The Intricacies of the Normal Distribution

Samuel Greeman

The usefulness of normality tests cannot be understated. There are many times as the junior statisticians we are when we are presented with a data set, and asked for, say, a confidence interval, but we can't start right away, since we don't know where this data comes from. In this paper, we will learn how to make sure the data you have comes from a normal distribution.

1 Introduction

When assessing normality, there are two typical methods that are useful: graphs, and tests. Much like other hypothesis tests, graphs are helpful for confirming an assumption, but tests are more definitive, and give indisputable results, and we will go over both.

2 Techniques for Assessing Normality

Normal Probability Plot Method

The Normal Probability Plot is also known as a QQ Plot, but this QQ Plot is on the Normal distribution. The x-axis of the plot is the sorted values of the sample data, and the y-axis is the z-scores for the values of the data, but it can work vice-versa as well. Typically, if the data follows a Normal distribution, the plotted points will have a linear trend. If the data is not Normal, then the plot will look like something else, often a logarithmic function.

Histogram Plot Method

Histograms are very easy to interpret. They essentially tell you how many points of your data fall into different numerical ranges, with a taller bar corresponding to a higher density of points in that region. If you want to assess normality of a dataset, then you can use histograms. Simply take the histogram of your sample data! If the data is normally distributed, then the histogram will resemble the bell-shaped curve of the normal distribution.

3 Normality Tests

Shapiro-Wilk

The Shapiro-Wilk Test for normality is exceptionally simple, at least once you have the calculation part done. To perform the test, you must calculate a value, W, as follows:
$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
And the parameters are:

$a_i = (a_1, ..., a_n)$ are constants that are generated from the sample mean, variance, and covariance of a sample of size n that is normally distributed.

$x_{(i)} = (x_{(1)}, ..., x_{(n)})$ is the i'th order statistic of the sample.

$\bar{x}$ is the sample mean.

$x_i = (x_i, ..., x_n)$ is the i'th point in the sample.

The values of a_i aren't usually calculated by hand or calculator. These are calculated using software like R. The cutoff value for determining normality is defined case-by-case, so in the example, we will show how to use the value.

Anderson-Darling

Weird theme of having 2 names in each test, as we see the Anderson-Darling test here. In this case, the statistic is even harder to calculate:
$AD = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)(ln(F(X_i) + ln(1 - F(X_{n-i+1}))$
Our parameters are as follows:

n is the sample size.

$F(X_i)$ is the cumulative distribution function for the normal distribution.

An important note is that in this case, the data need to be ordered, so the $i = 1$ observation is the smallest.

Even more tediously, computing the significance is dependent on the value of AD, so we leave it to R to calculate the associated p-values in our examples.

Jarque-Bera

Another test that is only doable with software, the test statistic for the Jarque-Bera Test is computed using the formula:

$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4})$

Our parameters are tricky, and again we use software for their calculations:

K is the kurtosis of the sample, which we described in the previous project.

S is the skewness of the sample, which we also touched on previously.

4 Example

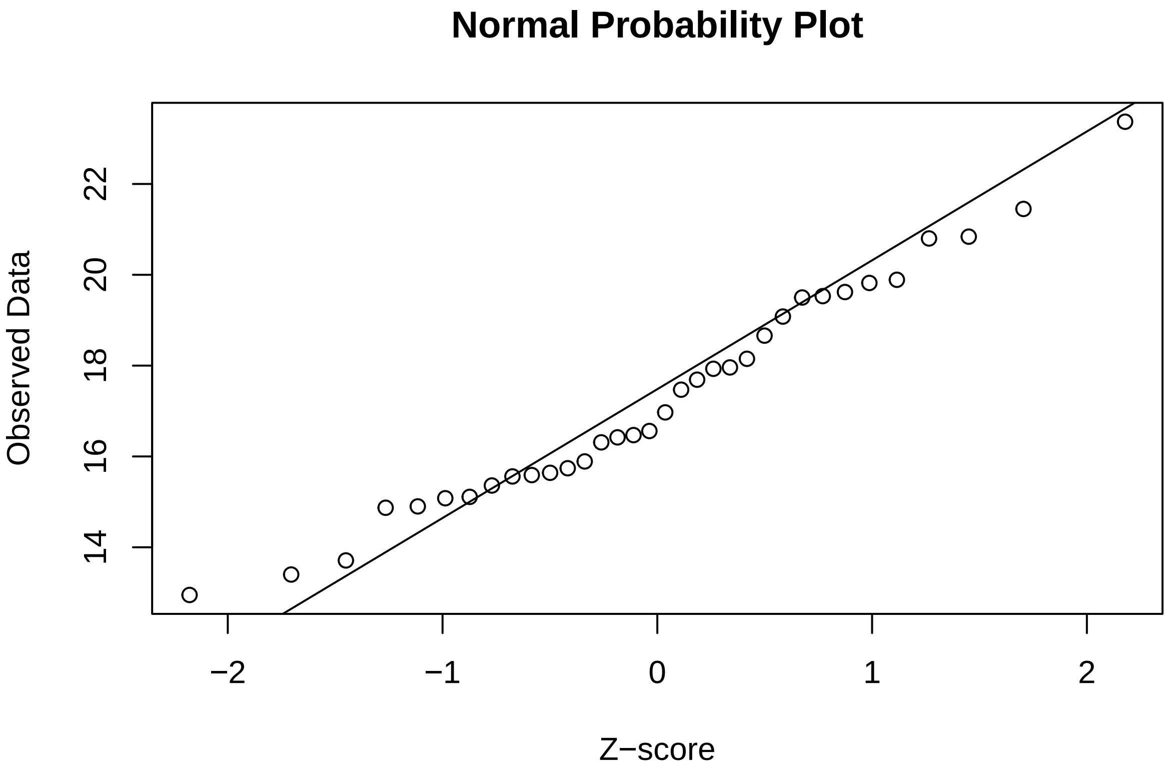
Here, we present all of these normality assessments in action with an example data set that I produced.

