





# DIAGNOSING A DISTRIBUTION

We often want to know what theoretical distribution best represents some observed empirical data. You may need to do this to select the appropriate parametric statistical test, analysis method, or validate some assumptions. In this activity, we will explore a new plot: the *quantile-quantile plot* (or q-q plot). This is an excellent diagnostic tool to identify whether a sample of data fits a particular distribution. While you may be interested in any number of distributions (e.g., normal or Gaussian distribution, exponential, Poisson, t distribution, binominal distribution, or Chi-Squared distribution), this activity will focus on the most common q-q plot for comparing data to the normal (Gaussian) distribution. The Keen text shows you how to construct a q-q plot for other distributions. You already have a good sense of how to visualize the distributions based on the diagnostic graphics you have studied to date (histogram, KDE, boxplot, empirical cumulative distribution function) and calculations of the descriptive statistics.

Today’s activity will also introduce you to R’s ability to synthesize random distributions easily. We will “fake” some empirical data using these procedures then apply the graphical diagnostics to them. Since we know the underlying distribution, we can easily see what the diagnostic graphs should look like if we have “normally distributed” data. In this way, when we apply this method to “real” data it should be clear what we are trying to diagnose. This is a way to explore some statistical concepts that cannot be overstated. In fact, many of my flashes of insight (those “ah-ha” moments) have come from comparing graphically and quantitatively “real” random distributions to ones that have been observed. The activity includes branches for you to explore for other distributions once you understand the basics.

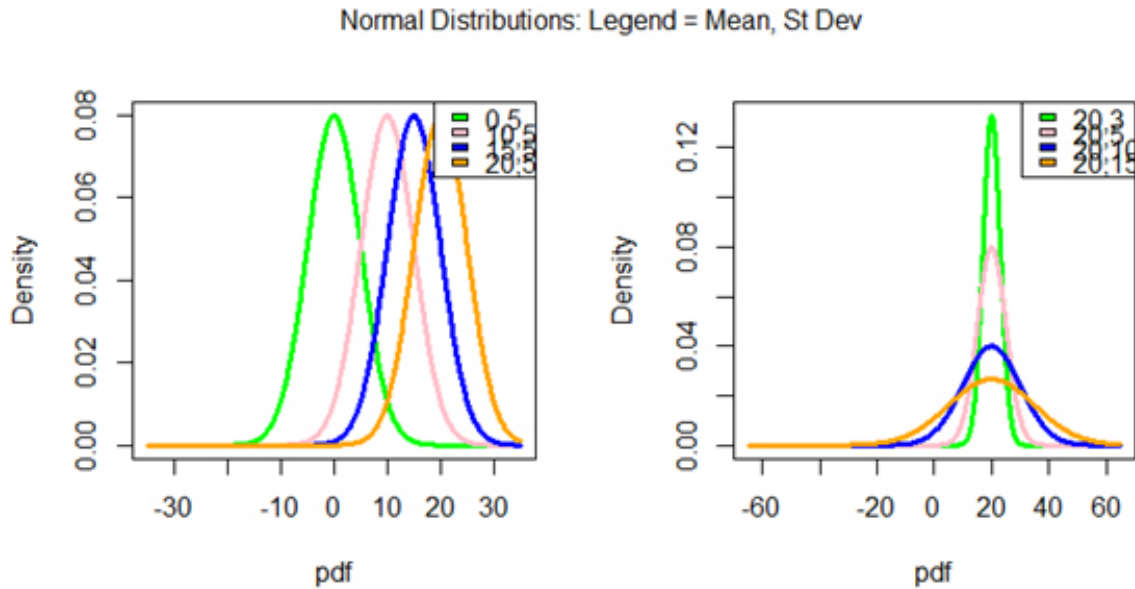
 <p><b>PURPOSE</b></p> <p>This activity will give you the opportunity to construct q-q plots for the purpose of graphically diagnosing distributions</p>	 <p><b>LEARNING OBJECTIVE</b></p> <p>Become familiar with R code technique and principles of distributions</p>
 <p><b>REQUIRED RESOURCES</b></p> <ul style="list-style-type: none"> <li>o R, R Studio</li> <li>o PostgreSQL ODBC driver, direct connection to PostgreSQL with R,</li> <li>o Sample R script for activity</li> <li>o Excel file showing the construction of Normal Q-Q plot</li> </ul>	 <p><b>TIME ALLOCATED</b></p> <p>90 minutes in class</p>

## TASKS



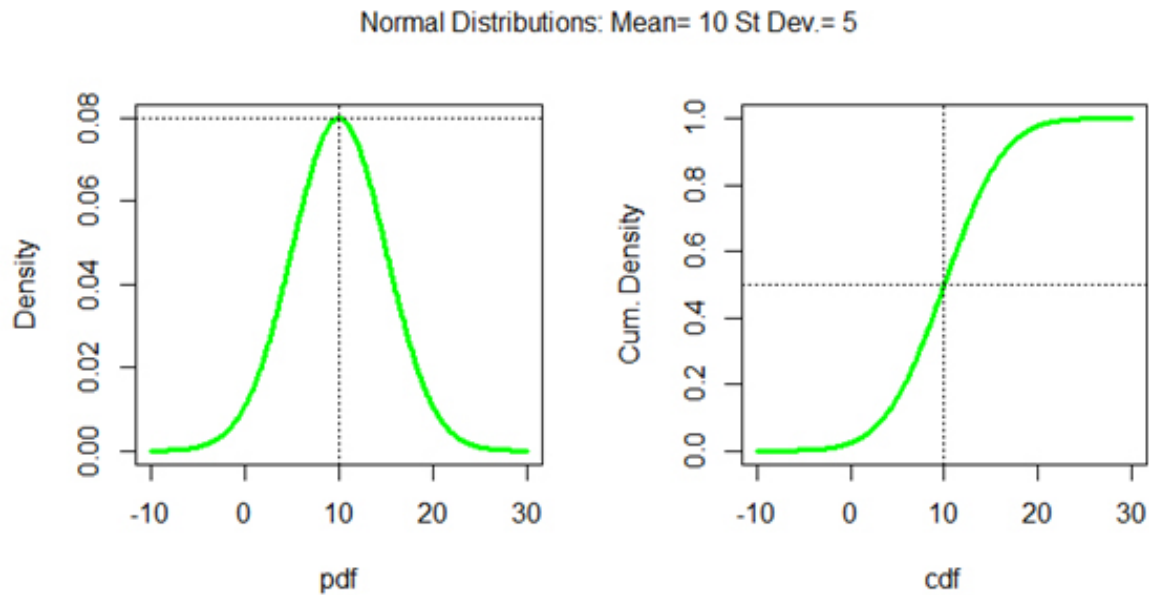
*First, the instructor will present a short overview PPT that will explain some of the key concepts and show you some sample code. Follow along with the R script for Activity 30 provided on the class website and lecture PowerPoints.*

- A. In the R script, section #2, there is example code for showing 2 plots. The first shows a set of theoretical normal distributions with 4 means and the same standard deviations. The second shows the maximum mean in first with 4 different standard deviations. Inspect the code line by line. The plots are created by using the `dnorm` function which returns the familiar “bell curve.” Change `mu` and `stdev` in the code until you are satisfied that you understand the differences in the shape, spread, and the mean of the distributions. Select one of the combinations to explore further (it doesn’t matter which one).



**Figure 53** Normal distributions with varying means and equal standard deviations (left), and varying standard deviations and equal means (right)

- a. Modify the code to produce the cumulative distribution curves. But before you do this, sketch out what you think the two plots will look like. For example, in left plot, how will the cumulative density function that you plot be arranged? Will they have the same slope/shape? What about in the right plot?
  - b. Now execute the code you modified (hint: just change the `dnorm` to the appropriate function).
  - c. If the plots don't look the way you expected, see if you can figure out why. Include both sets of plots in your write up and discussion. Write a short observation about what you know or don't know.
- B.** In the R script, section #3, there is code that is very similar to the code in section #2 but we will explore in more detail the density (`dnorm`) and cumulative density (`pnorm`) and quantile (`qnorm`) functions. Review the code line by line. Pay special attention to the `abline` functions that plot the dashed lines. Be sure that you can understand what is being plotted.



