

Part 1: Foundational Exercises

Ken Trinh, Kesha Julien, David Corrales Garcia

2/27/2022

Contents

1 Part 1: Foundational Exercises	1
1.1 Professional Magic	1
1.2 Wrong Test, Right Data	3
1.3 Test Assumptions	4
1.3.1 World Happiness	4
1.3.2 Legislators	4
1.3.3 Wine and health	5
1.3.4 Attitudes toward the religious	7

1 Part 1: Foundational Exercises

1.1 Professional Magic

Based on the information provided, we know that the two coins are represented by Bernoulli Random Variables. To determine whether or not it is indeed magic, we designed a test to flip the two coins three times and record the test statistics of the number of heads shown. In the case of a fair coin toss, our Null Hypothesis (H_0) is $p = 1/2$. Since we are flipping 2 coins three times, we would have at most 6 heads shown. So we plan to reject our Null Hypothesis if our test statistics is 0 or 6.

- a. Here we want to calculate the type 1 error rate. Type 1 error rate refers to the probability of rejecting the Null Hypothesis (H_0) given the Null Hypothesis is True.

$$P(\text{Rejecting } H_0 \mid H_0)$$

Let the test statistics be $T = (X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3)$, X_i and Y_i takes on values $[0, 1]$. Taking the rejection regions as 0 and 6. We expand the type 1 error rate as follow:

$$P(\text{Rejecting } H_0 \mid H_0) = P[(T = 0) \cup (T = 6) \text{ when } T \text{ is Bernoulli}(p/2)]$$

Since each X, Y coin toss is independent:

$$P(\text{Rejecting } H_0 \mid H_0) = P[(T = 0)] + P[(T = 6)] \text{ Independence}$$

Expanding the probability, and substituting

$$p = 1/2$$

into the system since we are given H_0 , we apply multiplication for three coin tosses here since each (X, Y) pairs are IID, and we have three pairs.

$$P[(T = 0)] = [X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 = 0, X_3 = 0, Y_3 = 0] = (p/2)^3 = (1/4)^3 = (1/4)^3 = 1/64$$

$$P[(T = 6)] = [X1 = 1, Y1 = 1, X2 = 1, Y2 = 1, X3 = 1, Y3 = 1] = (p/2)^3 = (1/4)^3 = (1/4)^3 = 1/64$$

$$P(\text{Rejecting } H_o \mid H_o) = 1/64 + 1/64 = 1/32 \approx 0.031$$

b. Here we want to calculate the statistical power of our test for the alternative hypothesis. Statistical Power is related to Type 2 error rate and refers to the probability of supporting the Alternative Hypothesis (H_a) given that the Alternative Hypothesis is True, and the statistical power equals 1 - type 2 error rate.

$$P(\text{Supporting } H_a \mid H_a)$$

We are given that H_a :

$$p = 3/4$$

, and that our rejection region is $[0, 6]$:

$$P(\text{number of heads in the rejection region} \mid H_a) = P(\text{number of heads in } [0, 6] \mid H_a)$$

Since each X, Y coin toss is independent:

$$P(\text{Supporting } H_a \mid H_a) = P[(T = 0)] + P[(T = 6)] \text{ Independence}$$

Expanding the probability, and substituting

$$p = 3/4$$

into the system since we are given H_a , we apply multiplication for 3 coin tosses here since each (X, Y) pairs are IID, and we have three pairs:

$$P[(T = 0)] = [X1 = 0, Y1 = 0, X2 = 0, Y2 = 0, X3 = 0, Y3 = 0] = (p/2)^3 = (3/8)^3 = 27/512$$

$$P[(T = 6)] = [X1 = 1, Y1 = 1, X2 = 1, Y2 = 1, X3 = 1, Y3 = 1] = (p/2)^3 = (3/8)^3 = 27/512$$

Thus, the statistical power, or the probability of not making a type 2 error is computed as follows:

$$\text{Statistical Power} = 27/512 + 27/512 = 27/256 \approx 0.105$$

1.2 Wrong Test, Right Data

Let X be the Random Variable describing how much customers like the regular website. Let Y be the Random Variable showing how much customers like the mobile website. The scale of measurement is on the 5 points Likert scales. Given that the Likert scale is an ordinal scale, if we are using the paired t-test, we would violate the assumption that the t-test uses a metric scale. To calculate t-test statistics in the paired t-test, we know that we need to calculate an expected value and the sample variation for both random variable X , Y . For that, the scale has to be continuous. The Likert scale, being an ordinal scale, has no objective variation between one value to another. We believe this would cause one to compute a bias test statistics, likely larger, making our test sample to fall into the rejection region, which would increase the type 1 error rate. To correct for this, we suggest Wilcoxon Sign-Rank Test because the two samples are dependent.

