Lab One, Part One

Kevin Lustig, Rebecca Nissan, Anuradha Passan, Giorgio Soggiu

3/1/2022

Contents

1 Foundational Exercises

1.1 Professional Magic

1.1.1 Type I Error of the test

The type I error rate (i.e. false positive) is the probability of rejecting the null hypothesis when it is correct. In this case, the type I error would be the probability of getting 0 or 6 for your test statistic (and therefore rejecting the null) given that the null is true, i.e. p = 1/2.

$$\text{Let } Z = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$$

$$P(Z = 0 \text{ or } Z = 6 | p = \frac{1}{2}) =$$

$$P(Z = 0 | p = \frac{1}{2}) + P(Z = 6 | p = \frac{1}{2}) =$$

Each flip of the pair is independent from all other flips of that pair.

$$[P(X_1 = 0 \text{ and } Y_1 = 0 | p = \frac{1}{2}) * P(X_2 = 0 \text{ and } Y_2 = 0 | p = \frac{1}{2}) * P(X_3 = 0 \text{ and } Y_3 = 0 | p = \frac{1}{2})] + [P(X_1 = 1 \text{ and } Y_1 = 1 | p = \frac{1}{2}) * P$$

$$[\frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2}] + [\frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2}] =$$

$$[\frac{1}{4} * \frac{1}{4} * \frac{1}{4}] + [\frac{1}{4} * \frac{1}{4} * \frac{1}{4}] =$$

$$\frac{1}{32}$$

1.1.2 Power of test given p = 0.75

The power of the test is equal to the probability of correctly rejecting the null hypothesis when the null is false and the alternative is true. In this case, the power would be the probability of getting 0 or 6 for our test statistic (and therefore rejecting the null) given that the alternative is true, i.e. p = 3/4.

$$\text{Let } Z = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$$

$$P(Z = 0 \text{ or } Z = 6 | p = \frac{3}{4}) =$$

$$P(Z = 0 | p = \frac{3}{4}) + P(Z = 6 | p = \frac{3}{4}) =$$

Each flip of the pair is independent from all other flips of that pair.

$$[P(X_1 = 0 \text{ and } Y_1 = 0 | p = \frac{3}{4}) * P(X_2 = 0 \text{ and } Y_2 = 0 | p = \frac{3}{4}) * P(X_3 = 0 \text{ and } Y_3 = 0 | p = \frac{3}{4})] + [P(X_1 = 1 \text{ and } Y_1 = 1 | p = \frac{3}{4}) * P$$

$$[\frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2}] + [\frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2}] =$$

$$[\frac{3}{8} * \frac{3}{8} * \frac{3}{8}] + [\frac{3}{8} * \frac{3}{8} * \frac{3}{8}] =$$

$$\frac{27}{512} + \frac{27}{512} =$$

$$\frac{27}{256}$$

1.2 Wrong Test, Right Data - Kevin

In the Likert scale, the meaningful distance between the different scale points is not consistent. That is, assuming the Likert scale for the websites survey includes five points from 1 = "Very Unsatisfied" to 5 = "Very Satisfied," with 3 being "Neutral," we cannot say that a change from 1 to 3 and from 2 to 4 are equivalent quantifiable changes in opinion [1]. In fact, the change in quality of experience necessary for a given respondent to go from Very Unsatisfied with one site to Neutral with the other may be considerably less than the change needed to go from Unsatisfied to Satisfied, though these each consist of a difference of two points. Therefore, though the values produced are numeric, these data violate one of the assumptions for a paired t-test – the use of metric, rather than ordinal, data.

A paired t-test relies on metric data because, like other related tests including the z-test, it is fundamentally a calculation of the difference of means between reference groups. A paired t-test would ask of our survey data: is the mean difference between paired opinion scores different than what we would expect if there were no preference for either website (mean difference within pairs = 0)? Stated otherwise, on average across all respondents, how likely is it that there is really a preference for one site or the other, and how large a preference? However, because of the aforementioned limitation of Likert scale values, we cannot meaningfully parse a mean paired disparity of e.g., +2, because to calculate this requires assuming that non-comparable changes from any one Likert scale point to another are equivalent. The mean of the paired differences is thus meaningless. It is even difficult to trust the directionality of the mean difference across all pairs (respondents like the mobile website more or less than the regular website without regard to how much), as in calculating a mean value purely from the raw Likert scale scores, we may calculate an incorrect value by not correctly taking into account the "weights" of the differences of opinion in, again, Very Unsatsified and Neutral versus Unsatisfied and Satisfied. It's possible to conceive of a scenario in which even the sign of the mean difference is therefore incorrect.

