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 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 $H_a : p = \frac{3}{4}$.

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

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.

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

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?*

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, meaning they are continuous and numeric. 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.

Optional Supplement to our Response

> **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, meaning they are continuous and numeric.
2. $\{X_i\}_{i=1}^n$ is an **independent and identically distributed** random sample.
3. $\{X_i\}_{i=1}^n$ is **normally distributed**, or $\{X_i\}_{i=1}^n \sim N(\mu_x, \sigma_x^2)$

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).

Do we want to propose converting the data from Likert-scale to binary?

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.*

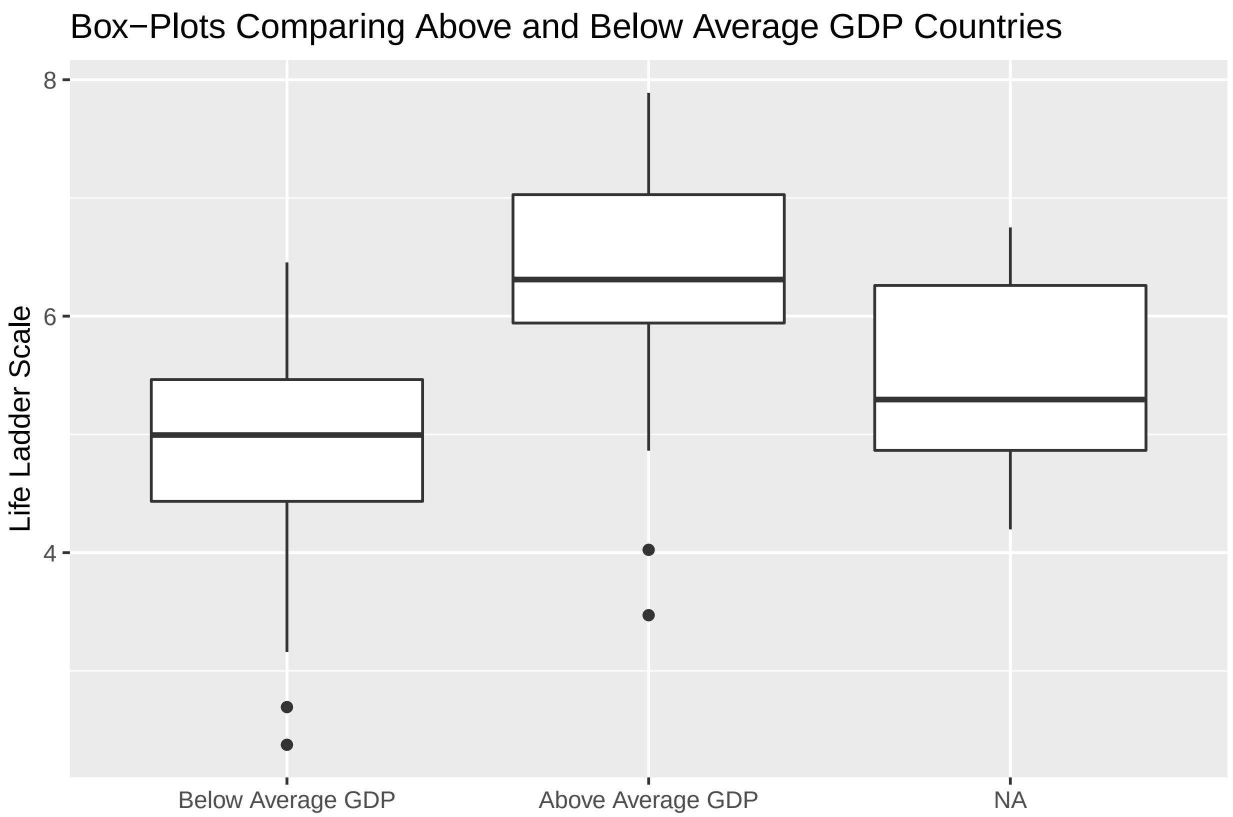
Our data comes from the World Happiness Report taken between 2019-2020 for a sample of countries, and specifically we are interested in the Cantril Life Ladder and GDP fields.

```
happiness_WHR <-
  'happiness_WHR.csv' %>%
  read_csv(col_types = 'ciddddddddd') %>%
  clean_names() %>%
  mutate(above_average_gpd =
           (log_gdp_per_capita >= mean(log_gdp_per_capita, na.rm = TRUE)) %>%
           ifelse('Above Average GDP', 'Below Average GDP') %>%
           factor(levels = c('Below Average GDP', 'Above Average GDP'))) %>%
    select(country_name, year, life_ladder, above_average_gpd)

happiness_WHR %>%
  group_by(above_average_gpd) %>%
  summarise(mean_life_ladder = mean(life_ladder),
            observations = n())
```

```
## # A tibble: 3 x 3
##   above_average_gpd mean_life_ladder observations
##   <fct>                        <dbl>        <int>
## 1 Below Average GDP             4.92          105
## 2 Above Average GDP             6.36          121
## 3 <NA>                          5.46           13
```

```
happiness_WHR %>%
  ggplot(mapping = aes(x = above_average_gpd, y = life_ladder)) +
  geom_boxplot() +
  labs(title = 'Box-Plots Comparing Above and Below Average GDP Countries',
       x = '',
       y = 'Life Ladder Scale')
```