1.3 Test Assumptions

1.3.1 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. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

```
# load the happiness dataset
happiness_dataset <- read.csv('datasets/happiness_WHR.csv')
```

Assumptions for a two-sample t-test: 1. Metric scale.

```
summary(happiness_dataset$Life.Ladder)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.375   4.971   5.768   5.678   6.428   7.889
```

Even though the life ladder scale is ordinal in nature, the data presented could be considered metric as it includes decimal averages for every individual in a country. For this reason, we believe that the assumption that we are using metric data is met.

2. IID data. Sample 1: Happiness in countries with high GDP per capita. Sample 2: Happiness levels in countries with low GDP per capita. Given the two samples, we believe that the happiness level of one individual would not provide information about the happiness of another individual, making us conclude that these values are independent. Provided that each individual's happiness level is derived from the same scale, we agreed that the values of the samples can also be considered identically distributed. For these two reasons, we believe the assumptions of IID are met.

3. Considering the sample size, we agree that there would be no major deviations from normality.

```
summary(happiness_dataset$Log.GDP.per.capita)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.966   8.827   9.669   9.584  10.527  11.648      13
```

```
happiness_dataset$is_high_gdp <- happiness_dataset$Log.GDP.per.capita > 9.584
sum(!is.na(happiness_dataset$Life.Ladder[happiness_dataset$is_high_gdp == T]))
```

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

```
sum(!is.na(happiness_dataset$Life.Ladder[happiness_dataset$is_high_gdp == F]))
```

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

Assuming that the samples are split based on the GDP per capita mean value of 9.584, we found that both samples exceed size of 30, so the Central Limit Theorem is met, expecting approximation of a normal distribution. For this reason, we conclude that there are no major deviations from normality.

1.3.2 Legislators

```
legislators_current <- read.csv('datasets/legislators-current.csv')
```

Assumptions for a Wilcoxon rank-sum test (Hypothesis of Comparison):

1. Our data is an ordinal scale: The original data was given as date of birth. To make the sample usable, the senator age is calculated using the current year 2022 minus the birth year, as a discrete value from 0 to 100. There is no continuous relationship between two age interval (ie. 89, 90 are two distinct ages). For that reason, we believe the assumption that the age is on an ordinal scale is met.

```
# make a column for age and plot it
legislators_current$age <- 2022 - strtoi(substr(legislators_current$birthday, 1, 4))
hist(legislators_current$age, main = 'Senator Age', breaks=20)
```

2. IID: The age of the senator is independent because a birth year of one senator does not provide any information to determine the birth year of another senator. The senator age is also identically distributed because it is derived from the same birth year convention on a regular calendar scale. For those reasons, the IID assumption is met!

1.3.3 Wine and health

```
# Installing pre-req packages
# load the wine dataset
install.packages("wooldridge")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
library(wooldridge)
```

The wine dataset contains observations of variables related to wine consumption for 21 countries. We would like to use this data to test whether countries have more deaths from heart disease or from liver disease.

Assumptions for a Wilcoxon Signed-Rank Test:

- 1) The data is a metric scale.
 - In particular, X and Y are both measured on the same metric scale.

By viewing the format of the wine dataset, the columns heart and liver are denoted as heart disease deaths per 100,000 and liver disease deaths per 100,000, respectively. Since the two columns both measure death per 100,000 individuals, it is measuring a continuous average number of death and is measuring from the same scale. Therefore the assumption that both samples are metric is met.

```
summary(wine$heart)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      36.0   131.0   191.0   183.3   220.0   300.0
```

```
summary(wine$liver)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      6.50   11.20   19.00   21.03   23.90   45.60
```

- 2) IID data.

- In particular, each pair (X_i, Y_i) is drawn from the same distribution, independently of all other pairs.

