Lab One, Part One

PUT YOUR GROUP NAMES HERE

PUT THE SUBMISSION DATE HERE

Contents

1 Part 1: Foundational Exercises

1.1 Professional Magic

1. **0.0312**
2. **0.105**

1.2 Wrong Test, Right Data

Using a paired test is correct, because our null hypothesis should be that there is no difference between how much a person likes the mobile website and the regular website. Additionally, each person answers the questions, so the data is dependent by person.

However, Likert Scales have ordinal data, and the t-test assumes metric data since it seeks to find the differences between the means of the samples. This means that we would need to make the assumption that the Likert Scale's variations in answers all mean the same thing, which they don't. This can be seen easily when we examine the difference between a 2 and a 3, versus a 4 and a 5 on the Scale. Additionally, people rank their emotions differently, so we care more about what the difference between their two rankings is.

Since we don't want to measure for the mean, and we have ordinal data, we should use Wilcoxon Paired Test instead. This will measure for paired differences between the samples instead of paired means, which is a better way to compare customer's sentiment. It also can handle signed-rank data, which is important for this "difference" calculation.

1.3 Test Assumptions

1.3.1 World Happiness

IID

Independent: Each country is independent because each country does not provide enough information that influences another country's perception of happiness based on GDP.

Identical: Each country is identical because they are each drawn from the same world population.

Metric

Although the Cantril ladder uses numeric values, we believe this data is ordinal because it is a ranking like a Likert Scale.

(Not too un-)normal

Our distribution can be assumed to be normal due to CLT and our samples being greater than 30.

```
# Read from csv
happy <- read.csv('./datasets/happiness_WHR.csv')

# Remove null values of GDP
filtered_happy <- happy[!is.na(happy$Log.GDP.per.capita),]

# Find mean gdp
mean_gdp <- mean(filtered_happy$Log.GDP.per.capita)

# High GDP Samples - 121
sum(filtered_happy$Log.GDP.per.capita > mean_gdp)
```

```
## [1] 121
```

```
# Low GDP Samples - 105
sum(filtered_happy$Log.GDP.per.capita <= mean_gdp)
```

```
## [1] 105
```

1.3.2 Legislators

The Wilcoxon rank-sum test does not require normality, so we removed it from our test assumptions

IID

Independent: Each sample has little to no correlation to another sample besides party which is what we want to use to validate against the distribution of each party's ages.

Identical: All of the democrats and republicans are sampled from the population of legislators. We have also checked that the number of senators is around 50 for each party, so the data is near identical to the population. If the sample size is greater than 50 for either group, then the sample would not be identical because there would be invalid senators that could possibly originate from previous terms of the same state.

```
# Read from csv
legislators <- read.csv('./datasets/legislators-current.csv')

# Keep senators only with valid parties
senators <- legislators[legislators$type == 'sen', ]
```

```
senators <- senators[!is.na(senators$party), ]

# Democratic Sample Size - 48
sum(senators$party == 'Democrat')
```

```
## [1] 48
```

```
# Republican Ssample Size - 50
sum(senators$party == 'Republican')
```

```
## [1] 50
```

Ordinal

Some arguments for using ordinal values is that age is an integer value, not continuous. We can specify the older party from year down to day and the data would still be discrete – the ages would only change the granularity of each bin.

```
head(senators$birthday)
```

```
## [1] "1952-11-09" "1958-10-13" "1943-10-05" "1947-01-23" "1960-04-13"
## [6] "1933-06-22"
```

There are also arguments against using ordinal. Age is a metric and the distance between each age is valuable and consistent. By using ordinal, we would be "weakening" the value of age since non-parametric tests are weaker than parametric tests. Additionally, age 0 has the same meaning as the metric 0 compared to an ordinal 0.

1.3.3 Wine and health

```
library(wooldridge)
?wine
```

IID

Independent: Each country's deaths are independent of each other. There are no variables that can give information about the size of another country's death.

Identical: The sample is **not** identical. We do not know if the 21 countries were randomly chosen from the same world population of around 200 countries. All of the countries chosen have very GDP, which is not representative of all countries including those with lower GDPs like those in Africa.

