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.

$$[P(X_1 = 0 \text{ and } Y_1 = 0 | p = \frac{1}{2}) * P(X_2 = 0 \text{ and } Y_2 = 0 | p = \frac{1}{2}) * P(X_3 = 0 \text{ and } Y_3 = 0 | p = \frac{1}{2})] +$$
$$[P(X_1 = 1 \text{ and } Y_1 = 1 | p = \frac{1}{2}) * P(X_2 = 1 \text{ and } Y_2 = 1 | p = \frac{1}{2}) * P(X_3 = 1 \text{ and } Y_3 = 1 | p = \frac{1}{2})] =$$
$$[\frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2}] + [\frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2} * \frac{\frac{1}{2}}{2}] =$$
$$[\frac{1}{4} * \frac{1}{4} * \frac{1}{4}] + [\frac{1}{4} * \frac{1}{4} * \frac{1}{4}] =$$
$$\frac{1}{32}$$

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.

$$[P(X_1 = 0 \text{ and } Y_1 = 0 | p = \frac{3}{4}) * P(X_2 = 0 \text{ and } Y_2 = 0 | p = \frac{3}{4}) * P(X_3 = 0 \text{ and } Y_3 = 0 | p = \frac{3}{4})] +$$
$$[P(X_1 = 1 \text{ and } Y_1 = 1 | p = \frac{3}{4}) * P(X_2 = 1 \text{ and } Y_2 = 1 | p = \frac{3}{4}) * P(X_3 = 1 \text{ and } Y_3 = 1 | p = \frac{3}{4})] =$$
$$[\frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2}] + [\frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2} * \frac{\frac{3}{4}}{2}] =$$
$$[\frac{3}{8} * \frac{3}{8} * \frac{3}{8}] + [\frac{3}{8} * \frac{3}{8} * \frac{3}{8}] =$$
$$\frac{27}{512} + \frac{27}{512} =$$
$$\frac{27}{256}$$

1.2 Wrong Test, Right Data

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 people in countries with 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 the distances 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.

Assuming that data collection and transformation protocols were followed in similar fashion throughout the data collection and transformation activities, it can be assumed that the random variables are independent and identically distributed.

Now moving on to investigating normality. A quick summary below of both "Life.Ladder" and "Log.GDP.per.capita" variables reveals that "Log.GDP.per.capita" contains 13 Na(s) values. These will be omitted for the rest of this investigation. Additionally, the summary below reveals the mean GDP per capita which will be used as the point to separate low and high GDP per capita countries.

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
```

Now that the data has been split between low and high GDP per capita countries, a quick look is taken at their respective distribution summaries for the LLS.

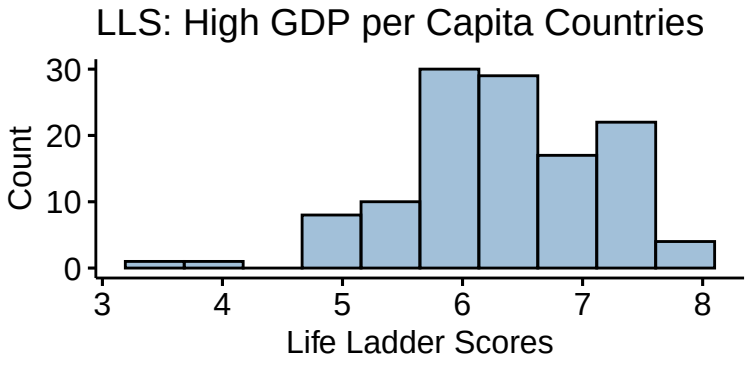
Summary: Life Ladder Score for High GDP per Capita Countries

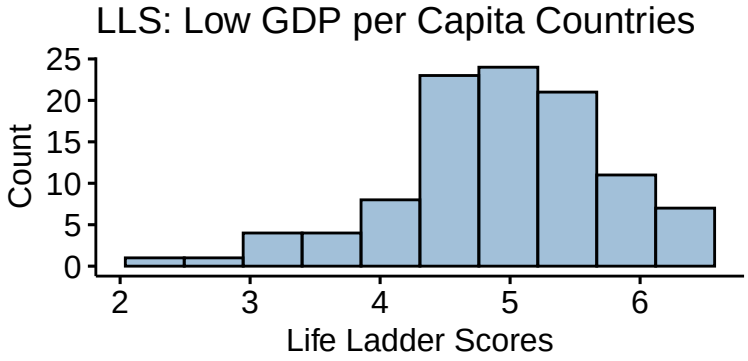
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.471   5.917   6.291   6.349   7.027   7.889
```

