# Week 3 – Assignment 1 – Statistical Testing

By: Zach Adair

Regis University

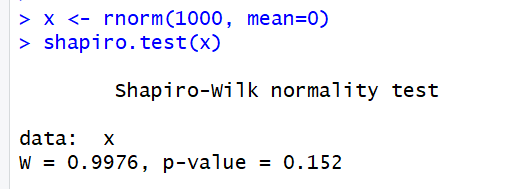
MSDS 650 – Data Analytics

## Introduction

For this assignment we will be looking at statistical testing and one in particular, the Shapiro-Wilk test. In a Shapiro-Wilk test the data is normally distributed and the null of the data is normally distributed while the alternative for this test, is that the data is not normally distributed. When the p-value is less than 0.05 then in typical hypothesis testing fashion the null hypothesis is rejected because there would be evidence that the data is not normally distributed like the null hypothesis stated. When the p-value is greater than 0.05, then the null is not rejected and the data of the normality will assumed. In these two examples the we will see how the Shapiro-Wilk test creates a p-value for a normal distribution and the t distribution and compare the frequencies.

## Exercise 1: Normal Distribution

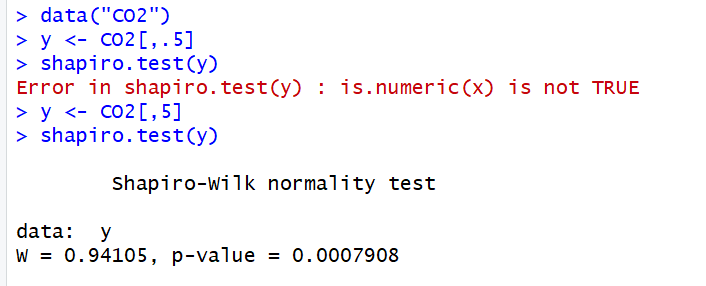
In the normal distribution we will generate a 1000 records with a mean of zero, and test the distribution. If the p-value generate is greater than 0.05 than the distribution generated will be normal.



After running the Shapiro-Wilk normality test we can determine that with a p-value of 0.152 then we would be able to assume the null hypothesis that the data is normally distributed.

## Exercise 2: CO2 Data

In exercise 2 I will be running the Shapiro-Wilk normality test on the CO2 dataset which is a preloaded dataset within RStudio. Once again if the p-value is greater than 0.05 the distribution will be assumed as normal.



From this rendition of the Shapiro-Wilk normality test we calculated a p-value of 0.007908 which is less than 0.05, thus the null hypothesis will be rejected that the distribution of the CO2 dataset is normal and the alternative hypothesis of the data not being normally distributed will take its place.

## Summary

I found this a wonderful assignment in learning about the Shapiro-Wilk normality test and how useful it can be to tell how a dataset is distributed. Knowing this information can be good for analysis of the data and understanding what to expect from further data analytics. Figuring out the distribution of the data can be terrific for not only descriptive analytics and what to expect from your mean and median, but also to figure out what further tests can be run so you can do predictive analytics.

