Lab One, Part One

Sophie, Matt, Torrey

03/01/2022

Contents

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.

$p \in [0, 1]$ 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.

1.1.1 What is the type 1 error rate of your test?

Type 1 error rate is the probability that a null hypothesis is rejected when in fact it should not have been rejected. In the above example, the null hypothesis is that our aunt does not have magical powers meaning the coins should function with normal probability ($p = 1/2$).

The type 1 error rate in this scenario is the probability that we will reject the null hypothesis of our aunt having no magical powers when she does *not* actually have magical powers.

The rejection criteria is when the number of heads flipped is 0 or 6 ($S \in \{0, 6\}$) and the probability of producing a result within the rejection criteria given our null hypothesis is $P(S) \in \{0, 6\}$.

Therefore, the type 1 error rate is:

$P(S) \in \{0, 6\} = P(S = 0) + P(S = 6) = 1/64 + 1/64 = 1/32 \approx 0.031$

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

The power of the test is the probability that we will correctly reject the null hypothesis when an alternate hypothesis is true. This can be expressed as $1 - \beta$ where β is the type 2 error rate.

In the scenario where the alternate hypothesis is $p = 3/4$, the power is probability that the null hypothesis will be rejected.

Therefore, the power of the test for the alternate hypothesis is:

$P(S) \in \{0, 6\} = P(S = 0) + P(S = 6) = 27/512 + 27/512 = 27/256 \approx 0.105$

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?

A paired t-test is used when we are interested in the difference between two variables for the same subject. For the two dependent samples, regular and mobile website, the measured values are available in pairs. Rather using the mean of a sample, the paired t-test uses $\bar{d}$, the sample mean difference, as the point of estimate of μ_d. The formula for the t test is

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

In this case, we would be measuring customer preference and the two samples are regular and mobile website responses. If we ignore the metric assumption and apply the paired t-test on the likert scale, the paired t-test formula would calculate respondent's difference in preference between the websites and average that to obtain $\bar{d}$.

The likert scale is an ordinal variable. This means that rankings and order matter, but there is no measurable distance between values. Rather, the numbers 1 to 5 are mere representations of the range "strongly dislike" to "strongly like" and are not objective numeric measurements such as temperature and distance. Additionally, the scale is subjective to each person so what one considers as "slightly disagree" given their opinion might be determined as "strongly disagree" for another. As a result, the paired t-test difference calculations would become meaningless.

Instead, I propose using the Wilcoxon Signed Rank Sum Test. After taking the difference between regular website likability and mobile website likability for each customer, the absolute values of the difference scores are assigned ranks and signs. By assigning ranks and signs, the test accounts for the likert scale subjectivity between customers and normalizes the distribution values.

1.3 World Happiness - Two-Sample T-Test

Assumption 1: Metric Scale

Not metric scale, as the data is time-series data. The same country's gdp, across different years. Also, the definition of independence/iid becomes difficult since countries are closely related for their GDP, especially in EU. This makes using a t-test difficult and cannot be used. The structure is ordinal. Normality can be assumed since the plot shows this as normal and this is non-parametric. The grouping variable is high and low GDP, with the grouping occurring at the median. The GDP is independent by person, but when being considered as a whole, GDP is not independent as countries trade with other countries closely and have numerous factors to consider. While the graph below shows normality, it does not show iid nor independence. As that cannot be seen solely from graphing or would need to show a much more precisely shifted median which this does not show.

```
library(ggplot2)
data <- read.csv('./datasets/happiness_WHR.csv')
head(data$Country.name)
```

```
## [1] "Afghanistan" "Albania"     "Algeria"     "Argentina"   "Armenia"
## [6] "Australia"
```

Countries are closely related and affect each others' GDPs as shown in the head. Also, GDP in the data is across different years for the same regions, for 2019 and 2020. This is time-series data preventing iid/independence.

```
library(ggplot2)
data <- read.csv('./datasets/happiness_WHR.csv')
table(data$year)
```

```
##
## 2019 2020
##  144   95
```

The test also when applied shows negative numbers, showing that the data does not conform for the test. Below is a preview of the Cantril ladder data. The values are a continuous range between 0 and 10, and mathematical operations can be applied. However, it is important to consider the fact that the Cantril ladder is a self-anchoring value and thus the meaning of the values are subjective.

```
library(ggplot2)
# load data
data <- read.csv('./datasets/happiness_WHR.csv')
head(data[,c("Life.Ladder", "Log.GDP.per.capita")])
```

```
##   Life.Ladder Log.GDP.per.capita
## 1       2.375              7.697
## 2       4.995              9.544
## 3       4.745              9.337
## 4       6.086             10.000
## 5       5.488              9.522
## 6       7.234             10.815
```

```
summary(data[,c("Life.Ladder", "Log.GDP.per.capita")])
```

```
##   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
```