It is this last point on directionality that suggests an alternative approach to this analysis. A non-parametric paired sign test allows us to analyze our ordinal data provided the observations are independent and identically distributed. It does not attempt, like the t-test, to quantify the size of the difference in opinion within pairs, if any. Rather, it treats all positive changes in opinion as equivalent, and does likewise with all negative changes. This alternative test has two main drawbacks. First, it does not have the statistical power of a paired t-test. Second, it loses substantial information present in the original survey responses in the form of the exact values within each paired set of responses. However, in doing so, it allows us to avoid the inaccurate mean calculation of the t-test, and focus on a more accurate analysis of a simpler question: do respondents prefer one website over the other? In looking solely at increases or decreases in opinion score, the paired sign test therefore gives us a reasonable expectation of finding such an effect if one is present in the data.

[1] At least, not with only five scale points; see, e.g.: Huiping Wu and Shing-On Leung, "Can Likert Scales Be Treated as Interval Scales?—a Simulation Study," Journal of Social Service Research 43, no. 4 (June 2017): pp. 527-532, https://doi.org/10.1080/01488376.2017.1329775.

1.3 Test Assumptions

1.3.1 World Happiness

Scenario: We have two variables: Life.Ladder and Log.GDP.per.Capita, and we want to see whether countries in high GDP per capita are more or less happy than people in countries with low GDP per capita.

Proposed test: Two Sample t-Test

Test Assumptions:
1. Metric variables
2. Random variables are independent and identically distributed (hereby referred to as i.i.d.)
3. Normality of random variables

The "Life Ladder" score (LLS) variable is composed of continuous values. Data can be classified and distance between values make sense. This point is important considering that the Two-Sample t-Test aims at comparing means of two random variables. The "Life Ladder" seems to correspond to a metric variable which satisfies the first assumption. The "Log GDP per capita" is used to divide the studied population into two groups and is not directly used by the test.(Thus, the data type considerations related to the "Log GDP per capita" are not useful.)

Validating the Independence of the random samples seems hazardous assuming that countries and people might be linked somehow. In other word, knowing one information about one country might provide information/insights of neighboring countries. The second assumption is not completely verified.

A data exploration of both "Life.Ladder" and "Log.GDP.per.capita" variables has revealed that "Log.GDP.per.capita" contains 13 Na(s) values. Countries having Na values for the GDP will be omitted for the rest of the study. Two histograms are plotted to help visualizing the distribution of the data. The x-axis displays the Life Ladder Scores and the y-axis the frequency of each Life Ladder Score values.These distributions are presented for countries from the "high GDP per capita" and "Low GDP per capita" groups.

Summary: Life Ladder Score and GDP per Capita

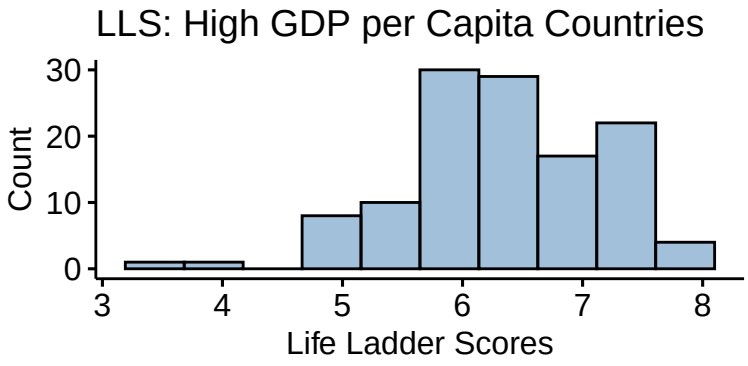
```
  Life.Ladder    Log.GDP.per.capita
 Min.   :2.375   Min.   : 6.966
 1st Qu.:4.971   1st Qu.: 8.827
 Median :5.768   Median : 9.669
 Mean   :5.678   Mean   : 9.584
 3rd Qu.:6.428   3rd Qu.:10.527
 Max.   :7.889   Max.   :11.648
                 NA's   :13
```

Summary: Life Ladder Score and GDP per Capita > Sample Mean

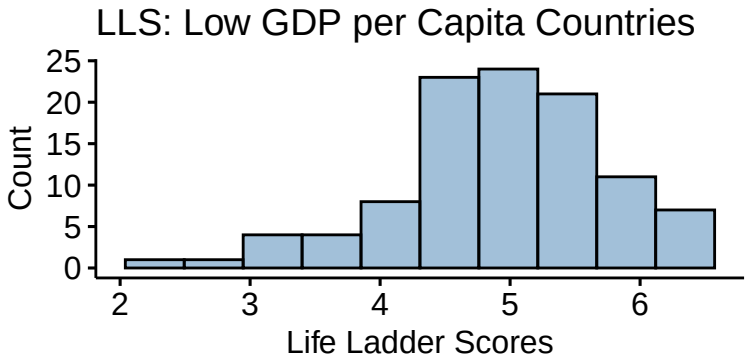
```
  Life.Ladder    Log.GDP.per.capita
 Min.   :3.471   Min.   : 9.583
 1st Qu.:5.917   1st Qu.: 9.993
 Median :6.291   Median :10.483
 Mean   :6.349   Mean   :10.415
 3rd Qu.:7.027   3rd Qu.:10.768
 Max.   :7.889   Max.   :11.648
```

Summary: Life Ladder Score and GDP per Capita < Sample Mean

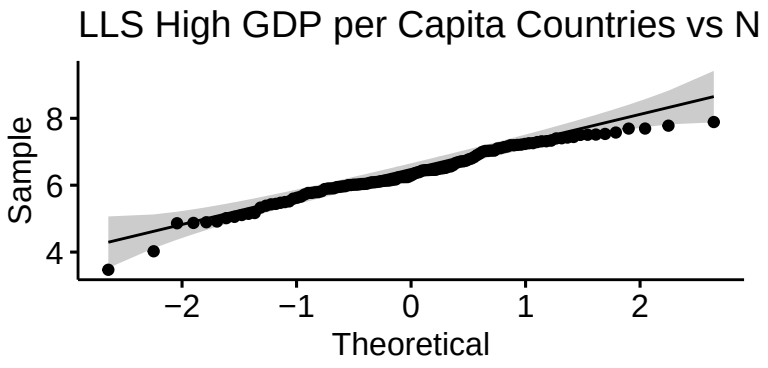
```
 Life.Ladder    Log.GDP.per.capita
Min.   :2.375   Min.   :6.966
1st Qu.:4.433   1st Qu.:8.100
Median :4.992   Median :8.592
Mean   :4.919   Mean   :8.609
3rd Qu.:5.463   3rd Qu.:9.234
Max.   :6.455   Max.   :9.575
```

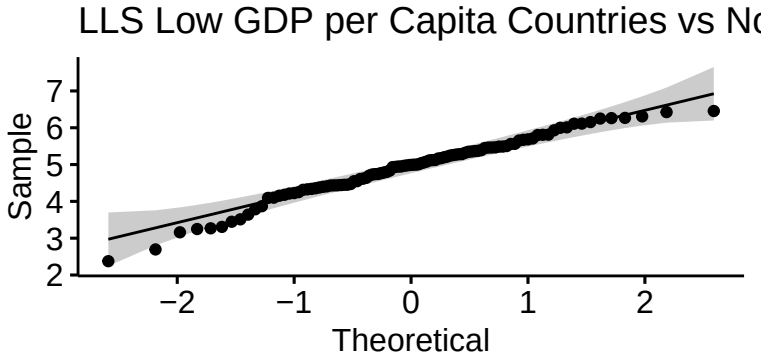


The distribution is left skewed and contains two peaks. The overall shape does not represent a normal distribution. The second distribution (below) contains one single peak closer to the middle of the distribution and shows a symmetric behavior. In that sense, the second distribution seems to be a better approximation of the normal distribution.



Density plots are used below to better assess the correlation between the samples and the normal distribution. Black dotes closely positioned to the 45 degrees reference line would suggest a data distribution close to the normal distribution.





The positions of black dotes fluctuates around the reference line without forming a straight line. The visual study of distribution does not provide clear yes/no answer. To quantitatively assess the distribution, a Shapiro-Wilk normality test has been performed on both Ladder scores random variables from Low and high GDP countries.

```
    Shapiro-Wilk normality test

data:  low_gdp_cap$Life.Ladder
W = 0.97549, p-value = 0.05055
```

$p > 0.05$ (barely) –> normal-ish

```
    Shapiro-Wilk normality test

data:  low_gdp_cap$Log.GDP.per.capita
W = 0.93951, p-value = 0.0001319
```

$p < 0.05$ –> not normal

```
    Shapiro-Wilk normality test

data:  high_gdp_cap$Life.Ladder
W = 0.97281, p-value = 0.01434
```

$p < 0.05$ –> not normal

```
    Shapiro-Wilk normality test

data:  high_gdp_cap$Log.GDP.per.capita
W = 0.96977, p-value = 0.00764
```

$p < 0.05$ –> not normal

The application of the test over the Life Ladder r.v from the Low GDP countries provides a p value barely superior to 0.05. The distribution may be normally distributed. Nevertheless, it is not the case for High GDP countries with p value inferior than 0.05. The second distribution can not be considered as normally distributed. Hence, the third assumption is not verified.

Conduct the test?

Considering the facts that IID is hazardous and random variables not normally distributed, the Two Sample t-Test is not applicable in this context.

1.3.2 Legislators

Scenario: We want to test whether Democratic or Republic senators are older, with two variables party and age (age needs to be calculated from DOB).

Proposed test: Wilcoxon Rank Sum Test

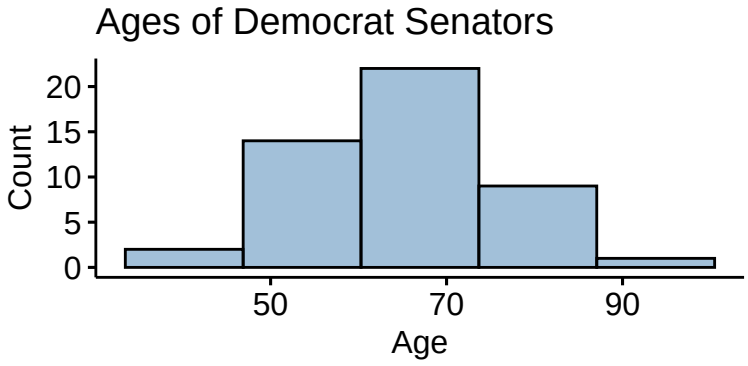
Test Assumptions:
1. Metric variable //Ordinal variables
2. i.i.d.
3. Same shape and spread of the the two variables

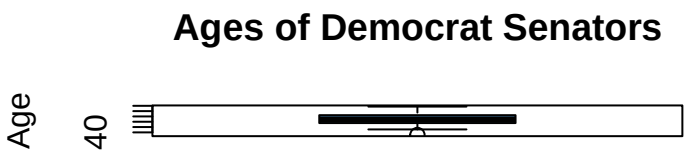
Summary: Democrat Senators

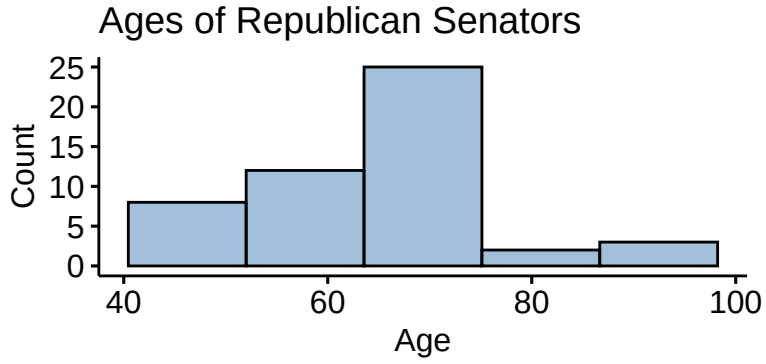
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.99   57.68   64.21   64.19   72.30   88.57
```

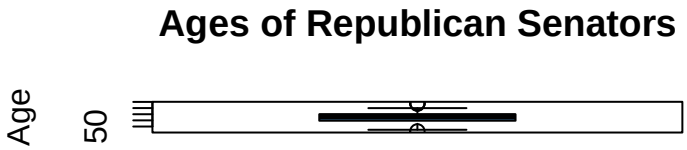
Summary: Republican Senators

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.11   59.64   66.36   64.83   69.89   88.33
```









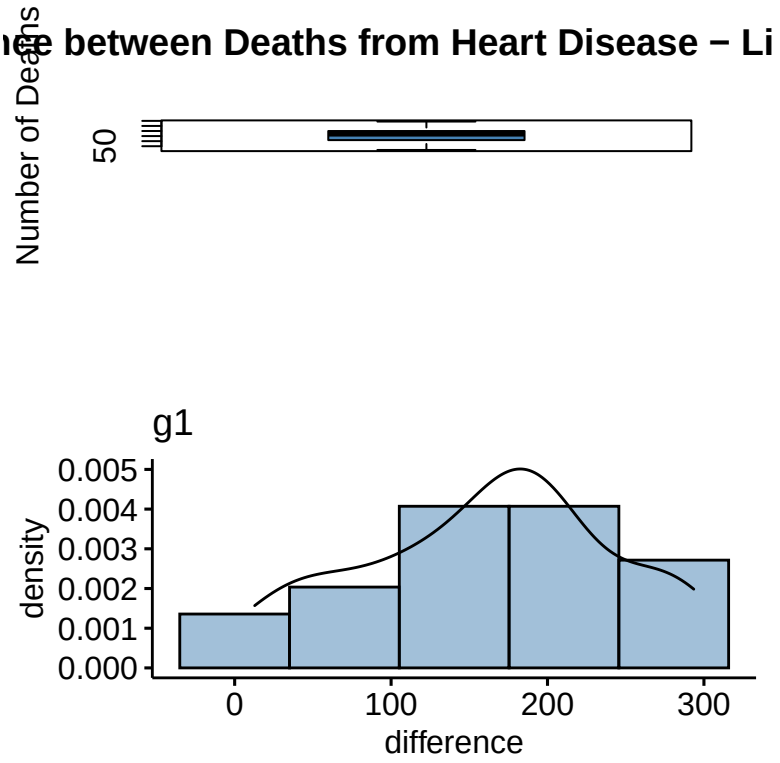
Conduct the test?
The Wilcoxon Rank Sum Test is less restrictive than the Two Sample t-Test. The metric assumption and IID are still required but not the normality of the data anymore. The legislators'age variable is made of continuous Values. Data can be classified and distance between values make sense. The variable is indeed a metric variable which satisfies the first assumption. Considering IID; All Republicans and Democrats'age are from the same population, namely policians'age.It satisfies the identically distributed consideration. Moreover, None of the politicians age provides information on the age of any other politician which satisfies the mutually independent condition and thus the IID. Considering these elements, the Wilcoxon Rank Sum test is applicable.

1.3.3 Wine and Health

Scenario: We want to test whether these countries have more deaths from heart disease or liver disease.

Proposed test: Wilxocon Signed Rank Test

Test Assumptions:
1. Metric Variables
2. i.i.d.
3. Paired data
4. Difference is symmetric



Conduct the test?
The "heart" and "liver" variables are composed of values obtained by counting which make these variable discrete variables. Additionally, the values of both observations can be compared,ranked as they share the same zero-base reference and counting granularity. Both rv are measured on a similar metric scale. The first assumption is verified.

The distribution of the difference between the paired samples "heart" and "liver" is slightly left skewed. Moreover the tight and left heights of the density curve at a similar distance to the median are not equivalent. This distribution is not perfectly symmetric around some median. However, we should not expect the differences distribution to be perfectly symmetric especially when the sample size is small. Thus, the assumption of symmetry is difficult to assess here but might be considered as verified.

The Wilcoxon Signed Rank Test can be applied in that context, with some cautions.

First box whisker done with advice from https://www.youtube.com/watch?v=Y4-wAT4SNM4&ab_channel=Dr.ToddGrande go to around 6 min

Second chart done with advice from https://www.datanovia.com/en/lessons/wilcoxon-test-in-r/

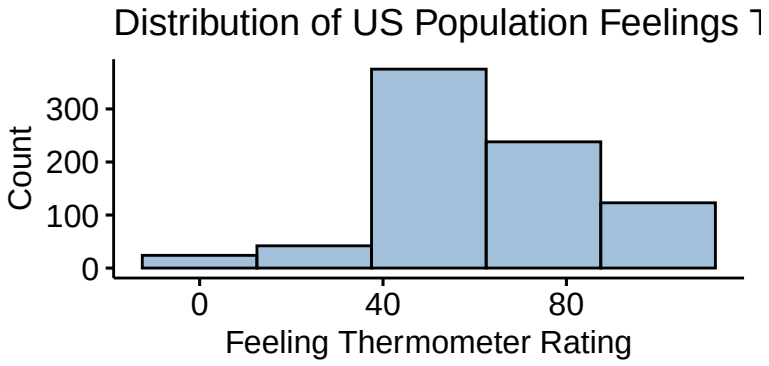
1.3.4 Attitudes Toward the Religion

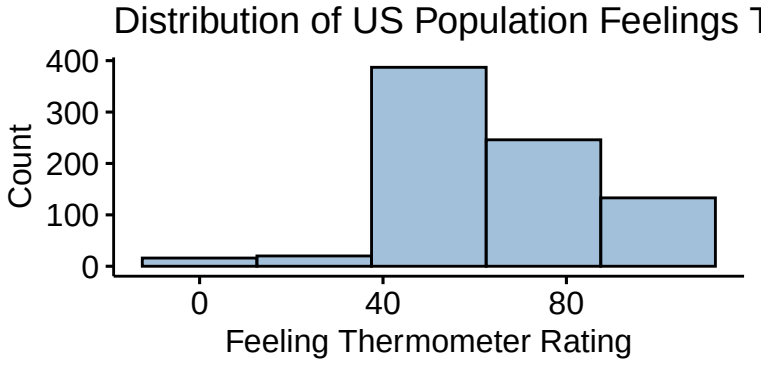
Scenario: We would like to know whether the U.S. population feels more positive towards Protestants or Catholics.

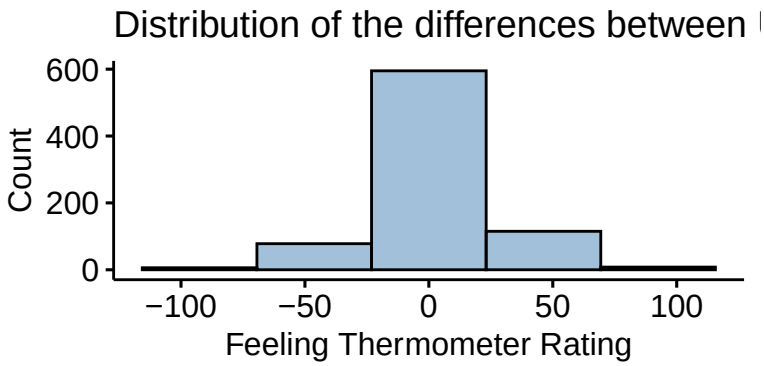
Proposed Test: Paired t-Test

Test Assumptions:
1. Metric Variables
2. i.i.d.
3. Paired
4. Normality

```
   cathtemp         prottemp
Min.   :  0.00   Min.   :  0.00
1st Qu.: 50.00   1st Qu.: 50.00
Median : 60.00   Median : 60.00
Mean   : 63.16   Mean   : 65.56
3rd Qu.: 85.00   3rd Qu.: 85.00
Max.   :100.00   Max.   :100.00
```








```
    Shapiro-Wilk normality test

data:  rel_data$cathtemp
W = 0.93377, p-value < 2.2e-16
```

p-value < 0.05 –> not normal

```
    Shapiro-Wilk normality test

data:  rel_data$prottemp
W = 0.89479, p-value < 2.2e-16
```

p-value < 0.05 –> not normal

```
    Shapiro-Wilk normality test

data:  diff_prottemp_cathtemp
W = 0.89433, p-value < 2.2e-16
```

Conduct the test?
The distribution of the differences between US population Feelings toward Catholics and Protestants is not normally distributed (considering both graph and Shapiro test value). The normality distribution of differences is one of the key criteria to apply a Paired t-test. The normality assumption is not verified, hence the paired t-test can not be applied in this context.

Note on Shapiro wilks test The Shapiro-Wilk test is a statistical test used to check if a continuous variable follows a normal distribution. The null hypothesis (H0) states that the variable is normally distributed, and the alternative hypothesis (H1) states that the variable is NOT normally distributed. So after running this test:

If p <= 0.05: then the null hypothesis can be rejected (i.e. the variable is NOT normally distributed). If p > 0.05: then the null hypothesis cannot be rejected (i.e. the variable MAY BE normally distributed).

source: https://quantifyinghealth.com/report-shapiro-wilk-test/