Lab 1: Hypothesis Testing

Adam Weintraut

This is a team-based lab. Your instructor will divide you into teams of three, or possibly four students. To maximize learning, we would like all students to engage with every lab component, discussing strategy with teammates, reviewing solutions, and iterating on text and code.

The lab consists of two parts. Part one consists of foundation exercises similar to a homework. Part two is a written statistical analysis.

This lab is due before your Unit 9 live session. You will find a separate place on Gradescope to submit each part. Please submit your `.Rmd` file, as well as your `.pdf` file. This is a team submission, so one person from your team should submit the teams' work, and associate the other members of the team with the submission (this is intuitive on Gradescope.)

Contents

1 Part 1: Foundational Exercises **2**
- 1.1 Professional Magic . . . 2
- 1.2 Wrong Test, Right Data . . . 3
- 1.3 Test Assumptions . . . 4
  - 1.3.1 World Happiness . . . 4
  - 1.3.2 Legislators . . . 6
  - 1.3.3 Wine and health . . . 8
  - 1.3.4 Attitudes toward the religious . . . 9

2 Part 2: Statistical Analysis **11**
- 2.1 Data . . . 11
- 2.2 The research question . . . 14
- 2.3 Guidance from political scientists . . . 14
- 2.4 Report guidelines . . . 14
- General guidance . . . 14
- Introduction . . . 14
- Conceptualization and Operationalization . . . 14
- Visual Design . . . 15
- Data wrangling . . . 15
- Hypothesis testing . . . 15
- Test, results and interpretation . . . 16

1 Part 1: Foundational Exercises

1.1 Professional Magic

Your aunt (who is a professional magician), claims to have created a pair of magical coins that share a connection to each other that makes them land in the same way. The coins are always flipped at the same time. For a given flip $i \in \{1, 2, 3...\}$, let X_i be a Bernoulli random variable representing the outcome of the first coin, and let Y_i be a Bernoulli random variable representing the outcome of the second coin. You assume that each flip of the pair is independent of all other flips of the pair. For all i, you also assume that X_i and Y_i have the joint distribution given in the following table.

	$X_i = 0$	$X_i = 1$
$Y_i = 0$	$p/2$	$(1-p)/2$
$Y_i = 1$	$(1-p)/2$	$p/2$

$p \in [0, 1]$ is a parameter.

Each flip of the pair is independent of all other flips of the pair. This means that whatever happens the first time that you flip both of the coins tells you nothing about the second time that you flip both of the coins, and so on. Essentially, this is a statement about the limits of your aunt's magic.

You design the following test to evaluate your aunt's claim: You flip the coins three times, and write down that your test statistic is the sum $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$. That is, your test statistic is essentially the number of heads that are shown.

Your null hypothesis is that $p = 1/2$, and you plan to reject the null if your test statistic is 0 or 6.

1. What is the type 1 error rate of your test? .03125
2. What is the power of your test for the alternate hypothesis that $p = 3/4$? .105

1.2 Wrong Test, Right Data

Imagine that your organization surveys a set of customers to see how much they like your regular website, and how much they like your mobile website. Suppose that both of these preference statements are measured on 5-point Likert scales.

A Likert scale is one where a person is provided ordered categories that range from lowest to highest. You can read more about them in this seminal research design text by Fowler, or this brief overview. If you were to run a paired t-test using this data, what consequences would the violation of the metric scale assumption have for your interpretation of the test results? What would you propose to do to remedy this problem?

Using a paired test is correct, because our null hypothesis should be that there is no difference between how much a person likes the mobile website and the regular website. Additionally, each person answers the questions, so the data is dependent by person.

However, Likert Scales have ordinal data, and the t-test assumes metric data since it seeks to find the differences between the means of the samples. This means that we would need to make the assumption that the Likert Scale's variations in answers all mean the same thing, which they don't. This can be seen easily when we examine the difference between a 2 and a 3, versus a 4 and a 5 on the Scale. Additionally, people rank their emotions differently, so we care more about what the difference between their two rankings is.

Since we don't want to measure for the mean, and we have ordinal data, we should use Wilcoxon Paired Test instead. This will measure for paired differences between the samples instead of paired means, which is a better way to compare customer's sentiment. It also can handle signed-rank data, which is important for this "difference" calculation.

1.3 Test Assumptions

For the four following questions, your task is to evaluate the assumptions for the given test. It is not enough to say that an assumption is met or not met; instead, present your evidence in the form of background knowledge, visualizations, and numerical summaries. If you produce a histogram as part of your evaluation, be sure to consider what the most appropriate bin width is. The test that we ask you to evaluate may or may not be the most appropriate test for the scenario. Because the goal of this task is to evaluate whether the data satisfies the assumptions necessary for the test to provide meaningful results, you do not need perform the test (you may perform the test, but we will not be marking for the test results).