Assumption 2: Independently and identically distributed data* IID means that data points are mutually independent and their probability distributions are the same. Definition of IID:

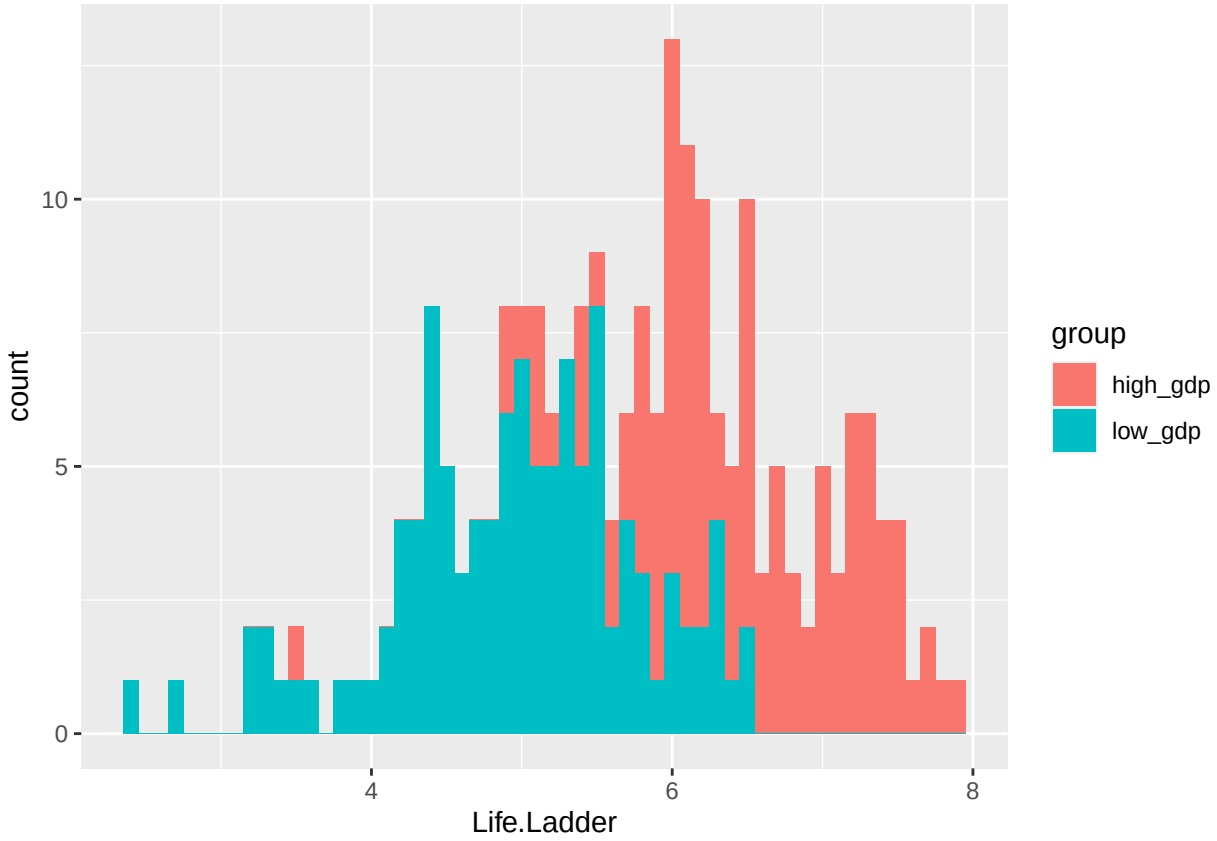
- undertake some process an arbitrary number of times (sampling) to create a value from a phenomenon.
- if each instance of sampling draws from the same probability distribution, we say the collection of values are identically distributed.

- if none of the instances of sampling provide information about other instances of sampling, then we say the collection of values are independent.

Also, iid is difficult to test even if graphing, thus it is easier to look at what destroys/prevents independence, to assume that independence likely does not occur due to these preventing attributes. Space/Time Series data would destroy iid, the data shows gdp across various years, thus this is not iid nor independent. The World Happiness dataset collects data from the Gallup World Poll surveys. This data is a sample from the world population and so translates to having each instance of sampling draws from the same probability distribution and is distributed on a per capita basis. The histogram below demonstrates that all the Life Ladder data follows a normal distribution. However, from the graphs it is not precise enough to tell if iid or independent.

```
# log GDP:  Median : 9.669
high_gdp <- data[data$Log.GDP.per.capita > 9.669,]
low_gdp <- data[data$Log.GDP.per.capita < 9.669,]
high_gdp$group <- 'high_gdp'
low_gdp$group <- 'low_gdp'
gdp_combo <- rbind(high_gdp, low_gdp)
gdp_fig <- ggplot(gdp_combo, aes(Life.Ladder, fill = group))+
           geom_histogram(binwidth=0.1)
gdp_fig
```

```
## Warning: Removed 26 rows containing non-finite values (stat_bin).
```