Metric

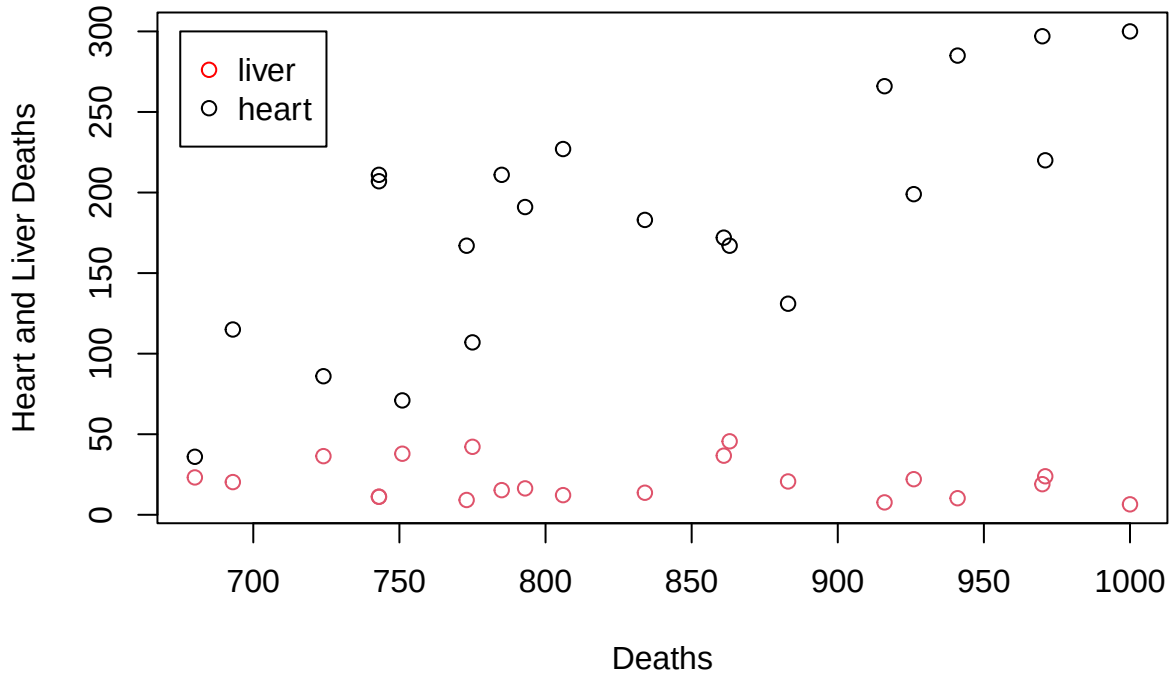
Death counts is a continuous value. Therefore, it s metric.

Difference is Symmetric

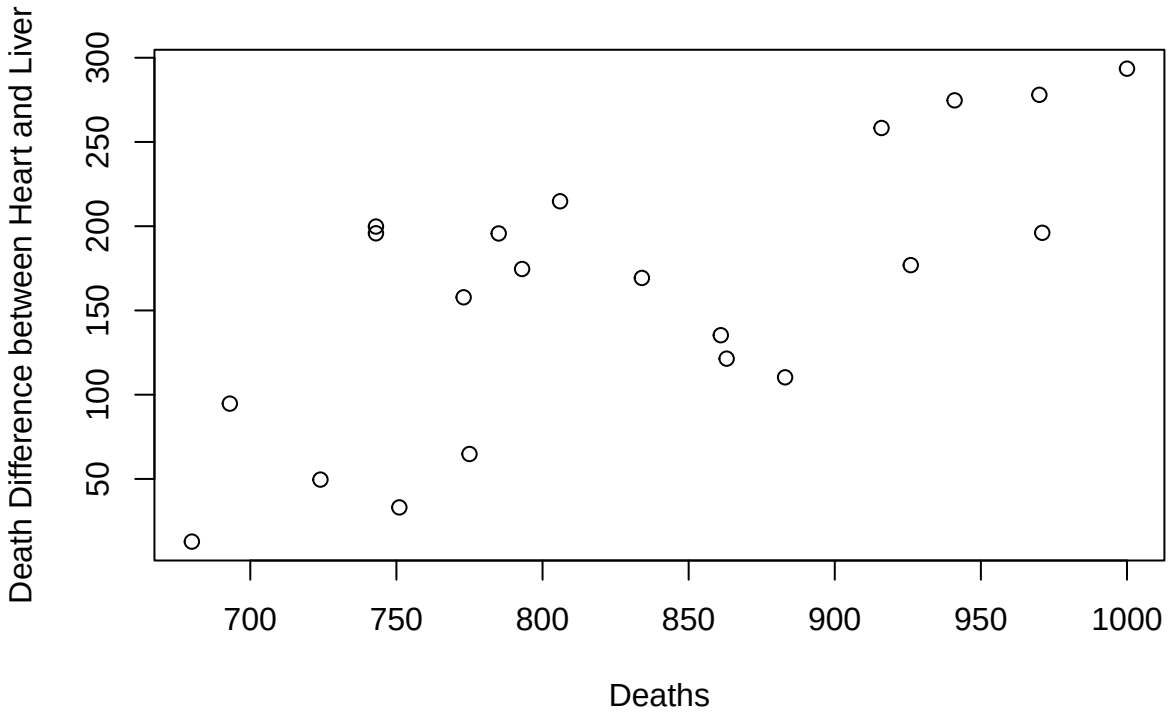
Upon inspection, we can see that the difference is **not** symmetric.

```
wine_df <- data.frame(
  deaths = c(wine$deaths),
  heart_liver_deaths = c(wine$heart, wine$liver),
  category = c(rep('heart', length(wine$heart)), rep('liver', length(wine$liver)))
)

# Plot deaths vs heart and liver deaths
group <- as.factor(wine_df$category)
```

```
plot(heart_liver_deaths ~ deaths, data=wine_df, col = group, xlab="Deaths", ylab="Heart and Liver Deaths
legend(675, 300, legend=c("liver", "heart"), col=c("red", "black"), pch=c(1,1))
```




```
diff_wine_df <- data.frame(
  deaths = c(wine$deaths),
  heart_liver_diff = c(wine$heart - wine$liver)
)

# Plot the difference between heart and liver death given deaths
plot(heart_liver_diff ~ deaths, data=diff_wine_df, xlab="Deaths", ylab="Death Difference between Heart a
```




```
# Tabular version of the death difference
# diff_wine_df[order(diff_wine_df$deaths),]
```