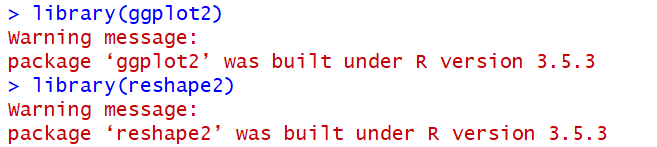
# Experiment from r-bloggers Website

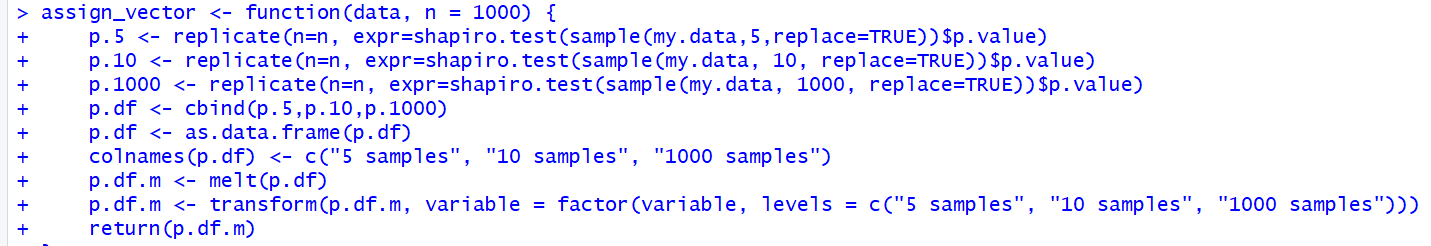
## Introduction

The extension of this assignment is on the website: <http://www.r-bloggers.com/normality-and-test-for-normality/>. In this experiment I will be walking through some tests of normality in RStudio and discovering the different ways this can be done.

## A Function to test normality many times

To start the experiment we need to generate some data. First I’ll load some libraries that will be needed and then form a function to run normality tests upon.

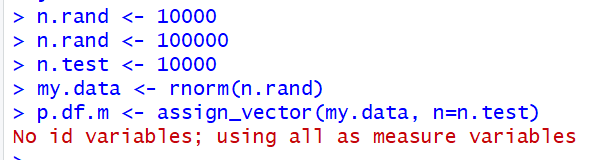




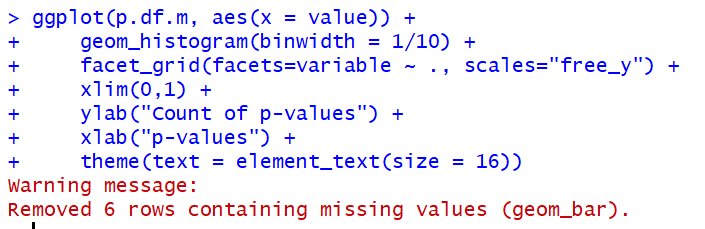
In this step we loaded in the libraries of ggplot2 and reshape2 and then created a vector of size 1000 along with several vectors.

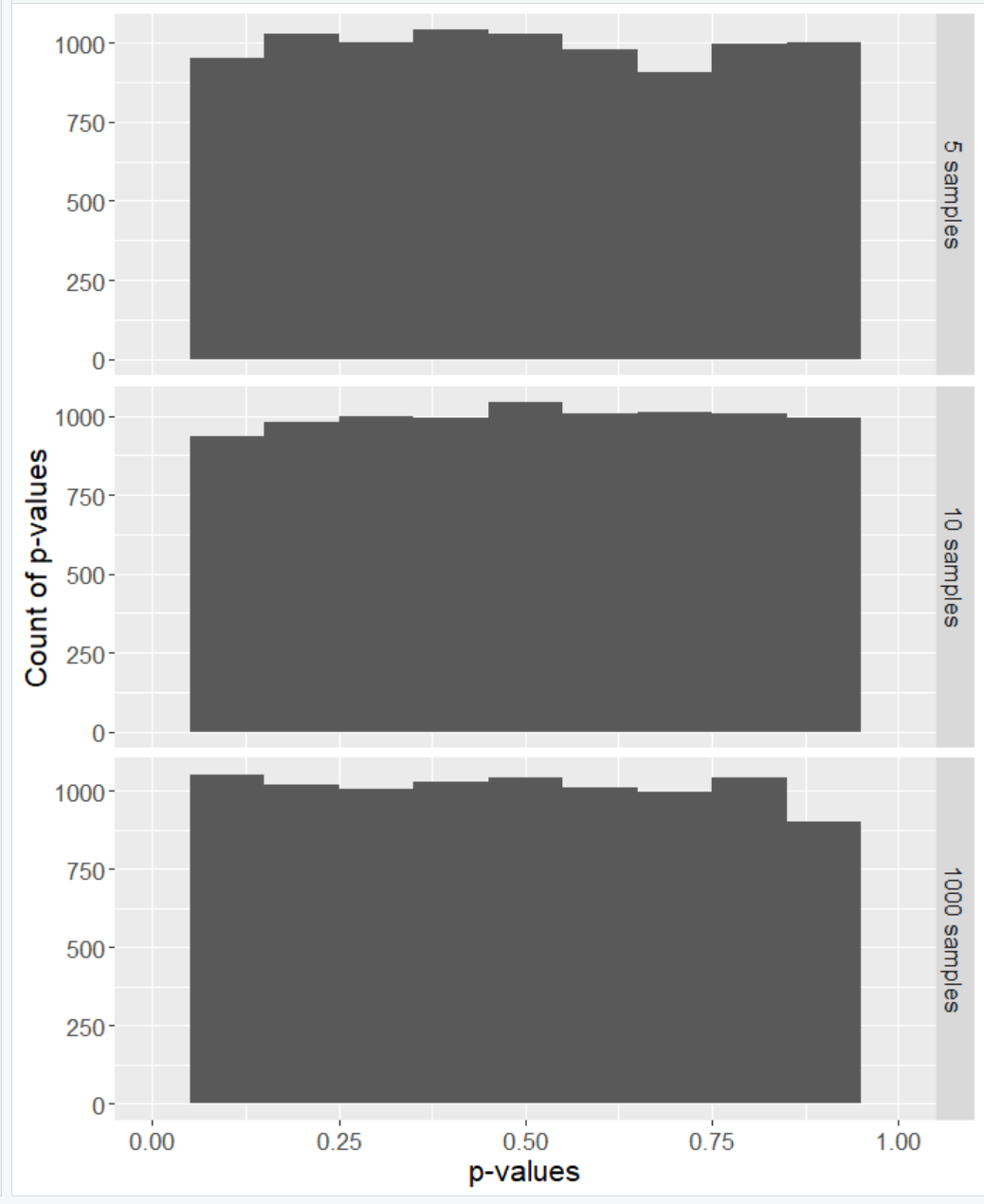
## Clean, random data

In this next step I’m going to simulate real-world conditions, where we have an underlying population from which we sample a limited number of times. After I’ll use my assign\_vector to test for normality.