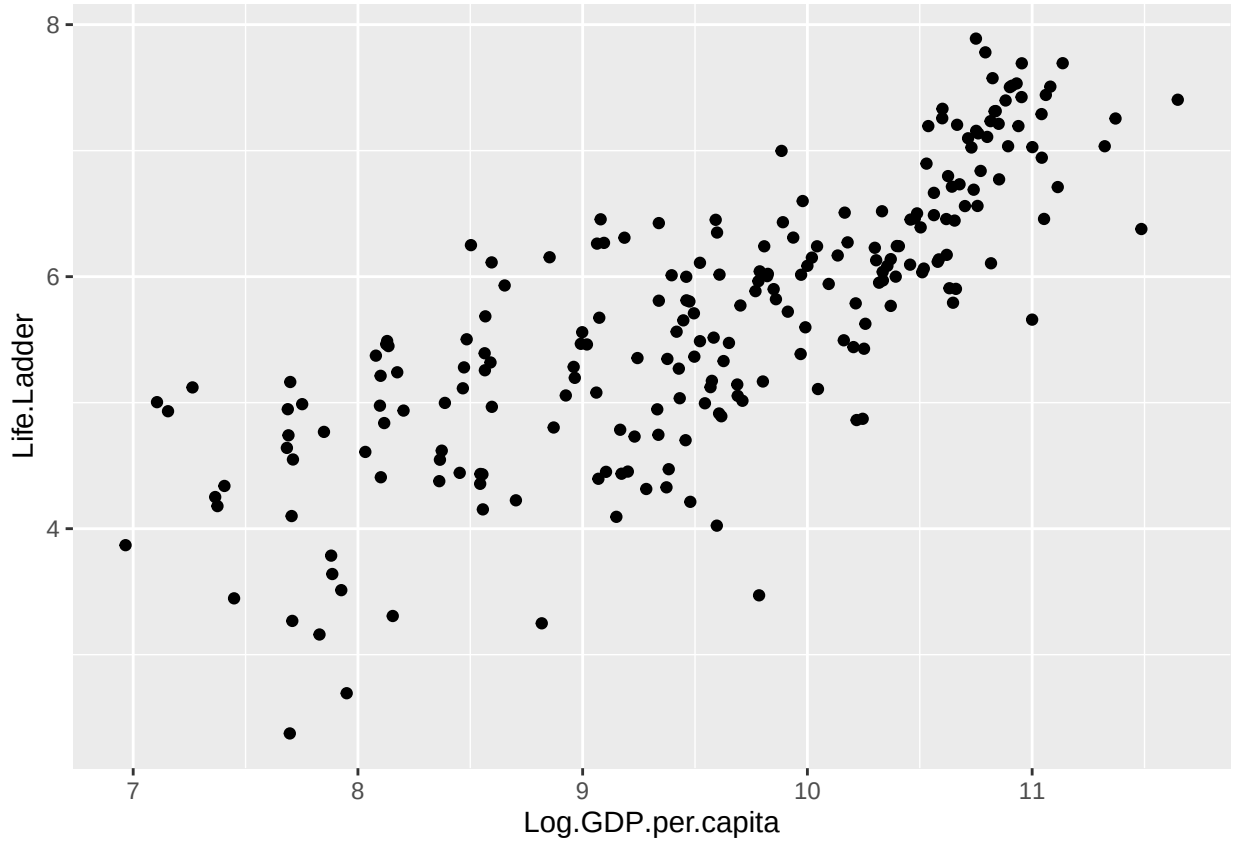


Assumption 3: No major deviations from normality Normality likely occurs, considering the sample size. In particular, the t-test is invalid for highly skewed distributions when sample size is larger than 30 and

