UCB W203 - Lab 1: Hypothesis Testing

Kevin Lustig, Rebecca Nissan, Anuradha Passan, Giorgio Soggiu

3/1/2022

Part 1: Foundational Exercises

1.1 Professional Magic

1.1.1 Type I Error of the test - Rebecca

The type I error rate (i.e. false positive) is the probability of rejecting the null hypothesis whe it is correct. The type I error would be the probability of getting 0 or 6, with the assumption that p = 0.5 (null).

This will be the alpha

1.1.2 Power of test given p = 0.75 - Rebecca

1.2 Wrong Test, Right Data - Kevin

In the Likert scale, the meaningful distance between the different scale points is not consistent. That is, assuming the Likert scale for the websites survey includes five points from 1 = "Very Unsatisfied" to 5 = "Very Satisfied," with 3 being "Neutral," we cannot say that a change from 1 to 3 and from 2 to 4 are equivalent quantifiable changes in opinion [1]. However, in fact the change in quality of experience necessary for a given respondent to go from Very Unsatisfied with one site to Neutral with the other may in fact be considerably less than the change needed to go from Unsatisfied to Satisfied, though these each consist of a difference of two steps. Therefore, though the values produced are numeric, these data violate one of the assumptions for a paired t-test – the use of metric, rather than ordinal, data.

A paired t-test relies on metric data because, like other related tests including the z-test, it is fundamentally a calculation of the difference of means between reference groups. A paired t-test would ask of our survey data: is the mean difference between paired opinion scores different than what we would expect if there were no preference for either website (mean difference within pairs = 0)? Stated otherwise, on average across all respondents, how likely is it that there is really a preference for one site or the other, and how large a preference? However, because of the aforementioned limitation of Likert scale values, we cannot meaningfully parse a mean paired disparity of e.g., +2, because to calculate this requires assuming that non-comparable changes from any one Likert scale point to another are equivalent. The mean of the paired differences is thus meaningless. It is even difficult to trust the directionality of the mean difference across all pairs (respondents like the mobile website more or less than the regular website without regard to how much), as in calculating a mean value purely from the raw Likert scale scores, we may calculate an incorrect value by not correctly taking into account the "weights" of the differences of opinion in, again, Very Unsatsified and Neutral versus Unsatisfied and Satisfied. It's possible to conceive of a scenario in which even the sign of the mean difference is therefore incorrect.

[1] At least, not with only five scale points; see, e.g.: Huiping Wu and Shing-On Leung, "Can Likert Scales Be Treated as Interval Scales?—a Simulation Study," Journal of Social Service Research 43, no. 4 (June 2017): pp. 527-532, https://doi.org/10.1080/01488376.2017.1329775.

It is this last point on directionality that suggests an alternative approach to this issue. A non-parametric paired sign test allows us to analyze our ordinal data provided the observations are independent and identically distributed. It does not attempt, like the t-test, to quantify the size of the difference in opinion within pairs, if any. Rather, it treats all positive changes in opinion as equivalent, and likewise with all negative changes. This alternative test has two main drawbacks. First, it does not have the statistical power of a paired t-test. Second, it loses substantial information present in the original survey responses in the form of the exact values within each paired set of responses. However, in doing so, it allows us to avoid the inaccurate mean calculation of the t-test, and focus on a more accurate analysis of a simpler question: do respondents prefer one website over the other? In looking solely at increases or decreases in opinion score, the paired sign test therefore gives us a reasonable expectation of finding such an effect if one exists in the data.

1.3 Test Assumptions - Giorggio and Annie

1.3.1 World Happiness

Scenario: We have two variables: Life.Ladder and Log.GDP.per.Capita, and we want to see whether countries in high GDP per capita are more or less happy than people in countries with low GDP per capita.

