 10.49, 13.66, 18.05, 20.57,
       1.12, 1.81, 1.86, 3.21, 3.55, 5.49, 5.70, 6.24, 14.60, 19.96)
n <- length(x)
k <- sum(x)

# Find sample mean for control and treatment group
mean.ctrl <- mean(y[x==0])
mean.trt <- mean(y[x==1])

# Calculating sample ACE
hat.ace <-  mean.trt - mean.ctrl
```

Our sample estimate of the average causal effect of the change is $\hat{\text{ACE}}$ =-3.453.

(b) Find an estimate of the variance of your estimate of the average causal effect.

```
# Find sample variance for control and treatment group
var.ctrl <- var(y[x==0])
var.trt <- var(y[x==1])
```

```
# Calculating an estimate of the variance of our estimate of the average causal effect
hat.var.hat.ace <- var.trt/k + var.ctrl/(n-k)
```

Our sample estimate of the variance of the estimate of the average causal effect of the change is $\hat{Var}(\hat{\text{ACE}}) =$ 7.8583028.

(c) Construct a (conservative) 95% confidence interval for the ACE using your answers from (a) and (b).

```
#Constructing a conservative 95% confidence interval for the smaple ACE
lower <- hat.ace - 1.96*sqrt(hat.var.hat.ace)
upper <- hat.ace + 1.96*sqrt(hat.var.hat.ace)
```

A conservative 95% confidence interval for the sample average causal effect is (-8.9474022, 2.0414022).

Extra Credit: Confidence Intervals for Tighter Upper Bounds

1.) Use $\frac{1}{n-1}[\frac{n-k}{k}\hat{\sigma}_1^2 + \frac{k}{n-k}\hat{\sigma}_2^2 + 2\sqrt{\hat{\sigma}_1^2\hat{\sigma}_2^2}]$ for $\hat{Var}(\hat{ACE})$ (with finite population correction):

```
# Find the variances with finite population correction
var.trt2 = (n - 1)/n * var.trt
var.ctrl2 = (n - 1)/n * var.ctrl

# Calculate new estimate for variance
hat.var.hat.ace2 <- (1/(n-1)) * (var.trt2*(n-k)/k + var.ctrl2*k/(n-k) + 2*sqrt(var.trt2 * var.ctrl2))

# Calculate the 95% confidence interval for ace.hat
lower2 <- hat.ace - 1.96*sqrt(hat.var.hat.ace2)
upper2 <- hat.ace + 1.96*sqrt(hat.var.hat.ace2)
```

Our new, and tighter, 95% confidence interval for $\hat{\text{ACE}}$ is (-8.9464258, 2.0404258).

2.) Use Aronow's estimator for the special case where $k = n - k$ (with finite population correction):

```
# Calculate covariance under Aronow's estimator
covxy <- sum(y[0:10]*y[11:20])/10 - sum(y[0:10])*sum(y[11:20])/100

# Calculate new estimate for variance
hat.var.hat.ace3 <- (1/(n-1)) * (var.trt2*(n-k)/k + var.ctrl2*k/(n-k) + 2*covxy)

# Calculate the 95% confidence interval for ace.hat
lower3 <- hat.ace - 1.96*sqrt(hat.var.hat.ace3)
upper3 <- hat.ace + 1.96*sqrt(hat.var.hat.ace3)
```

Our new, and even tighter, 95% confidence interval for $\hat{\text{ACE}}$ is (-8.7628491, 1.8568491).

Now consider Fisher's approach:

(d) Perform a test of the null hypothesis that no customer would have had a different response time had they been in the other group. Report the p-value from a two-sided test, using the estimator ACE as your statistic. [Hint: See Lecture 6, slides 39 and 40].

H_0: No customer would have had a different response time had they been in the other group. This is the same as saying $y_i(1) - y_i(0) = 0, \ \forall i$.

Using $D = \hat{\text{ACE}}$ as our test statistic, in order to report the p-value from a two-sided test, we will run a permutation test. Since we ran a completely randomized experiment, the randomization-based distribution of D is computed by considering all $\binom{n}{k}$ possible treatment assignments. In this case there are $\binom{20}{10}$ possible treatment assignments. Since there are 184756 possible treatment assignments, instead of calculating by hand, we will use the exactRankTests package to find the p-value.

```
# Calculating the p-value from a two-sided test using ace.hat as the test statistic
pval <- perm.test(y[0:10], y[11:20])$p.value
```

Based on the permutation test, our two-sided p-value was: $P(|D| \geq 3.453) = 0.2359328$. Thus we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level.

Conclusion: As calculated in part 3d, the null hypothesis rejection p-value is 0.2359. Since the p-value is greater than the significance level of $\alpha = 0.05$ we fail to reject the null hypothesis. Since we have failed to reject the null hypothesis, there is insufficient evidence to suggest that customers would have had a different response time had they been in the other group.

4. A company sends out e-mails offering a special offer to new subscribers for an online news service. The company tries two different versions, each sent to 5000 customers, and records the number of customers who respond.

	Subscribe	No Response
Control E-mail	41	4959
Treatment E-mail	62	4938

Carry out Fisher's Randomization test for this dataset.

(a) Carefully state the null hypothesis that you are testing.

The null hypothesis we are testing when running a Fisher's Randomization test for this dataset is called the "sharp null hypothesis" which states: $\text{ICE}_i = 0$ for all i units. This is equivalent to saying that every unit is either of type always good or never good. Thus we are saying subscribers wouldn't have had a different response had they been in the other email group. In particular: $H_0 : y_i(0) = y_i(1)\ \forall i = 1, ..., n$.

(b) Compute the p-value. Hint: See Slide 34 in Lecture 6.

If we assume H_0 to be true all potential outcomes are observed, thus it follows that $Y \equiv y_i(0)(1 - X_i) + y_i(1)X_i = y_i(0) = y_i(1)$. In this setting, $\sum_i X_i(1 - y_i(0))$ is the total number of zeros in a random sample without replacement of size k from a population of size n with $\sum_i y_i(0)$ ones and $\sum_i (1 - y_i(0))$ zeros. Thus, $\sum_i X_i(1 - y_i(0))$ has a hypergeometric distribution and we can use this fact to compute the p-value.

```
# Setting up the data

# population size
N <- 10000

# total number in treatment
k <- 5000

# total number who subscribe
m <- 103

# total number who don't subscribe
n <- N-m

# calculating the one-sided p-value
pval2 <- phyper(4938,n,m,k)
```

The one-sided p-value using Fisher's Randomization test for this dataset is $P(\sum_i X_i(1 - y_i(0)) \leq 4938) = 0.0235431$.

(c) Give an interpretation of your answer to (b), as a probability.

Conclusion: As calculated in part 4b, the null hypothesis rejection p-value is 0.0235. Since the p-value is less than the significance level of $\alpha = 0.05$ we reject the null hypothesis. Since we rejected the null hypothesis,

there is sufficient evidence to suggest that customers would have had a different response had they been in the other email group. In terms of a probability, the p-value of 0.0235431 tells us that the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the sharp null hypothesis is true, is 0.0235431. In particular, under the assumption of the sharp null hypothesis, the probability that we see 4938 or less individuals not respond after given the treatment email is 0.0235431.