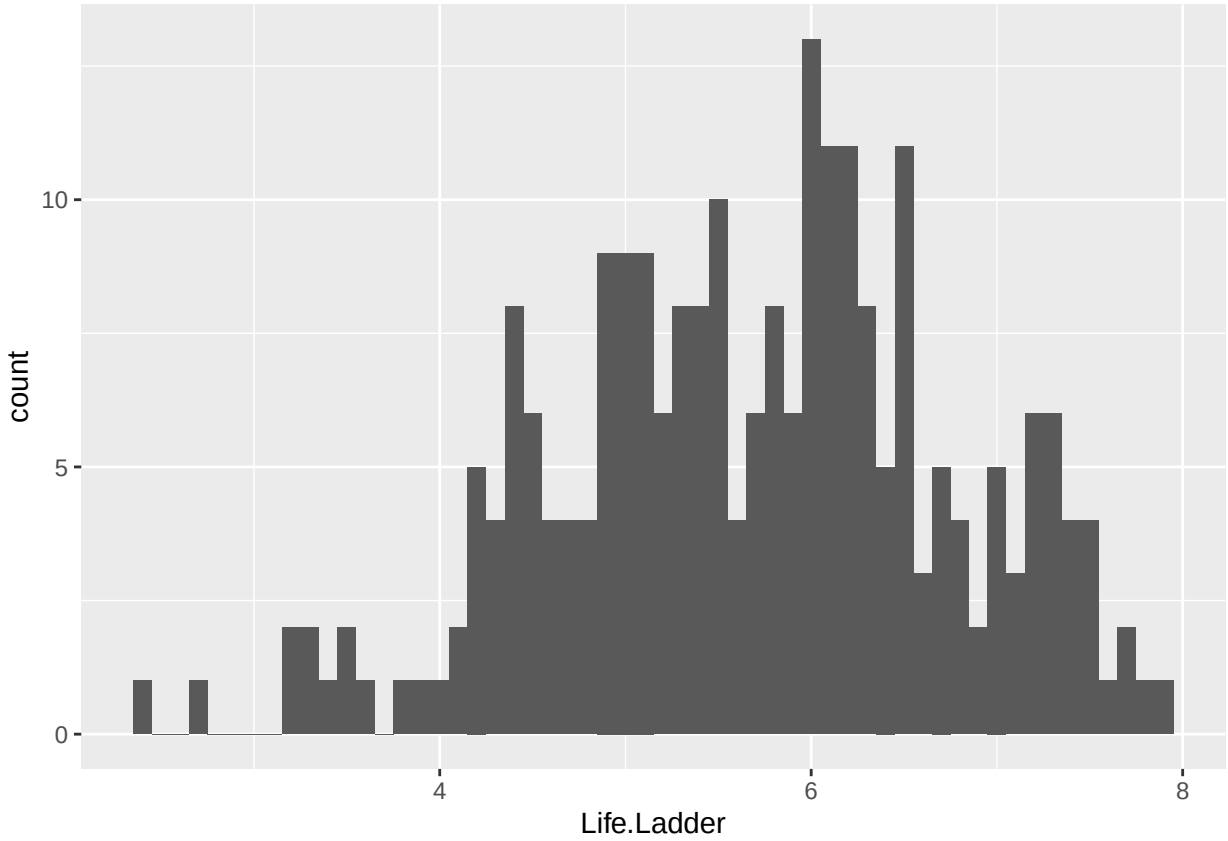
the other assumptions above failing. It may also be invalid for very highly skewed distributions at higher sample sizes. T-tests are usually used for sample sizes less than 30. This is because the t-test and normal distribution will not be distinguishable if the sample size is too big.

```
ggplot(data, aes(Log.GDP.per.capita, Life.Ladder)) + geom_point()
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```




```
p <- ggplot(data, aes(Life.Ladder))+
  geom_histogram(binwidth=.1)
p
```



7	Austria	2019	7.195	10.939	0.96

Showing 1 to 8 of 239 entries, 11 total columns

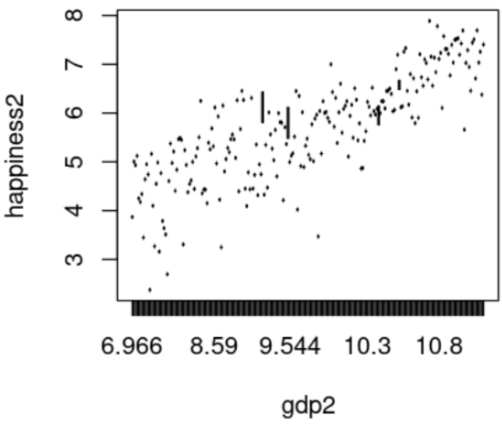
Console Terminal Render Jobs

R 4.1.2 · /cloud/project/

```
        Welch Two Sample t-test

data:  happiness2 and gdp2
t = -39.228, df = 460.93, p-value < 2.2e-16
alternative hypothesis: true difference in means is n
ot equal to 0
95 percent confidence interval:
 -4.101412 -3.710093
sample estimates:
mean of x mean of y
 5.678084  9.583836

