Lab 1 Part 1 - Foundational Excercises

Team03: Savita Chari, Tymon Silva, Denny Lehman

```
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(knitr)
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1

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

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   1.4.0

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

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. You also assume that

$$P(X_i = 0) = P(X_i = 1) = P(Y_i = 0) = P(Y_i = 1) = 1/2,$$

and write,

$$P(X_i = Y_i) = p.$$

Your aunt claims that $p > 1/2$.

You design a test to evaluate your aunt's claim. You flip the coins 3 times and define your test statistic to be the sum $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$

Your null hypothesis is that $p = 1/2$. 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?
2. What is the power of your test for the alternate hypothesis that $p = 3/4$?

1. What is the type 1 error rate of the test?

The type 1 error occurs when we reject the null hypothesis when it is actually true. Type I error, a false positive, is typically denoted as alpha

$$\alpha = P(rejectHo|Ho)$$

We are given that Ho is when p = 1/2. p is the probability that Xi = Yi

$$H_o : p = \frac{1}{2}, p = P(X_i = Y_i)$$

We know that each flip of the pair has the following potential outcomes:

$$P(X_i = Y_i) = P(X = 0, Y = 0) + P(X = 1, Y = 1)P(X_i = Y_i|H_o) = 0.5 = P(X = 0, Y = 0) + P(X = 1, Y = 1)P(X = 0, Y =$$

We are also given that our test statistic, theta_hat, will be equal to

$$\theta = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$$

From the problem definition, we reject the null when our test statistic theta is 0 or 6 after 3 flips of pairs of coins. Type I error is thus

$$\alpha = P(rejectHo|Ho)\alpha = P(\theta \in \{0, 6\}|p = \frac{1}{2})\alpha = P(\theta = 0|p = \frac{1}{2}) + P(\theta = 6|p = \frac{1}{2})$$

Solve for each term using the binomial distribution formula

$$binomialdistributionformulaP_x = nCxp^x(1-p)^{n-x}solveforP(\theta = 0|p = \frac{1}{2})P(X_i = 0) + P(Y_i = 0) = \frac{1}{4}P(\theta = 0|p = \frac{1}{2}) = 0C$$

result of a pair of flips both coming up tails is

$$0 = X_i + Y_iP(trial = 0) = P(X_i = 0) * P(Y_i = 0)P(onetrial = 0) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}P(\theta = 0) = \frac{1}{4}^3$$

$$2 = X_i + Y_iP(trial = 2) = P(X_i = 1) * P(Y_i = 1)P(onetrial = 2) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}\theta = 3trialsP(\theta = 6) = \frac{1}{4}^3$$

now sub back in for alpha

$$\alpha = P(\theta = 0|H_o) + P(\theta = 6|H_o)\alpha = \frac{1}{4}^3 + \frac{1}{4}^3\alpha = \frac{2}{64} = \frac{1}{32}$$

2. What is the power of your test for the alternate hypothesis that $p = 3/4$?

c

$$Power = 1 - \beta$$

Where

$$\beta$$

is the type 2 error, or the false negative rate. Type 2 error occurs when we fail to reject the null hypothesis when we should.

$$Power = P(rejectnull|H_a)P(X_i = Y_i) = p = \frac{3}{4}\beta = P(failtorejectH_o|H_a)$$

## Study question 2: Wrong Test, Right Data ### Our Analysis: The response categories in Likert scales creates rank order/ordinal data, but the intervals between values cannot be presumed equal.Therefore, standard statistical analysis such as the mean, standard deviation etc. are inappropriate for ordinal data. The data generated by this survey is not continuous so it's distribution cannot be measured,So A-B is not a normal distribution. Hence, this data is not appropriate for a paired t-test.

Study question 3: Legislator age

You would like to test whether Democratic or Republican senators are older. List all assumptions for a Wilcoxon rank-sum test and evaluate them.

Hypothesis

H0: Median age of Democrat is same as mean age of Republican senators. Our null hypothesis is that there is no difference in the mean age between all the Democrat legislators and the Republican legislators. The expectation of X equals the expectation of Y. In other words, $\Delta = 0$.

What is Wilcoxon rank-sum test?

It is a way of examining the relationship between a numeric outcome variable (Y) and a categorical explanatory variable(X, with 2 levels) when the groups are Independent. It is a nonparametric test for unpaired data. It is used to compare one random variable X against another random variable Y.

The Wilcox test for comparing two population means makes the following assumptions:

1. The two samples are independent of one another and are Metric scale.
2. The two populations have equal variance or spread
3. The two populations are normally distributed (iid)

The following rules must be observes

1. The first assumption must be satisfied for a two-sample t-test.
2. But when assumptions #2 and #3 (equal variance and normality) are not satisfied, the samples size must be larger then 30 for the results are approximately correct
3. But when our samples are small and our data skew or not normal, we probably should not perform two-sample t-test

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   birthday = col_date(format = ""),
##   district = col_double(),
##   senate_class = col_double(),
##   cspan_id = col_double(),
##   govtrack_id = col_double(),
##   votesmart_id = col_double(),
##   washington_post_id = col_logical(),
##   icpsr_id = col_double()
## )
## i Use `spec()` for the full column specifications.
```