1.3.4 Attitudes toward the religious

IID

Independent: We cannot assume independence. There is no information about the sampling technique. Even though each id is unique and anonymized, we do not know if there is a possiblity that each id is correlated. Are the people from the same household? Are there any demographic constraints?

Identical: We cannot assume that each sample is pulled from the same population. Similar with independence, this has many unknown information. Is each sample randomly grabbed from the US population? There is just not enough data.

Metric

The data based on the feeling thermometer is ordinal instead of metric. This is similar to the Cantril Ladder or Likert Scale, only with a larger scale. How much can we explain the "difference" between values such as the difference between 0-1 and 50-51? Understanding how much one is in favor of a certain religion is hard to quantify. Additionally, there is a upper and lower bound limit from 0 to 100 – answers surpassing those boundaries are invalid.

(Not too un-)normal

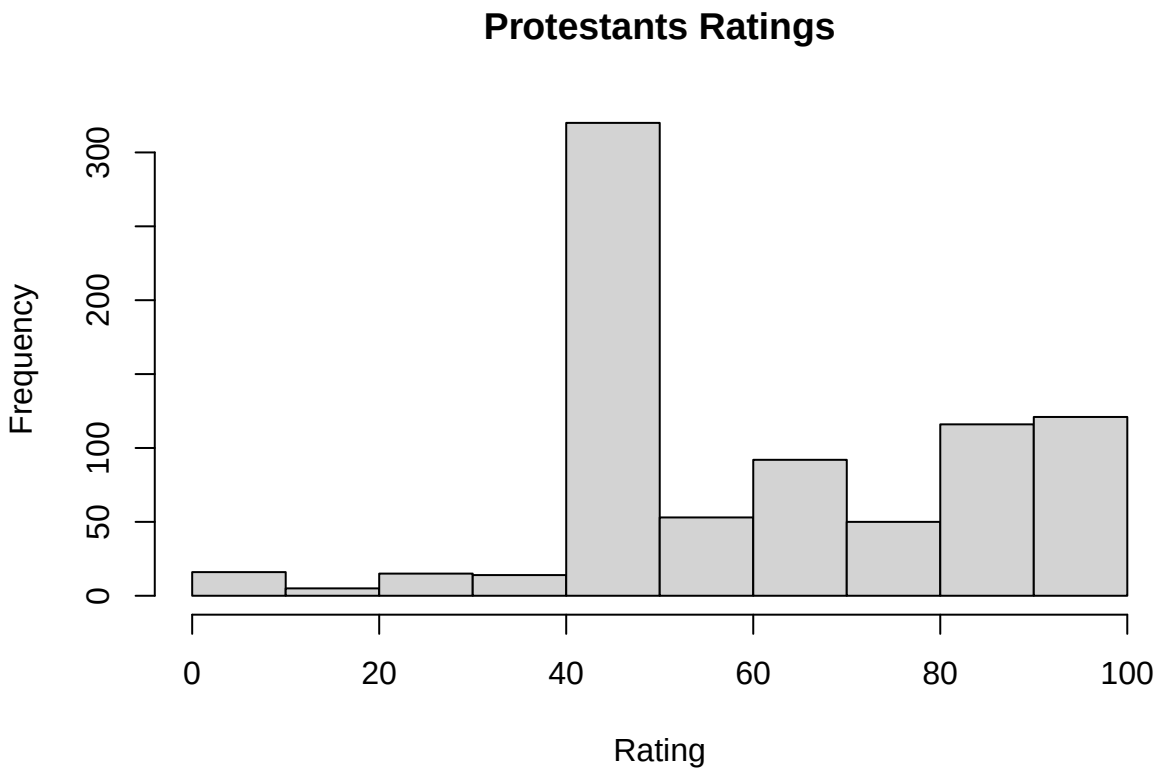
We could state that the distribution is normal. It is not too un-normal because the histogram does vaguely resemble a bell curve. Additionally, the sample size is 802, thus we could utilize CLT to assume normality.

On the other hand, we could argue that the data is skewed with a large sample size, thus the population may also be skewed.

```
religion = read.csv('./datasets/GSS_religion.csv')
nrow(religion)
```

```
## [1] 802
```

```
hist(religion$prottemp, xlab="Rating", main="Protestants Ratings")
```




```
hist(religion$cathtemp, xlab="Rating", main="Catholic Rating")
```