>
```



The T-Test above being run fails as well, showing that the data does not fit.

1.4 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.

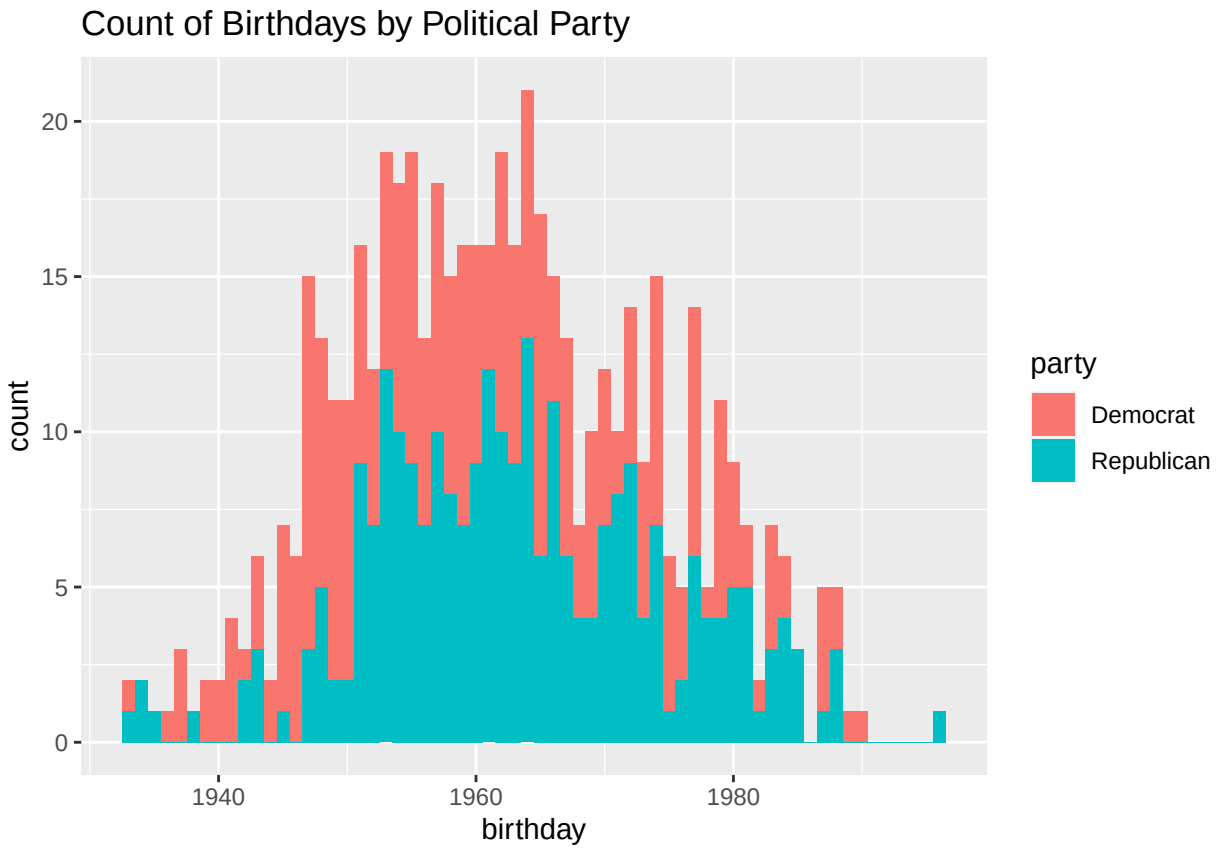
The Wilcoxon Rank-Sum Test using Hypothesis of Comparison is a nonparametric test for two independent samples. Its null hypothesis, $P(X < Y) = P(Y > X)$ compares two groups using only 3 operations ($<$, $=$, $>$) and thus is suitable for comparing ordinal variables.

Assumption 1: Ordinal Scale An ordinal variable is a categorical variable for which the possible values are ordered. Mathematical operations such as addition, subtraction, division, and multiplication cannot be applied to ordinal variables. Instead, ordinal variables can be compared to see what's greater than, less than, or equal to each other. The variable age is represented by "birthdays" in the Legislators dataset. We cannot directly apply mathematical operations to birthday because it is a character datatype, but it can be compared as per the definition of ordinal variables.

```
typeof(data$birthday)
```

```
## [1] "character"
```

Assumption 2: Each pair (X_i, Y_i) is drawn independent of other pairs from the same distribution Each pair of birthday-political party is independent because drawing one does not affect the result of the other. Thus, each pair is drawn independent of other pairs. Below is a graph of the Democrat and Republican sample continuous distributions, which have the same shape and spread. This demonstrates that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_m$ are two independent random samples from continuous distributions with means μ_1 and μ_2.



1.5 Wine and health

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

```
head(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
```

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.

Assumption 1: Metric Scale - Is the data in a metric scale?