1.3.1 World Happiness

The file `datasets/Happiness_WHR.csv` is subsetted from the World Happiness Report, a yearly publication that uses data from the Gallup World Poll surveys. The variable life ladder is a measure of happiness, described in the FAQ as follows:

> This is called the Cantril ladder: it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The rankings are from nationally representative samples, for the years 2018-2020.
>
> This tries to make it so that this is a metric, but it still isn't.

You would like to know whether people in countries with high GDP per capita (higher than the mean) are more happy or less happy than people in countries with low GDP (lower than the mean).

List all assumptions for a two-sample t-test. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

> iid paired or unpaired We should use an unpaired test, we have split into "high gdp" and "low gdp", based on the mean. Since these sets are disjoint, we
>
> metric data Although their preface aims to make it metric, I believe this data is ordinal because it is a ranking. Using a t-test is to test the mean. normal distribution n<30 -> must be normalish n>30 -> CLT
>
> Mann-Whitney Test: compares the medians of the two samples, more applicable for ranking.

```r
happy <- read.csv('./datasets/happiness_WHR.csv')

summary(happy)
```

```
##  Country.name            year       Life.Ladder    Log.GDP.per.capita
##  Length:239         Min.   :2019   Min.   :2.375   Min.   : 6.966
##  Class :character   1st Qu.:2019   1st Qu.:4.971   1st Qu.: 8.827
##  Mode  :character   Median :2019   Median :5.768   Median : 9.669
##                     Mean   :2019   Mean   :5.678   Mean   : 9.584
##                     3rd Qu.:2020   3rd Qu.:6.428   3rd Qu.:10.527
##                     Max.   :2020   Max.   :7.889   Max.   :11.648
##                                                    NA's   :13
##  Social.support   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices
##  Min.   :0.4200   Min.   :48.70                    Min.   :0.3850
##  1st Qu.:0.7590   1st Qu.:62.00                    1st Qu.:0.7360
##  Median :0.8480   Median :67.20                    Median :0.8220
##  Mean   :0.8256   Mean   :65.84                    Mean   :0.8038
##  3rd Qu.:0.9175   3rd Qu.:70.45                    3rd Qu.:0.8910
##  Max.   :0.9830   Max.   :77.10                    Max.   :0.9700
##                   NA's   :8                        NA's   :2
##    Generosity       Perceptions.of.corruption Positive.affect  Negative.affect
```

```
##  Min.   :-0.28900   Min.   :0.0700            Min.   :0.3220   Min.   :0.0830
##  1st Qu.:-0.11900   1st Qu.:0.6470            1st Qu.:0.6450   1st Qu.:0.2250
##  Median :-0.04500   Median :0.7780            Median :0.7330   Median :0.2830
##  Mean   :-0.01524   Mean   :0.7168            Mean   :0.7149   Mean   :0.2891
##  3rd Qu.: 0.07400   3rd Qu.:0.8480            3rd Qu.:0.7900   3rd Qu.:0.3440
##  Max.   : 0.56100   Max.   :0.9630            Max.   :0.8910   Max.   :0.5320
##  NA's   :14         NA's   :14                NA's   :2        NA's   :2
```