Assumptions for two-sample t-test 1. Normal - CHECK! 2. i.i.d. - yes (assuming same data collection and data transformation procuedures), explain 3. Metric variables - yes, explain

https://www.datanovia.com/en/lessons/normality-test-in-r/#check-normality-in-r

```
# Get a quick look at the two relevant variables
summary (wh_data)
```

```
##   Life.Ladder    Log.GDP.per.capita
##  Min.   :2.375   Min.   : 6.966
##  1st Qu.:4.971   1st Qu.: 8.827
##  Median :5.768   Median : 9.669
##  Mean   :5.678   Mean   : 9.584
##  3rd Qu.:6.428   3rd Qu.:10.527
##  Max.   :7.889   Max.   :11.648
##                  NA's   :13
```

```
# Calculate the mean GDP per capita (will be useful in sorting high and low GDP per capita countries)
gdp_capita_mean <- round(mean(wh_data$Life.Ladder, na.rm = TRUE), digits = 2)
print(paste0('Mean of GDP per capita: ', gdp_capita_mean))
```

```
## [1] "Mean of GDP per capita: 5.68"
```

```
# Split the data into low and high gdp per capita countries
## But first remove rows that have NAs
wh_data <- na.omit(wh_data)
```

```
## Now split into the high and low gdp per capita country groups
high_gdp_cap <- wh_data %>% filter(Log.GDP.per.capita > gdp_capita_mean)
print('Quick look at the variables for high GDP per capita countries:')
```

```
## [1] "Quick look at the variables for high GDP per capita countries:"
```

```
summary(high_gdp_cap)
```

```
##   Life.Ladder    Log.GDP.per.capita
##  Min.   :2.375   Min.   : 6.966
##  1st Qu.:4.979   1st Qu.: 8.827
##  Median :5.779   Median : 9.669
##  Mean   :5.691   Mean   : 9.584
##  3rd Qu.:6.442   3rd Qu.:10.527
##  Max.   :7.889   Max.   :11.648
```

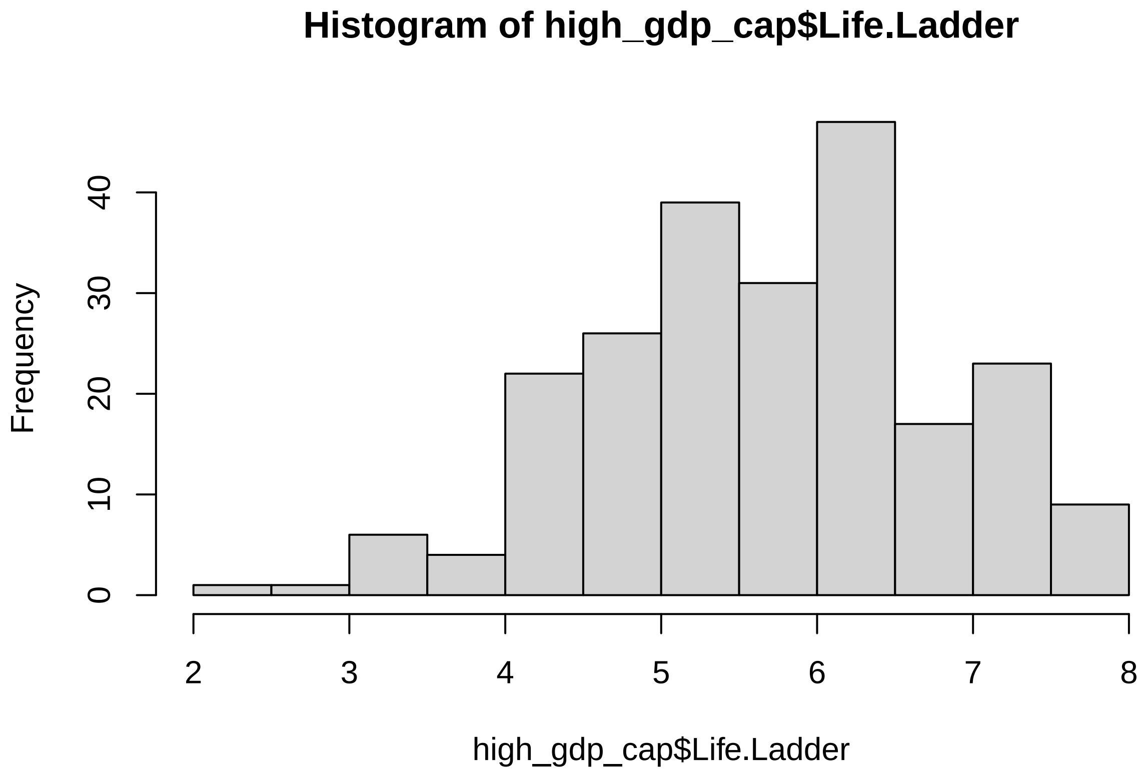
```
low_gdp_cap <- wh_data %>% filter(Log.GDP.per.capita < gdp_capita_mean)
print('Quick look at the variables for low GDP per capita countries:')
```

```
## [1] "Quick look at the variables for low GDP per capita countries:"
```

```
summary(low_gdp_cap)
```

```
##   Life.Ladder  Log.GDP.per.capita
##  Min.   : NA   Min.   : NA
##  1st Qu.: NA   1st Qu.: NA
##  Median : NA   Median : NA
##  Mean   :NaN   Mean   :NaN
##  3rd Qu.: NA   3rd Qu.: NA
##  Max.   : NA   Max.   : NA
```

```
# Observe the histograms of both variables
hist(high_gdp_cap$Life.Ladder) #make pretty
```



```
#hist(low_gdp_cap$Life.Ladder) #fix #make pretty
```

```
# Can do a Shapiro test as an extra source of evidence to test for normalcy
shapiro_test1 <- wh_data %>% shapiro_test(Life.Ladder, Log.GDP.per.capita)
as.data.frame(shapiro_test1)
```

```
##             variable statistic            p
## 1        Life.Ladder 0.9879751 5.480127e-02
## 2 Log.GDP.per.capita 0.9600202 5.900690e-06
```

1.3.2 Legislators

Scenario: We want to test whether Democratic or Republic senators are older, with two variables party and age (age needs to be calculated from DOB).

Assumptions for Wilcoxon Rank-Sum test: 1. Ordinal variables - yes, explain 2. i.i.d. - yes, explain 3. Same shape and spread of the the two variables - box whisker plot, histogram

```
# Load in Data set
leg_data <- read.csv('datasets/legislators-current.csv')

# Select necessary variables and calculate age
leg_data  <- leg_data %>%
  select(birthday, party, type) %>%
  filter(type == 'sen') %>%
  mutate(age = as.numeric(difftime(Sys.Date(), as.Date(birthday), unit = "weeks"))/52.25)
leg_data <- leg_data %>% select(party, age)
```

```
# Split the data by party
dem_sen <- leg_data %>% filter(party == 'Democrat')
rep_sen <- leg_data %>% filter(party == 'Republican')
indep <- leg_data %>% filter(party == 'Independent')
## We see all Senators have been accounted for as the number of rows of above 3 data frames adds up to
```

Summary: Democrat Senators

```
# Get a quick look at the relevant variables (Democrats and Republicans only)
summary(dem_sen$age)
```

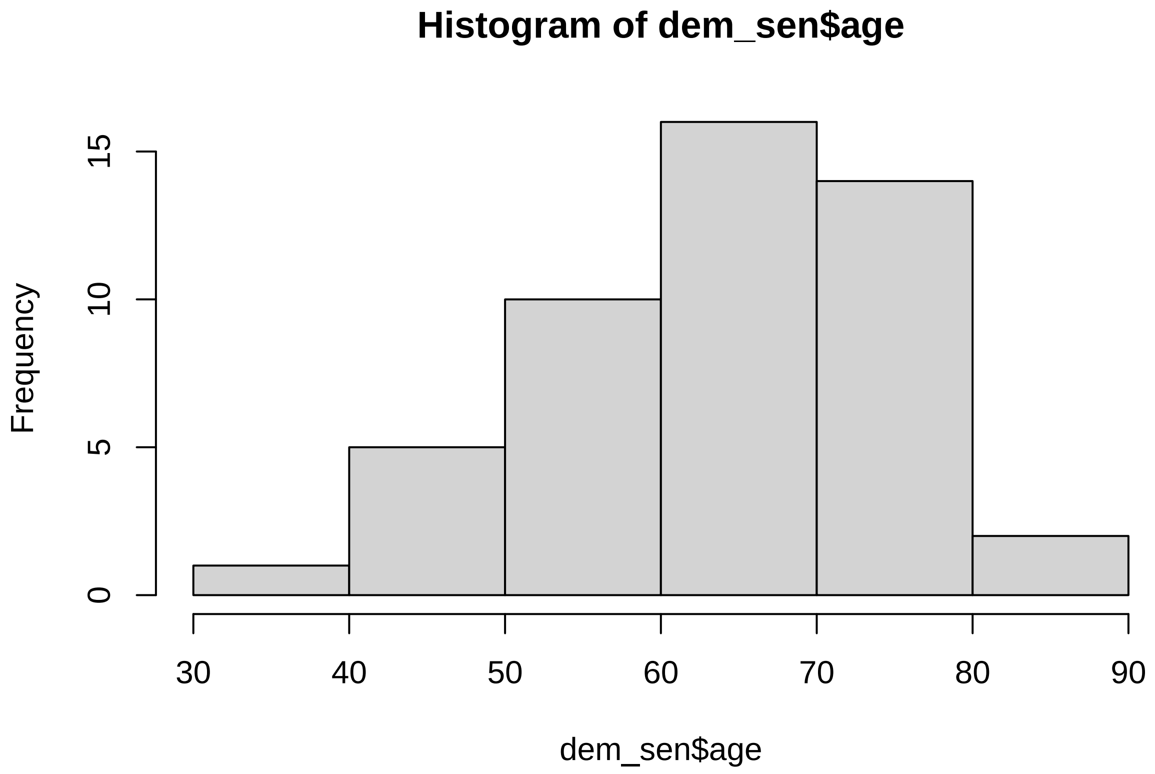
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.97   57.66   64.19   64.17   72.28   88.55
```

Summary: Republican Senators

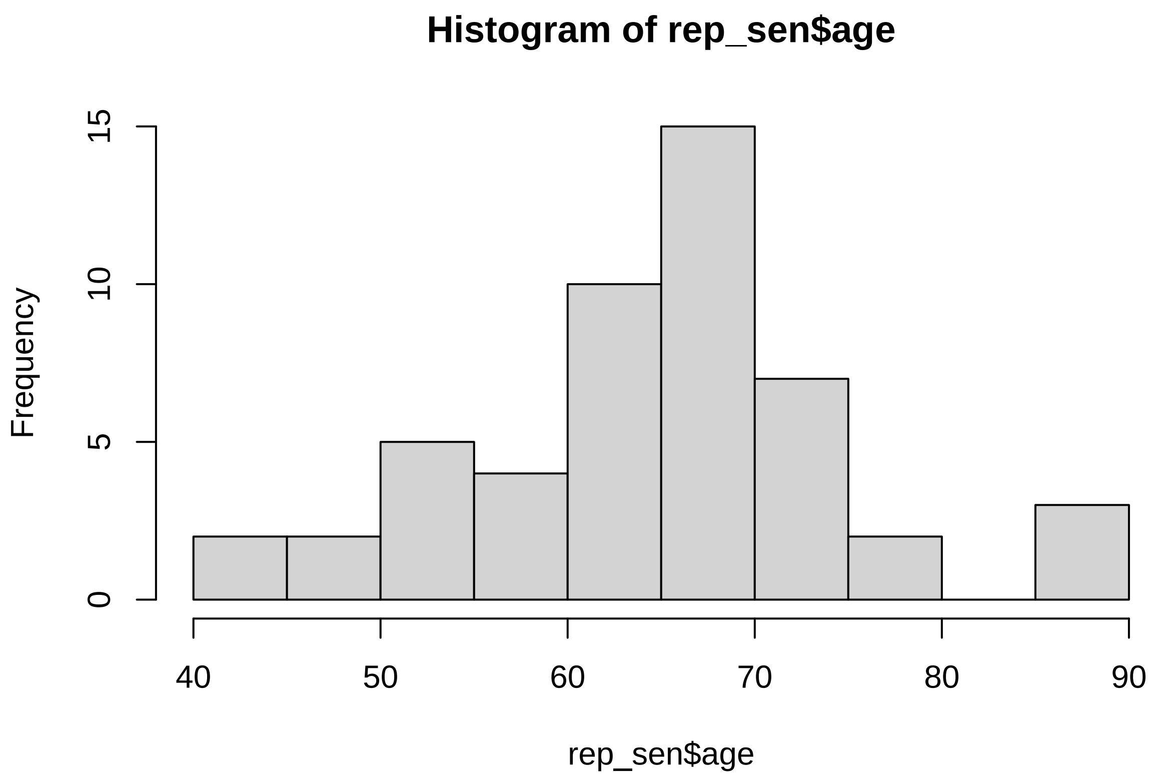
```
# Get a quick look at the relevant variables (Democrats and Republicans only)
summary(rep_sen$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   42.09   59.62   66.34   64.81   69.88   88.31
```

```
# Check whether the variables have the similar shape/spread - box whisker plot or histogram
hist(dem_sen$age)
```



```
hist(rep_sen$age)
```



```
# Check for some metric on similarities between the two variable distributions
```

1.3.3 Wine and Health

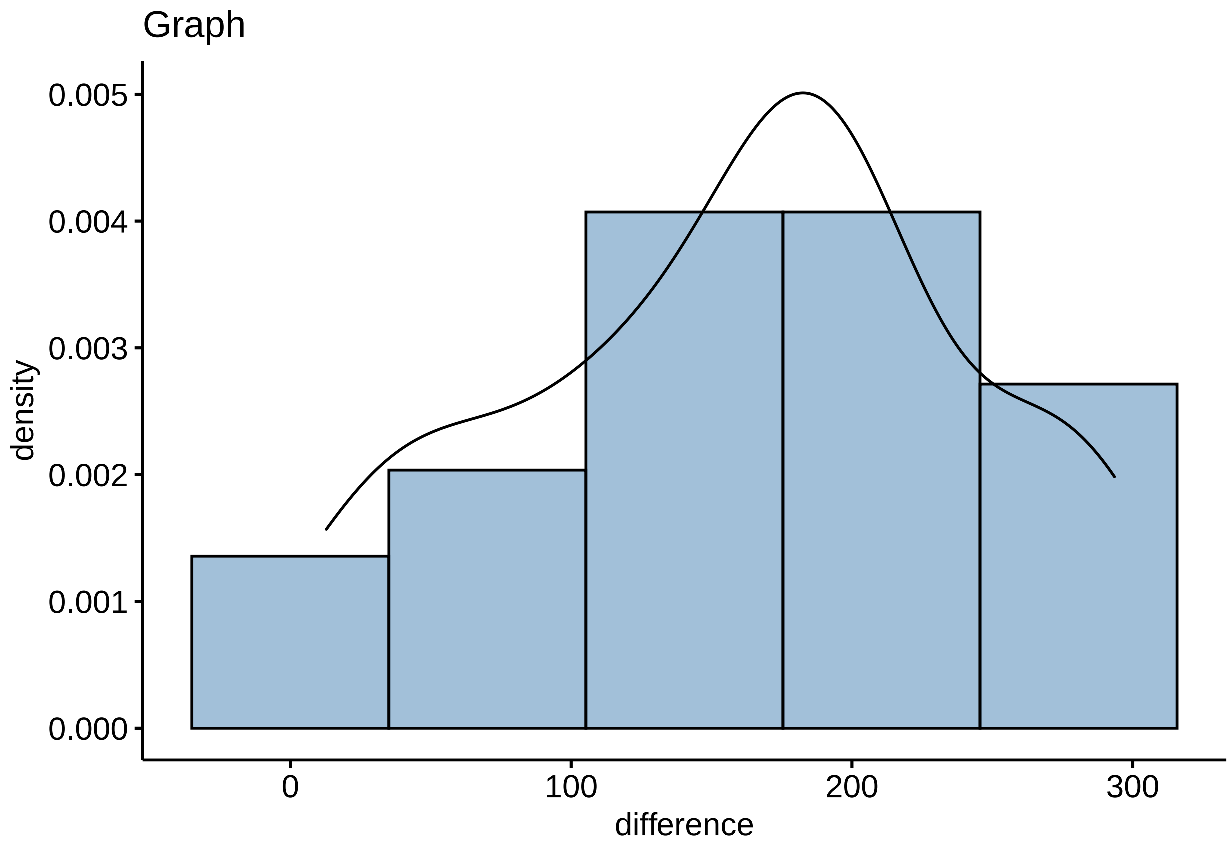
Scenario: We want to test whether these countries have more deaths from heart disease or liver disease.

Assumptions for Wilcoxon Signed Rank Test: 1. Metric Variables - yes, explain 2. i.i.d. - yes, explain 3. Difference is symmetric - need to test for this 4. Paired - explain why/why not