Example

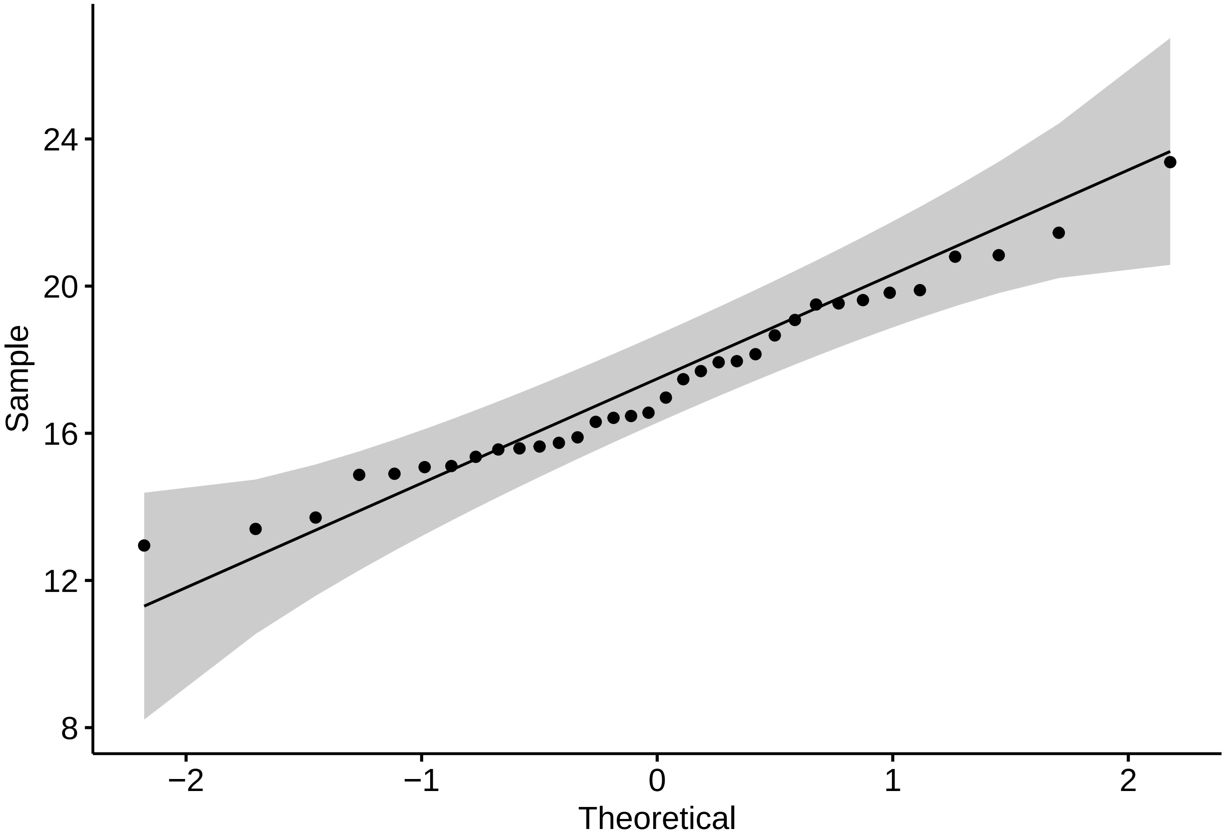
Samuel Greeman

Our first example involves a data set of size n = 34 with population mean 17.21 and population standard deviation of 2.65. We will use all of our tests to see if our data set comes from a normal distribution.