```r
non_null_gdp <- !is.na(happy$Log.GDP.per.capita)

# find mean --
## is our mean_gdp calculated  accurately? log(mean) or mean(log) ?
mean_log_gdp <- mean(happy$Log.GDP.per.capita[non_null_gdp])
print(paste0("mean_log_gdp: ", mean_log_gdp))
```

```
## [1] "mean_log_gdp: 9.58383628318584"
```

```r
# create flag for high gdp
happy$high_gdp_f <- happy$Log.GDP.per.capita > mean_log_gdp

# high gdp sample size - 121
high_gdp_sample_size <- sum(!is.na(happy$Log.GDP.per.capita[happy$high_gdp_f==T]))
print(paste0("high gdp sample size: ", high_gdp_sample_size))
```

```
## [1] "high gdp sample size: 121"
```

```r
# low gdp sample size - 105
low_gdp_sample_size <- sum(!is.na(happy$Log.GDP.per.capita[happy$high_gdp_f==F]))
print(paste0("low gdp sample size: ", low_gdp_sample_size))
```

```
## [1] "low gdp sample size: 105"
```

```r
## is our mean_gdp calculated  accurately?
# For 2 sampled t-test, we need to assume
# - iid
#       - disjoint samples based on mean gdp cutoff
# - continuous
#       - gdp, log, seems continuous
# - normally distributed
#       - n < 30 : need normal distribution
#       - n > 30 : CLT
# - parametric or non-parametric?
# - paired or unpaired?

happy$Life.Ladder
```

```
##   [1] 2.375 4.995 4.745 6.086 5.488 7.234 7.195 5.173 7.098 5.114 5.821 6.772
##  [13] 4.976 5.674 6.016 3.471 6.451 5.108 4.741 4.998 4.937 7.109 4.251 5.942
##  [25] 5.144 6.350 4.609 5.213 6.998 5.626 6.137 7.693 6.004 5.809 4.328 6.455
##  [37] 6.035 4.100 7.780 6.690 4.914 5.164 4.892 7.035 4.967 5.952 6.262 4.768
##  [49] 5.930 5.659 6.000 7.533 3.249 5.347 5.006 7.255 7.332 6.445 5.392 6.309
##  [61] 5.908 4.453 6.272 4.619 6.425 6.106 5.685 5.197 5.970 4.024 3.512 5.121
##  [73] 5.330 6.064 7.404 4.339 3.869 5.428 4.988 6.733 4.153 6.241 6.432 5.803
##  [85] 5.563 5.386 5.057 4.932 4.434 4.436 5.449 7.425 7.205 6.113 5.004 4.356
##  [97] 5.467 5.015 7.442 4.443 4.483 6.086 5.653 5.999 6.268 6.242 6.095 6.130
## [109] 5.441 3.268 6.561 5.489 6.241 3.447 6.378 6.243 6.665 5.035 5.903 6.457
## [121] 4.213 4.396 7.398 7.694 6.537 5.464 3.640 6.022 4.179 4.315 4.872 5.474
```

```
## [133] 4.948 4.702 6.711 7.157 6.944 6.600 6.154 5.081 5.467 4.197 3.307 2.694
## [145] 5.365 5.901 7.137 7.213 6.173 5.280 6.839 4.408 5.559 5.516 6.110 5.598
## [157] 4.377 5.241 7.025 6.151 5.771 5.709 6.508 6.260 6.897 7.515 5.168 5.354
## [169] 4.472 5.462 6.453 4.549 7.889 6.714 5.123 7.312 5.319 5.788 5.295 6.038
## [181] 7.575 4.225 4.865 4.785 7.035 7.195 6.488 5.257 6.118 4.094 6.168 4.547
## [193] 6.294 6.250 5.284 6.229 6.391 6.157 6.015 5.964 5.812 6.011 5.722 4.803
## [205] 4.431 4.451 7.504 7.257 5.503 5.054 7.290 5.080 6.139 5.768 5.495 6.560
## [217] 6.042 6.519 6.462 4.947 5.793 6.502 7.314 7.508 6.751 5.373 3.786 5.885
## [229] 4.731 4.862 4.641 5.270 6.458 6.798 7.028 6.310 4.574 4.838 3.160
```

```r
afg_ss <- sum(!is.na(happy$Life.Ladder[happy$Country.name=='Afghanistan']))
```

