Lab 1: Evaluating Assumptions
World Happiness

2023-03-05

There are three assumptions for the two-sample T-test:

1. **Metric Scale**: The data points must be continuous and metric in nature.
2. **IID data**: Each data point must be IID.
3. **No major deviations from normality, considering the sample size**- 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.

Assumption 1: Metric scale

In particular, the t-test is not valid for variables which only have an ordinal structure. The variables in question, the happiness life ladder and GDP per capita, are both continuous and metric. The summary below shows a 5 figure summary of each variable showing the quantitative nature of these variables. Therefore, this metric scale assumption is met.

```
summary(happiness$`Life Ladder`)
```

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

```
summary(happiness$`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
```

Assumption 2: IID Data

Each data point must be independent and individually distributed in the test. The countries' responses do not affect one another's choices, as each country's data point were gathered from a nationally representative sample of THAT country. Therefore, each country's happiness life ladder response and GDP per capita value are independent of one another's. The countries are also identically distributed, as each country's data in this dataset were all pulled from the larger World Happiness Report. Therefore, the IID assumption is met.
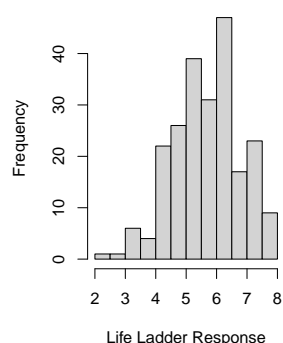
Assumption 3: No major deviations from normality, considering the sample size

As shown below, the distribution of responses for both variables, the happiness life ladder and GDP per capita, are both relatively normally distributed. There is not a heavy skew any direction, with only a slight skew to the left. However, this can be dismissed as it is a high sample size of 226(after removing countries with empty GDP per country), so the Central Limit Theorem kicks in. Therefore, there are no major deviations from normality, and this assumption holds.

```
happiness = happiness[!is.na(happiness$`Log GDP per capita`),] # removing NA values for countries that don't report
nrow(happiness) ## There are a total of 226 complete responses from countries.
```

```
## [1] 226
```