```
typeof(wine$deaths)
```

```
## [1] "integer"
```

```
typeof(wine$heart)
```

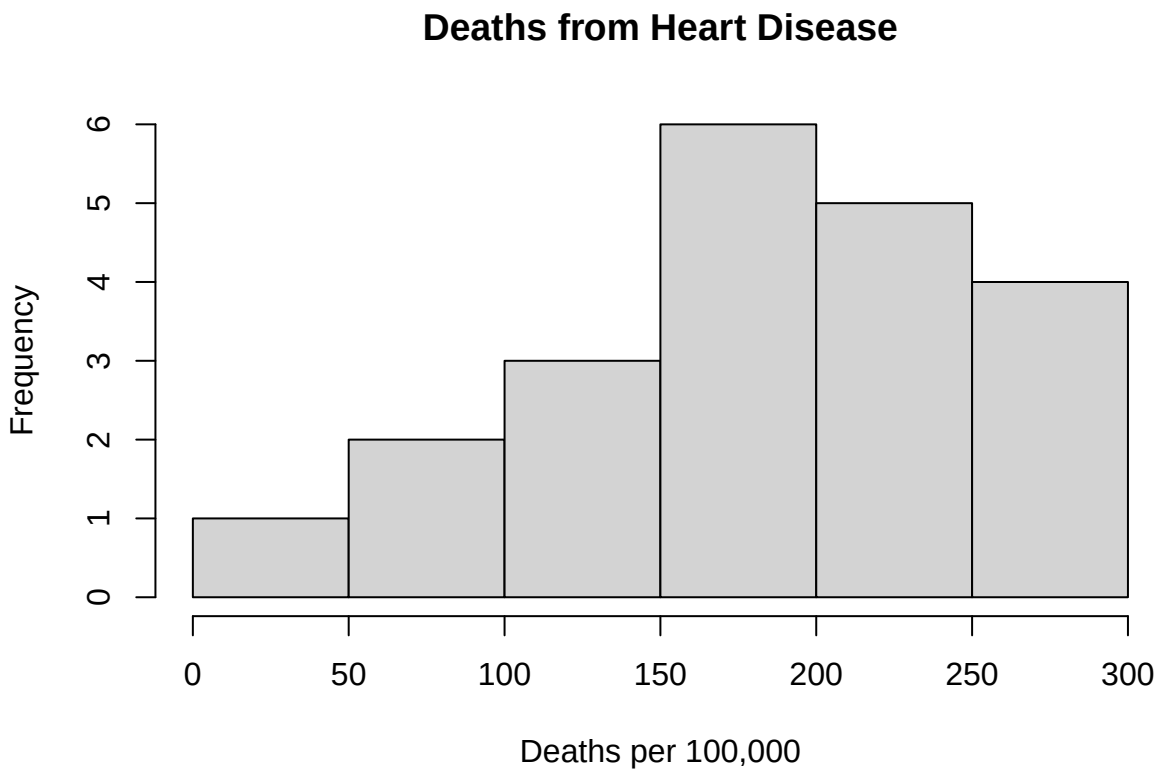
```
## [1] "integer"
```