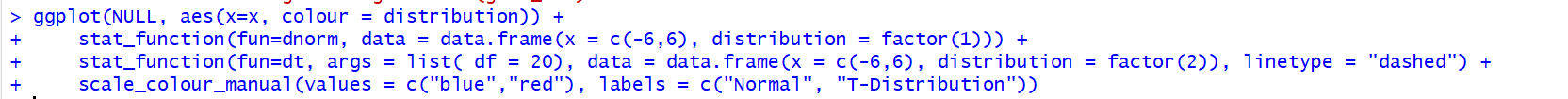


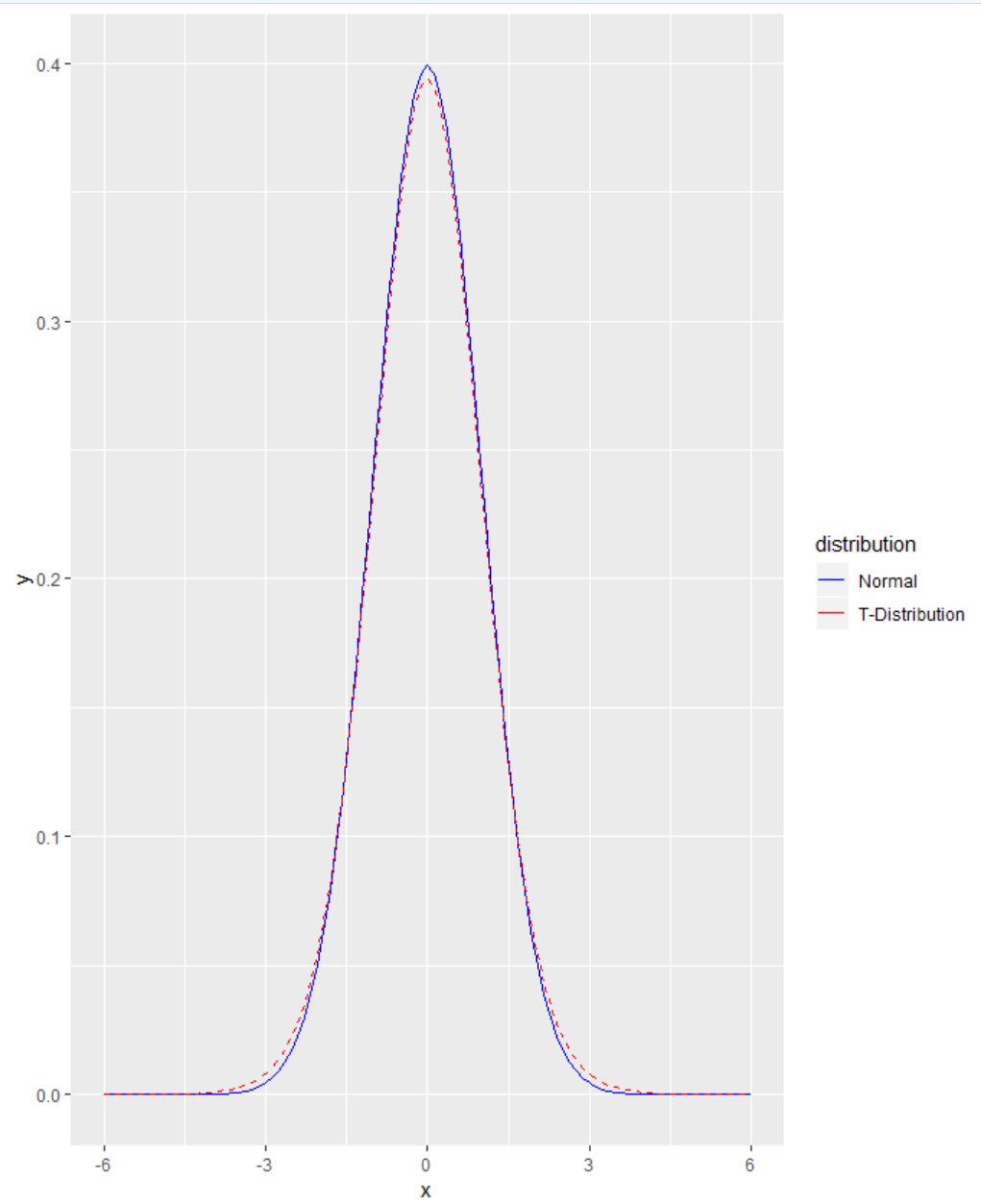
Here I assigned some variables, I did some research concerning the “No id variables; using all as measure variables” but I couldn’t find anything about the issue so I treated it as a warning and moved on. Now I will plot my first graph, a histogram for the 5, 10, and 1000 samples vectors made earlier.