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 $\bar{Y}$.

Assumptions: 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, meaning they are continuous and numeric.

   - 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.

```
happiness_WHR %>%
  pull(life_ladder) %>%
  summary()
```

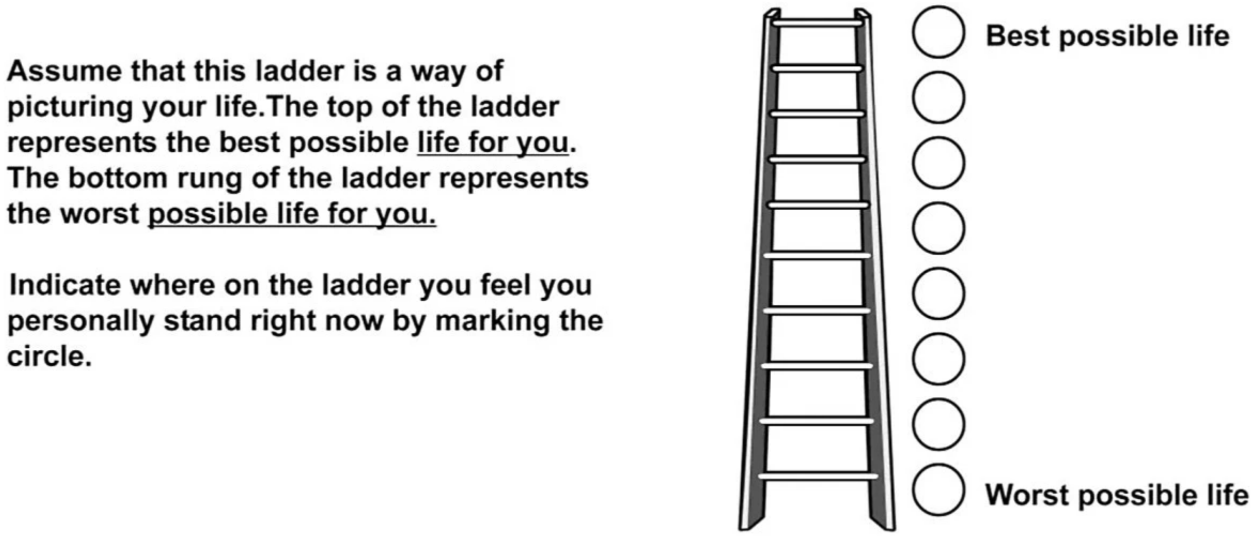
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.375   4.971   5.768   5.678   6.428   7.889
```

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

```
happiness_WHR %>%
  summarise(pct_distinct_life_ladder = (n_distinct(life_ladder) / n()) %>%
                                        scales::percent(0.1))
```

```
## # A tibble: 1 x 1
##   pct_distinct_life_ladder
##   <chr>
## 1 97.9%
```

- However, we note that individual Cantril ladder responses are based on subjective 11-point scale. Though the respondants 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.



Assume that this ladder is a way of picturing your life.The top of the ladder represents the best possible life for you. The bottom rung of the ladder represents the worst possible life for you.

Indicate where on the ladder you feel you personally stand right now by marking the circle.

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.

```
happiness_WHR %>%
  select(-above_average_gpd) %>%
  pivot_wider(names_from = year, values_from = life_ladder) %>%
  filter(!is.na(`2020`)) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 3
##   country_name ‘2019‘ ‘2020‘
##   <chr>          <dbl>  <dbl>
## 1 Albania         5.00   5.36
## 2 Argentina       6.09   5.90
## 3 Australia       7.23   7.14
## 4 Austria         7.20   7.21
## 5 Bahrain         7.10   6.17
```

- 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.

