Problem Set 2

Alex, Micah, Scott, and David

12/01/2021

Contents

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

1.1 State a null hypothesis

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

Answer: The treatment effect is zero for all subjects. In other words, there is no change in beliefs of subjects as a result of attending the Hajj.

1.2 Group by average

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 variables.

```
# the result should be a data.table with two columns and two rows
hajj_group_mean <- d[, .(mean_views = mean(views)), keyby = .(success)]
hajj_group_mean
```

```
##    success mean_views
## 1:       0   1.868304
## 2:       1   2.343137
```

```
# from the `hajj_group_mean` produce a single, numeric vector that is the ate.
# check that it is numeric using `class(hajj_ate)`
hajj_ate <- d[ , mean(views), keyby = .(success)][ , diff(V1)]
hajj_ate
```

```
## [1] 0.4748337
```

Answer: Those who won the lottery have generally a more positive view of 2.34 on average compared to those who did not win the lottery with a view of 1.87. Average treatment effect is 0.475.

```
## 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
## this should be a numeric vector that has a length equal to the number
## of RI permutations you ran
hajj_ri_distribution <- function(simulations = 10000) {

  res <- NA
  for(sim in 1:simulations) {
    res[sim] <- d[ , .(mean_views = mean(views)),
                   keyby = .(randomize(control, treatment))][ , diff(mean_views)]
  }
  return(res)
}
```

1.3 Randomization inference: At least as large

C. 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.

```r
# length 1 numeric vector from comparison of `hajj_ate` and `hajj_ri_distribution`
hajj_count_larger <- sum(dist_sharp_null >= hajj_ate)
hajj_count_larger
```

```
## [1] 19
```

Answer: 19 are at least as large as the actual ATE

1.4 Randomization inference: one-sided p-value

If there are `hajj_count_larger` (19) randomizations that are larger than `hajj_ate` (0.4748337), what is the *one-tailed* p-value? Both write the code in the following chunk, and include a narrative description of the result following your code.

```r
# length 1 numeric vector
hajj_one_tailed_p_value <- mean(dist_sharp_null < hajj_ate)
hajj_one_tailed_p_value
```

```
## [1] 0.9981
```

Answer: there is 0.9981 probability of observing a treatment effect smaller (in absolute scale) than what was observed, given that the sharp-null hypothesis were true

1.5 Randomization inference: two-sided p-value

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.

```r
# length 1 numeric vector
hajj_two_tailed_p_value <- mean(abs(dist_sharp_null) > abs(hajj_ate))
hajj_two_tailed_p_value
```

```
## [1] 0.0033
```

Answer: The p-value of 0.0033 is < 0.5. This leads us to the conclusion to reject the Sharp Null Hypothesis.

2 Sports Cards

2.1 t-test and confidence interval

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.)

```
# this should be the t.test object. Extract pieces from this object
# in-text below the code chunk.
t_test_cards <- d[, t.test(bid ~ uniform_price_auction)]
t_test_cards
```

```
##
##  Welch Two Sample t-test
##
## data:  bid by uniform_price_auction
## t = 2.8211, df = 61.983, p-value = 0.006421
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   3.557141 20.854624
## sample estimates:
## mean in group 0 mean in group 1
##        28.82353        16.61765
```

```
str(t_test_cards)
```

```
## List of 10
##  $ statistic  : Named num 2.82
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named num 62
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0.00642
##  $ conf.int   : num [1:2] 3.56 20.85
##   ..- attr(*, "conf.level")= num 0.95
##  $ estimate   : Named num [1:2] 28.8 16.6
##   ..- attr(*, "names")= chr [1:2] "mean in group 0" "mean in group 1"
##  $ null.value : Named num 0
##   ..- attr(*, "names")= chr "difference in means between group 0 and group 1"
##  $ stderr     : num 4.33
##  $ alternative: chr "two.sided"
##  $ method     : chr "Welch Two Sample t-test"
##  $ data.name  : chr "bid by uniform_price_auction"
##  - attr(*, "class")= chr "htest"
```

Narrative Analysis: ... The p-value is that leads us to reject the null hypothesis that there is no difference between the means of two groups. Additionally, the confidence interval is between [3.56, 20.85]. Which means that 95% of the time, the difference in means would be in range [3.56, 20.85].