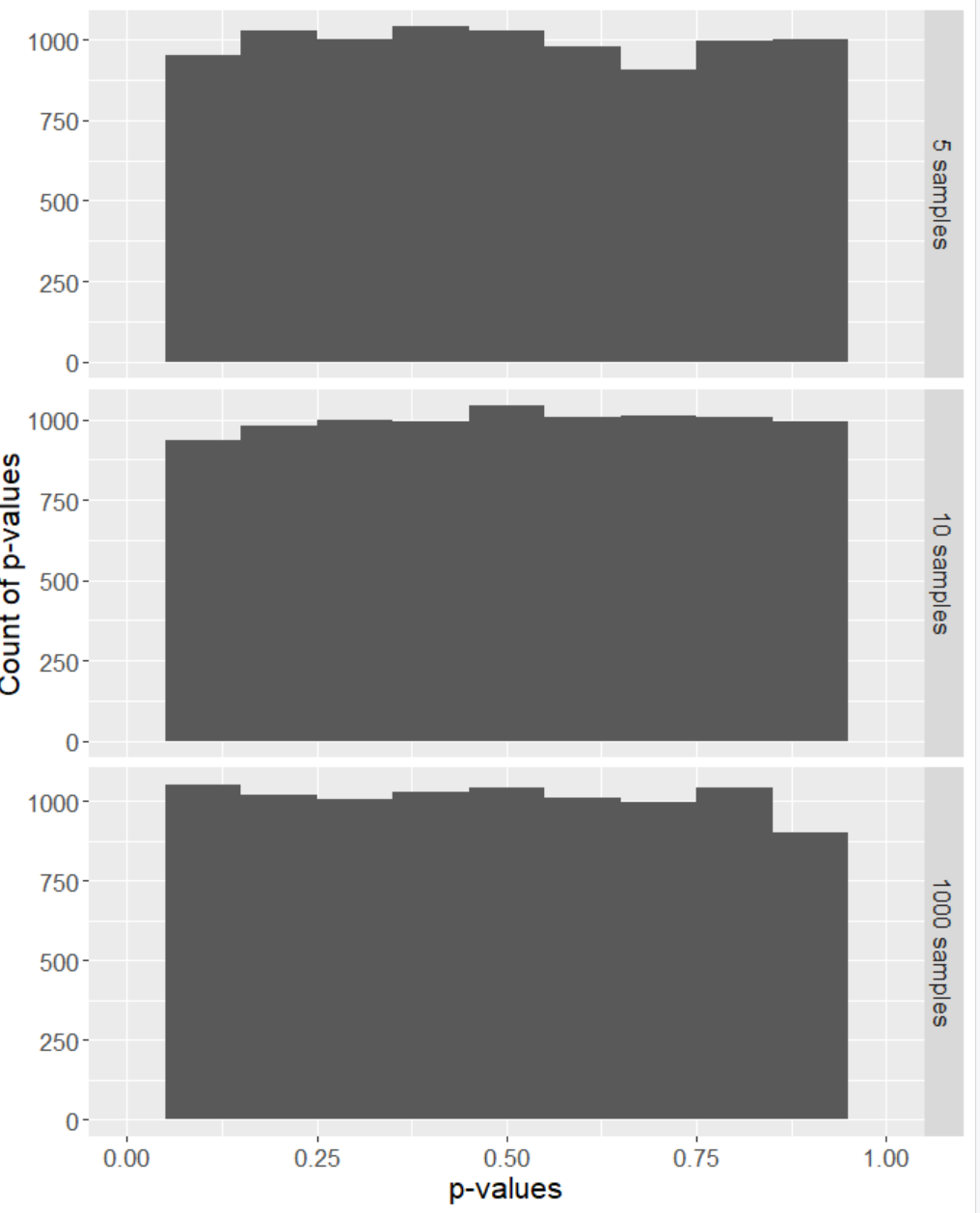
Here are the histograms for our three samples, the distribution is hard to track just by our eye but we have more graphs to plot so I’m going to keep moving forward with those next ones.





In our last plot it was the normal distribution mapped upon the T-distribution and almost seamlessly on top of each other from the look of it. From the experiment write up it mentions the “fat pencil” test and I would say the T-distribution certainly does pass that test.