Summary: Life Ladder Score for Low GDP per Capita Countries

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.375   4.433   4.992   4.919   5.463   6.455
```

A quick glance gives the impression that the LLS for both low and high GDP per capita countries may be skewed a bit to the left. This can be confirmed by looking at a visualization of the distribution via histograms (done below). The x-axis displays the Life Ladder Scores and the y-axis the frequency of each Life Ladder Score values.





The distribution for high GDP per capita countries is left skewed and contains two peaks. The overall shape does not seem to represent a normal distribution. The second distribution for low GDP per capita countries contains a single peak closer to the middle of the distribution and shows a symmetric behavior, indicating perhaps that there may be some normality.

But since a visual investigation does not seem to be the most convincing of the LLS' normality, a Shapiro-Wilk normality test will be performed to quantitatively assess the distribution of the LLS for both high and low GDP per capita countries. The Shapiro Wilk normality test compares the sample distribution to a normal distribution to determine whether or not the data shows serious deviation from normality. The null hypothesis of this test is that our "sample distribution is normal". If the test is significant, the distribution is not normal.[2]

```
    Shapiro-Wilk normality test

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

The Shapiro Wilk test for the LLS distribution for low GDP per capita countries (above) has a p-value of 0.05055, indicating it barely made the cut for us to accept the null that there is no significant difference between the sample and a normal distribution - i.e. the sample is normal (barely).

[2]Source: "Statistical Tests and assumptions", https://www.datanovia.com/en/lessons/normality-test-in-r/#check-normality-in-r.

```
    Shapiro-Wilk normality test

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

The Shapiro Wilk test for the LLS distribution for high GDP per capita countries has a p-value of 0.01434, indicating that the null should be rejected and that the sample is not normal.

Given that the results from the Shapiro Wilk test barely indicated normality for the life ladder scores for low GDP per capita countries and did NOT indicate normality for the life ladder scores for high GDP per capita countries, the third assumption of normality is not validated.

Conduct the test?
Considering that the normality assumption is not met, the Two Sample t-Test would not be best suited for this context.

1.3.2 Legislators

Scenario: 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 or ordinal variables
2. i.i.d.
3. Same shape and spread of the two samples

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 age variable (of the legislators) is made of continuous values in which the distances between the values make sense. The variable is metric and satisfies the first assumption.

The ages of the legislators come from the same distribution (politicians' ages). Moreover, a politician's age does not provide information about the age of any other politician, which satisfies the independence condition and thus the IID, satisfying the second assumption.

A summary of the two sample distributions alongside visual aids (histograms and box and whisker plot) will help determine whether tthe third assumption (same shape and spread) is satisfied.

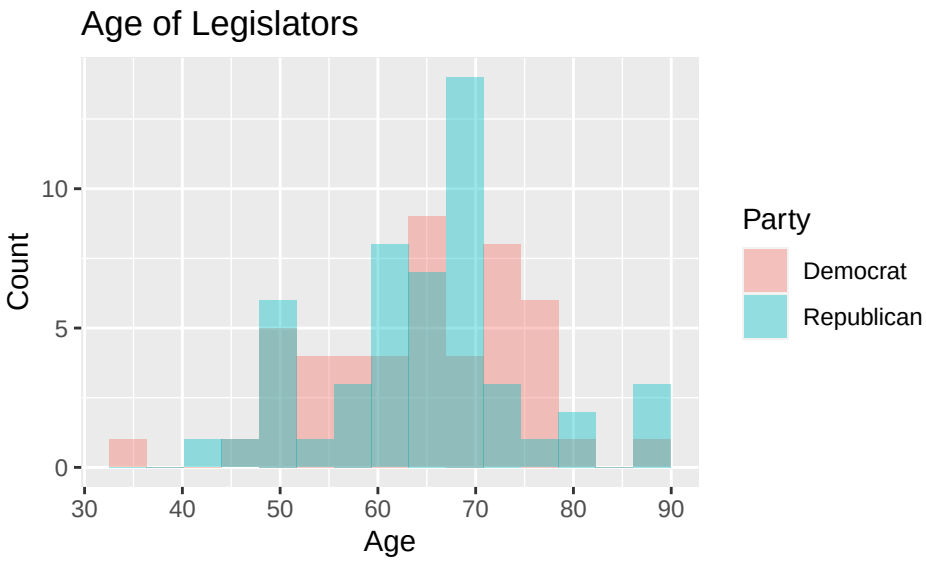
Summary: Democrat Senators

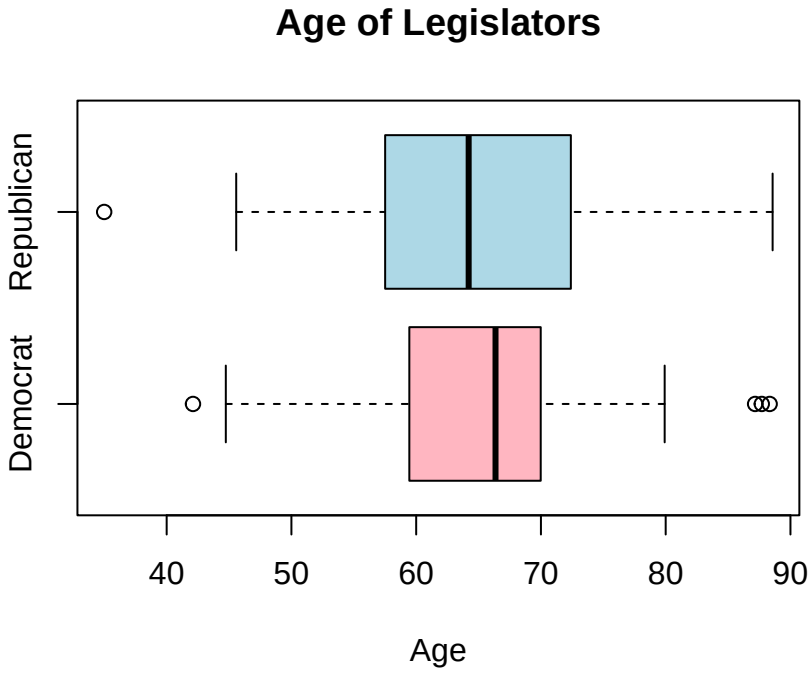
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.99   57.68   64.21   64.19   72.30   88.57
```