**Figure 54** Normal distribution density function plot (left) and cumulative density function plot (right)

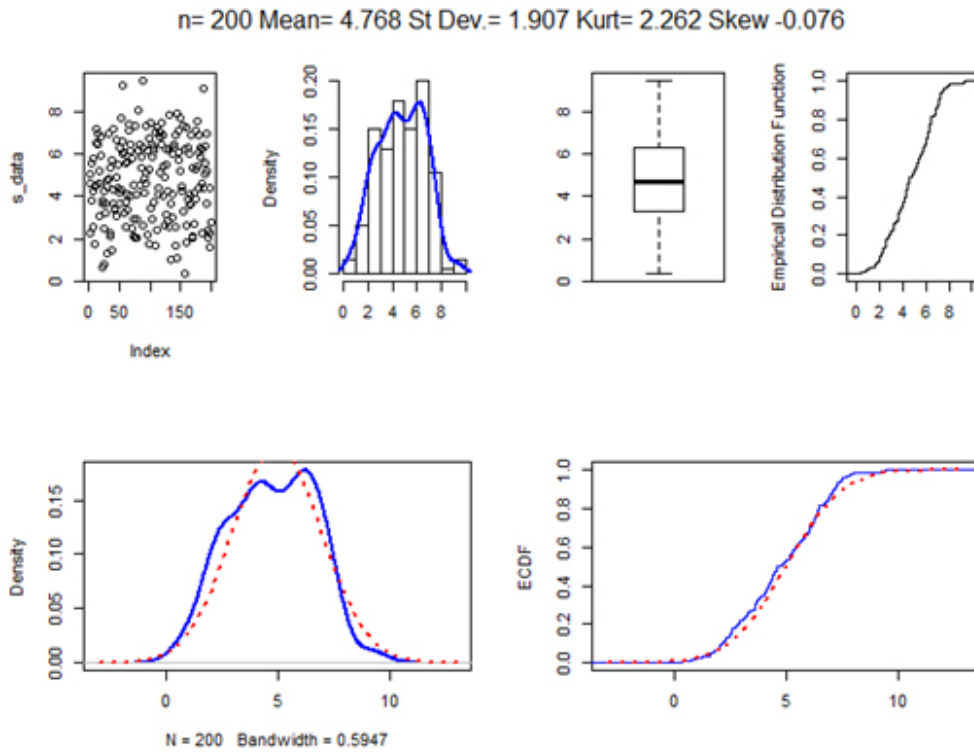
- a. As mentioned previously, the special case of the normal distribution is the standard normal where the mean is 0 and standard deviation is 1. This is also called the *Z distribution*. Can you create your own set of plots, with the mean equal to zero and the standard deviation equal to one? Add lines for the 0.025, 0.05, 0.50, 0.95, and 0.975 percentiles and the corresponding quantile, probability density, and cumulative density. Have you seen these values before? Find a standard normal Z-table in a statistics textbook or online and think about how they are related.

C. In the R script, section #4 just produces all of the diagnostic plots that you are seen so far for a synthetic distribution that you generated with the following code:

```
mu_r <- 5 ; stdev_r <- 2; n <-200
set.seed (10)
s_data <- rnorm(n, mean=mu_r, sd=stdev_r)
```

The call to `rnorm` generates a vector of `n` elements that will have a mean and standard deviation as specified. Note that every call to `rnorm` will produce a different vector unless we set the seed for the random number generator. The call to `set.seed` sets the seed for the random number generator. The `rnorm` will return the same sequence when called in this session. You know these plots represent data that are normally distributed (since `s_data` is randomly generated to be normally distributed). The bottom plots include the KDE plotted over with the theoretical distribution that our sample data was drawn from. Inspect all of these plots so that you get a sense of what a “real” normal distribution looks like.

- a. First, try a few different sizes of sample draws, small to large such as `n` (10, 100, 1000, 10000) without changing the mean and standard deviation. Before running the code, how do you expect the plots to change?
- b. Comment out the `set.seed` command and rerun the code for the same `n`, mean, and standard deviation. After creating 5 or so plots, play them back with the arrow button in R Studio and watch the data change. Now, do the same thing but include the execution of the `set.seed` command. Describe what you see.