2.2 Interpretation of confidence interval

In your own words, what does this confidence interval mean? This can be simple language, but it has to be statistically appropriate language.

Answer: In simple language, we can see that the 95% of the time the difference in means that between our control and treatment groups falls in between 3.5 dollars and 20.9 dollars. In other words, uniform_price_auction increases increases bids between 3.5 to 20.9 dollars. While useful, this large of a confidence suggests to have an experiment with a larger sample to gain more precise results.

2.3 Randomization inference, and confidence interval?

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.

```r
## first, do you work for the randomization inference

#decided not to use these method.
# n_ri_loops <- 1000
#
# cards_ate              <- 'fill this in'
# cards_ri_distribution <- 'fill this in' # numeric vector of length equal
#                                          # to your number of RI permutations
# cards_ri_quantiles    <- 'fill this in' # there's a built-in to pull these.
# cards_ri_p_value      <- 'fill this in'
```

```r
# length 1 numeric vector
cards_two_tailed_p_value <- mean(abs(dist_sharp_null) > abs(cards_ate))
cards_two_tailed_p_value
```

```
## [1] 0.004
```

Narrative: ... 2.5% of simulated ATEs is smaller than -8.26 and 2.5% of simulated ATEs are larger 9.26. We can see that the actual data ATE of -12.2058824 falls below the 2.5% of the quantile, hinting to us to reject the sharp null hypothesis.

2.4 Compare regression and randomization inference

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.

```r
# this should be a model object, class = 'lm'.

mod <- lm(bid ~ uniform_price_auction, data = d)
summary(mod)
```

```
##
## Call:
## lm(formula = bid ~ uniform_price_auction, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.824 -11.618  -3.221   8.382  58.382
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             28.824      3.059   9.421 7.81e-14 ***
## uniform_price_auction  -12.206      4.327  -2.821  0.00631 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 17.84 on 66 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.09409
## F-statistic: 7.959 on 1 and 66 DF,  p-value: 0.006315
```

Narrative: ... he p-value of for t.test () and regression (0.00631) methods are very close and so are the confidence intervals. The two tailed p-value for randomization inference, on the other hand, seems to be smaller with a value 0.004. The p-value for randomization inference, as expected, varies with each simulation. (Note: I have set a seed)

2.5 Regression with robust confidence interval

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.

```
# this should be a numeric vector of length 2
cards_robust_ci <- coefci(mod, "uniform_price_auction", vcov. = vcovHC(mod))
cards_robust_ci
```

```
##                           2.5 %    97.5 %
## uniform_price_auction -20.97407 -3.437696
```

```
#?coefci
```

Narrative: ... As we can see above, the robust confidence interval is similar but not the same to the confidence interval created by t-test.

2.6 Compare and contrast results

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

Answer: …The p-value of for t.test () and regression (0.00631) methods are very close and so are the confidence intervals. The two tailed p-value for randomization inference, on the other hand, seems to be smaller with a value 0.004. The p-value for randomization inference, as expected, varies with each simulation. The sample is not big, in this case randomization inference should give us more accurate result since we are simulating the data a 1000 times. If the sample size was large enough, the p-values from regression and randomization inference would be very similar.

Point to note: if the data meets the assumptions of all 3 methods, then we should be able to produce similar models. However, if the data does not meet the assumptions of one model, then other methods should be taken into consideration.

3 Power Analysis

3.1 Describe your testing procedure

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.)

Answer: For my experiment, the null hypothesis would be that there is zero difference in the average bids placed for control and treatment auction groups. Since this experiment would be for a class project, I am increasing the alpha level as 0.13 instead of traditional 0.05. If my p-value is below 0.13, then I would reject the null hypothesis. I am willing to take weak evidence to reject my null so demonstrate that the alpha level can changed quite easily and that in return changes our ability to reject or fail to reject the null.