Summary: Republican Senators

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  42.11   59.64   66.36   64.83   69.90   88.33
```

Looking at the summaries of the two samples, the distributions do not seem to be too different, as evidenced by the close means, 1st quartile, and maximum.





Looking closer at the sample distributions in the above figures, it still doesn't seem clear whether the spread and shape of the distributions are the same. To double check, the Ansari Bradley test can be done to test this. The null hypothesis in this test is that the two population distribution functions corresponding to the two samples are identical against the alternative hypothesis that they differ.[3]

```
Ansari-Bradley test
```

[3] "Ansari-Bradley Test", https://www.quality-control-plan.com/StatGuide/ansari.htm

```
data:  dem_sen$age and rep_sen$age
AB = 1113, p-value = 0.2162
alternative hypothesis: true ratio of scales is not equal to 1
```

The results from the Ansari-Bradley test shows a p-value of 0.2162, indicating that the null should not be rejected and that it is likely that the spread and shape of the distribution functions corresponding to the two samples are the same. The third assumption seems to be satisfied.

Conduct the test?
Given that all the assumption seems to be satisfied, the Wilcoxon Rank Sum test would be applicable in this case.

1.3.3 Wine and Health

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

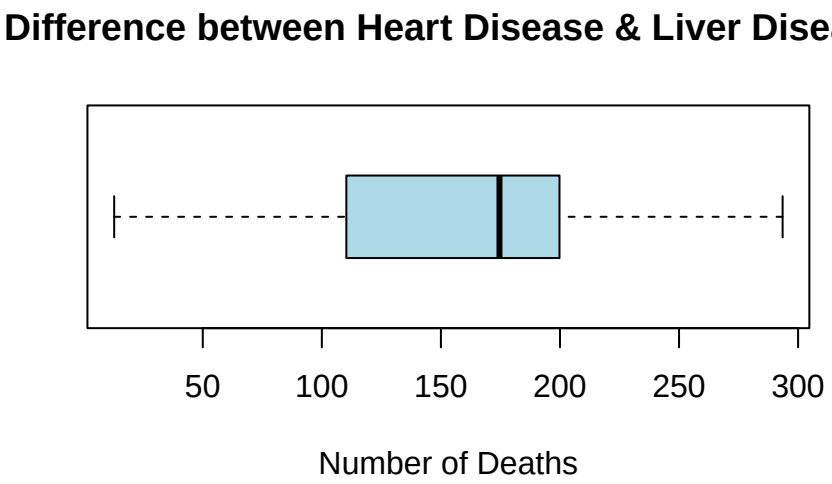
Proposed test: Wilxocon Signed Rank Test

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

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

The heart disease and liver disease variables come from their respective distributions and with the assumption that the data was collected and transformed following the same protocol, the second assumption of independent and identically distributions can be satisfied.

Another assumption of the Wilcoxon signed rank test is that it is conducted on paired data. In this case since the data is coming from the same test subject (the selected group of countries), it is confirmed that this assumption is satisfied.



The distribution of the difference between the paired samples "heart" and "liver" is a bit left skewed. Moreover this distribution does not seem to be symmetric around the median. However, one should not expect the differences in distribution to be perfectly symmetric especially when the sample size is small (n=21). Thus, the assumption of symmetry is difficult to assess here but could be considered as verified.

Conduct the test?
The Wilcoxon Signed Rank Test can be applied in that context, with some caution.

