Lab 1: Question 3.3

Bailey Kuehl, Erik Sambrailo, Jenna Sparks, Roshni Tajnekar

March 7, 2023

We would like test whether countries have more deaths from heart disease or from liver disease. To examine this, a **Wilcoxon signed-rank test** is proposed. We define two random variables H (Heart Disease) and L (Liver Disease), where both $h, l \in \mathbb{N}$. Because we are using a Wilcoxon signed-rank test, the null hypothesis becomes $H_0 : P(W_+) = P(W_-)$, meaning that the the probability that the sum of the ranks is positive, W_+, is equal to the probability that the sum of the ranks is negative, W_-.

Assumption 1: Metric Data (Interval Scale)

In this case, we are working with counts of death per country, which can be considered metric data. Metric data can be divided up into interval scale data, thus confirming that **the data is valid for this test.**

Assumption 2: Independence

All $(H_1, H_2, ...H_n)$ and $(L_1, L_2, ...L_n)$ need to be I.I.D. Because we are examining heart and liver disease per country, there are potential violations to this assumption in the data. First, we have a concern related to geographical clustering. Sampling different countries for health-related data is not independent because countries tend to have trends in health, access to medical resources, etc. Additionally, both heart disease and liver disease can coexist, and it's entirely possible that patients who died of heart disease had some form of liver disease, which potentially affected their heart. Because of these reasons, there are several potential **violations of the independence assumption.**

Assumption 3: The population distribution of the difference scores is symmetric about the mean μ

To test this assumption, we want to test whether the difference between pairs is symmetric.