3.2 Suppose you only had ten subjects, what would you learn

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 should be a test object

set.seed(223)
t_test_ten_people <- function(data =d, size_in_each_group = 5) {
  d[, .(bid = sample(bid, 5)), by = uniform_price_auction] %$%
    t.test(bid ~ uniform_price_auction) %$%
    p.value
}

t_test_ten_people()
```

```
## [1] 0.8546946
```

Answer: The p-value is 0.855 which is less than 1.0 threshold that we set above (Note: I have set a seed and without a seed, each time you run above function, we would get a different p-value)

3.3 With only ten subjects, what is your power?

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.*

```
# fill this in with the p-values from your power analysis
#t_test_p_values <- rep(NA, 1000)
## you can either write a for loop, use an apply method, or use replicate
## (which is an easy-of-use wrapper to an apply method)


t_test_p_values <- replicate(1000, expr = t_test_ten_people())
```

```
#?replicate

head(t_test_p_values)
```

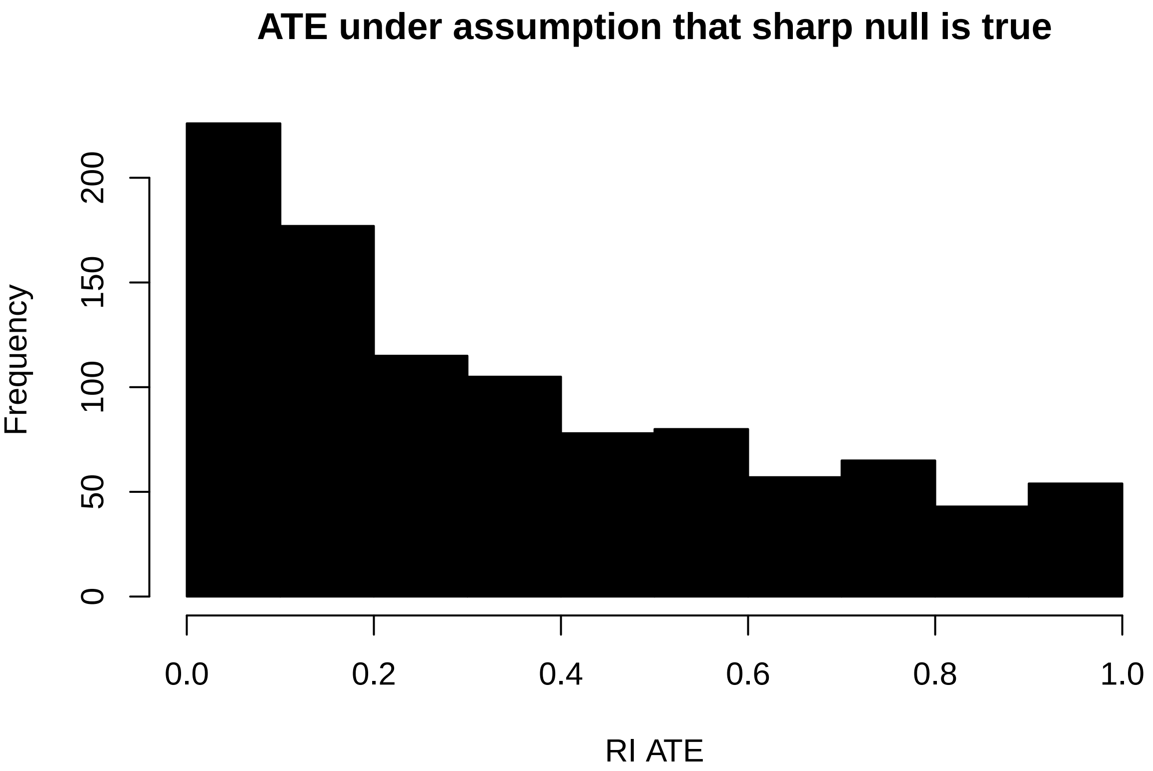
```
## [1] 0.3437506 0.4988245 0.2272396 0.5554555 0.8589445 0.5347150
```

Answer: Each replication spits out a p-value which are stored in t_test_p_values table.

3.4 Visual analysis

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?*