1.3.4 Attitudes Toward the Religion

Scenario: Investigate whether the U.S. population feels more positive towards Protestants or Catholics.

Proposed Test: Paired t-Test

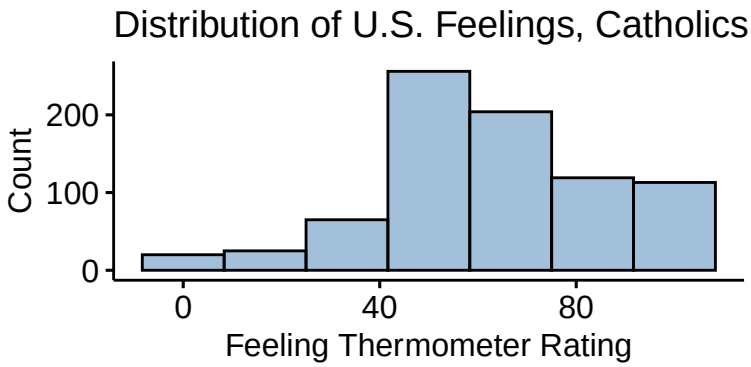
Test Assumptions:
1. Metric Variables
2. i.i.d.
3. Paired data
4. Normality

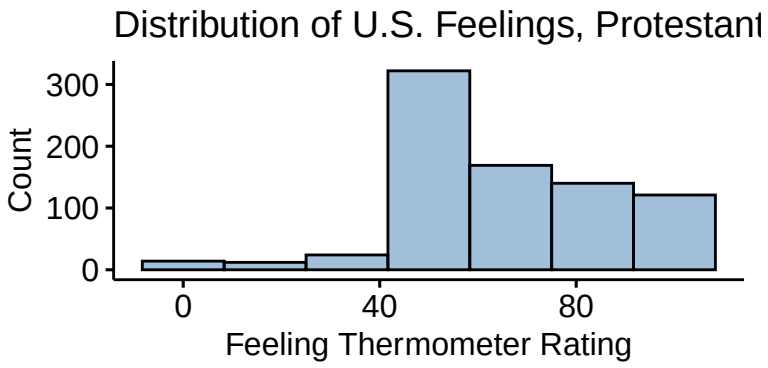
The feeling thermometer variables towards both Protestants and Catholics is continuous and metric in nature since the respondents can choose their rating on a continuous scale from 0 to 100. Thus the first condition is satisfied

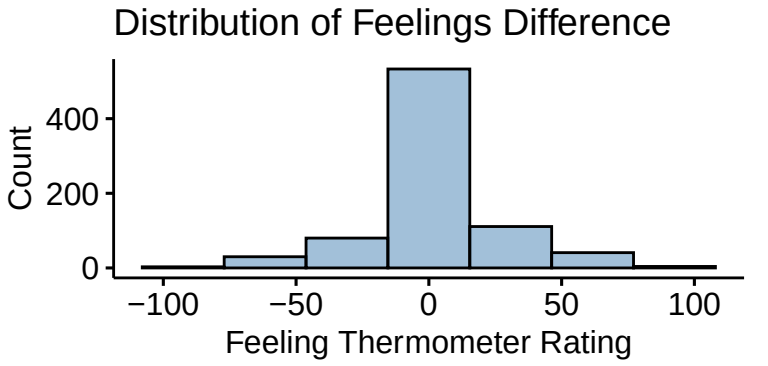
The two feeling thermometer variables seem to also be independent and identically distributed assuming that the conductors of the 2004 General Social Survey were following the same protocol in collecting and transforming these data - satsifying the second condition.

The paired t-test requires the data be "paired." This is true in this case because the test-subjects, i.e. the respondents, for both the feeling thermometer towards Protestants and Catholoics are the same - satisfying the third condition.

```
   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
```







Looking at the above to histograms it does not look like the distributions are normal. The differences histogram looks symmetrical but not necessarily normal. The Shapiro-Wilk test conducted below will confirm these observations.

```
    Shapiro-Wilk normality test

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

The results of the Shapiro-Wilk test on the feelings thermometer towards Catholics, has a p-value that is close to 0, indicating this variable does not have a normal distribution.

```
    Shapiro-Wilk normality test

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

The results of the Shapiro-Wilk test on the feelings thermometer towards Protestants, has a p-value that is close to 0, indicating this variable does not have a normal distribution.

```
    Shapiro-Wilk normality test

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

For the difference, the results also show here a p-value less than 0.05, indicating that the difference distribution is not normal.

Based on the results of the visual investigation and Shapiro-Wilk tests, the third condition of normality is not met.

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 values). The normality distribution of differences is one of the key criteria to apply a Paired t-test. Even though the other criteria regarding metric variables, independent and identical distributions, and paired data are met, the normality assumption is not verified. Hence the paired t-test can not be applied in this context.