```
# Load in dataset
library(wooldridge)
wine_data  <- wine

# Select necessary variables
wine_data <- wine_data %>%
  select(country, heart, liver)
```

```
# Test for symmetry
## First calculate the difference between Heart and Liver
wine_data <- wine_data %>% mutate(difference = heart - liver)
## Look at the histogram too see if there is symmetry
gghistogram(wine_data, main = 'Graph', x = 'difference', y = '..density..', fill= 'steelblue', bins = 5
```



1.3.4 Attitudes Toward the Religion

Scenario: We would like to know whether the U.S. population feels more positive towards Protestants or Catholics.

Assumptions for a Paired t-Test 1. Metric Variables - yes, explain 2. i.i.d. - yes, explain 3. Paired - yes, explain 4. Normalcy - Check for this

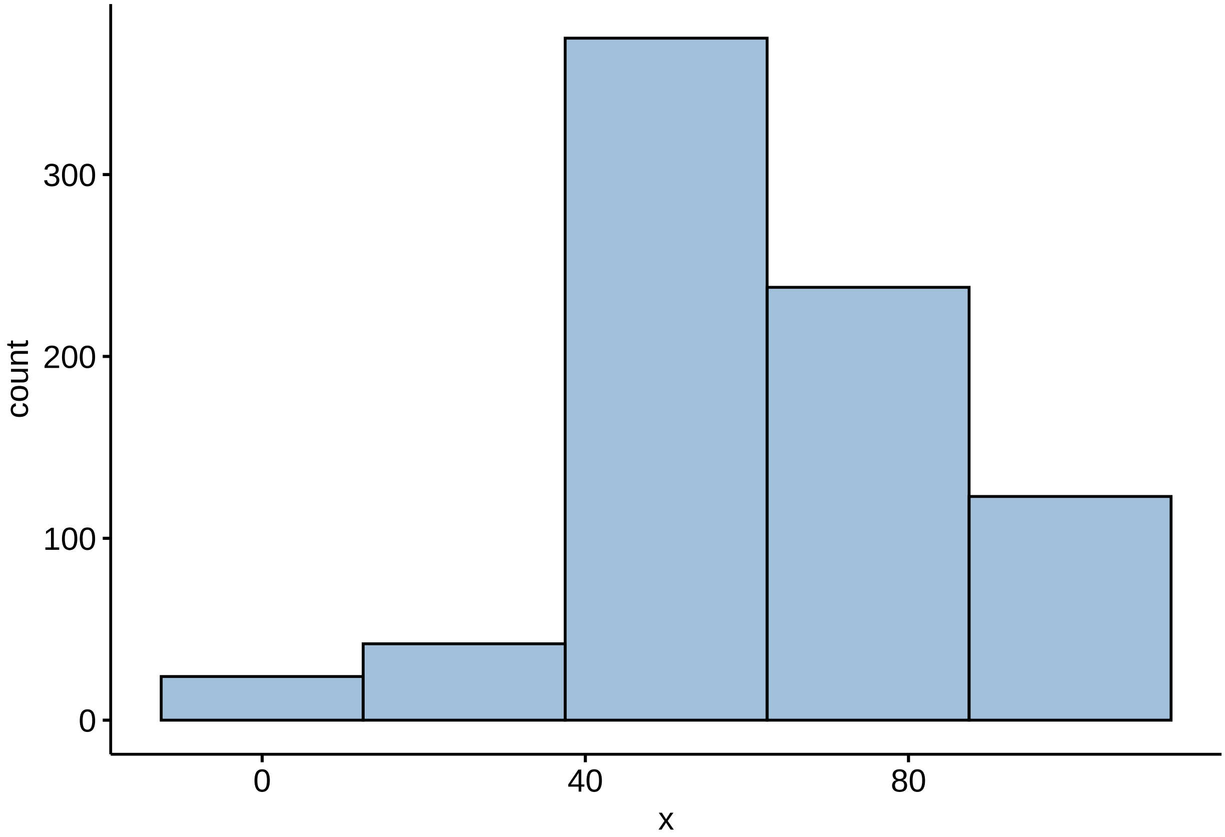
```
# Load in dataset
rel_data <- read.csv('datasets/GSS_religion.csv')

# Select necessary variables
rel_data <- rel_data %>% select(cathtemp, prottemp)
```

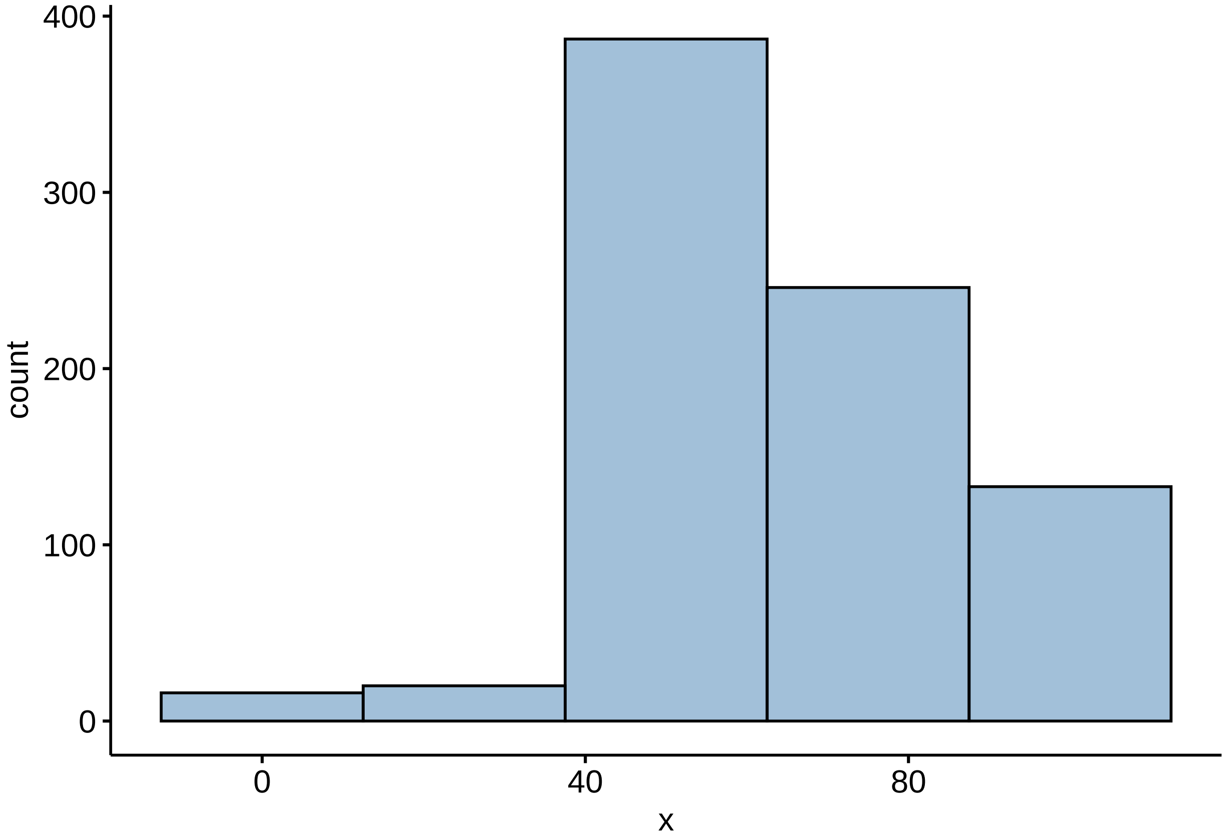
```
# Get a quick look at the relevant variables
summary(rel_data)
```

```
##     cathtemp         prottemp     
##  Min.   :  0.00   Min.   :  0.00  
##  1st Qu.: 50.00   1st Qu.: 50.00  
##  Median : 60.00   Median : 60.00  
##  Mean   : 63.16   Mean   : 65.56  
##  3rd Qu.: 85.00   3rd Qu.: 85.00  
##  Max.   :100.00   Max.   :100.00
```

```
#Observe the histograms for both variables
gghistogram(rel_data$cathtemp, fill= 'steelblue', bins = 5)
```



```
gghistogram(rel_data$prottemp, fill= 'steelblue', bins = 5)
```



```
# Conduct a Shapiro test to also help determine normalcy
shapiro_test2 <- rel_data %>% shapiro_test(cathtemp, prottemp)
as.data.frame(shapiro_test2)
```

```
##   variable statistic            p
## 1 cathtemp  0.933774 2.248818e-18
## 2 prottemp  0.894794 5.562661e-23
```

Part 2: Statistical Analysis

Data source: 2020 American National Election Studies (ANES) - 2020 Time Series Study

Research question: Did Democratic voters or Republican voters experience more difficulty voting in the 2020 election?

Define the following: 1. Democratic voters 2. Republican voters 3. Factors contributing to difficulty in voting –> create an index?

Data cleaning and wrangling - Kevin

Form final dataset

EDA and basic information on the variables and proof we can make assumptions needed for tests we will run

Run tests

Interpret tests

Create visualizations (if not done before)

Write report

```r
anes_data <- read.csv('datasets/anes_timeseries_2020_csv_20220210.csv')
```