```
hist(
  t_test_p_values,
  col = 'black',
  xlab = 'RI ATE',
  main = 'ATE under assumption that sharp null is true'
  )
```



Answer: From the graph above, we can see that the distribution is skewed to the left and that most p-values generated are less 0.2.

3.5 Interpret your results, given your power

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?

```
#counting how many of the p-values in the distribution would be below my alpha level of 0.13
counter <- sum(t_test_p_values < 0.13 )
counter
```

```
## [1] 279
```

Answer: This seems like a false positive result. Given my experiment design and alpha level, ~28% of the time, I would have rejected null hypothesis and about ~82% of the time I would have failed to reject the null. Based on these values, the results seem quite uninformative.

3.6 Conduct a power analysis

F. 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.

```
t_test_rejects <- sum(t_test_p_values < 0.13 )
t_test_rejects
```

```
## [1] 279
```

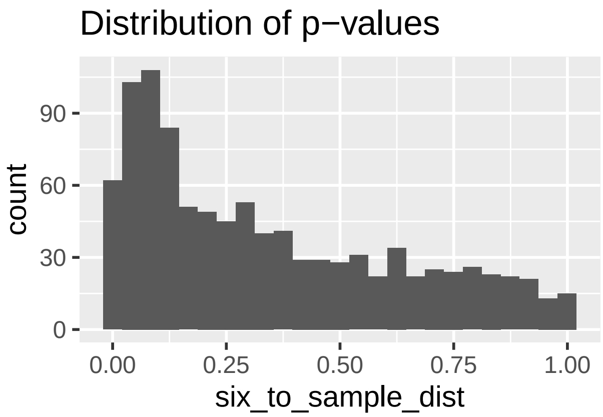
Answer: In 279 out of a 1000 simulations, I would have rejected the null.

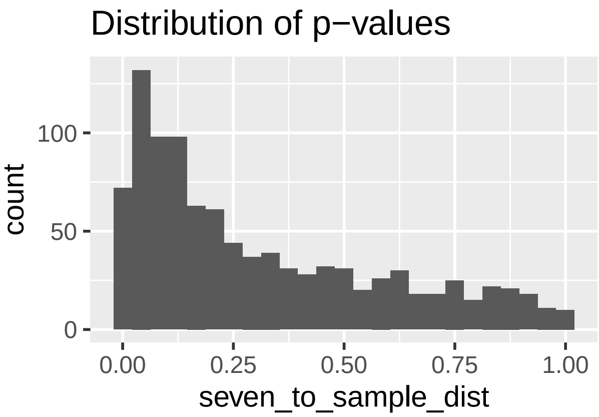
3.7 Moar power!

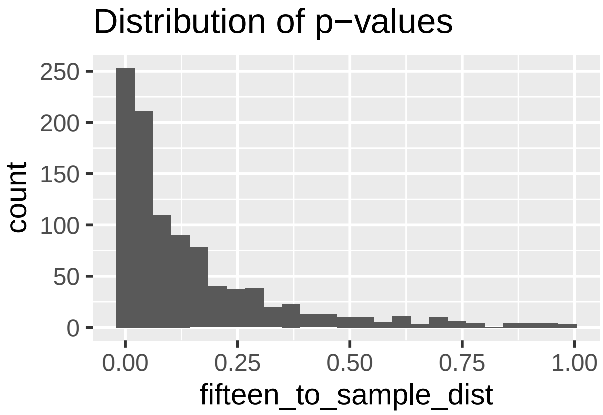
Does buying more sample increase the power of your test? Apply the algorithm you have just written onto different sizes of data. Namely, conduct the exact same process that you have for 10 people, but now conduct the process for every 10% of recruitment size of the original data: Conduct a power analysis 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).

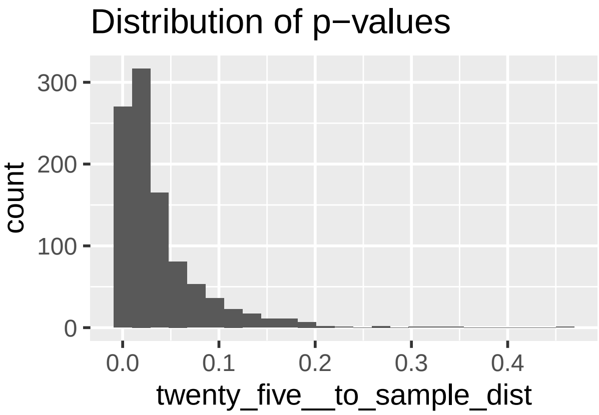
```
library(gridExtra)
#increasing the sample size, running the 1000 simulations and then graphing p-value distributions

