Week 12

Rebecca Lindgren

2025-04-10

# Project Description

The purpose of my final project is to create an R package that allows user to quickly and comprehensively evaluate the normality of their data, a *critical* first step in many statistical analyses. I will include functions for the following “checks” in my package:

* Statistical Tests
  + Shapiro-Wilk Test
  + Komogorov-Smirnov Test
  + Anderson-Darling Test
* Visual Checks
  + QQ Plot
  + Density Plot
  + Histogram with normal curve overlay
* Run **ALL** checks

## Statistical Tests

The purpose of the statistical tests function(s) will be to provide users with a statistical evaluation of their data’s normality. One issue with visual checks of data normality is that the evalution is subjective. Users will have to ask themselves: “Is this normal *enough*?” With statistical tests, you can make a decision based on the p-value, an objective cut-off.

The stat\_test function will default to running all statistical tests (test = all; Shapiro-Wilk, Komogorov-Smirnov, and Anderson-Darling). However, users could specify if they want just one test run: ex. test = sw.

The input for the stat\_test function needs to be a numeric vector. Missing values in the data are allowed and will be automatically removed (although a warning will be displayed that there are missing values).

The output will be the results of each statistical test. I also want the output to be friendly for beginners, therefore, I will include some sort of message that tells the user whether their data follows a normal distribution or not based on the p-values of the statistical tests. For example:

if (sw$pvalue < sig\_level)  
 "A significant p-value suggests non-normality" else "A non significant p-value suggests normality"

## Visual Checks

The purpose of providing visual checks of data normality is to allow users to gain a better understanding of the distribution of their data. Although statistical tests tell you, objectively, whether or not the data follows a normal distribution, if it doesn’t, it is still valuable for the researcher to know *how* the data deviates from normality. Perhaps it is bi-model or extremely right-skewed. Having that extra knowledge about the data’s distribution allows for a better understanding leading to better interpretations.

The input for the visual checks also needs to be a numerical vector. The output will default to all the plots, but users may specify if they only want one specific plot (plot = qq rather than plot = all). To aid users in interpreting their visualizations, I will include descriptions of what a “normal” output would look like. For example:

## [1] "In a QQplot, data following a normal distribution would align closely along the diagonal line."

## ALL Checks

Rather than having to run multiple functions to evaluate your data statistically and visually, I want to include a function that runs **all** the checks at once for ease. This function is the main purpose of my package. I include the other functions just so that users wouldn’t have to load other packages should they want to run specific checks, but the idea behind my package is to quickly provide a well-rounded evaluation of the user’s data.

The input for this function also needs to be a numeric vector.

The output for this function will be a lot. I am still trying to figure out how to present it to the user in a succinct and digestible way initially, but also provide the specifics should they want to dive deeper. The output will include the output for all the previously described checks.