UCB W203 - Lab 1: Hypothesis Testing

Kevin Lustig, Rebecca Nissan, Anuradha Passan, Giorgio Soggiu

3/1/2022

```
knitr::opts_chunk$set(echo = FALSE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v readr   2.1.1
## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.1.4     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(readxl)
library(rstatix)
```

```
##
## Attaching package: ’rstatix’

## The following object is masked from ’package:stats’:
##
##     filter
```

```
library(ggpubr)
```

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 “distance” between the different options is not consistent, thus violating the metric scale assumption needed to run a paired t-test. We could suggest a paired sign test where the only only assumptions needed are: ordinal variables and i.i.d.

1.3 Test Assumptions - Giorggio and Annie

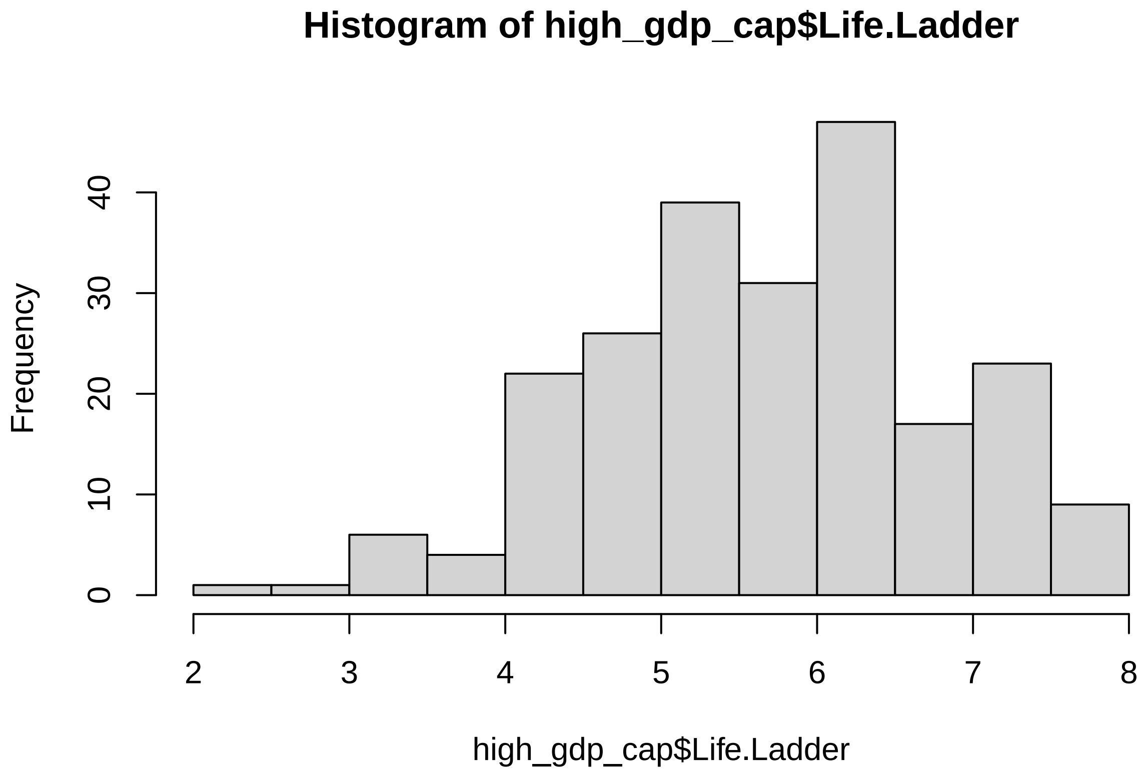
1.3.1 World Happiness

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

```
##   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
```



