

Math 214

Applied Statistics

Laboratory Project #5

Due: Wednesday April 11

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 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?

Lets concentrate on the Wind measurements. Since we want to compare these measurements to the Standard Normal Distribution, the first thing we'd like to do is normalize the data. This is easy, all we want is the *z-scores* of the Wind data. In R, define a new measurement:

```
> zwind = (Wind - mean(Wind))/sd(Wind)
> hist(zwind, prob=T)
```

Now we can compare the distribution of *zwind* to the Standard Normal Distribution. First, we know that 68% of Normally Distributed data lies between ± 1 standard deviation of the mean. To see if this is true for the Wind data, we need to compute the percentage of data in *zwind* that lies between ± 1 . In R:

```
> 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?

Questions:

1. Compare the Wind data distribution to the Standard Normal curve for 5 values of z . IE: Compare the numbers in the Table to the percentages found using R.
2. Compute the following Probabilities using (a) the Wind Data and (b) the Normal Distribution.
 - (a) Prob(Wind \geq 10 mph)
 - (b) Prob(Wind \geq 15 mph)
 - (c) Prob(Wind \geq 20 mph)
 - (d) Prob(Wind \leq 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.

Note: To get mean and sd of Ozone, use the following:

```
> mean_Ozone=mean(Ozone,na.rm=T)
> sd_Ozone=sd(Ozone,na.rm=T)
```

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.

qqnorm

Of course, R already has a table of the Standard Normal Distribution stored someplace in its vast memory. An easy way to compare a given distribution to a normal distribution with the same

mean and standard deviation is provided by the graphical command `qqnorm` which produces a quantile-quantile plot. A q-q plot compares the quantiles of one distribution to those of another. If the distributions have similar shapes, the quantiles will align with each other.

Try the following:

```
> qqnorm(zwind)
```

This compares the quantiles of the z scores of Wind to those of a theoretical normal distribution. If the data lies on a straight line, then we can conclude that the Wind measurements are normally distributed. Are they?

Questions:

1. For each variable, Wind, Temp, and Ozone, construct a boxplot and a qq-plot of the variable against a Normal Distribution. From the qqnorm plots, which of the three is most 'normally distributed'? Which is least 'normally distributed'?
2. For Ozone, compare the boxplot and the qqnorm plot. Use the boxplot to explain, clearly, why the qqnorm plot is NOT a straight line.
3. Find some other data in R, either in the existing data sets (type `data()` to see a list of data sets) or from Kitchens' book that is NOT normally distributed. Construct both a boxplot of this data and a qqnorm plot. Use the boxplot to explain why the qqnorm is shaped the way it is.