Lab 1 Part 1 - Foundational Excercises

Team03: Savita Chari, Tymon Silva, Denny Lehman

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

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

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.

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

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.

Test Assumptions

For the four following questions, your task is to evaluate the assumptions for the given test using your background knowledge and examining the data.

World Happiness

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 and evaluate them.

Two-sample t-test assumptions

1. Metric variables
   The Life Ladder variable is a metric variable. We can justify this by reviewing the FAQ for this dataset that states, "... 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." From this description, we know the responses fall on a 0 to 10 scale, where the distances or gaps between each point on the scale are the same, since each rung on a ladder is the same distance for the next.
2. IID
   Here, we assume IID is upheld since Gallup[1] is a reputable research firm, and we expect their sampling and data collection processes maintain the independence and identically distributed assumption.
3. Data is normally distributed (check sample size)
   Since we have sufficient sample sizes (greater than 30) in each of the Low (n=105) and High (n=121) groups for GDP, we can assume normality by the Central Limit Theorem.

Thus, all the assumptions for a two-sample t-test are valid and the test would be appropriate.

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(),
```

[1] https://www.gallup.com/corporate/212381/who-we-are.aspx

```
##   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

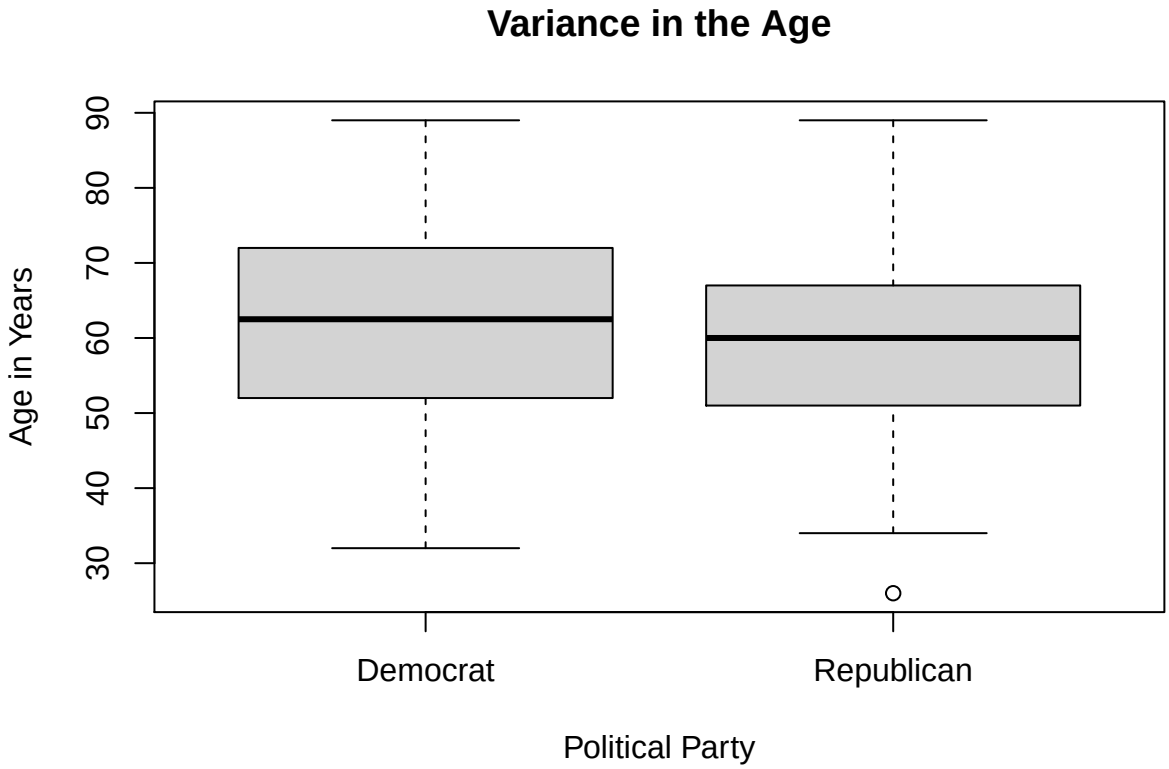
```
legislator_DemRep_Age_data %>% count(party)
```