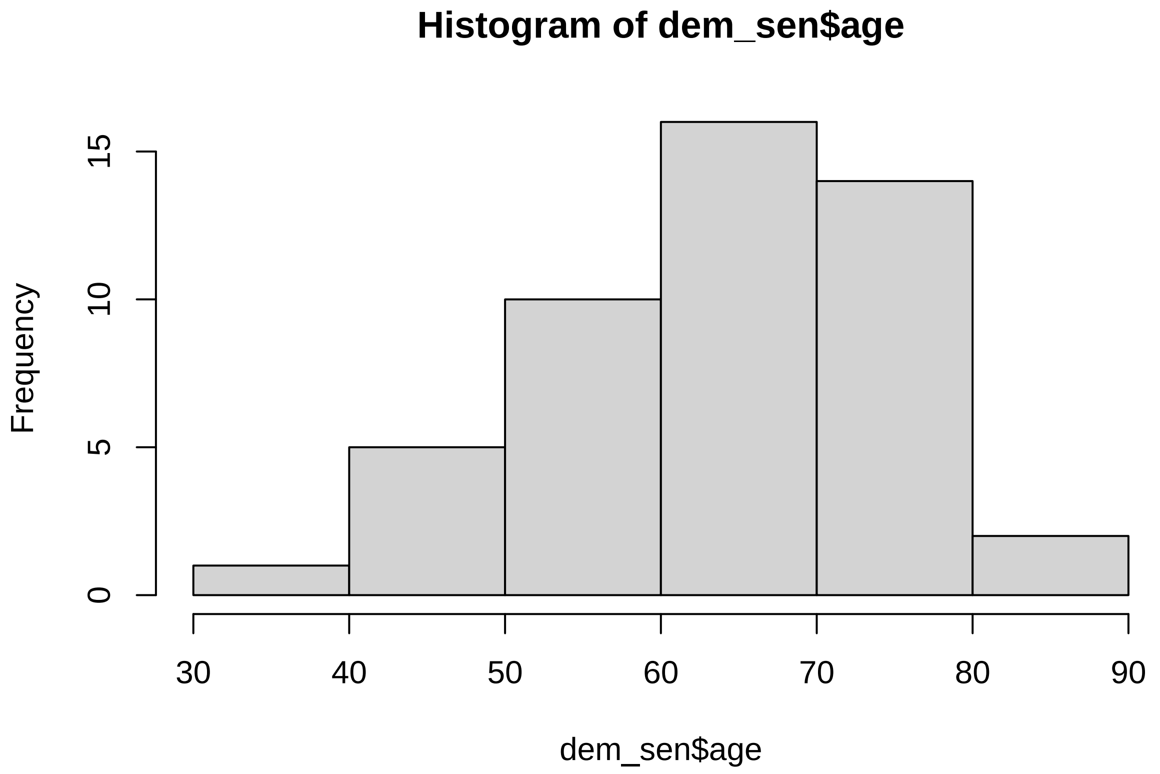
1.3.2 Legislators

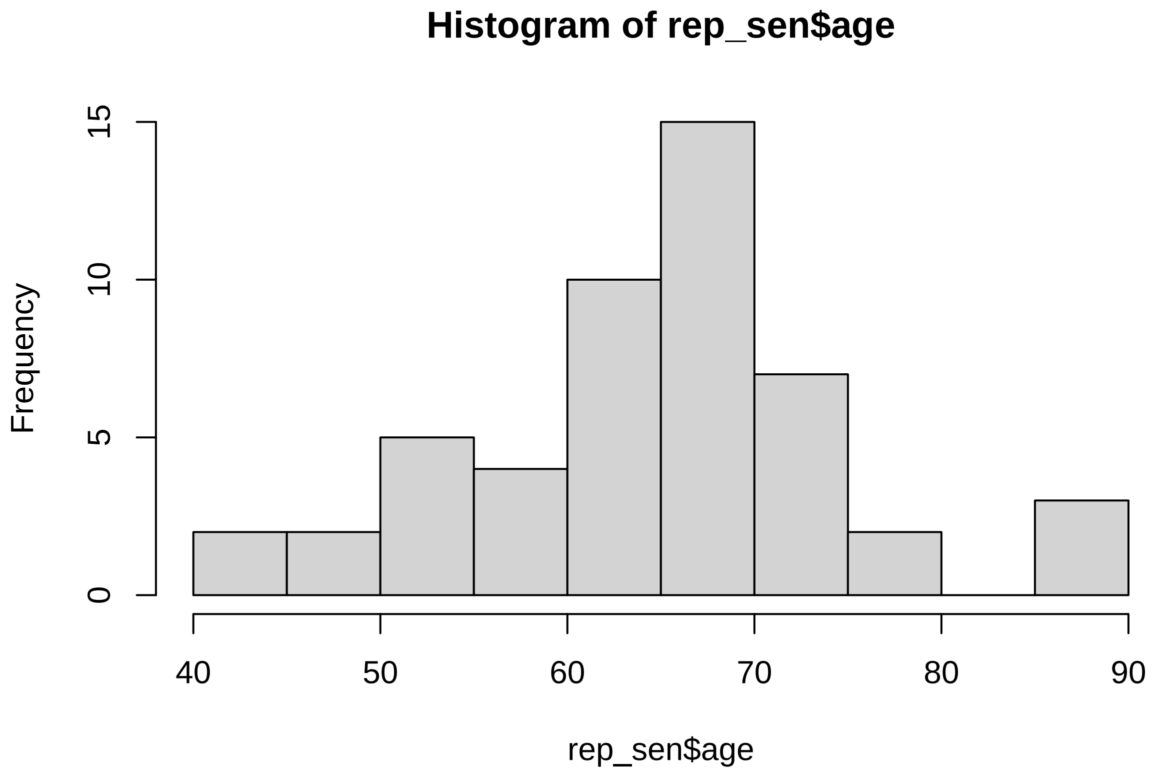
Scenario: We want to test whether Democratic or Republic senators are older, with two variables party and age (age needs to be calculated from DOB).