```
happiness_WHR %>%
  mutate(continent =
           case_when(country_name %in% c('France', 'United Kingdom',
                                         'Germany', 'Belgium',
                                         'Finland') ~ 'Europe',
                     country_name %in% c('Brazil', 'Bolivia',
                                         'Chile', 'Colombia',
                                         'Uruguay') ~ 'South America')) %>%
  filter(!is.na(continent)) %>%
  group_by(continent) %>%
  summarise(mean(life_ladder))
```

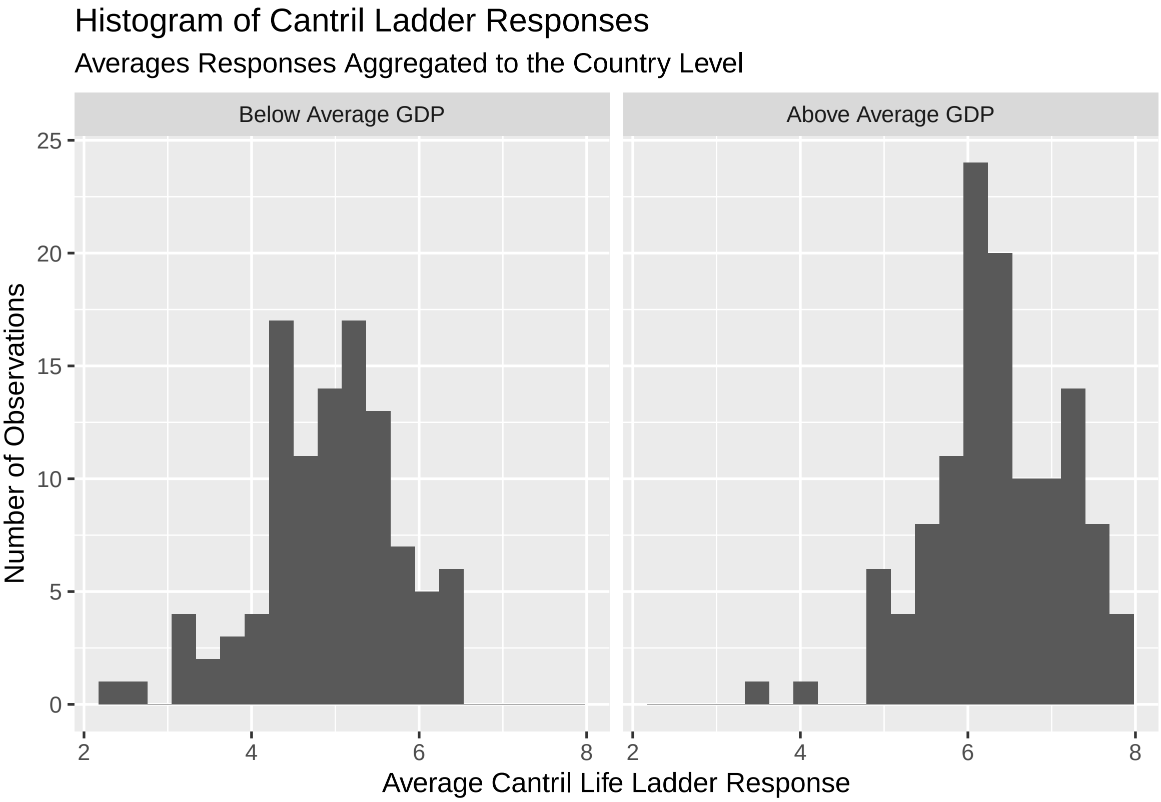
```
## # A tibble: 2 x 2
##   continent     `mean(life_ladder)`
##   <chr>                       <dbl>
## 1 Europe                       7.10
## 2 South America                6.09
```

- 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.

```
happiness_WHR %>%
  filter(!is.na(above_average_gpd)) %>%
  ggplot(mapping = aes(x = life_ladder)) +
  geom_histogram(bins = 20) +
  labs(title = 'Histogram of Cantril Ladder Responses',
       subtitle = 'Averages Responses Aggregated to the Country Level',
       x = 'Average Cantril Life Ladder Response',
       y = 'Number of Observations') +
  facet_wrap(~above_average_gpd)
```



- 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

```
2 + 2
```

```
## [1] 4
```

1.3.3 Wine and Health

```r
2 + 2
```

```
## [1] 4
```

1.3.4 Attitudes Toward the Religious

```r
2 + 2
```

```
## [1] 4
```

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, meaning they are continuous and numeric.
2. $\{X_i\}_{i=1}^n$ is an **independent and identically distributed** random sample.
3. $\{X_i\}_{i=1}^n$ is **normally distributed**, or $\{X_i\}_{i=1}^n \sim N(\mu_x, \sigma_x^2$
    - With sufficiently large n (e.g., surveying at least 30 customers), this assumption is met asymptotically since the Central Limit Theorem will cause a normal distribution of the mean.

Violations: