Is My Data Normally Distributed?

Making a Decision Based on Visualizing Data, Finding Skewness and Kurtosis, and Performing Formal Tests for Normality

Prepared by Allison Horst
Bren School of Environmental Science & Management, UCSB

Assessing Normality

When we are deciding whether or not we can assume normality when performing statistical comparisons, there are three main ways in which we can assess the data (which should be used in combination):

a) Visualization of the data (histogram, Q-Q plots)
b) Skewness and kurtosis (see handout)
c) Formal tests for normality

In this document, these three approaches are summarized with an example for ‘non-normal’ and ‘relatively normal’ data in each case.

1) Visualize the Data

The first thing that you usually want to do is look at your data. Two common approaches are using a histogram (by the hist() function in R) and using a quantile-quantile (Q-Q) plot (which might employ the qqnorm() function in R). For example, let’s say that I have a sample ‘A,’ which contains the following 17 observations:

\[ A \leftarrow \text{c(1.1, 2.0, 1.4, 1.0, 1.1, 2.3, 3.0, 3.5, 2.2, 1.8, 4.3, 5.6, 7.0, 8.1, 8.6, 10.9, 12.2)} \]

The first thing you might do is look at the data graphically.

\[ > \text{hist(A)} \]
Just from the histogram, you might begin to think that the sample (and underlying distribution from which it was sampled) may not be normally distributed. But that is not enough for us to make a decision about a statistical test. How else can we look at the data?

One other option is using a quantile-quantile (Q-Q) plot. In a Q-Q plot, the percentiles of a theoretical standard normal distribution are plotted against the actual percentiles of your sample data. If your sample is relatively normal, you would expect a nearly linear positive relationship between the two.

For example, if I have a sample ‘B’ that contains the following 25 observations:

```r
B <- c(1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 5, 6, 7, 7, 8, 7, 7, 8, 9, 10, 11)
```

We can see from a histogram of ‘B’ that the data appears relatively normal (left graph, below). Then, if we create a Q-Q plot using the following command:

```r
> qqnorm(B)
```

We see that the Q-Q plot (right graph, below) is indeed highly linear.

Here, both the histogram and Q-Q plot indicate that the data in ‘B’ is relatively normally distributed.
Let’s contrast these graphs with the histogram and Q-Q plot for ‘A’. The histogram for ‘A’ appeared relatively non-normal (shown again below, on the left). What does the Q-Q plot reveal (below, on the right)?

We can see above that, as expected, the Q-Q plot reveals a non-linear relationship between the theoretical and sample quantiles. Combined, these should raise a red flag if you are considering moving forward with a normal assumption.

So we have looked at our data in a couple of different ways that can help give us a sense of whether or not our data might be relatively normal. But how can I quantify that information? The first thing you might do is determine the skewness and kurtosis.

2) **Determine the Skewness and Kurtosis**

Skewness and kurtosis are measures for how asymmetric (skewness) and ‘pointed’ (kurtosis) a distribution is, compared to a normal or standard normal distribution. For a thorough summary, and for information regarding how to interpret skewness and kurtosis values, please see the “Skewness and Kurtosis” document on the course GauchoSpace site.

Skewness and kurtosis are easily computed in R once the ‘moments’ package has been installed (see document).

Let’s again look at Samples A and B that were visualized above, where ‘A’ looked positively skewed, and ‘B’ looked relatively normal in both the histogram and the Q-Q plot. If we have installed the ‘moments’ package, skewness for ‘A’ and ‘B’ are simply found by:

```R
> skewness(A)
[1] 0.8873918
> skewness(B)
[1] 0.125439
```

The values reveal that ‘A’ is moderately (bordering on highly) skewed in the positive direction, which is consistent with what we see in the histogram. Sample ‘B’, however, has a very low skewness (0.125), and we would conclude based on a rule of thumb (see document) that ‘B’ is relatively symmetrically distributed – again, consistent with what we saw in the histogram and Q-Q plot for ‘B’.
We can similarly determine kurtosis:

```r
> kurtosis(A)
[1] 2.482633
> kurtosis(B)
[1] 2.620641
```

While somewhat less informative, the kurtosis shows us that both ‘A’ and ‘B’ are slightly ‘flatter’ than a standard normal distribution, though not to a great extent (a standard normal has a kurtosis = 3. See document on GauchoSpace for more information).

So now we know a few more things about ‘A’ and ‘B’:

- **Sample A (n = 17):** Appears positively skewed, the Q-Q plot is non-linear, there is a moderate to high positive skewness, and it is slightly flatter than a standard normal.

- **Sample B (n = 25):** Appears relatively normal, the Q-Q plot is linear, there is low skewness, and it is slightly flatter than a standard normal.