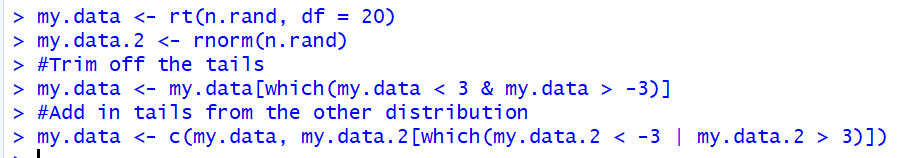


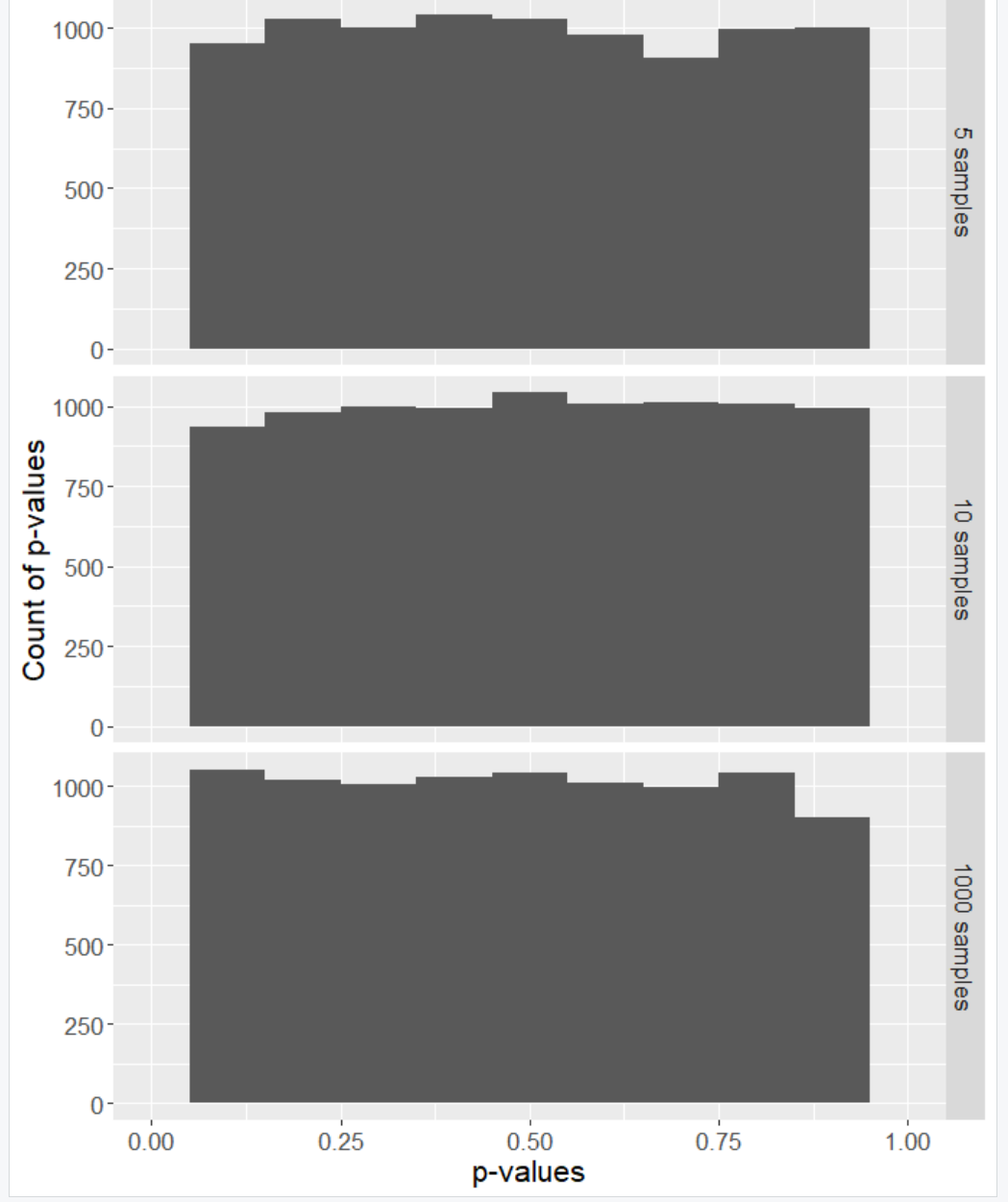


## Testing the Tails

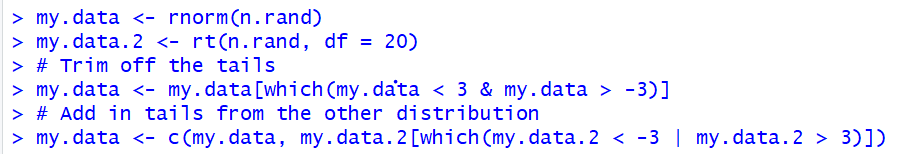
So far within this experiment we haven’t learned a whole lot about the tails and the distributions are hard to notice a real difference of distribution between them either, so our next step here is to test the tails.