Let X_i be the number of deaths caused by heart disease in country i and Y_i be the number of deaths caused by liver disease in the same country i , such that, both, X_i and Y_i are pairs of the same total distribution of death of the same country, making the pair naturally dependent by virtue of being part of the same country, and independent from other pairs, belonging to other countries. In this way, we find that we can generate pairs (X_i, Y_i) from the same distribution and independently of other pairs, thus, meeting this assumption as well. This can be seen in the following distribution @ref(plot:histogram-dist)

```
par(mfrow=c(1,2))
hist(wine$heart, xlab="Heart Disease Related Death", main="Heart Disease Related Death")
hist(wine$liver, xlab="Liver Disease Related Death", main="Liver Disease Related Death")
```

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

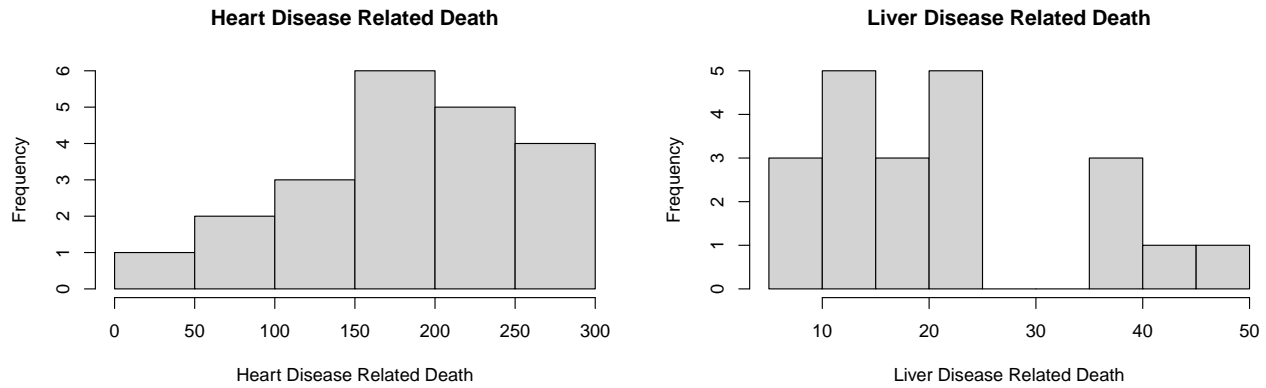


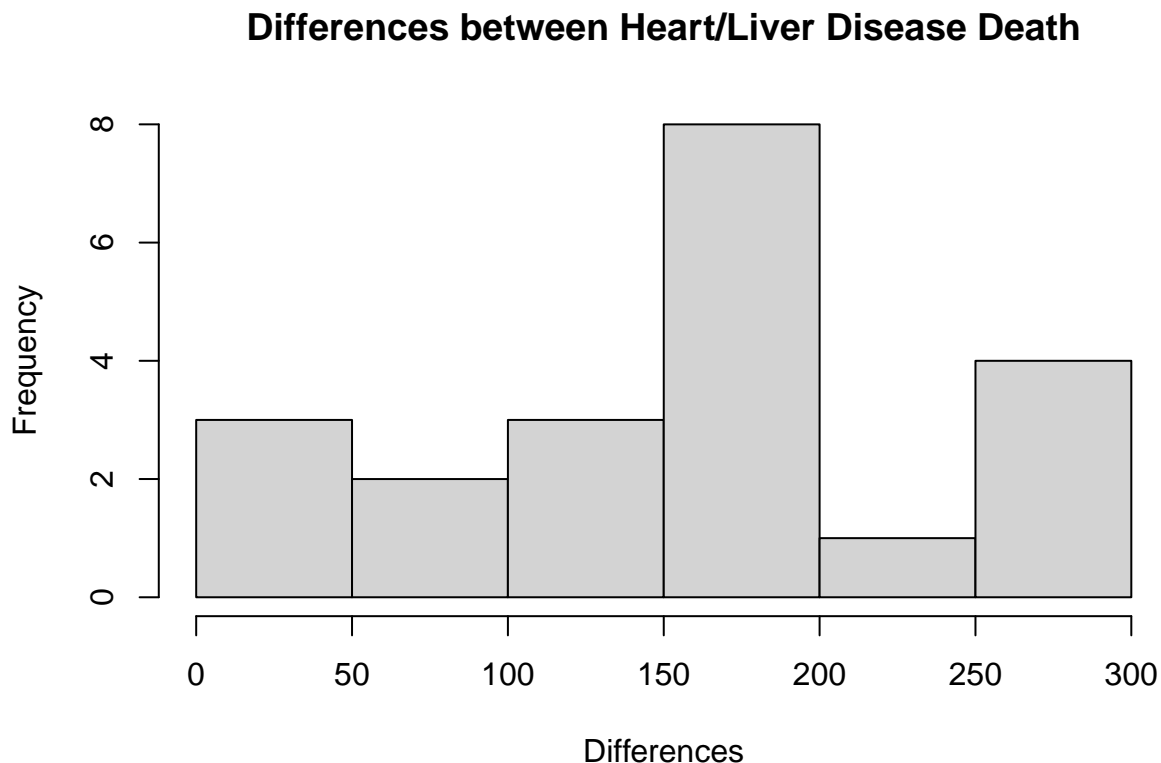
Figure 1: Distribution of Heart and Liver Disease Related Death