Data Cleaning

The Data set has 538 rows and 34 columns. But we need a Subset of this data to answer the research question

```
Num_Of_Rows <- nrow(legislator_full_data)
Num_Of_Columns <- ncol(legislator_full_data)
sprintf('The Dataset had %i Rows and %i Columns', Num_Of_Rows ,Num_Of_Columns)
```

```
## [1] "The Dataset had 538 Rows and 34 Columns"
```

In order to perform Wilcoxon test we need 2 types of columns from our dataset

1. [X] A categorical column with 2 distinct groups : We chose the 'party' column for this requirement but the data set includes more than two categories. We filtered only the Republican and Democrat data in order to perform Wilcoxon rank-sum test as we are interested in only those categories. Thus we fulfill the requirement of having only 2 categories.
2. [Y] A column with numeric outcome: We choose the 'birthday' field for this requirement. Though this field is of class 'Date', some manipulation will be needed to make it a numeric outcome

Data Transformation

The birthday field is of 'Date' class and is in YYYY/MM/DD format In order to find out the age of the legislator we calculate the current age of the legislator by doing a datediff between current date and the birthday

```
currentDate <- Sys.Date()
legislator_DemRep_Age_data  <- legislator_subset_data %>%
  mutate(AgeInYears = round((as.numeric(difftime(currentDate , birthday,units = "weeks")))/52))
```

Exploring Assumption 1 : Are the data sets independent?

Well, yes, a legislator cannot belong to 2 parties simultaneously, so the data is independent but still we can very easily check the count of rows for each party in the data sets

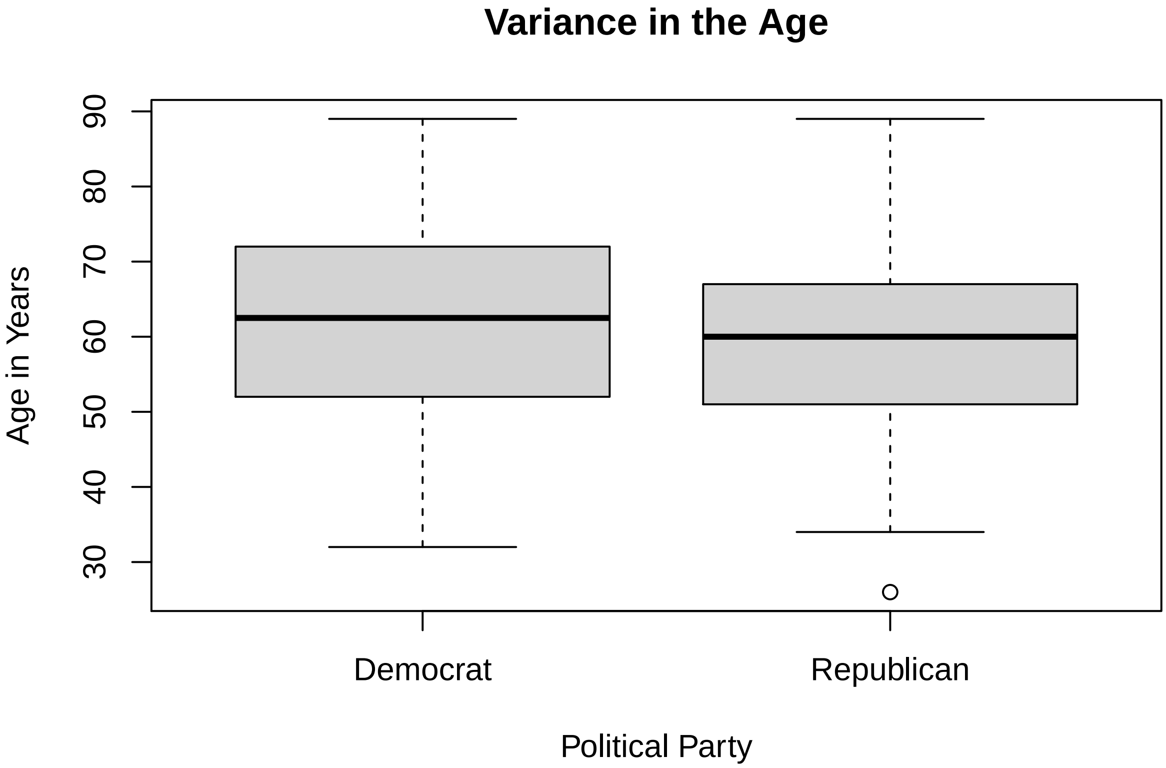
```
legislator_DemRep_Age_data %>% count(party)
```

```
## # A tibble: 2 x 2
##   party          n
##   <chr>      <int>
## 1 Democrat     272
## 2 Republican   264
```

