How Normal is a Distribution?

A large part of statistical analysis is based on the properties of normally distributed random variables. There’s a famous (infamous?) LAW that states that data coming from a large number of independent experiments will produce a Normal Distribution and as such, a lot of statistical models assume ‘Normality’.

In this exercise, we will take what we know about normal distributions and compare the theoretical distribution to some real data.

Begin an R session. First, lets clear out any old junk that’s lying around in the workspace. (Warning, this command will delete any work you have done!)

> rm(list=ls()) ## Removes (rm) all variables (good for saving space)

Next, lets load a data set containing information on the air quality in New York.

> data(airquality)
> attach(airquality) ## To allow us to access the named variables by name.

The data set contains measurements of

> names(airquality)
[1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"

Ozone is the primary ingredient in ‘smog’; the more ozone in the air, the worse the air quality. Too much ozone is dangerous to one’s health and poses real hazards for the elderly and those with lung ailments. A ‘bad air’ day is one with high ozone concentrations.

Preliminary Question: Take a look at histograms of the three main measurements, Ozone, Wind and Temp. From the histograms, which could best be described by a normal distribution? Why?
> sum(zwind > -1 & zwind < 1)  # Number of data points in 1 sd of mean
[1] 100
> sum(zwind > -1 & zwind < 1)/length(zwind)  # Percentage of data in 1 sd
[1] 0.6535948

Pretty close. Using the table of areas for the Standard Normal Distribution, we find that 86.64% of Normally Distributed data lies within ±1.5 standard deviations of the mean. Check this for the Wind data:

> sum(zwind > -1.5 & zwind < 1.5)/length(zwind)
[1] 0.869281

Again, the Normal Distribution prediction is quite close. Is this true for other values of zwind? Is this true for the Temperature and the Ozone Distributions as well? Answers the questions below and hand in your answers.

Questions:

1. Compare the Wind data distribution to the Standard Normal curve for 5 values of $z$. i.e. Compare the percentages in the Table in the book (for the standard normal distribution) to the percentages found using R.

2. Compute the following Probabilities using (a) the Wind Data and (b) the Normal Distribution.
   (a) Prob(Wind > 10 mph)
   (b) Prob(Wind > 15 mph)
   (c) Prob(Wind > 20 mph)
   (d) Prob(Wind < 5 mph)

3. Rescale the Ozone and Temperature data to get zscores. Re-do the comparisons in Question 1 above for these distributions. Which of the three is the 'most normal'? Why? Give numbers to support your conclusion. Which of the three is 'least' normal.

4. A (lazy) statistician decides to build a model for Ozone concentration in New York based on a normal distribution. Assume ozone concentration is normally distributed with mean and standard deviation given by the data in Ozone. What does this Normal model predict for the chances of finding an Ozone Concentration (1) Greater than 70, (2) Greater than 110? How do these predictions compare to the actual data? Explain how and why a normal model might underpredict the chances of 'bad air' days.