Here the number of observations for the differences between death by heart disease and death by liver disease is calculated to be $n=21$. This number is not large enough to meet the 30 minimum as ascribed by the Central Limit Theorem. Additionally, when plotted in a histogram, we see that the distribution does not show a normal distribution either. For those reasons, we agree that the assumption that the distribution of the difference $X - Y$ is symmetric around some mean μ is not met.

```
wine$diff <- wine$heart-wine$liver
sum(!is.na(wine$diff))
```

```
## [1] 21
```

```
hist(wine$heart-wine$liver,
     xlab="Differences",
     main='Differences between Heart/Liver Disease Death',
     breaks=5)
```



1.3.4 Attitudes toward the religious

```
GSS_religion <- read.csv('datasets/GSS_religion.csv')
head(GSS_religion)
```

```
##   X year id prottemp cathtemp
## 1 1 2004 4      50      50
## 2 2 2004 7       5      85
## 3 3 2004 9      50      50
## 4 4 2004 14     60      60
## 5 5 2004 21    100       0
## 6 6 2004 24      5       5
```

Assumptions for a Paired t-Test:

1) Metric scale.

- In particular, the t-test is not valid for variables which only have an ordinal structure.

```
summary(GSS_religion$prottemp)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.00  50.00   60.00   65.56  85.00   100.00
```

```
summary(GSS_religion$cathtemp)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.00  50.00   60.00   63.16  85.00   100.00
```

To assess whether the two prottemp and cathtemp samples are on a metric scale, we evaluated if the data is presented with a degree of difference between categories. From the description of the feeling thermometer (a 0-100 rating), a rating between 50 degrees and 100 degrees means that one feels favorable and warm toward the group and a rating between 0 degrees and 50 degrees means that one does not feel favorable toward the group. Theoretically, if one looks at the interval [49,51] one can see that any values between 50 to 51 degrees fall into the favorable group while any values between 49 to 50 fall into the unfavorable group. Using this logic, we find that there is information between interval for these two samples. Therefore, the assumption of metric scale is met.

2) IID data.

- In particular, each pair of measurements (X_i, Y_i) is drawn from the same distribution, independently of all other pairs.

The description of the GSS questions suggested that the feeling thermometer is a rating from 0 to 100. The rating is obtain by asking the participant to give a rating. Keeping this in mind when viewing the summary of the two samples in this dataset, prottemp and cathtemp, we see that both have a minimum value of 0 and a maximum value of 100. Additionally, the inner quartile range, computed by taking the difference between the 3rd quartile and the 1st quartile, is equal to 35 for both samples. From this standpoint, each sample pair (X_i, Y_i) is drawn from the same distribution ranging from a scale of 0 to 100 shown in Figure @ref(fig:box-plot). Therefore they are identically distributed. In terms of independence, because a given sample pair (X_i, Y_i) is an individual's rating on a protestant group and catholic group, it provides no information on another user rating for the two groups, we conclude that each sample pair is independent. For these reason, we believe that the assumption of IID data is met.

```
par(mfrow=c(1,2))
b1 <- boxplot(GSS_religion$prottemp, main="Protestant Temperature")
b2 <- boxplot(GSS_religion$cathtemp, main="Catholic Temperature")
```