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

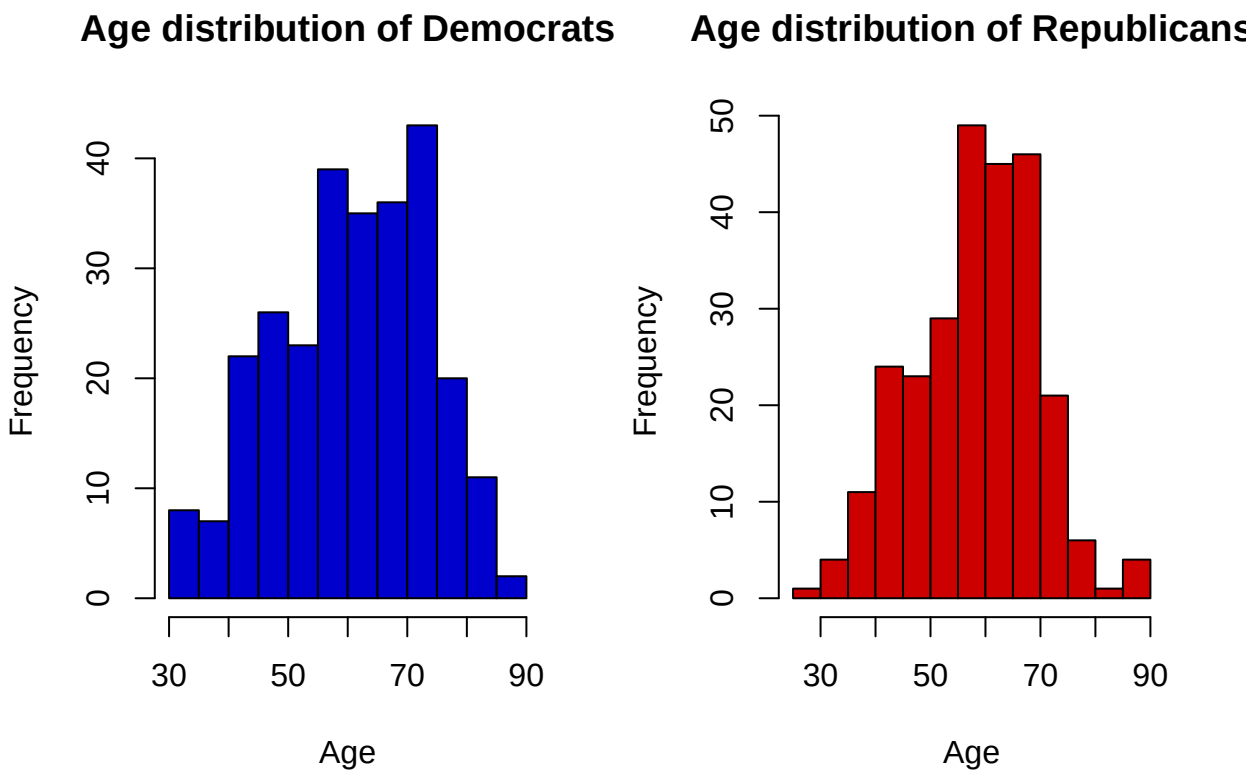
Hence, Assumption 1 is True, the sample-size for Democrats and Republicans is not same. Also, we have 272 rows for Democrats and 264 rows for Republicans. This sample size is larger than the minimum required 30 rows in order to avoid data skews and to make this a valid Wilcoxon rank-sum test.

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 = 40782, p-value = 0.006492
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  0.9999745 5.0000091
## 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

It's for your health!

