ENV 710 – Applied Data Analysis
Fall 2013
Lab 2: Descriptive/Exploratory Statistics

The goals of this lab are to become familiar with discrete probability functions, including the binomial and Poisson distributions. The first part of the lab provides some more useful **R** functions that should make your computing experience more enjoyable. The second part of the lab provides some practice with exploratory statistics. The third part of the lab focuses on discrete probability functions.

At the end of each section, there are a few questions to answer. Please type your answers to each problem, including any requested graphs, in a Word document. *Submit your answers and your R-code to the class Sakai site under the folder Assignments before 5 pm on either Mon., Sep. 8 (Section 07) or Wed., Sep., 10 (Sections 01 and 02).*

More functions in R

In this lab, we introduce a few new **R** commands.

`ls()` - Provides a list of all the objects in the workspace
`rm()` - Removes individual objects from the workspace
`attach()` - Attaches a database to the R search path, making it possible to refer to a variable in the data.frame by their names alone
`detach()` - Removes databases, usually a data.frame that has been attached with `attach()` or a package attached by `library()` or `require()`
`is.na()` - Logical function that identifies all NA's within a vector, returning either TRUE or FALSE for each value
`dbinom()` - Returns the height of the probability density function of the binomial distribution
`pbinom()` - Returns the cumulative density function of the binomial distribution; given a number, it computes the probability that a random number from a binomial distribution will be less than that number
`rbinom()` - Generates a random number from the binomial distribution defined by the probability (`prob`) of "successes" in a specified number (`size`) of trials
`dpois(), ppois(), rpois()` - Same functions as above, but for the Poisson distribution defined by a single parameter, lambda, which specifies the mean and variance of the distribution

Download the database AfrPlots.csv from Sakai and attach it to your workspace.

```
afdat <- read.csv("AfrPlots.csv", header = T)
attach(afdat)
```

The database consists of tree plot data from Africa. In 30 1-ha plots, all the trees ≥ 10 cm dbh (diameter-at-breast height) were measured, and from this data the biomass, basal area, and other statistics were calculated for each plot. Note also that there were two census periods: the initial census and then a census 4 years later.

Because afdat is attached, you can call up the different columns of data without specifying the dataframe (e.g. `ChaveMoist`, rather than `afdat$ChaveMoist`).

Create a couple of variables (e.g., `var1 <- c(3, 5, 7)`) and try removing them individually using rm(). You can remove all the objects in your workspace by leaving the list blank: `rm(list=ls())`. Be careful! That last bit of code will remove *all* the variables and objects in your workspace. Removing objects is useful when you multiple defined variables and worry about confusing them.

Another bit of code that will be useful for this lab is the `is.na()` function. Oftentimes you may want to remove NA's from a vector. Some functions, like the `mean()` function below have options to remove NA's from the operation.

```
vec <- c(0, 4, NA, 2, NA, 7)
is.na(vec)

mean(vec)
mean(vec, na.rm = T)
```

But you may want to remove NA's from a vector before completing any other operations. This line of code basically reads: define `vec` as all the values of `vec` where `vec` is not `NA`. The exclamation mark mean "not".

```
vec[!is.na(vec)]
```

Exploratory data analysis in R

Skewness and kurtosis

We have talked extensively about measures of location and spread like the mean and standard deviation. Skewness and kurtosis are measures of shape. A histogram can give you the general idea of the shape of a distribution of data, but skewness and kurtosis give more precise, numerical evaluations. Why do we care? Many classical statistical tests and intervals depend on assumptions of a normal distribution (a bell-shaped curve), with skewness and *excess* kurtosis of 0. Significant skewness and kurtosis indicate that data are not normally distributed. Let's examine the skewness and kurtosis of the distribution of numbers of trees (*Trees*) in the 30 field plots.

Let's take a look at the distribution of numbers of trees in the plots.

```
hist(Trees, breaks=10, col="tomato")
boxplot(Trees)
```

Don't like the choice of colors - tomato? See the *Colors in R* document under R Shortcuts on Sakai that gives a list of all the possible colors.

Skewness is a measure of the extent to which a probability distribution of a random variable leans to one side of the mean. If a distribution has a negative skew (left-skewed or left-tailed), the left tail is longer and the mass of the distribution is concentrated to the right. As a rule, the mean of the data is less than the median for a left-skewed distribution. For positive skew or right-skewed distribution, it would be just the opposite. Skewness of 0 means the distribution is symmetrical around the mean and the mean is equal to the median.