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

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   42.09   59.62   66.34   64.81   69.87   88.31
```

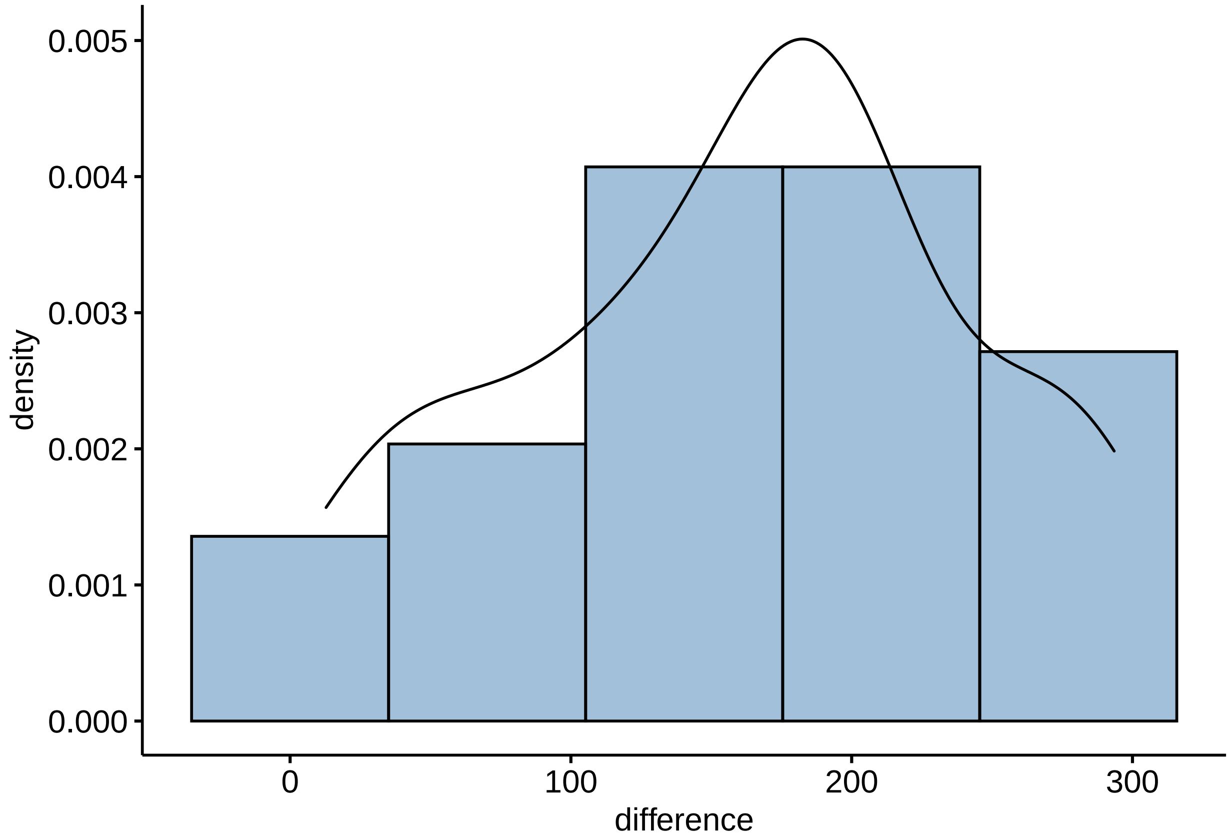

Histogram of dem_sen$age