At this point, we might be thinking that we shouldn’t use a normal assumption for ‘A,’ but a normal assumption is probably okay for ‘B.’ But is that enough for us to make a decision, particularly since our sample sizes for both are below the ‘n = 30’ cut-off to use parametric tests (i.e. Student’s t) even if the distribution is non-normal?

The next step is to use a formal test for normality.

### 3) Formal Tests for Normality

There are several statistical tests designed specifically to help you make a decision about the ‘normality’ of your data.

Three of the most common, and widely used, formal tests for normality are: the Anderson-Darling test, the Shapiro-Wilk test, and the Lilliefors (Kolmogorov-Smirnov) test.

In the three tests, the null and alternative hypotheses are the same (though the theory behind each is very different):

- **H₀:** The random observations are sampled from a normally distributed population (i.e., the data is normally distributed)

- **H₁:** The observations are from a non-normally distributed population (i.e., the data is not normally distributed)
Shapiro-Wilk Test for Normality

The Shapiro-Wilk test may be the most commonly used test for normality. In R, the test is straightforward: simply use the `shapiro.test()` function with your sample. Make sure to explore the function (`?shapiro.test`) before using to familiarize yourself with the arguments.

For example, let's look at the report of the Shapiro-Wilk test for normality with samples ‘A’ and ‘B’ from above:

```r
> shapiro.test(A)

Shapiro-Wilk normality test

data:  A
W = 0.8568, p-value = 0.01366

> shapiro.test(B)

Shapiro-Wilk normality test

data:  B
W = 0.9775, p-value = 0.8326
```

What would we conclude for each sample, based on the Shapiro-Wilk test results?

For Sample A, the p-value of 0.013 would lead us to **reject the null hypothesis** (using a significance level of 0.05). We would therefore retain the alternative hypothesis that the data is **not** normally distributed.

For Sample B (p = 0.833), we would **retain the null hypothesis** (for α = 0.05) that the data is normally distributed.

Thus, based on **three different assessments** (visualization, skewness and kurtosis, and a formal test for normality), we would probably decide the following:

- **Sample A: not normally distributed**
- **Sample B: normally distributed**

Review the process above for going through the three steps to convince yourself that they are generally in agreement.

Lilliefors Test for Normality

The Lilliefors test (a unique version of the Kolmogorov-Smirnov test, specifically for normal distributions) is easily performed in R once the ‘nortest’ package is installed.

Install the ‘nortest’ package in RStudio (Packages >> Install Packages >> search for ‘nortest’ >> Install), then load into your active RStudio workspace (make sure that the box next to ‘nortest’ is selected).
Explore the function (?lillie.test()). Perform the Lillefors test for normality as follows:

```r
> lillie.test(A)
Lilliefors (Kolmogorov-Smirnov) normality test
data:  A
D = 0.1969, p-value = 0.07887

> lillie.test(B)
Lilliefors (Kolmogorov-Smirnov) normality test
data:  B
D = 0.1063, p-value = 0.6595
```

Here, we would retain the null hypothesis (that the data is normally distributed) for both Samples, ‘A’ and ‘B’. Notice that this is different from our conclusion using the Shapiro-Wilk tests (where the null was rejected for ‘A’). The results of the Lilliefors test (compared to the Shapiro-Wilk test) are consistent with previous findings, where the Shapiro-Wilk test was more powerful (i.e. more likely to correctly reject a null hypothesis) than variations of the Kolmogorov-Smirnov test, like Lilliefors. Thus, Lilliefors may be considered a more conservative (though less precise) test for normality. Still, notice that the p-value is close to indicating non-normality.

**The Anderson-Darling Test for Normality**

Also in the ‘nortest’ package, you can find another test for normality: the Anderson-Darling test (function ad.test()). The test is similar to those above, and yields the following:

```r
> ad.test(A)
Anderson-Darling normality test
data:  A
A = 0.9331, p-value = 0.01368

> ad.test(B)
Anderson-Darling normality test
data:  B
A = 0.2265, p-value = 0.795
```

Notice that the results from the Anderson-Darling test are similar to the Shapiro-Wilk test: we would reject the null for Sample A (and therefore conclude that A is not normally distributed), but retain the null for Sample B (concluding that B is normally distributed).
Concluding Statement

The three approaches above (visualization, quantification of skewness and kurtosis, and formal tests for normality) should be used in combination to inform your use of a statistical test. However, in some cases the results might not be definitive (as for the Lilliefors test for Sample A, above) – in those cases, you should use your judgment based on a combination of all of the data.

This is a perfect time to remember: “Statistics are no substitute for judgment.” –Henry Clay

Keep in mind that there are a number of other tests for normality (Cramer-von Mises, Pearson’s, Shapiro-Francia, etc.) that can be found in the ‘nortest’ package. For more information, see the R documentation for the various tests:

?shapiro.test
?cvm.test
?pearson.test
?sf.test
?ad.test
?lillie.test