Normal Probability Plot



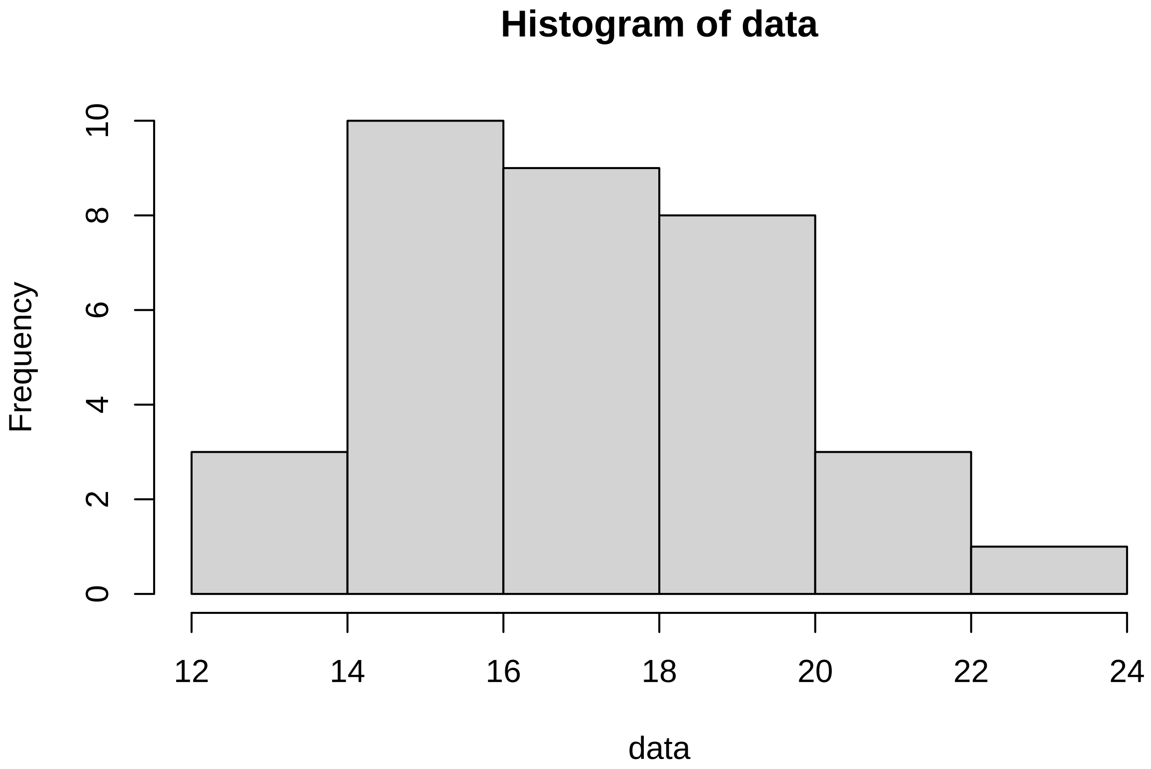
We can see from this plot that our sample observations are linearly related with the theoretical Z-scores, but to make it more clear, consider this plot:



This is the same plot except there is a cone surrounding a general area around the line. This is to help us see that all of our observations are within the 'cone of certainty' that gets wider as we stray further away from the mean, meaning this sample looks to be a sample from a normal distribution.

Next, we will look at a histogram:

Histogram



Unlike the previous plots, this histogram is not as clearly normal-looking. Sure, there are more observations closer to the mean, but it's not as clear-cut. If I were to look at just this plot, I may not be convinced that this sample is from a normal distribution.

Moving onto the tests, we will start with the Shapiro-Wilk Test, which R will run and give us the test statistic and its p-value.

Shapiro-Wilk Test

```
shapiro.test(data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.9725, p-value = 0.5336
```

Our test statistic, W, is 0.9725, which is correspondent to a p-value of 0.5336. For this test, the null hypothesis is that our sample does indeed come from a normal distribution, so we would need a p-value less than, say, 0.05 if we wanted to reject the null hypothesis, so normality passes this test.

On to the Anderson-Darling Test, where we will hopefully get the same results.

Anderson-Darling Test

```
ad.test(data)
```

```
##
##  Anderson-Darling normality test
##
## data:  data
## A = 0.3736, p-value = 0.398
```

Our test statistic A = 0.3736 carries with it a similarly high p-value of 0.398, which is again not close to small enough to consider rejecting the null hypothesis of normality.

Our final test is the Jarque-Bera Test.

Jarque-Bera Test

```
jarque.bera.test(data)
```

```
##
##  Jarque Bera Test
##
## data:  data
## X-squared = 1.1271, df = 2, p-value = 0.5692
```

The test statistic for the Jarque-Bera Test is basically a chi-squared test statistic, and as we can see, the test statistic, 1.1271 only has a p-value of 0.5692 with 2 degrees of freedom, once again telling us there is no evidence of a departure from normality.

From these tests and plots, we got 4 definitive, supportive results of normality, and just 1 neutral result. This lets us say that it is very likely that the sample that we have follows a normal distribution.

5 Conclusion

At the end of the day, it's obvious what we learned: alone, these tests definitely help confirm or subdue doubts about normality a little bit, but the more tests you run, the more evidence you get one way or another because it is highly unlikely that two tests will tell you different things. In essence, the more tests for normality and diagnostics you run, the more confident you will be in your result.

References

- R, RStudio

- Notes from my high school AP Statistics course