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

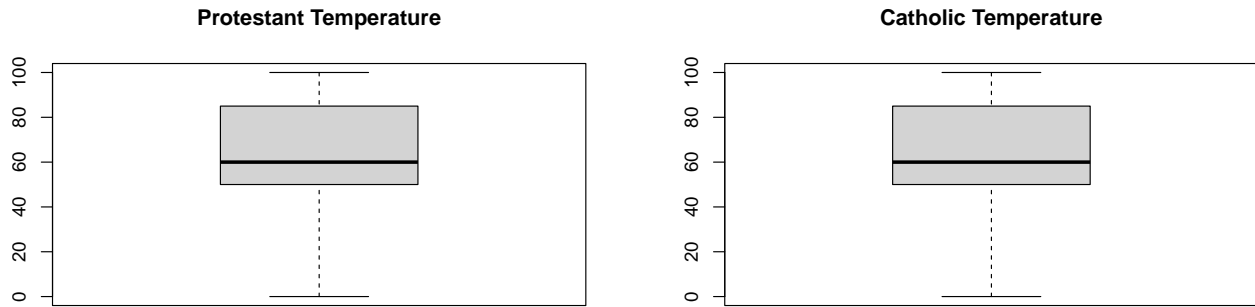
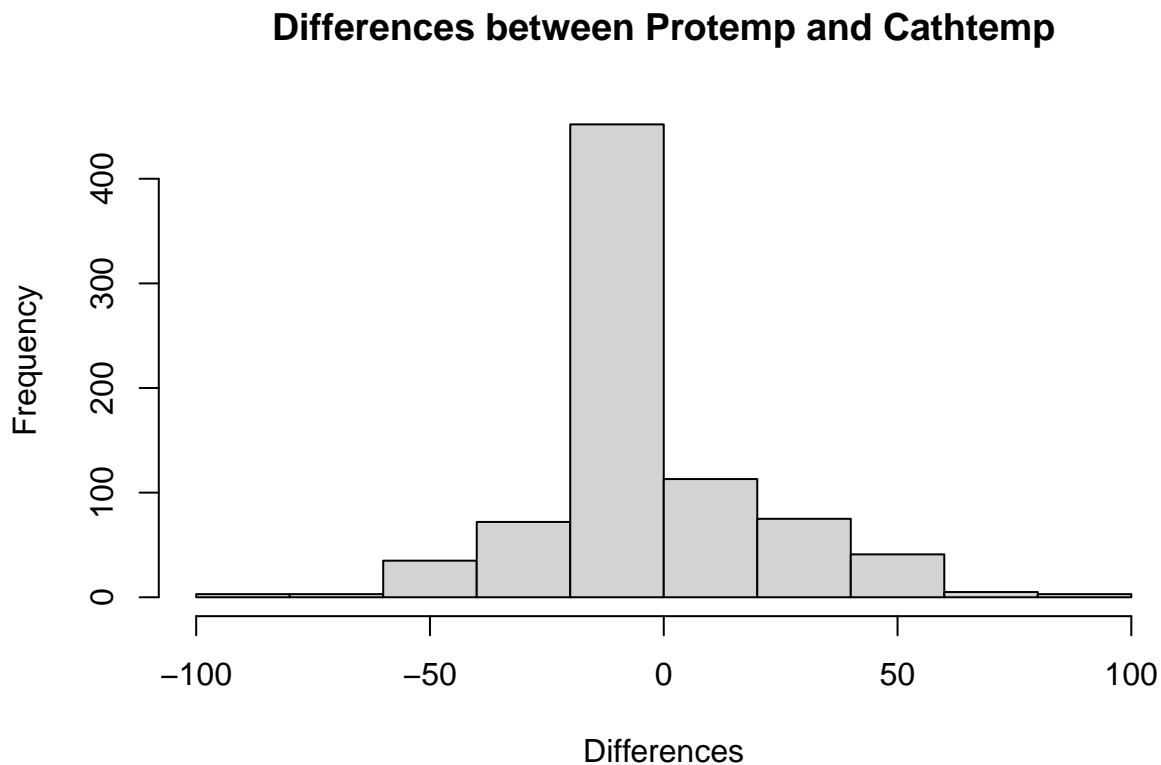


Figure 2: Distribution of Protestant and Catholic Temperature

- In particular, the t-test is invalid for highly skewed distributions when sample size is larger than 30. It may also be invalid for very highly skewed distributions at higher sample sizes.

```
hist(GSS_religion$prottemp-GSS_religion$cathtemp,
     xlab="Differences",
     main='Differences between Prottemp and Cathtemp', breaks=10)
```



To assess if the difference between the two measurements has no major deviation from normality, we first need to compute the differences between the two samples and count the number of non-NAN values in a results column. Here we see that the sample size is $n=802$, which exceeds a minimum bound of $n=30$. For this reason, we agree that the Central Limit Theorem assumption is met, approximating a normal distribution. Viewing the histogram with number of bins($\text{breaks} = 10$), we also see that the sample is relatively distributed with center at approximately 0. For these reasons, the assumption of no deviation from normality is met.