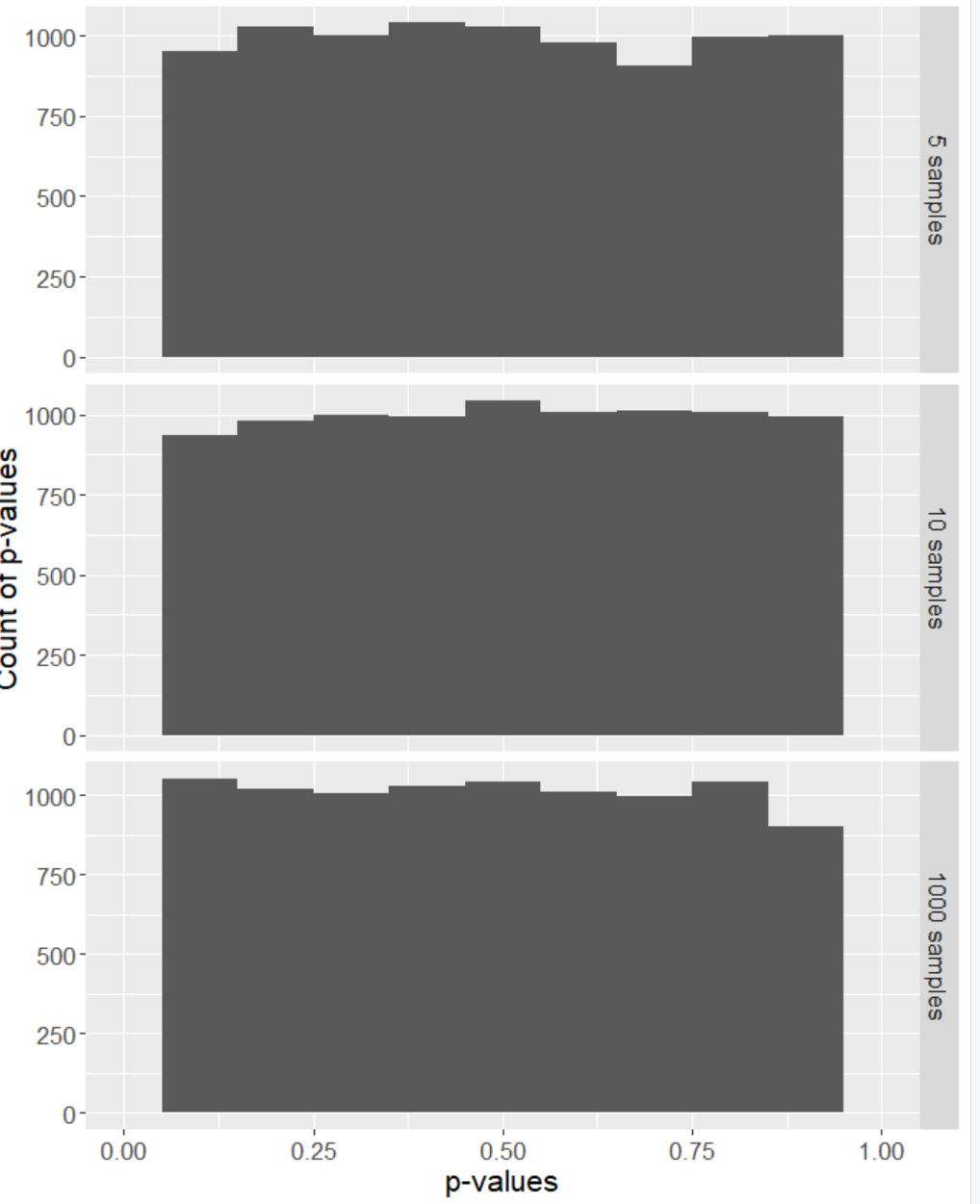
To test the tails, I will construct a data set that uses the t-distribution for 99%, leaving the 1% of the tails outside this distribution.





