Part 1: Foundational Exercises

Lab 1: Hypothesis Testing - W203 Section 8

Team Anglerfish: Aruna Bisht, Jing Lange, Chase Madson, Maria Manna

Contents

1.1 Professional Magic **1**

1.1.1 Type I Error Rate 1

1.1.2 Power 2

1.2 Wrong Test, Right Data **3**

1.2.1 Violating t-test Assumptions 3

1.2.2 Remedial Measure 3

1.3 Test Assumptions **4**

1.3.1 World Happiness 4

1.3.2 Legislators 7

1.3.3 Wine and Health 9

1.3.4 Attitudes Toward the Religious 11

1.1 Professional Magic

1.1.1 Type I Error Rate

Question: *What is the type I error rate of the test?*

a. Let's denote the test statistic t with the given formula $t = X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$.

b. The type I error rate α is the probability that we incorrectly reject the null hypothesis $H_0 : p = \frac{1}{2}$. In our test, we reject the null hypothesis when the test statistic is either zero or six.

$$\begin{aligned}\alpha &= Pr(H_0 \text{ is rejected when it is true}) \\ &= Pr(t \in \{0, 6\} \,|\, p = \frac{1}{2})\end{aligned}$$

c. We can rearrange $t = 0$ as 3 consecutive flips of $(X_i, Y_i) = (0, 0)$, and similarly $t = 1$ as $(X_i, Y_i) = (1, 1)$. We note that there is $\binom{6}{0} = 1$ way of observing $t = 0$ and $\binom{6}{6} = 1$ way of observing $t = 1$.

$$\begin{aligned}&= \binom{6}{0} Pr([(X_1, Y_1) = (0, 0)] \cap [(X_2, Y_2) = (0, 0)] \cap [(X_3, Y_3) = (0, 0)] \,|\, p = \frac{1}{2}) \\ &+ \binom{6}{6} Pr([(X_1, Y_1) = (1, 1)] \cap [(X_2, Y_2) = (1, 1)] \cap [(X_3, Y_3) = (1, 1)] \,|\, p = \frac{1}{2})\end{aligned}$$

d. Given that each flip of the pair is independent of all other flips of the pair, we can restate as the product of probabilities.

$$\begin{aligned}&= Pr([(X_1, Y_1) = (0, 0)] \,|\, p = \frac{1}{2}) \times Pr([(X_2, Y_2) = (0, 0)] \,|\, p = \frac{1}{2}) \times Pr([(X_3, Y_3) = (0, 0) \,|\, p = \frac{1}{2}) \\ &+ Pr([(X_1, Y_1) = (1, 1)] \,|\, p = \frac{1}{2}) \times Pr([(X_2, Y_2) = (1, 1)] \,|\, p = \frac{1}{2}) \times Pr([(X_3, Y_3) = (1, 1) \,|\, p = \frac{1}{2})\end{aligned}$$

e. We express this in terms of the joint distribution functions we are given.

$$= (f_{X,Y|p=\frac{1}{2}}(0, 0))^3 + (f_{X,Y|p=\frac{1}{2}}(1, 1))^3$$

f. When $p = \frac{1}{2}$, then $f_{X_i,Y_i}(0, 0) = f_{X_i,Y_i}(1, 1) = \frac{\frac{1}{2}}{2} = \frac{1}{4}$, and thus we plug that probability into the above.

$$\begin{aligned}&= (\frac{1}{4})^3 + (\frac{1}{4})^3 \\ &= \mathbf{\frac{1}{32} = 0.03125}\end{aligned}$$

1.1.2 Power

Question: *What is the power of your test for the alternative hypothesis that* $p = \frac{3}{4}$

a. The power of this test is $1 - \beta$, where β is the type II error rate (i.e., the probability that we fail to reject H_0 when it is false). Specifically, we want to calculate the power under the alternative hypothesis $H_a : p = \frac{3}{4}$.

$$\begin{aligned} Power &= 1 - \beta \\ &= 1 - Pr(\text{Fail to reject } H_0 \text{ when } H_a \text{ is true instead}) \\ &= 1 - Pr(t \notin \{0, 6\} \mid H_a : p = \frac{3}{4}) \\ &= 1 - [1 - Pr(t \in \{0, 6\} \mid H_a : p = \frac{3}{4})] \\ &= Pr(t \in \{0, 6\} \mid H_a : p = \frac{3}{4}) \end{aligned}$$

b. We repeat our steps c, d, and e from our solution to 1.1.1 above to reach the joint probability density function when $p = \frac{3}{4}$.

$$= (f_{X,Y|p=\frac{3}{4}}(0,0))^3 + (f_{X,Y|p=\frac{3}{4}}(1,1))^3$$

c. When $p = \frac{3}{4}$, then $f_{X_i,Y_i}(0,0) = f_{X_i,Y_i}(1,1) = \frac{\frac{3}{4}}{2} = \frac{3}{8}$, and thus we plug that probability into the above.

$$\begin{aligned} &= (\frac{3}{8})^3 + (\frac{3}{8})^3 \\ &= \mathbf{\frac{27}{256} = 0.10546875} \end{aligned}$$

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.

1.2.1 Violating t-test Assumptions

> **Question**: *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?*
>
> **Assumptions**: In order to run a paired t-test on some sample $\{X_i\}_{i=1}^n$ and receive reliable results, there are 3 assumptions that must be met.

1. Each X_i comes from an underlying random variable that is **metric** in scale.
2. $\{X_i\}_{i=1}^n$ is an **independent and identically distributed** random sample.
3. $\{X_i\}_{i=1}^n$ is **normally distributed**, or $\{X_i\}_{i=1}^n \sim N(\mu_x, \sigma_x^2)$

For the results of a t-test on some sample to be reliable, then each draw from the sample must come from an underlying random variable that is **metric** in scale. As the preference statements in this survey are each measured on a 5-point Likert-scale, we know that this assumption is violated.

Running a paired t-test on Likert-scale data means calculating the test statistic $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$, which gives us the difference from the mean in units of standard deviation. These units lose their meaning when calculated using Likert-scale data; it doesn't make sense to subtract or take a mean of these values. This is because data from Likert-scale have **non-standard intervals**, are **categorical instead of continuous**, and have **incomparable magnitudes** as values are subjective to each customer in the sample (a "very good" from one person can mean something different than another person's "very good"). Therefore, our interpretation of the test statistic would be invalid and so would be any conclusions we draw from it.

1.2.2 Remedial Measure

> **Question**: *What would you propose to do to remedy this problem?*

Since Likert-scale data contains paired values that are non-metric, it is better to use a **hypothesis of comparisons under the Wilcoxon Rank-Sum Test**. If we want to compare Likert-scale responses about customers' sentiment about our regular website (X) and our mobile website (Y), we would want to test the null hypothesis that $H_0 : P(X < Y) = P(X > Y)$, meaning equal likelihood. If the evidence suggests that we can reject this null hypothesis, then we can provide a useful conclusion to our organization about how our customers view these websites differently.

The assumptions required under this test are that the data must follow an ordinal scale (which Likert-scale does) and that each pair is drawn from the same distribution (which the organization can accomplish through random sampling).

1.3 Test Assumptions

Evaluate the assumptions for each of the following tests.

1.3.1 World Happiness

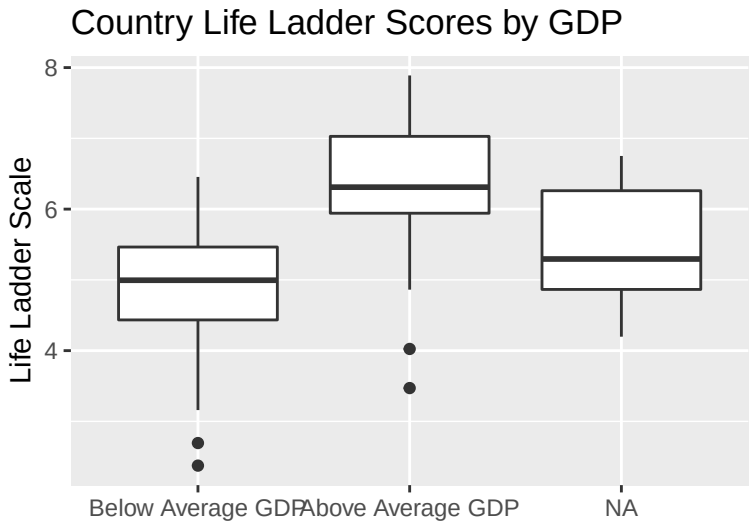
> **Instructions**: *List all assumptions for a **two-sample t-test**. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.*

Let $\{X_i\}_{i\in[1,105]}$ represent the 105 *life_ladder* observations for **below average** GDP countries in the sample, giving us $\bar{X} = 4.924581$.