Frequency
0
5
10
15
30
40
50
60
70
80
90
dem_sen$age



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

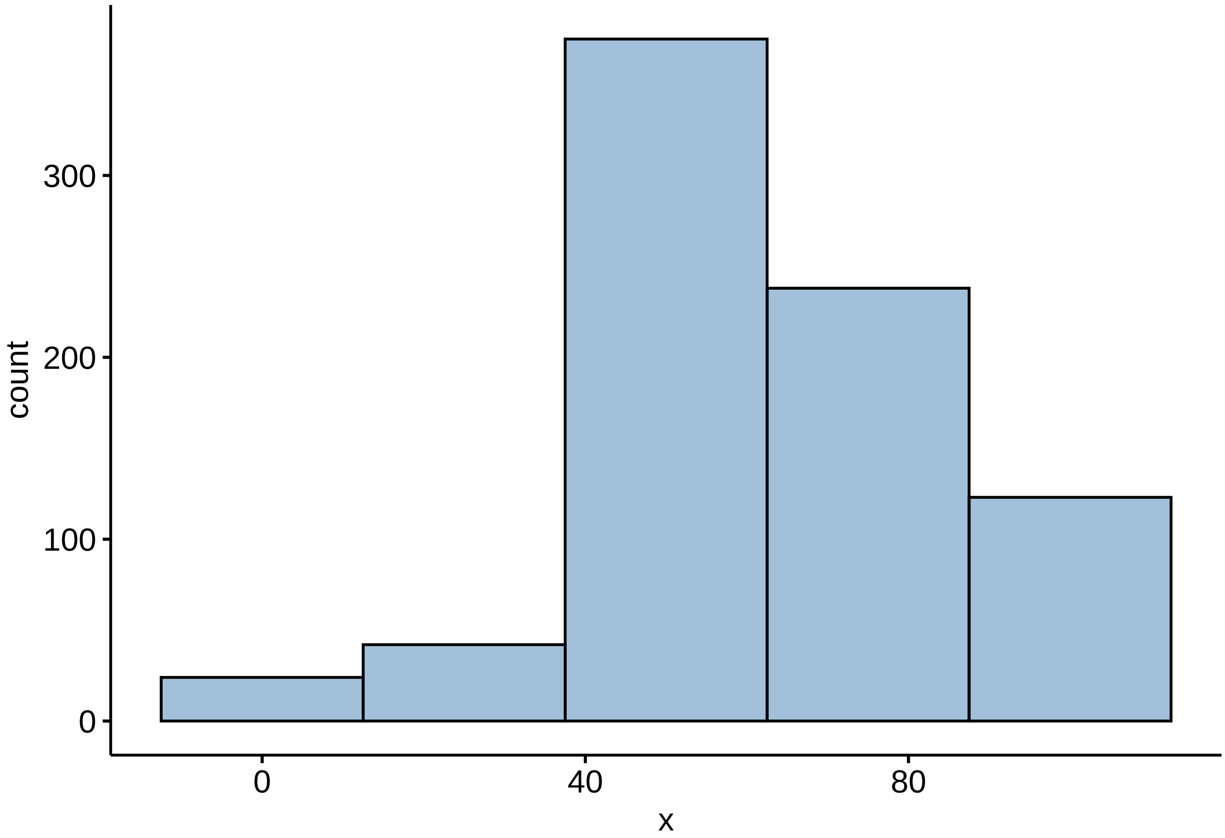


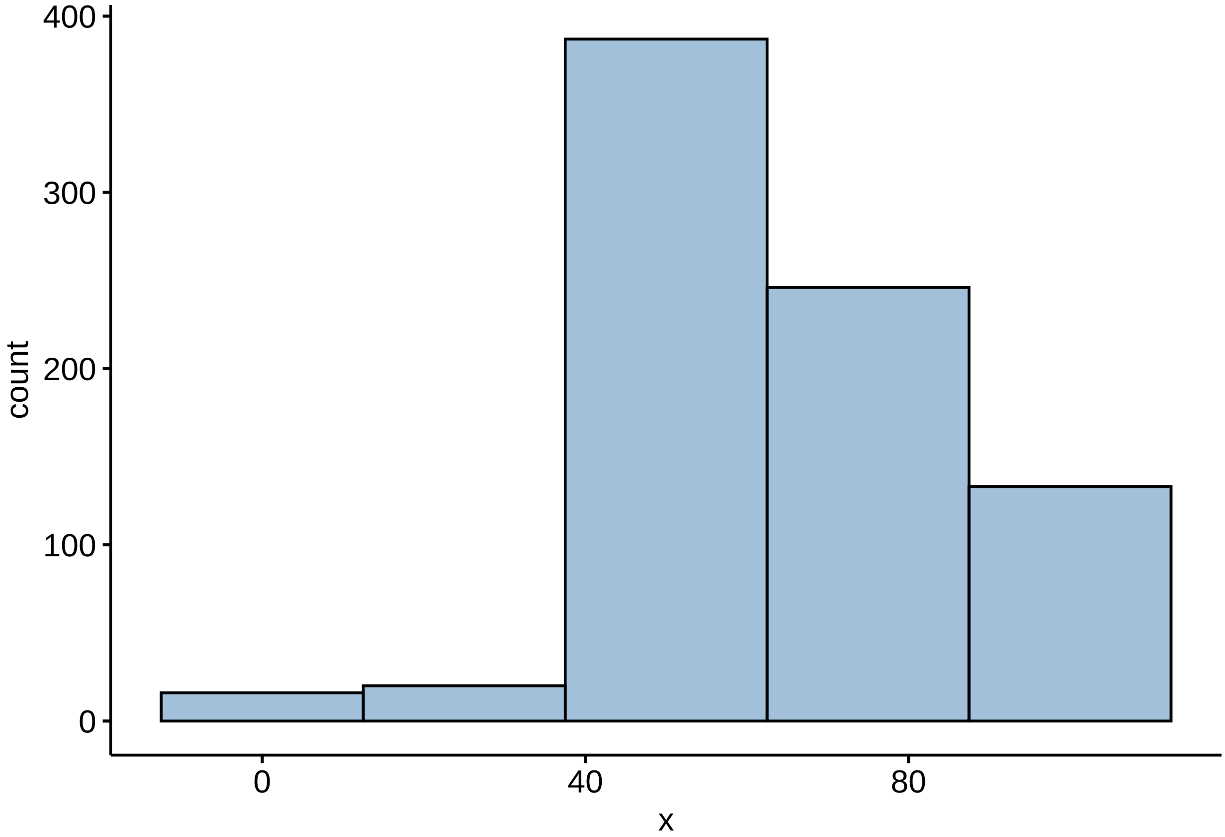
1.3.4 Attitudes Toward the Religion

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

```
##     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  
```


count
300
200
100
0
0
40
80
x



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