```
typeof(wine$liver)
```

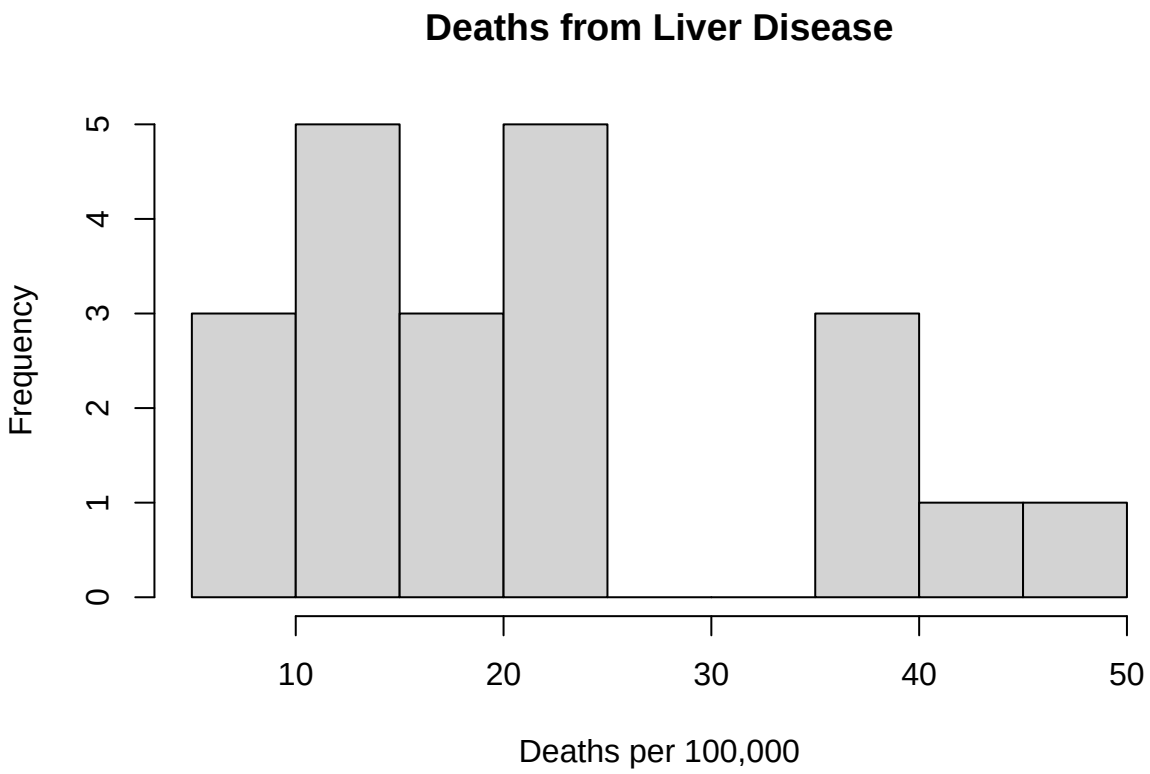
```
## [1] "double"
```

The "deaths" and "deaths from heath disease" fields are both integer data types but the "deaths from liver disease" is a double data type so they will need to be standardized in some way. Outside of that, if the data is being analyzed at a country-level, there are only 21 data points so the central limit theorem may not "kick in" since it is below the recommended 30 samples.

```
hist(wine$heart, main="Deaths from Heart Disease", xlab="Deaths per 100,000")
```




```
hist(wine$liver, main="Deaths from Liver Disease", xlab="Deaths per 100,000")
```

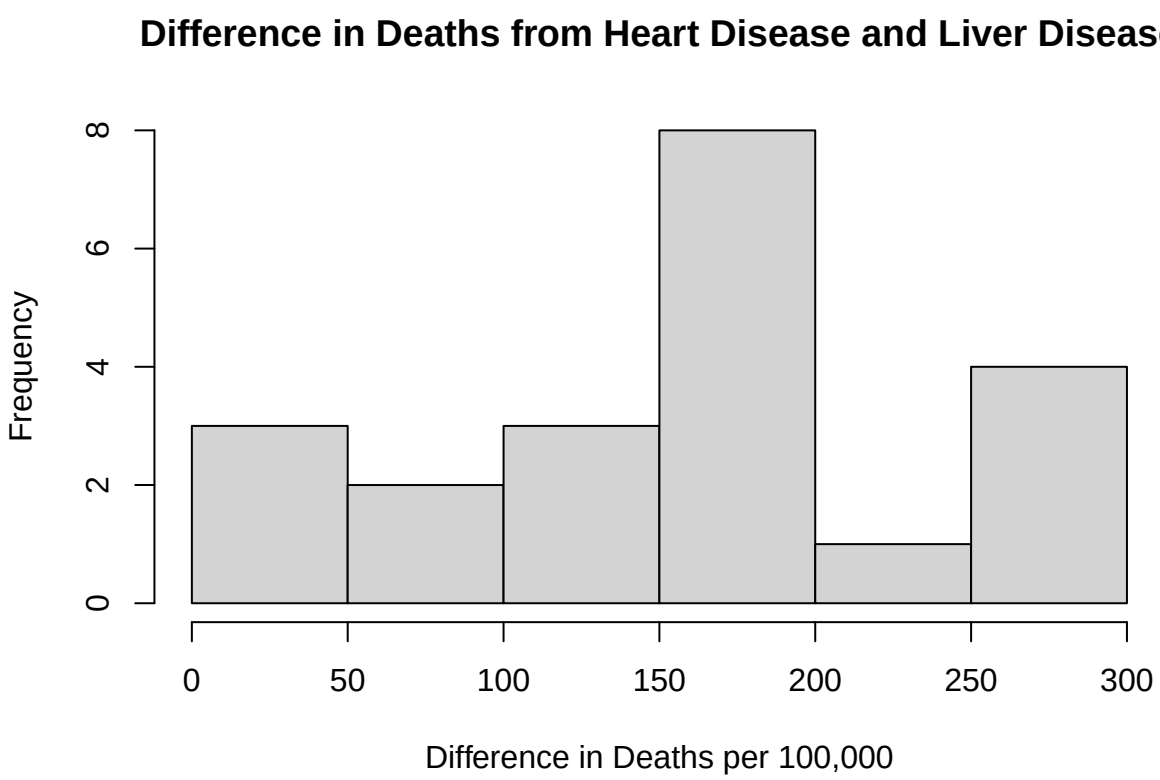


Assumption 2: IID Data - Is the data independent + identically distributed?

The data between countries appears to be independent but there may be other factors caused by cultural differences that could cause an impact in mortality rate.

Assumption 3: The distribution of the difference $X - Y$ is symmetric around some mean μ