Let $\{Y_i\}_{i\in[1,121]}$ represent the 121 *life_ladder* observations for **above average** GDP countries in the sample, giving us $\bar{Y} = 6.355488$.

We want to consider running a Welch's two-sample t-test that compares the mean of sample $\bar{X}$ against $\bar{Y}$.

above_average_gpd	mean_life_ladder	observations
Below Average GDP	4.924581	105
Above Average GDP	6.355488	121
NA	5.459000	13



Assumptions for the Two-Sample t-Test: In order to get reliable results from this test, there are 3 assumptions that must be met.

1. Each X_i and Y_i comes from underlying random variables that are **metric** in scale.

   - From the summary table below, we see that life_ladder is precise to the third decimal point, and thus does not consist of whole-numbered values. This is characteristic for continuous variables.
     - Note: The life_ladder variable we see in the data appears to be averages of Cantril ladder responses aggregated to the country-level.

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

- We see below that we observe duplicate values in less than 5% of the life_ladder observations. This is characteristic for continuous variables.

pct_distinct_life_ladder
97.9%

- However, we note that individual Cantril ladder responses are based on subjective 11-point scale. Though the respondents are asked to imagine fixed-interval spacing between each step of the ladder, it is not clear that this satisfactorily approximates the standard interval requirement of metric scales. Even though we are working with country-level averages, we conclude that **the life_ladder variable does not meet the metric scale requirement** since the distance between each scalar value cannot be linearly measured.

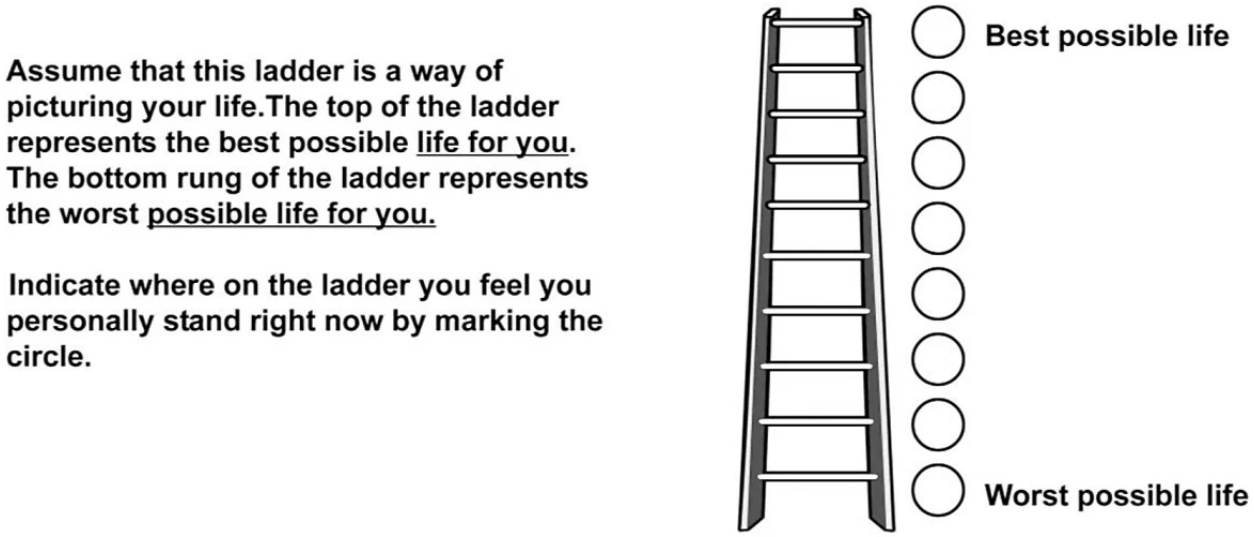


Figure 1: Cantril Scale Visualized. Source: Sawatzky et al. (e.g., 2010), http://www.hqlo.com/content/8/1/17

2. $\{X_i\}_{i\in[1,105]}$ and $\{Y_i\}_{i\in[1,121]}$ are **independent and identically distributed** random samples.