You would like to use it [wine dataset] to test whether countries have more deaths from heart disease or from liver disease. List all assumptions for a signed-rank test and evaluate them.

```
##         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
```

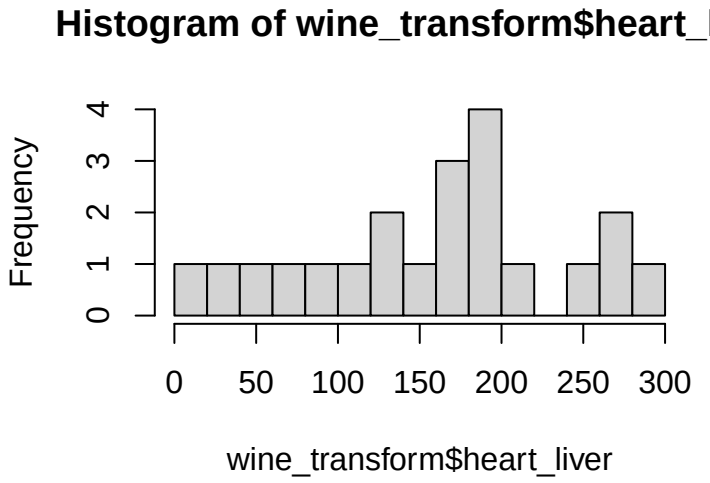
Wilcoxon signed-rank test assumptions

1. Metric variables We need metric variables for a paired Wilcoxon signed-rank test, and the heart and liver variables satisfy this assumption. The Heart and Liver variables represent the number of deaths from heart and liver disease in each country.

2. IID
   There are some concerns with IID for this data. Since there dataset only contains 21 countries, we do not know how these countries were drawn, particularly if they were randomly select. Furthermore, we do not know how the data was collected, which means the data draw could not have been identically distributed.

3. Difference is symmetric The last assumption requires the difference between pairs follows a symmetric distribution. If you look at the histogram below of the difference between heart and liver disease deaths, one could argue that the distribution is skewed left, which violates our difference is symmetric assumption.

Due to major concerns with assumptions 2 and 3, the Wilcoxon signed-rank test would not be a valid test to perform to answer the question.



Positive vibes

You would like to test whether the US population feels more positive towards Protestants or towards Catholics. List all assumption for a paired t-test and evaluate them.

```
## spec_tbl_df [802 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ X1      : num [1:802] 1 2 3 4 5 6 7 8 9 10 ...
##  $ year    : num [1:802] 2004 2004 2004 2004 2004 ...
##  $ id      : num [1:802] 4 7 9 14 21 24 25 28 29 31 ...
##  $ prottemp: num [1:802] 50 5 50 60 100 5 50 40 50 50 ...
##  $ cathtemp: num [1:802] 50 85 50 60 0 5 60 60 50 50 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   year = col_double(),
##   ..   id = col_double(),
##   ..   prottemp = col_double(),
##   ..   cathtemp = col_double()
##   .. )
```

Paired two-sample t-test assumptions

1. Metric variables
   The protemp and cathtemp variables . We can justify this by reviewing the FAQ for this dataset that states, "... 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." From this description, we know the responses fall on a 0 to 10 scale, where the distances or gaps between each point on the scale are the same, since each rung on a ladder is the same distance for the next.

2. IID
   Here, we assume IID is upheld since Gallup[2] is a reputable research firm, and we expect their sampling and data collection processes maintain the independence and identically distributed assumption.

3. Data is normally distributed (check sample size)
   Since we have sufficient sample sizes (greater than 30) in each of the Low (n=105) and High (n=121) groups for GDP, we can assume normality by the Central Limit Theorem.

[2]https://www.gallup.com/corporate/212381/who-we-are.aspx

Thus, all the assumptions for a two-sample t-test are valid and the test would be appropriate.