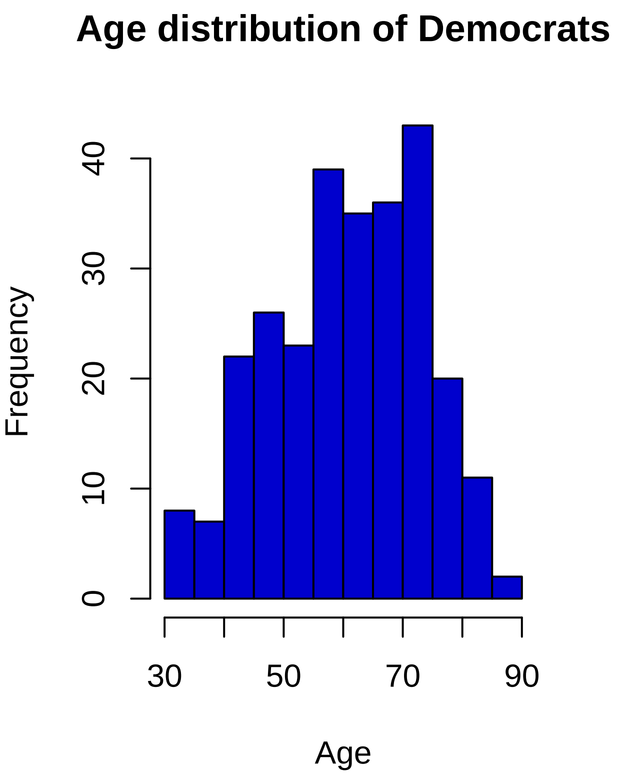
Hence, Assumption 1 is True, the sample-size for Democrats and Republicans is not same. Also, we have 272 rows for Democrats and 264 rows for Republicans. This sample size is larger than the minimum required 30 rows in order to avoid data skews and to make this a valid Wilcoxon rank-sum test.

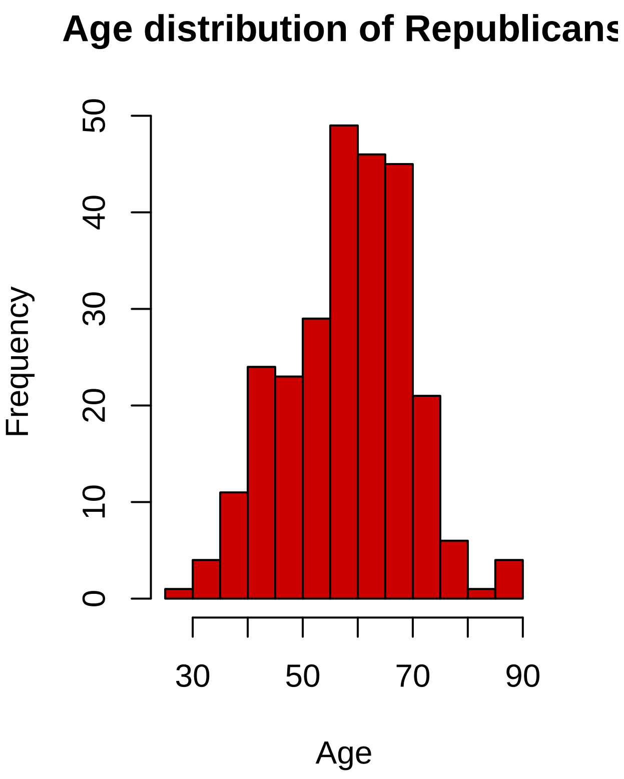
Exploring Assumption 2 : Does the population have equal variance?

With our large data set we do not need to fulfill this assumption but from the box plot below, it is very clear that the Democrates have a bigger variance than the Republicans wrt. their age. In the Republican data set there is an outliar too.



### Exploring Assumption 3 : Is the populations normally distributed(iid) ? Our sample size is larger than 30 so we do not have to fulfill this assumption, but from the graph below we can tell that the age distribution between republicans and democrats is not normal, but not too skewed too.





Performing two sided Wilcoxon rank-sum test

In order to prove the null hypothesis we perform the Wilcoxon rank-sum test. our data satisfies the requirements for conducting this test.

```
wilcox.test(legislator_DemRep_Age_data$AgeInYears~legislator_DemRep_Age_data$party, mu=0, alt = "two.si
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  legislator_DemRep_Age_data$AgeInYears by legislator_DemRep_Age_data$party
## W = 40789, p-value = 0.006416
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  0.999954 5.000031
## sample estimates:
## difference in location
##                      3
```

Test Outcome and it’s correctness

1. We used mean to create our test statistics which is well accepted distributions for statistical testing.
2. we followed the decision rules of a hypothesis testing which gives us the guarantee that our false positive rejection rate (the type 1 error rate) is bounded.

3. Our test is significant as our p-value is greater than 0.0025

With 95% confidence interval, We are rejecting the null hypothesis