- One reason to doubt the independence of these samples is that **many countries appear twice in this data** for their annual values observed in 2019 and 2020. It is safe to assume that a country's observed value for one year is correlated with its values from previous years (i.e., **autocorrelation**). Therefore, we are not working with a true random sample.

country_name	2019	2020
Albania	4.995	5.365
Argentina	6.086	5.901
Australia	7.234	7.137
Austria	7.195	7.213
Bahrain	7.098	6.173

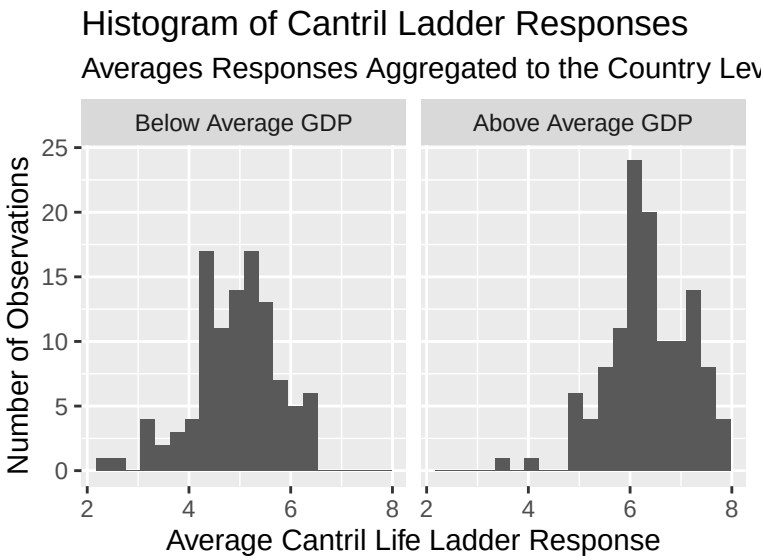
- Another reason to doubt the independence of these samples is that **some groups of countries could be correlated based on geographic cluster**. For example, we can select a few European and South American countries and then compare their average life_ladder to see what may be a significant difference in averages.

continent	mean(life_ladder)
Europe	7.0986
South America	6.0856

- As the data is ordinal and not metric, the CLT does not apply and there is not an obvious way to compare distributions from the variables to confirm whether they are identically distributed.

3. $\{X_i\}_{i\in[1,105]}$ and $\{Y_i\}_{i\in[1,121]}$ are **normally distributed**, or the sample size is large enough to approximate normality due to the CLT (assuming no severe skewness).

- We more than satisfy the $n > 30$ rule-of-thumb for minimum sample size.
- The raw distributions of life_ladder within each group show no noticeable deviations from normality.



- However, the data is ordinal so sample size does not matter; the CLT cannot be applied.

Additionally, data used for two-sample t-tests should:

- Have a grouping variable that is defined and present. In this case, the grouping variable is GDP (low GDP and high GDP) so this condition is met.
- Have distributions of X and Y that are both normal and have equal variance for each group. With ordinal data, normality cannot exist; the data is nonparametric. Additionally, even if the CLT was able to be applied to the data, our main concern would be that there is strong skewness with a small sample. We know that the CLT guarantees normality for large samples. However, the sample size in this study is less than 30 countries. Therefore, even if the data were metric, the CLT is not necessarily applicable and normality cannot be assumed.

1.3.2 Legislators

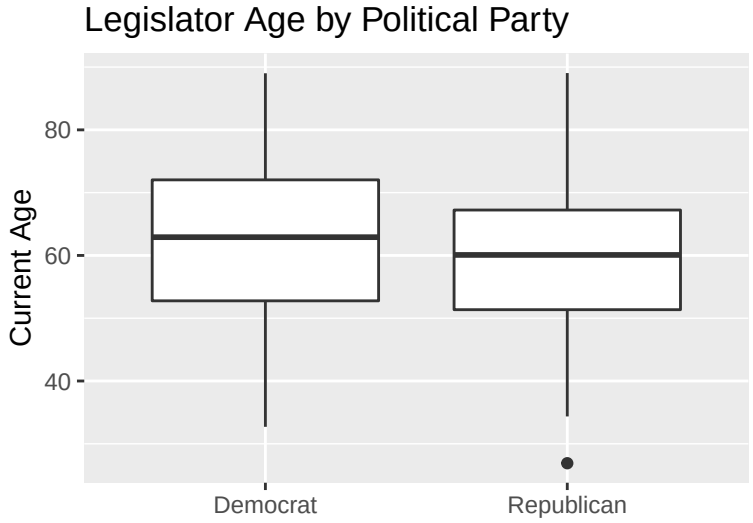
> **Instructions**: *List all assumptions for a **Wilcoxon rank-sum test (using the Hypothesis of Comparisons)**. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.*

