Problem Set 2

Code

Imports

1. What happens when pilgrims attend the Hajj pilgrimage to Mecca?

What happens when a diverse set of people are brought together toward a common purpose? Maybe it brings people together, or maybe instead it highlights the differences between groups. Clingingsmith, Khwaja and Kremer (2009) (https://dash.harvard.edu/handle/1/3659699) investigate the question. by asking Pakistani nationals to provide information about their views about people from other nations.

The random assignment and data is collected in the following way (detailed in the paper):

- Pakistani nationals apply for a chance to attend the Hajj at a domestic bank. Saudi Arabia agreed in the time period of the study (2006) to grant 150,000 visas.
- Of the 135,000 people who applied for a visa, 59% of those who applied were successful.
- The remainder of the visas were granted under a different allocation method that was not randomly assigned, and so this experiment cannot provide causal evidence about other allocation mechanisms.
- Among the 135,000 who apply, the authors conduct a random sample and survey about these respondents views about others.

Using the data collected by the authors, test, using randomization infernece, whether there is a change in beliefs about others as a result of attending the Hajj.

- Use, as your primary outcome the `views` variable. This variable is a column-sum of each respondent's views toward members of other countries.
- Use, as your treatment feature `success`. This variable encodes whether the respondent successfully attended the Hajj.

Hide

```
d <- fread("clingingsmith_2009.csv")
```

1. State the sharp-null hypothesis that you will be testing.

> 'For all Pakistani nationals' in the authors' sample, attending Hajj resulted in no change in beliefs about others as compared to before their trip.'

2. Using `data.table`, group the data by `success` and report whether views toward others are generally more positive among lottery winners or lottery non-winners. This answer should be of the form `d[ , .(mean_views = ...), keyby = ...]` where you have filled in the `...` with the appropriate functions and varaibles.

Hide

```
hajj_group_mean <- d[ , .(mean_views = mean(views)), keyby = success]

hajj_ate        <- 2.343137 - 1.868304

class(hajj_ate)
```

```
[1] "numeric"
```

But is this a "meaningful" difference? Or, could a difference of this size have arisen from an "unlucky" randomization? Conduct 10,000 simulated random assignments under the sharp null hypothesis to find out. (Don't just copy the code from the async, think about how to write this yourself.)

Hide

```
## do your work to conduct the randomization inference here.
## as a reminder, RI will randomly permute / assign the treatment variable
## and recompute the test-statistic (i.e. the mean difference) under each permutation

ate_v = c()

for (i in 1:10000) {

  assignment_vector <- rep(c(0,1), each = 958/2)
  ri = data.table(group = sample(assignment_vector), outcomes = d$views)
  ate = mean(ri[group==1]$outcomes) - mean(ri[group==0]$outcomes)
  ate_v[i]=ate

}


hajj_ri_distribution <- ate_v
```

3. How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE? Conduct your work in the code chunk below, saving the results into `hajj_count_larger`, but also support your coding with a narrative description. In that narrative description (and throughout), use R's "inline code chunks" to write your answer consistent with each time your run your code.

> 'Of the simulated random assignments, `18` are at least as large as the actual estimate of ATE.'

4. If there are `hajj_count_larger` randomizations that are larger than `hajj_ate`, what is the implied *one-tailed* p-value? Both write the code in the following chunk, and include a narrative description of the result following your code.

'The implied one-tailed p-value is `0.0018`, which is very small compared to the typical cutoff of p=0.05. It is unlikely that these results would arise due to random chance alone.'

5. Now, conduct a similar test, but for a two-sided p-value. You can either use two tests, one for larger than and another for smaller than; or, you can use an absolute value (`abs`). Both write the code in the following chunk, and include a narrative description of the result following your code.

'The two-tailed p-value is `0.0032`, which is still much smaller than the typical cutoff of p=0.05. It is unlikely that these results would arise due to random chance alone.'

2. Sports Cards

In this experiment, the experimenters invited consumers at a sports card trading show to bid against one other bidder for a pair trading cards. We abstract from the multi-unit-auction details here, and simply state that the treatment auction format was theoretically predicted to produce lower bids than the control auction format. We provide you a relevant subset of data from the experiment.