```
hist(wine$heart - wine$liver,
     main="Difference in Deaths from Heart Disease and Liver Disease",
     xlab="Difference in Deaths per 100,000")
```



As an extension from the previous point about the CLT not "kicking in", there does not appear to be a symmetric distribution around a mean because the deaths linked to liver disease are not symmetrically distributed.

1.6 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.

Assumption 1. Metric scale - Is the data in a metric scale?

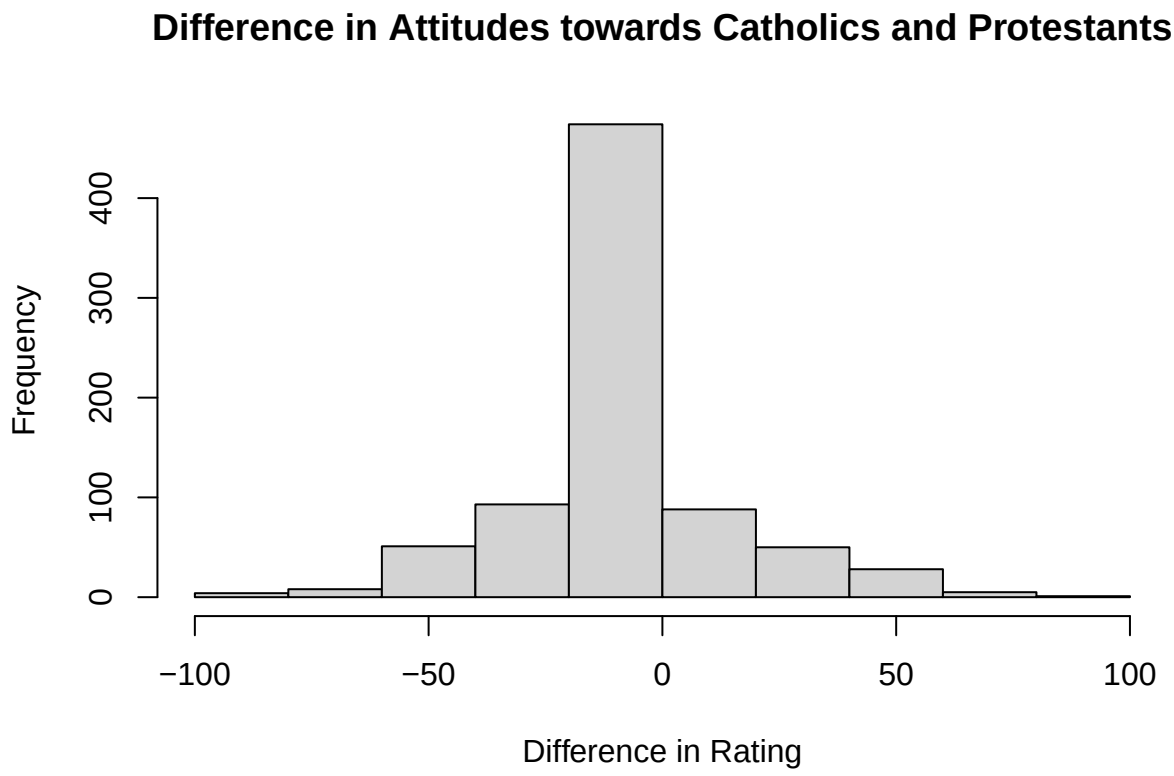
The prottemp and cathtemp variables are both integer values from 0 to 100, but the data is not metric because the values do not scale appropriately, meaning that a score of 50 does not correspond to feeling 2x "better" feeling than a score of 25.

Assumption 2. IID Data - Is the data independent and identically distributed?

It is hard to tell whether the data can be considered IID because we do not know much about the sample population. If the sample consisted of an over-sized number of Catholics or Protestants, it could produce misleading results when extrapolated to the entire US population.

Assumption 3. The distribution of the difference between measurements has no major deviations from normality, considering the sample size.

```
hist(gssReligion$cathtemp - gssReligion$prottemp,
     main="Difference in Attitudes towards Catholics and Protestants",
     xlab="Difference in Rating")
```



Based on the histogram produced by the differences in the measurements, there looks to be an approximately normal distribution so the third assumption passes.