#having 6 controls and 6 treatments
six_to_sample <- function(data =d, size_in_each_group =6) {
  d[, .(bid = sample(bid, 6)), by = uniform_price_auction] %$%
    t.test(bid ~ uniform_price_auction) %$%
    p.value
}

six_to_sample_dist <- replicate(1000, expr = six_to_sample())

six_to_sample_dist_plot <- ggplot() + aes(six_to_sample_dist) + geom_histogram(bins = 25) + labs(title =

#having 7 controls and 7 treatments
```

```
seven_to_sample <- function(data =d, size_in_each_group =7) {
  d[, .(bid = sample(bid, 7)), by = uniform_price_auction] %$%
    t.test(bid ~ uniform_price_auction) %$%
    p.value
}

seven_to_sample_dist <- replicate(1000, expr = seven_to_sample())

seven_to_sample_dist_plot <- ggplot() + aes(seven_to_sample_dist) + geom_histogram(bins = 25) + labs(ti



#having 15 controls and 15 treatments
fifteen_to_sample <- function(data =d, size_in_each_group =15) {
  d[, .(bid = sample(bid, 15)), by = uniform_price_auction] %$%
    t.test(bid ~ uniform_price_auction) %$%
    p.value
}

fifteen_to_sample_dist <- replicate(1000, expr = fifteen_to_sample())

fifteen_to_sample_dist_plot <- ggplot() + aes(fifteen_to_sample_dist) + geom_histogram(bins = 25) + labs


#having 25 controls and 25 treatments
twenty_five__to_sample <- function(data =d, size_in_each_group =25) {
  d[, .(bid = sample(bid, 25)), by = uniform_price_auction] %$%
    t.test(bid ~ uniform_price_auction) %$%
    p.value
}

twenty_five__to_sample_dist <- replicate(1000, expr = twenty_five__to_sample())

twenty_five__to_sample_dist_plot <- ggplot() + aes(twenty_five__to_sample_dist) + geom_histogram(bins =


grid.arrange(six_to_sample_dist_plot, seven_to_sample_dist_plot, fifteen_to_sample_dist_plot, twenty_fi
```









```
six_to_sample_t_test_reject         <- sum(six_to_sample_dist < 0.13 )
seven_to_sample_t_test_reject       <- sum(seven_to_sample_dist < 0.13 )
fifteen_to_sample_t_test_reject     <- sum(fifteen_to_sample_dist < 0.13 )
twenty_five_to_sample_t_test_reject <- sum(twenty_five__to_sample_dist < 0.13 )

six_to_sample_t_test_reject
```

```
## [1] 324
```

```
seven_to_sample_t_test_reject
```

```
## [1] 368
```

```
fifteen_to_sample_t_test_reject
```

```
## [1] 636
```

```
twenty_five_to_sample_t_test_reject
```

```
## [1] 948
```

Answer: As we can see above, increasing the sample size does need increase the power. In other words, increasing the sample size, increases statistical power and decrease type II errors.