Let $\{X_i\}_{i\in[1,272]}$ represent the 272 Democratic legislators in the data, giving us $\bar{X} = 61.76656$.

Let $\{Y_i\}_{i\in[1,264]}$ represent the 264 Republican legislators in the data, giving us $\bar{Y} = 59.18111$.

We want to consider running a Wilcoxon rank-sum test using the hypothesis of comparisons to $H_0 : P(X > Y) = P(X < Y)$.

party	mean_age	observations
Democrat	61.77751	272
Republican	59.19206	264



Assumptions for the Wilcoxon Rank-Sum Test (Hypothesis of Comparisons): In order to get reliable results from this test, there are 2 assumptions that must be met.

1. Each X_i and Y_i comes from underlying random variables that are **ordinal** in scale.

    - From the summary table below, we see that current_age_in_years is precise to the day and is characteristic of a continuous variable. Being continuous, we are able to use more restrictive alternative t-tests if we want. But a variable being continuous does not violate this assumption since continous variables have the requisite property of being ordered. Hence this requirement is met.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.90   51.80   61.23   60.50   69.19   89.05
```

2. Each X_i is drawn from the same distribution, each Y_i is drawn from the same distribution, and all X_i and Y_i are mutually independent.

- Each X_i is drawn from the Democratic party and each Y_i is drawn from the Republican party. While the ages of the representatives may be independent of each other, we do not believe that the political parties of the representatives are independent of each other. That is to say, if there are two senators from the same state, there is likely some covariance between their political parties. This then likely invalidates the assumption that the pairs too are independent. The table below shows that only a quarter of states have between 25% and 75% of representatives from the same party; most states have a strong preference for one party over the other. **Thus, this assumption may not be adequately met**.

sum(pct_dem > 0.25 & pct_dem < 0.75)/n()
0.25

- The sample data is a snapshot of the current Democrat and Republican legislators, and how each Democrat and Republican senator is elected may change through time.

Additionally, data used for Wilcoxon Rank-Sum Tests (Hypothesis of Comparisons) should:

- Have a grouping variable that is defined and present. In this case, the grouping variable is the political party (Democrat and Republican).
- Not have substantial differences in group sample sizes. While the number of Congressmen/women/representatives are not exactly equal based on the two main political parties, they are extremely close and therefore there are no substantial differences in group sample sizes.

gender	legislators	mean_age
F	146	60.67477
M	390	60.44018

- Not have too many ties. As shown below we see less than 5% of legislators share a birthday, and so this assumption is met.

pct_distinct_age
98.5%

1.3.3 Wine and Health

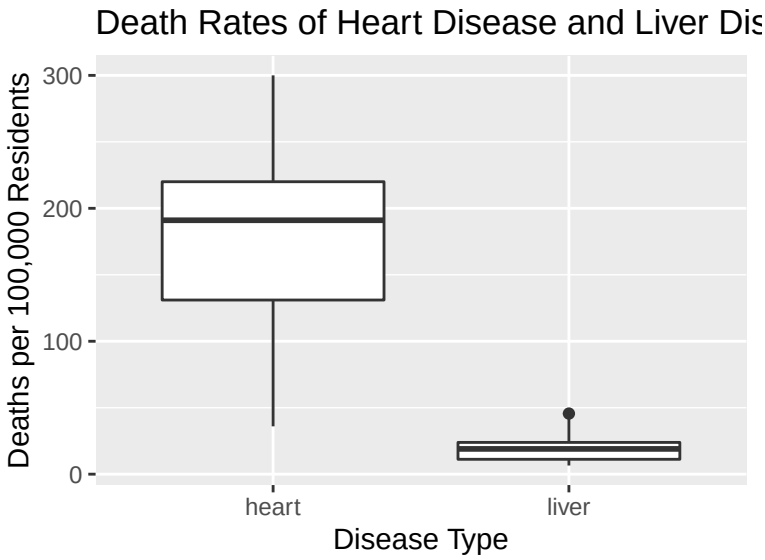