1.3.2 Legislators

The file `datasets/legislators-current.csv` is taken from the congress-legislators project on Github. You would like to test whether Democratic or Republican senators are older.

List all assumptions for a Wilcoxon rank-sum test. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

- get all senators and their ages (from birthday)
- split up into democrats / republicans
-
- Could test on bucketing on age groups to define this
- Wicoxon rank-sum
  - NO REQUIREMENT FOR NORMALITY, we don't care for this one.
- i.i.d.
  - CLT - test sample size for each group > 30
  - identical - distribution is same as what you test
  - independent - one sample has nothing to do with the other. you can't change their age
- ordinal
  - why is the age ordinal? is it?
    - argument for: age is an integer value, not continuous (unless we include months, is this natural?)
    - argument against: age is a metric and its distance is valuable.
- paired or unpaired?
  - unpaired,

```r
legs <- read.csv('./datasets/legislators-current.csv')
summary(legs)
```

```
##   last_name          first_name         middle_name           suffix
##  Length:538         Length:538         Length:538         Length:538
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    nickname          full_name           birthday            gender
##  Length:538         Length:538         Length:538         Length:538
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      type              state              district       senate_class
##  Length:538         Length:538         Min.   : 0.000   Min.   :1.00
##  Class :character   Class :character   1st Qu.: 3.000   1st Qu.:1.00
##  Mode  :character   Mode  :character   Median : 6.000   Median :2.00
##                                        Mean   : 9.984   Mean   :2.01
##                                        3rd Qu.:13.000   3rd Qu.:3.00
##                                        Max.   :53.000   Max.   :3.00
##                                        NA's   :100      NA's   :438
##     party               url              address             phone
##  Length:538         Length:538         Length:538         Length:538
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  contact_form         rss_url            twitter            facebook
##  Length:538         Length:538         Length:538         Length:538
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    youtube           youtube_id        bioguide_id          thomas_id
##  Length:538         Length:538         Length:538         Min.   :  91
##  Class :character   Class :character   Class :character   1st Qu.:1635
##  Mode  :character   Mode  :character   Mode  :character   Median :1922
##                                                           Mean   :1750
##                                                           3rd Qu.:2126
##                                                           Max.   :2296
##                                                           NA's   :209
##  opensecrets_id        lis_id            fec_ids            cspan_id
##  Length:538         Length:538         Length:538         Min.   :    260
##  Class :character   Class :character   Class :character   1st Qu.:  45591
##  Mode  :character   Mode  :character   Mode  :character   Median :  79718
##                                                           Mean   : 543374
##                                                           3rd Qu.:1003305
##                                                           Max.   :9275683
##                                                           NA's   :156
##   govtrack_id      votesmart_id    ballotpedia_id     washington_post_id
##  Min.   :300018   Min.   :   119   Length:538         Mode:logical
##  1st Qu.:412199   1st Qu.: 22411   Class :character   NA's:538
##  Median :412570   Median : 52964   Mode  :character
##  Mean   :412042   Mean   : 75411
##  3rd Qu.:412772   3rd Qu.:133024
##  Max.   :456862   Max.   :188334
##                   NA's   :62
```

```
##     icpsr_id      wikipedia_id
##  Min.   :14066   Length:538
##  1st Qu.:21106   Class :character
##  Median :21564   Mode  :character
##  Mean   :24264
##  3rd Qu.:21972
##  Max.   :94659
##  NA's   :77
```

- Wicoxon rank-sum ordinal var i.i.d.

1.3.3 Wine and health

The dataset `wine` can be accessed by installing the `wooldridge` package.

```
install.packages("wooldridge")
library(wooldridge)
?wine
wine
```

It contains observations of variables related to wine consumption for 21 countries. You would like to use this data to test whether countries have more deaths from heart disease or from liver disease.

List all assumptions for a signed-rank test. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

Wicoxon signed-rank - paired - this is true, each row has the same value

- metric var
  - continuous value, death counts.
- i.i.d.
  - each country's deaths are independent to one another, but it's not id distributed since we only have 21 countries out of all the countries in the world. If we are trying to use this data to test whether countries have more deaths from heart or liver, we don't have enough data to represent the other countries.
- difference is symmetric
  - not sure.
  - continuous symmetric distribution?
  - how do we prove this without actually doing the test???
  -

```
# install.packages("wooldridge")
library(wooldridge)

?wine
wine
```

```
##          country alcohol deaths heart liver
## 1      Australia     2.5    785   211  15.3
## 2        Austria     3.9    863   167  45.6
## 3       Belg/Lux     2.9    883   131  20.7
## 4         Canada     2.4    793   191  16.4
## 5        Denmark     2.9    971   220  23.9
## 6        Finland     0.8    970   297  19.0
## 7         France     9.1    751    71  37.9
## 8        Iceland     0.8    743   211  11.2
## 9        Ireland     0.7   1000   300   6.5
## 10        Israel     0.6    834   183  13.7
```

```
## 11       Italy     7.9    775   107  42.2
## 12       Japan     1.5    680    36  23.2
## 13  Netherlands     1.8    773   167   9.2
## 14  New Zealand     1.9    916   266   7.7
## 15      Norway     0.8    806   227  12.2
## 16       Spain     6.5    724    86  36.4
## 17      Sweden     1.6    743   207  11.2
## 18  Switzerland     5.8    693   115  20.3
## 19          UK     1.3    941   285  10.3
## 20          US     1.2    926   199  22.1
## 21 West Germany     2.7    861   172  36.7
```

```
summary(wine)
```

```
##    country             alcohol          deaths          heart      
##  Length:21          Min.   :0.600   Min.   : 680   Min.   : 36.0  
##  Class :character   1st Qu.:1.200   1st Qu.: 751   1st Qu.:131.0  
##  Mode  :character   Median :1.900   Median : 806   Median :191.0  
##                     Mean   :2.838   Mean   : 830   Mean   :183.3  
##                     3rd Qu.:2.900   3rd Qu.: 916   3rd Qu.:220.0  
##                     Max.   :9.100   Max.   :1000   Max.   :300.0  
##      liver      
##  Min.   : 6.50  
##  1st Qu.:11.20  
##  Median :19.00  
##  Mean   :21.03  
##  3rd Qu.:23.90  
##  Max.   :45.60  
```

1.3.4 Attitudes toward the religious

The file `datasets/GSS_religion` is a subset of data from the 2004 General Social Survey (GSS).