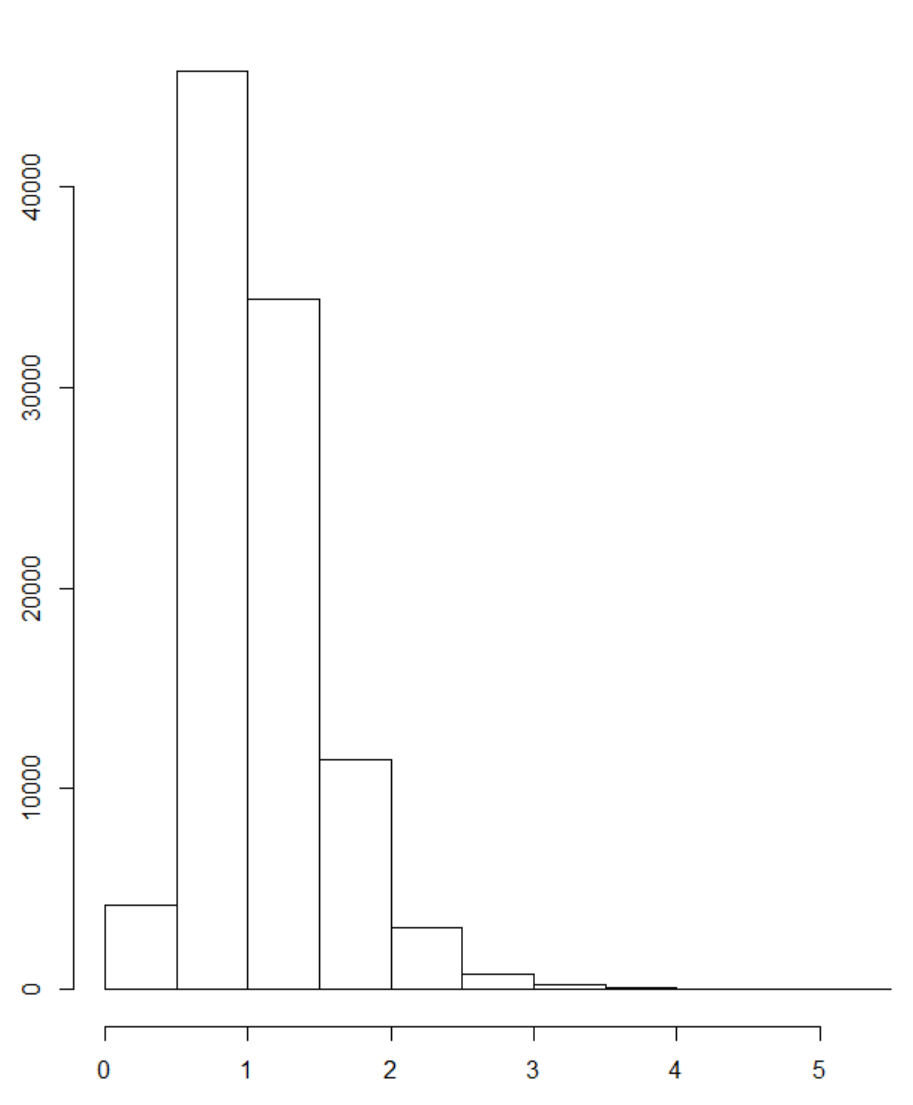
Now we will flip it around where the data is 99% normally-distributed, but using the t-distribution in the extreme tails.





Here is 99% of the data from the normal distribution, yet the normality test looks almost the same as the normality test for just the t-distribution





From the plot tests run above what I’ve learned is that small sample sizes make everything appear normal, and that normality tests are sensitive to what goes on in the extreme tails.

From what we’ve run we have the evidence to reject that the data is normal. The data being normally distributed isn’t something to worry too much about, when analyzing data it is more important to rely on statistics when needing to catch big errors. When making a prediction model, recognize the chance for errors in judgment and explicitly state any of your assumptions.