> **Instructions**: *List all assumptions for a **signed-rank test**. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.*

Let $\{X_i\}_{i\in[1,21]}$ represent the rate of heart disease deaths per 100,000 citizens for 21 countries, giving us $\bar{X} = 183.28571$.

Let $\{Y_i\}_{i\in[1,21]}$ represent the rate of liver disease deaths per 100,000 citizens for 21 countries, giving us $\bar{Y} = 21.03333$.

We want to consider running a Wilcoxon signed-rank test using the hypothesis of no difference $H_0 : \mu = 0$.

disease_type	mean_deaths	observations
heart	183.28571	21
liver	21.03333	21



Assumptions for the Wilcoxon Signed-Rank Test: In order to get reliable results from this test, there are 3 assumptions that must be met.

1. Each X_i and Y_i comes from underlying random variables that are **metric** in scale, especially since the operation of subtraction imposes a metric structure.

    - The rate of deaths by disease is a ratio of two count numbers, resulting in a discrete metric scale variable. Their differences, means, and variances can be meaningfully computed and interpreted. The difference between 5 and 6 deaths is the same as between 10 and 11, for example. Also, heart and liver diseases diagnosis have well established standards which countries listed in the dataset share. Thus we consider this assumption to be met.

2. Each pair (X_i, Y_i) is drawn from the same distribution, independently of all other pairs.

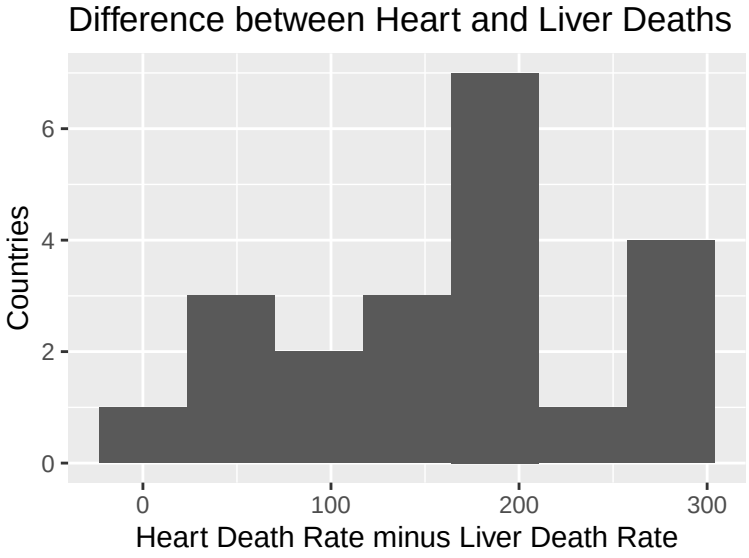
    - We might doubt the independence of these samples considering that **some groups of countries could be correlated based on geographic cluster**. Certainly, the cultural norms surrounding alcohol consumption are more prevalent some countries more than others.

- For example, when we group countries by general region and compare their average mean_deaths_per_100k, we see enough difference in averages to doubt the independence of the sample. Importantly, no counties from South America or Africa were sampled, and very few were sampled from Asia. Thus, this assumption is not met.

region	countries	heart	liver
Atlantic	3	265.3333	9.333333
North EU	3	243.6667	14.133333
Oceania	2	238.5000	11.500000
Americas	2	195.0000	19.250000
Other Region	8	148.8750	24.162500
Central EU	3	88.0000	38.833335

3. The difference between pairs $X - Y$ follows a symmetric distribution with mean μ.

- Graphing distribution of $X - Y$ on a histogram, we see a concerningly heavy right-tail that could disturb our assumption of normality. This difference is not symmetric about some mean μ.



- Because the sample size of our data is just 21 countries, we have not met the minimum $n \geq 30$ to rely on μ approximating normal due to the CLT. Therefore, this assumption does not hold true.
- Additionally, data used for Wilcoxon Signed-Rank Tests should not have too many ties. As shown below we see less than 10% of countries have the exact rates, and so this assumption is met.

disease_type	pct_distinct_rates
heart	90.5%
liver	95.2%

1.3.4 Attitudes Toward the Religious

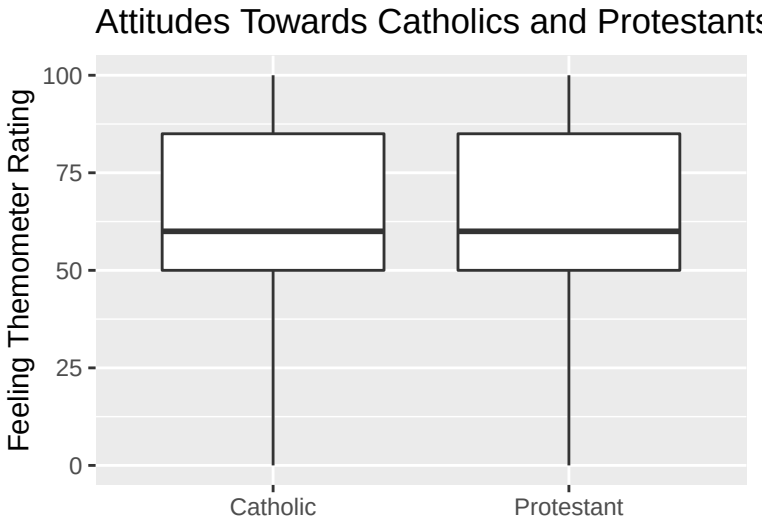
> **Instructions**: *List all assumptions for a **paired t-test**. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.*

Let $\{X_i\}_{i\in[1,21]}$ represent the rate of heart disease deaths per 100,000 citizens for 21 countries, giving us $\bar{X} = 183.28571$.

Let $\{Y_i\}_{i\in[1,21]}$ represent the rate of liver disease deaths per 100,000 citizens for 21 countries, giving us $\bar{Y} = 21.03333$.

We want to consider running a Wilcoxon signed-rank test using the hypothesis of no difference $H_0 : \mu = 0$.

religion	mean_temp	observations
Catholic	63.15835	802
Protestant	65.56110	802



Assumptions for a Paired t-Test: In order to get reliable results from this test, there are 3 assumptions that must be met.

1. Each X_i and Y_i comes from underlying random variables that are **metric** in scale.

    - We see below that we observe duplicate values in less than 2% of the responses, and the top 5 most popular responses accounted for 82% of the data. Effectively, this approximates something like a 5- or 10-point Likert scale, which is ordinal but not continuous. While the researchers may have wanted to appromixate continuous data by provinding such precision, it appears that respondents are prone to rounding to the nearest tenth, fifth, or half of 100, with few respondents making use of the full precision.

pct_distinct_temperature
1.7%

temperature	proportion_of_responses
50	35.5%
100	13.9%
85	12.2%
70	12.1%
60	7.9%

- Again, we note that individual responses are based on **subjective scale** with **non-standard intervals**. Though the respondants are asked to rate groups using the feeling thermometer with presumed fixed-interval spacing between each point, it is not clear that this satisfactorily approximates the standard interval requirement of metric scales. It would not make sense to calculate the mean and variance, which would make this test unviable. We conclude that **the temperature variable does not meet the metric scale requirement** since the distance between each scalar value cannot be linearly measured.

2. Each pair (X_i, Y_i) is **drawn from the same distribution, independently of all other pairs (i.i.d.)**.

   - According to the study owners (NORC)[https://gss.norc.org/About-The-GSS], "the General Social Survey (GSS) is a nationally representative survey of adults in the United States..." and provides no specific detail on how their samples are drawn beyond that. Taking them at face value, their study being nationally representative suggests their data was either conducted without sampling bias or that existing sampling bias is mitigated before analysis. **We don't have enough information to evaluate whether this data was i.i.d.**
   - Some reason to **doubt** the independence of the sampling is that the proximity to a place of religious worship and the general region within the United States of the individuals responding to the survey would likely impact their responses.

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

   - As this data is not metric, **CLT cannot be applied** and **this does not hold**.