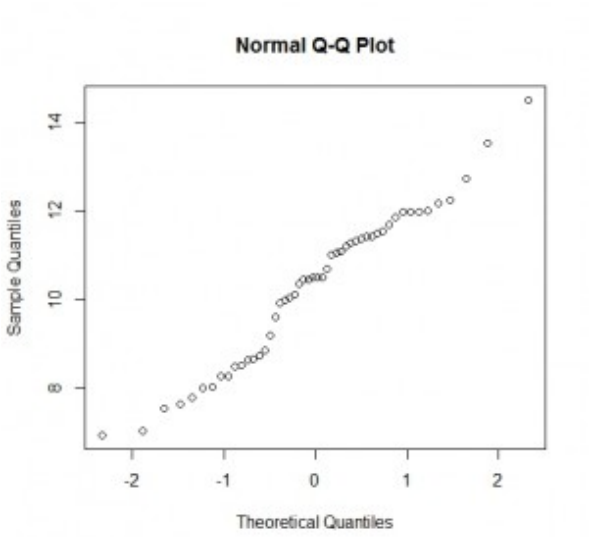
Understanding Q-Q Plots

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



(https://data.library.virginia.edu/files/example_qq.jpeg)

Now what are "quantiles"? These are often referred to as "percentiles". These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard Normal distribution from 0.01 to 0.99 by increments of 0.01:

Data Sources (/datasources/)

Research Data Management (/data-management/)

Research Software (/research-software/)

StatLab: Data Analytics (/statlab/)

Social, Natural, Engineering Sciences (/sne/)

Workshops (/training/)

People (/rds-staff/)

FAQs (/faq/)

Related Resources (/related-resources/)

Latest News

- Welcome Jenn!

```
qnorm(seq(0.05,0.95,0.05))
```

We can also randomly generate data from a standard Normal distribution and then find the quantiles. Here we generate a sample of size 200 and find the quantiles for 0.01 to 0.99 using the `quantile` function:

```
quantile(rnorm(200),probs = seq(0.01,0.99,0.01))
```

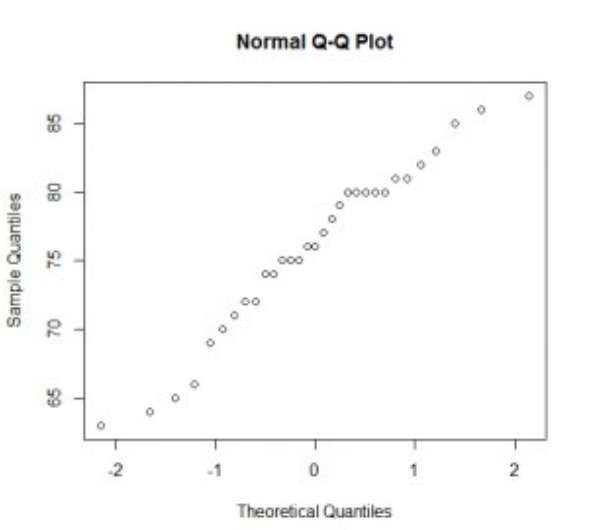
So we see that quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall. However it's worth noting there are many ways to calculate quantiles. In fact, the `quantile` function in R offers 9 different quantile algorithms! See `help(quantile)` for more information.

Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution.

In R, there are two functions to create Q-Q plots: `qqnorm` and `qqplot`.

`qqnorm` creates a Normal Q-Q plot. You give it a vector of data and R plots the data in sorted order versus quantiles from a standard Normal distribution. For example, consider the `trees` data set that comes with R. It provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. One of the variables is `Height`. Can we assume our sample of Heights comes from a population that is Normally distributed?

```
qqnorm(trees$Height)
```



(https://data.library.virginia.edu/files/trees.jpeg)

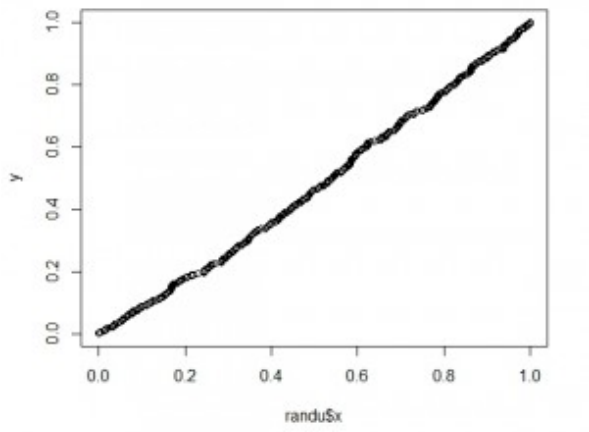
That appears to be a fairly safe assumption. The points seem to fall about a straight line. Notice the x-axis plots the theoretical quantiles. Those are the quantiles from the standard Normal distribution with mean 0 and standard deviation 1.

