Lab 1: Evaluating Assumptions

Wine and Health

There are three assumptions for a Wilcoxon signed-rank test:

1. **Metric scale.** In particular, X and Y are both measured on the same metric scale.
2. **IID data.** In particular, each pair X_i, Y_i is drawn from the same distribution, independently of all other pairs.
3. **The distribution of the difference X - Y is symmetric around some mean μ.**

Assumption 1: Metric scale

The variables representing the number of deaths from heart disease and liver disease are both quantities of people per 100,000 who died from each respective disease per country observed. This information is provided in the R documentation of the dataset, shared below, and therefore these variables meet the criteria of being measured on the same metric scale:

```
?wine
```

Format

A data.frame with 21 observations on 5 variables:

- **country:**
- **alcohol:** liters alcohol from wine, per capita
- **deaths:** deaths per 100,000
- **heart:** heart disease dths per 100,000
- **liver:** liver disease dths per 100,000

Assumption 2: IID data

This data is identically distributed because the amount of alcohol consumption, overall deaths, and heart & liver disease deaths per observation come from an individual country. This also means the data is independent, as the data cannot come from or reflect more than one country. The number of heart disease deaths in Finland cannot tell us anything about the number of those deaths in Austria, for example.

Assumption 3: The distribution of the difference X - Y is symmetric around some mean μ

As reflected below, the distribution of the difference between the number of heart disease deaths and liver disease deaths per country around $\mu = 162.3$ is not symmetric. Therefore, this assumption has not been met.

```r
diff_heart_liver <- wine$heart - wine$liver
```

```r
summary(diff_heart_liver)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   12.8   110.3   174.6   162.3   199.8   293.5
```

```r
hist(diff_heart_liver)
```