Lab 1: Evaluating Assumptions

Attitudes Toward the Religious

2023-03-05

There are three assumptions for the paired T-test:

1. **Metric Scale**. In particular, the t-test is not valid for variables which only have ordinal structure
2. **IID data**. In particular, each pair of measurements, (X_i, Y_i) is drawn from the same distribution, independently of all other pairs
3. **The distribution of the difference between measurements has 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

All three assumptions for the paired T-test are met, as reviewed below:

Assumption 1: Metric Scale

Both the prottemp and cathtemp variables are based on a 'feeling thermometer', which is on a scale of 0 to 100 and are as such, are interval/ratio data (thus meeting the criteria of being metric scale data).

Assumption 2: IID

The data is identically distributed, as each prottemp-cathtemp pair comes from an individual respondent that is drawn from the same GSS sample. The data used for GSS_religion is the 2004 General Social Survey (GSS). According to Wikipedia, *'The GSS sample is drawn using an area probability design that randomly selects respondents in households across the United States to take part in the survey. Respondents that become part of the GSS sample are from a mix of urban, suburban, and rural geographic areas.'* The sampling method suggests that the data are independent, given the randomization process used.

Assumption 3: No major deviations from normality, considering sample size

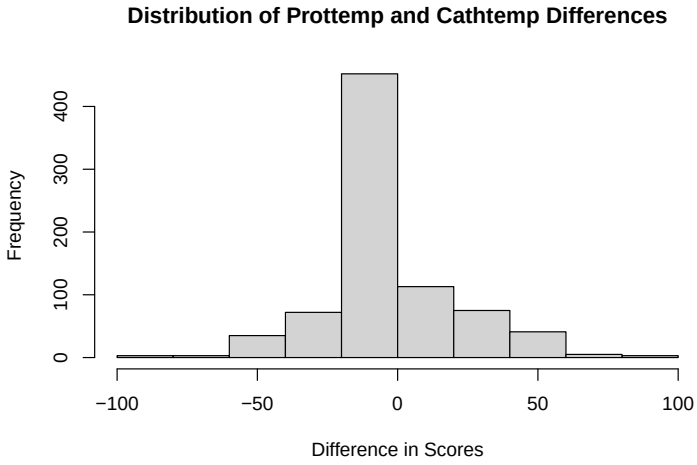
```
sum(!is.na(gss$prottemp) & (gss$prottemp >= 0) & (gss$prottemp <=100) &
    !is.na(gss$cathtemp) & (gss$cathtemp >= 0) & (gss$cathtemp <=100))
```

```
## [1] 802
```

The above code shows that there are 802 observations which have valid responses for both prottemp and cathtemp. There are 802 valid responses, which matches the number of observations in the gss dataset (802 observations). Thus, there is no need to subset the data any further.

```
diff_prottemp_cathtemp <- gss$prottemp - gss$cathtemp
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -85.000   0.000   0.000   2.403  10.000 100.000
```



The distribution of differences between the measurements (diff_prottemp_cathtemp) appears to be mostly normal, with a very slight right skew (mean > median, but the difference between the two is minor (2.4)). In addition, the sample size (802 pairs) is adequately large for the Central Limit Theorem to hold, particularly in light of the mostly normal distribution.