In this question, we are asking you to produce p-values and confidence intervals in three different ways:

1. Using a `t.test`;
2. Using a regression; and,
3. Using randomization inference.

1. Using a `t.test`, compute a 95% confidence interval for the difference between the treatment mean and the control mean. After you conduct your test, write a narrative statement, using inline code evaluation that describes what your tests find, and how you interpret these results. (You should be able to look into `str(t_test_cards)` to find the pieces that you want to pull to include in your written results.)

'We are 95% confident that the true ATE lies somewhere between `-20.8546241, -3.5571406`.'

2. In plain language, what does this confidence interval mean?

'The confidence interval indicates that if this experiment was performed many times, 95% of all randomizations would show a difference in means between `-20.8546241, -3.5571406`.'

3. Conduct a randomization inference process, with `n_ri_loops = 1000`, using an estimator that you write by hand (i.e. in the same way as earlier questions). On the sharp-null distribution that this process creates, compute the 2.5% quantile and the 97.5% quantile using the function `quantile` with the appropriate vector passed to the `probs` argument. This is the randomization-based uncertainty that is generated by your design. After you conduct your test, write a narrative statement of your test results.

Hide

```
n_ri_loops <- 1000
cards_ate             <- 'mean(treatment) - mean(control)''
cards_ri_distribution <- 'fill this in' # numeric vector of length equal
```

```
Error: unexpected string constant in:
"cards_ate             <- 'mean(treatment) - mean(control)''
cards_ri_distribution <- '"
```

'Narrative Statement'

4. Do you learn anything different if you regress the outcome on a binary treatment variable? To answer this question, regress `bid` on a binary variable equal to 0 for the control auction and 1 for the treatment auction and then calculate the 95% confidence interval using *classical standard errors* (in a moment you will calculate with *robust standard errors*). There are two ways to do this – you can code them by hand; or use a built-in, `confint`. After you conduct your test, write a narrative statement of your test results.

'We are 95% confident that the true ATE lies somewhere between `-20.844162` and `-3.5676027`.'

5. Calculate the 95% confidence interval using robust standard errors, using the `sandwich` package. There is a function in `lmtest` called `coefci` that can help with this. It is also possible to do this work by hand. After you conduct your test, write a narrative statement of your test results.

Hide

```
mod.vcovHC <- vcovHC(mod)
cards_robust_ci <- coefci(mod, vcov.=mod.vcovHC, level=0.95)
cards_robust_ci
```

```
                        2.5 %    97.5 %
(Intercept)           21.87899 35.768073
uniform_price_auction -20.97407 -3.437696
```

'We are 95% confident that the true ATE lies somewhere between `-20.9740683` and `-3.4376964`.'

6. Characterize what you learn from each of these different methods – are the results contingent on the method of analysis that you choose?

> 'While the values of the percentiles changed slightly from model to model (~0.1), the different methods of analysis did not result in meaningfully different results.'

Power Analysis

(Because there are a *lot* of ways to write this code, we're not going to write a tight testing suite against this question.)

Understanding whether your experiment design and data collection strategy are able to reject the null hypothesis *when they should* is valuable! And, this isn't theoretical value. If your design and data collection cannot reject the null hypothesis, why even run the experiment in the first place?

The classical formulation of power asks, "Given a test procedure and data, what proportion of the tests I *could conduct* would reject the null hypothesis?"

Imagine that you and David Reiley are going to revive the sports card experiment from the previous question. However, because it is for a class project, and because you've already spent all your money on a shiny new data science degree :raised_hands: :money_with_wings: , you're not going to be able to afford to recruit as many participants as before.

1. Describe a t-test based testing procedure that you might conduct for this experiment. What is your null hypothesis, and what would it take for you to reject this null hypothesis? (This second statement could either be in terms of p-values, or critical values.)

> 'Randomly divide a group of sports card enthusiasts into a treatment and control group. The treatment group will bid on sports cards in a uniform price auction, while the control will bid under some other common auction format. The null hypothesis will be that the average bid made by the control group will be equal to the average bid by the treatment group, indicating that the uniform price auction format did not produce an effect on bidding behavior. A two sample t-test will compare the mean bid values and reject the null hypothesis if the test yields a p-value less than 0.05.'

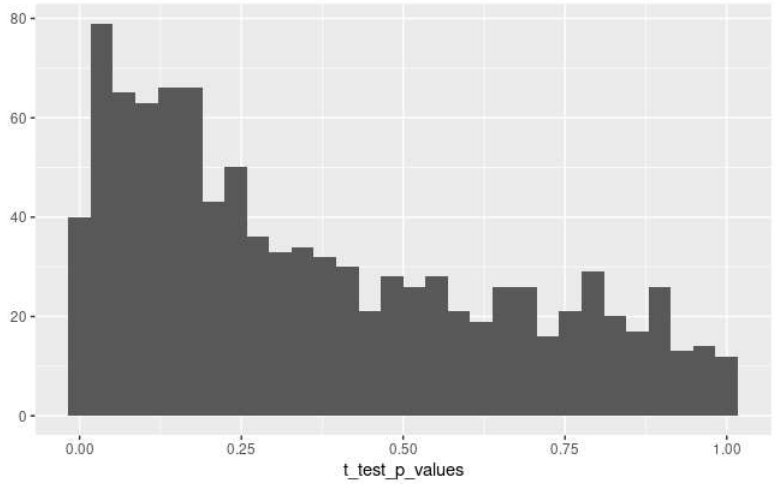
2. Suppose that you are only able to recruit 10 people to be a part of your experiment – 5 in treatment and another 5 in control. Simulate "re-conducting" the sports card experiment once by sampling from the data you previously collected, and conducting the test that you've written down in part 1 above. Given the results of this 10 person simulation, would your test reject the null hypothesis?

> 'This test yields a p= `0.2414274` , which would not reject the null.'

3. Now, repeat this process – sampling 10 people from your existing data and conducting the appropriate test – one-thousand times. Each time that you conduct this sample and test, pull the p-value from your t-test and store it in an object for later use. Consider whether your sampling process should sample with or without replacement.
4. Use `ggplot` and either `geom_hist()` or `geom_density()` to produce a distribution of your p-values, and describe what you see. What impression does this leave you with about the power of your test?

Hide

```
qplot(t_test_p_values, geom="histogram")
```



> 'The distribution of p-values is very spread out, which indicates that the test has low statistical power. I would be worried about my ability to identify a significant treatment and the risk of false discovery in this experiment.'

5. Suppose that you and David were to actually run this experiment and design – sample 10 people, conduct a t-test, and draw a conclusion. **And** suppose that when you get the data back, **lo and behold** it happens to reject the null hypothesis. Given the power that your design possesses, does the result seem reliable? Or, does it seem like it might be a false-positive result?

> ‘The broad distribution of p-values indicates that the experiment has low power. Low power increases the risk of false negatives and the risk of false discovery. Therefore, in this case I would be worried that the results we are seeing are false positive results.’

6. Apply the decision rule that you wrote down in part 1 above to each of the simulations you have conducted. What proportion of your simulations have rejected your null hypothesis? This is the p-value that this design and testing procedure generates. After you write and execute your code, include a narrative sentence or two about what you see.

> ‘ `11.3` % of simulations rejected the null hypothesis. Knowing that this experiment is low powered, I would have liked to see a higher rejection rate to be confident in the results of the experiment.’

7. Does buying more sample increase the power of your test? Apply the algorithm you have just written onto different sizes of data. Namely, conduct the exact same process that you have for 10 people, but now conduct the process for every 10% of recruitment size of the original data: Conduct a power analysis with a 10%, 20%, 30%, … 200% sample of the original data. (You could be more granular if you like, perhaps running this task for every 1% of the data).

Hide

```
#Apologies folks, ran out of time

percentages_to_sample <- 'fill this in'
```