(http://data.library.vi
jenn/)

- Fall 2016 Data Science Short Courses (http://data.library.vi 2016-data-science-short-courses/)
- UVA DataFest! (http://data.library.vi datafest/)
- Data Science Sponsored Courses (http://data.library.vi science-short-courses-2/)
- UVa Library Subscription to lynda.com (http://data.library.vi library-subscription-to-lynda-com/)

The `qqplot` function allows you to create a Q-Q plot for any distribution. Unlike the `qqnorm` function, you have to provide two arguments: the first set of data and the second set of data. Let's look at the `randu` data that come with R. It's a data frame that contains 3 columns of random numbers on the interval (0,1). Random numbers should be uniformly distributed. Therefore we can check this assumption by creating a Q-Q plot of the sorted random numbers versus quantiles from a theoretical uniform (0,1) distribution. Here we create a Q-Q plot for the first column numbers, called `x`:

```
y <- qunif(ppoints(length(randu$x)))
qqplot(randu$x,y)
```

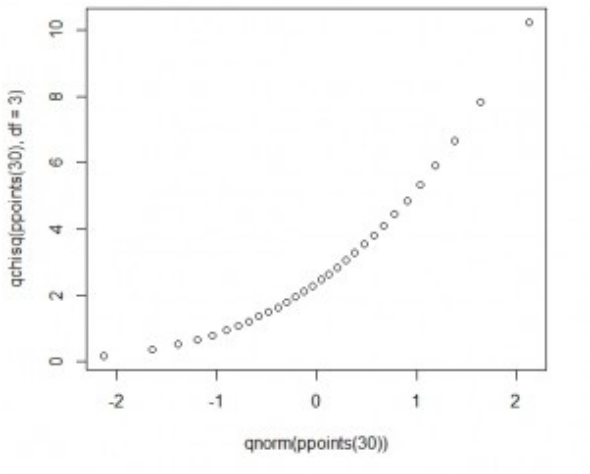


(https://data.library.virginia.edu/files/randu.jpeg)

The `ppoints` function generates a given number of probabilities or proportions. I wanted the same number of values in `randu$x`, so I gave it the argument `length(randu$x)`, which returns 400. The `qunif` function then returns 400 quantiles from a uniform distribution for the 400 proportions. I save that to `y` and then plot `y` versus `randu$x` in the `qqplot` function. Again, we see points falling along a straight line in the Q-Q plot, which provide strong evidence that these numbers truly did come from a uniform distribution.

What about when points don't fall on a straight line? What can we infer about our data? To help us answer this, let's generate data from one distribution and plot against the quantiles of another. First we plot a distribution that's skewed right, a Chi-square distribution with 3 degrees of freedom, against a Normal distribution.

```
qqplot(qnorm(ppoints(30)), qchisq(ppoints(30),df=3))
```

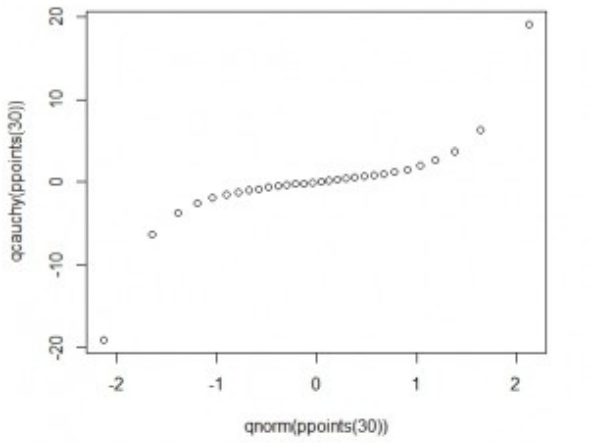


(https://data.library.virginia.edu/files/skew_right.jpeg)

Notice the points form a curve instead of a straight line. Normal Q-Q plots that look like this usually mean your sample data are skewed.

Next we plot a distribution with “heavy tails” versus a Normal distribution:

```
qqplot(qnorm(ppoints(30)), qcauchy(ppoints(30)))
```



(https://data.library.virginia.edu/files/heavy_tails.jpeg)

Notice the points fall along a line in the middle of the graph, but curve off in the extremities. Normal Q-Q plots that exhibit this behavior usually mean your data have more extreme values than would be expected if they truly came from a Normal distribution.

For questions or clarifications regarding this article, contact the UVa Library StatLab: statlab@virginia.edu (mailto:statlab@virginia.edu)

Clay Ford
Statistical Research Consultant
University of Virginia Library
August 26, 2015