The variables `prottemp` and `cathtemp` are measurements of how a respondent feels towards protestants and towards Catholics, respectively. The GSS questions are phrased as follows:

> I'd like to get your feelings toward groups that are in the news these days. I will use something we call the feeling thermometer, and here is how it works:
>
> I'll read the names of a group and I'd like you to rate that group using the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. If we come to a group whose name you Don't recognize, you don't need to rate that group. Just tell me and we'll move on to the next one. If you do recognize the name, but you don't feel particularly warm or cold toward the group, you would rate the group at the 50 degree mark.
>
> How would you rate this group using the thermometer?

You would like to test whether the US population feels more positive towards Protestants or towards Catholics.

List all assumptions for a paired t-test. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

paired t-test? - should be paired, each person has values for each (or gives the metrics) - same subject for each scenario

metric? - argue ordinal instead. - discrete - ranking (sentiment, feeling) - similar to ladder. - how do we explain the "difference" between values? - "how much in favor" hard to quantify i.i.d ? - conditional approval but not enough info bla bla bla - independent - we think yes? but they need to talk about how they sampled - are the people in the same household? - demographic constraints? - identically distributed - Does this represent the entire US Population? Look at the survey to see if we are accurately representing the US pop. - maybe, can't guarantee, not enough data here.

```r
religion = read.csv('./datasets/GSS_religion.csv')
summary(religion)
```

```
##        X              year            id            prottemp     
##  Min.   :  1.0   Min.   :2004   Min.   :   4.0   Min.   :  0.00  
##  1st Qu.:201.2   1st Qu.:2004   1st Qu.: 728.8   1st Qu.: 50.00  
##  Median :401.5   Median :2004   Median :1373.5   Median : 60.00  
##  Mean   :401.5   Mean   :2004   Mean   :1381.9   Mean   : 65.56  
##  3rd Qu.:601.8   3rd Qu.:2004   3rd Qu.:2053.5   3rd Qu.: 85.00  
##  Max.   :802.0   Max.   :2004   Max.   :2808.0   Max.   :100.00  
##     cathtemp     
##  Min.   :  0.00  
##  1st Qu.: 50.00  
##  Median : 60.00  
##  Mean   : 63.16  
##  3rd Qu.: 85.00  
##  Max.   :100.00
```

```r
# possible to recognize one group?
# did they not include them in the sample we see? - this makes sense
# are there any rows with a null catholic and a protestant
sum(is.na(religion$cathtemp))
```

```
## [1] 0
```

```r
sum(is.na(religion$prottemp))
```

```
## [1] 0
```

2 Part 2: Statistical Analysis

The American National Election Studies (ANES) conducts surveys of voters in the United States, with a flagship survey occurring immediately before and after each presidential election. In this part, you will use the ANES data to address a question about voters in the US. Your team will conduct a statistical analysis and generate a written report in pdf format.

This is an exercise in both statistics and professional communication. It is important that your techniques are properly executed; equally important is that your writing is clear and organized, and your argument well justified.

2.1 Data

Data for the lab should be drawn from the 2020 American National Election Studies (ANES). You can access this data at https://electionstudies.org. This is the official site of the ANES, a project that has been ongoing since 1948, and federally funded by the National Science Foundation since 1977.

To access the data, you will need to register for an account, confirm this account, and then login. The data that you need should come from the **2020 Time Series Study**.

You will note that there are two forms of data that are available, data that is stored in a `.dta` format, and data that is stored in a `.sav` format. Both of these are proprietary data formats (`.dta` for STATA, and `.sav` for SPSS). You will need to find an appropriate library to read this data into R; we recommend that you find a package that is within the "tidyverse".

While you're at the ANES website, you will also want to download the codebook, because all of the variables are marked as something like, `V200002` – which isn't very descriptive without the codebook.

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the codebook.

Like many modern surveys, the ANES includes survey weights, which are used to correct for situations in which members of one demographic group are more likely to respond to the survey than members of another demographic group. (The target proportions are ultimately based on US census data). The survey weights make it possible to generalize from the a population that represents people who take the survey to a population that represents the United States as a whole. These weights are beyond the scope of our class and you are not expected to utilize them. You will still be able to learn about a population model, even if applicability to the US population is limited.

```r
all_data <- read.csv("./datasets/anes_timeseries_2020_csv_20220210.csv")


all_data$how_voted <- all_data$V201024
all_data$reg_deadline <- all_data$V202114a # did not meet the registration deadline
all_data$reg_nowhere <- all_data$V202114b # did not know where or how to register
all_data$reg_residency <-  all_data$V202114c # did not meet residency reqs
all_data$reg_form <- all_data$V202114d # registration form not filed correctly
all_data$reg_id <- all_data$V202114e # did not have required identification
all_data$reg_no_interest <- all_data$V202114f # had no interest in the election
all_data$reg_no_difference <- all_data$V202114g # my vote wouldn't make a difference
all_data$reg_disability <- all_data$V202114h # had a permanent disability or illness
all_data$reg_no_english <- all_data$V202114i # has difficulty with english
all_data$reg_not_eligible <- all_data$V202114j # not eligible to vote
all_data$reg_other <- all_data$V202114k # other

all_data$party <- all_data$V201018
all_data$difficulty_scale <- all_data$V202119 # Likert Scale on Difficulty
```

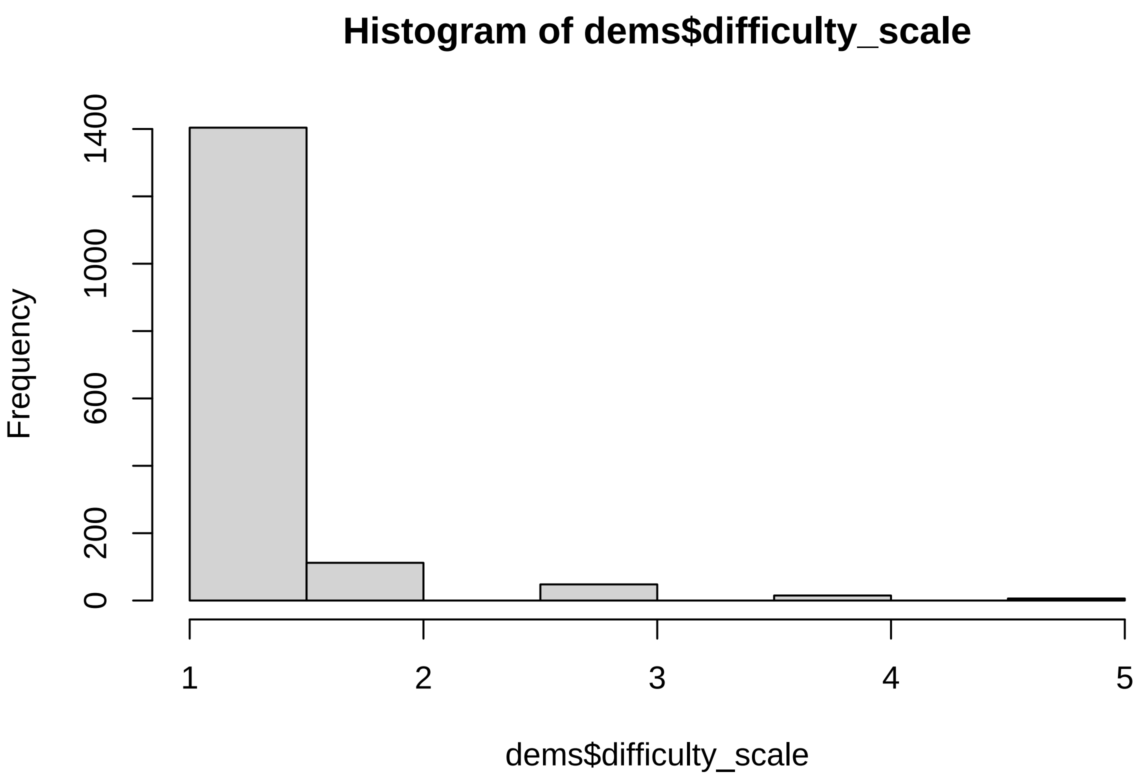
```r
data <- all_data[, c('party', 'difficulty_scale')]

# data

# democrat = 1
dems <- data[((data$party == 1) & (data$difficulty_scale > 0)),]
# republican
reps <- data[((data$party == 2) & (data$difficulty_scale > 0)),]

dems_sample_size <- sum(!is.na(dems$difficulty_scale))
print(paste("democrat sample size", dems_sample_size))
```

```
## [1] "democrat sample size 1585"
```

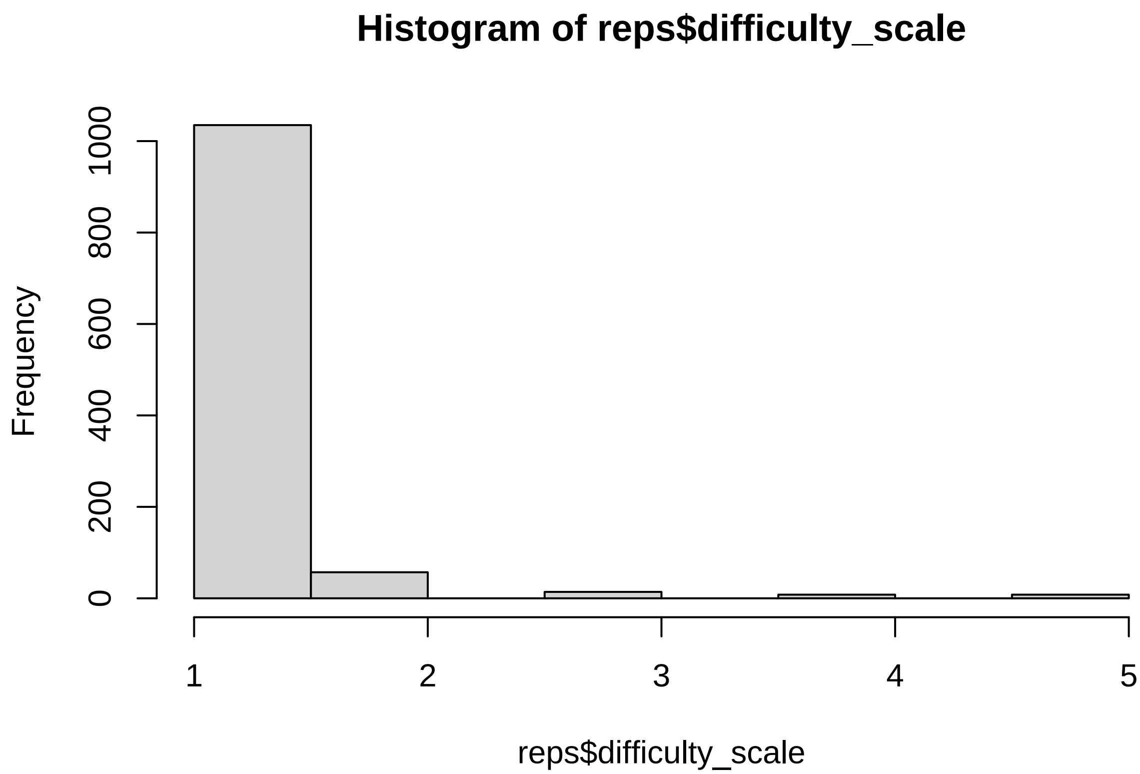
```r
reps_sample_size <- sum(!is.na(reps$difficulty_scale))
print(paste("reps sample size", reps_sample_size))
```

```
## [1] "reps sample size 1122"
```

```r
hist(dems$difficulty_scale)
```




```r
hist(reps$difficulty_scale)
```




```r
# assumptions
# dem and reps equally likely to have diffulty in voting
# iid
# unpaired (can't be dem and rep, each person has one )
# ordinal (likert scale basically)

# null hypothesis: reps and dems had the same difficulty in voting
# alternative: checking
wilcox.test(reps$difficulty_scale, dems$difficulty_scale, alternative = "two.sided") # their difficulty
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  reps$difficulty_scale and dems$difficulty_scale
## W = 856578, p-value = 0.001677
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Ha = reps had more difficulty than the dems
wilcox.test(reps$difficulty_scale, dems$difficulty_scale, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  reps$difficulty_scale and dems$difficulty_scale
## W = 856578, p-value = 0.9992
## alternative hypothesis: true location shift is greater than 0
```

```r
# Ha = reps had less diffulty than the dems
wilcox.test(reps$difficulty_scale, dems$difficulty_scale, alternative = "less")
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  reps$difficulty_scale and dems$difficulty_scale
## W = 856578, p-value = 0.0008384
## alternative hypothesis: true location shift is less than 0
```

2.2 The research question

Use the ANES data to address the following question:

> Did Democratic voters or Republican voters experience more difficulty voting in the 2020 election?

2.3 Guidance from political scientists

Political identification in the US is a complex phenomenon that is the topic of a large academic literature. See `./background_literature/petrocik_2009.pdf` for some guidance about how stated political identity might not match with revealed political identity at the ballot box.

As practical guidance:

1. Is it reasonable to use the vote that someone cast to identify their party preference in this case? What if someone had so difficult a time voting that they did not cast a ballot?
2. Please treat individuals who "lean" in one direction or another as members of that party. This means that someone who "Leans Democratic" should be classified as as Democrat; and someone who "Leans Republican" should be classified as a Republican.

2.4 Report guidelines

This section provides some guidance for you as you write your report. In `rubric.md` we provide you with specific statements of how we will evaluate your report.

General guidance

- You should knit an .Rmd file to create your pdf report.
- Your report should be no more than 3 pages in standard latex formatting (i.e. `output: pdf_document`)
- You should assume your reader is familiar with statistics, but has no special knowledge of the ANES survey.
- Follow the .Rmd template that we have created, using the prompts to guide you through the parts of an analysis. Make sure you fill in each prompt with all information requested.
- Your report should contain either a plot or a table that advances the argument.

Introduction

- Begin your report with an introduction to motivate the analysis.
- Introduce the topic area and explain why the research question is interesting.
- The introduction must "do work," connecting the general topic to the specific techniques in the report.

Conceptualization and Operationalization

How do you get from a research question to data? First, ensure that the concepts in your question are clear.

- Who or what is a voter?
- Who is a "Republican" and who is a "Democrat"?
- What is difficulty voting?

Only after you have informed your reader of what these concepts are can you then describe how you are going to *measure* these concepts.

- What would be the best **possible** method of measuring this concept? Is this method possible? Why or why not?
- What is the best **available** method of measuring this concept? Why have you opted to use this measurement instead of other possibilities? Map the concept definitions that you have written down onto the variables that you are going to use. Describe, precisely, how the variables were generated, if they come from survey data, provide the text of the question that the respondent is reacting to, not the variable name.
- What, if any, changes have you made to the dataset from how it was provided? Why did you make those changes, how much data was affected, and what are the consequences for any estimates that you produce?

Visual Design

- Any plots or tables that you include must follow basic principles of visual design.
  - A plot/figure must have a title that is informative.
  - Variables must be labeled in plain language. As an example, `v20002` does not work for a label.
  - A plot should have a good ratio of information to ink / space on the page. Do not select a large or complicated plot when a simple table conveys the same information directly.
- Do not include any plot (or R output in general) that you do not discuss in your narrative.
- The code that makes your plot/figure should be included in your report `.Rmd` file, but should not be shown in your final report. To accomplish this, you can use an `echo=FALSE` argument in the code chunk that produces the plot/figure.

Data wrangling

To answer your research question, you will have to clean, tidy, and structure the data (A.K.A. wrangle).

- The code to wrangle data should be included with your deliverable somehow. If you choose to include it in your report `.Rmd` file, then it should not be shown in the PDF of your final report. To accomplish this, you can use an `echo=FALSE` argument for the code chunk that does the wrangling.
  - A better practice – not strictly necessary for this lab – would be to write a function that loads and cleans *all* of the data that is being used by your team for its reports. This way, a single function can be run (and evaluated by your reader) for all the loading, cleaning, and manipulating.
- While we do not want to prohibit you from using additional tools for data manipulation, you should be able to complete this lab with no more than the base `stats` library, plus `dplyr` and `ggplot2` for data manipulation and plotting. Other tools within the tidyverse are available to use, but don't feel like you have to search them out.
- You will learn more by writing your own function than you would searching for a package that does one thing for your report.

Hypothesis testing

To answer your research question, you will have to execute one of the statistical tests from the course.

- The code that executes your test *should* be shown in your report, because it makes very clear the specific test that you're conducting.
- You need to argue, from the statistical principles of the course, why the test you are conducting is the *most appropriate* way to answer the research question.
- Although you might not do this for a report at your organization, for this class please list every assumption from your test, and evaluate it (assess whether the assumption is a reasonable reflection of the natural process that generated the data).
- If you identify problems with some assumptions for your test, that does not mean that you should abandon the analysis or hide the problem. If these "limitations" exist, please describe them honestly, and provide your interpretation of the consequences for your test.

- While you can choose to display the results of your test in the report, you also *certainly* need to write about these results. This should be accomplished using inline code chunks, rather than by hard-coding / hard-writing output into your written report. An example of this is included in `lab_1_example_solution.Rmd`.

Test, results and interpretation

Please discuss whether any statistically significant results that you find are of *practical significance*. There are many ways to do this, but the best will provide your reader enough context to understand any measured differences in a scale appropriate to your variables. Explain the main takeaway of your analysis and how it relates to the broader context you identified in the introduction.