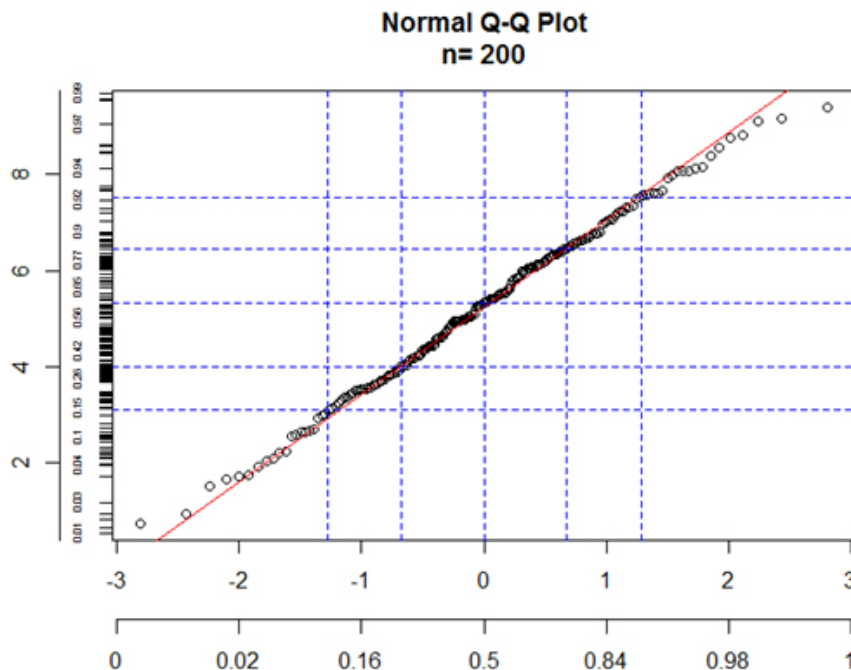


**Figure 55**

(top left to right) Scatter plot of random numbers generated for sample n=200, histogram of same random number generated sample with KDE overlaying histogram, boxplot of same data, and empirical cumulative distribution function of the sample.

(bottom left to right) Theoretical distribution overlain by KDE, and theoretical cumulative distribution function overlain by empirical distribution function.

- D.** The q-q plot is simple in concept and easy to interpret once you know the “rules” but perhaps a little confusing to construct. The R script in section #5 walks you through the construction of the plot. Spend some time working through the code line by line and comparing the output. R’s built in functions `qqnorm` and `qqline` do all the work for you but it may be more informative to look at the “by hand” code. Open up the accompanying Excel spreadsheet and look at the construction. Match the Excel functions to the R functions. Confirm in Section 4 that the computed quantiles match between Excel and R.



**Figure 56** Quantile-Quantile plot of the random number generated sample n=200

**E. APPLICATION**

You have been asked if the average speed observed in male and female cyclists is normally distributed and if so, what are the parameters that best describe this distribution from the class database. To do this, write a query to select all data from the bicycle performance data frame. Graphically diagnose whether the male and female observed average speed data are normally distributed. Show the empirical data in a histogram, KDE, and fit an approximate normal to these data, as well as show the empirical cumulative density function and corresponding normal cumulative density function. Of course, include the q-q plot. For each variable prepare a short discussion showing these figures. Annotate the figure with your observations.

**DELIVERABLE**

Prepare a short write up of your discovery. Submit a PDF copy to the class dropbox.

**ASSESSMENT**

This activity is a short response activity. Your score will be based on the quality and depth of discussion. The response expected differs by question as described in the following rubric:

**Activity 30 Grading Rubric**

	Excellent (10)	Good (8)	Poor (6)	NONE
Discussion	Insightful discussion or commentary relating to the question at hand demonstrating student understanding of the task.	Discussion was competent regarding the lessons, but lacked in insightful discussion.	The discussion did not address the lessons or applicability of the activity.	Did not submit
Graphic(s)	Graphic correctly demonstrated the distribution of the data set.	Graphic did not accurately demonstrated the distribution of the data.	No graphics were present in the discussion.	Did not submit
Quality	Document is typed, formatted, contains appropriate grammar and language.	Document has minor grammatical errors or inappropriate language.	Document was unorganized, contained inappropriate language and/or grammatical errors.	Did not submit