Following Gotelli and Ellison, skewness can be calculated as:

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(Y_i - \bar{Y})^3$$

where Y_i is the *i*th data point, $\bar{Y}$ is the mean, *s* is the standard deviation, and *n* is the number of data points.

This can be written as a function in **R**. A function is a part of a computer program that performs some specific action (like calculating skewness).

```
myskewness <-  function(y) {
       n <- length(y)
        skew <- 1/(n*sd(y)^3)*sum((y-mean(y))^3)
        skew
  }
```

Note the features of a function include the command `function()` that defines the object as a function. The function is then specified within the `{}` brackets and must correctly perform a task and return any results, which may be stored in other objects like vectors or data.frames.

To call this function to find the skewness of the variable `Trees`, type:

```
myskewness(Trees)
```

What is the skewness of the variable `BasalArea`?

Kurtosis is a measure of the flatness or peakedness of the probability distribution of a random variable. Distributions with negative kurtosis are called platykurtic, whereas distributions with positive kurtosis are called leptokurtic. See the following general rules.

- Kurtosis = 3 – a normal distribution has kurtosis of exactly 3 (excess kurtosis = 0). Any distribution with kurtosis ≈ 3 is called mesokurtic.
- Kurtosis > 3 (excess kurtosis > 0) - leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker

tails. This means high probability for extreme values.
- Kurtosis < 3 (excess kurtosis < 0)- platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.

Following Gotelli and Ellison, excess kurtosis can be calculated as:

$$g_2 = \left[\frac{1}{ns^4}\sum_{i=1}^{n}(Y_i - \bar{Y})^4\right] - 3$$

with the variables having the same meanings as above. The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, *excess* kurtosis is often calculated as here by subtracting 3 from the measurement.

In future labs we will estimate whether skewness or kurtosis is "significantly" different from a normal distribution – that is, they are too high to be explained as random variation from a normal distribution.

Measures of location and spread

We have talked about measures of location and spread in lecture. There are several measures of Central Tendency, including the median, mean, trimmed mean, harmonic mean, and geometric mean. Similarly, there are several measures of spread, including variation, standard deviation, and coefficient of variation.

The variance of a sample of data can be calculated as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The standard deviation is the square root of the variance. And, the coefficient of variation is the ratio of the standard deviation to the mean, reported as a percentage.

Problem #1: Summarize the data for mean growth (`MeanGr`) of the plots. Mean growth is the difference in average DBH for each plot between the first tree census and the second census; thus, there are only data for `MeanGr` during the second census. Include the following in your analysis:

(a) Plot a histogram and boxplot of `MeanGr`.
(b) Write a function for kurtosis (using the equation above) and calculate the skewness and kurtosis of the distribution of mean growth data.
(c) Write a function for variance in **R** and compare it to the `var()` command.
(d) Report the mean, median, variance, standard deviation and coefficient of variation of the data.

Discrete probability in R

To explore discrete probabilities distributions in **R**, we first need to generate some random numbers. It is, in fact, almost impossible to generate random numbers on a computer, so we actually use a pseudo-random sample. In order to reproduce the results, we need to set the seed for our pseudo-random number generator to a value (let's use 1001) using the `set.seed()` function. This sets the number generator to always start at 1001 so that you will get the reproducible results. For example, let's randomly choose a number from 1 to 10.

```
set.seed(1001)
sample(x = 1:10, size = 1)
```

If you run the two lines of code again, you should get the same number. But if you change the seed, or if you don't set the seed before sampling, you will get a different result.

Binomial distribution

Remember from lecture that a Bernouilli trial is an experiment with two possible outcomes, like flipping a coin or checking whether a species is present or absent, etc. The outcome is referred to as a success or failure. Most often in environmental sciences, we are interested in what happens over a sequence of Bernouilli trials. For our random variable *X*, we will count the number of successes: the number of times out of 10 that we flip "heads". The random variable *X* is a binomial random variable.

To simulate a Bernouilli trial, we can write:

```
set.seed(1001)
dbinom(1, size=1, prob=0.5)
```

Here `dbinom` is the density of the binomial distribution. This should become clear as we go along. Because the values of the distribution are discrete, the probability density function will not be continuous like a normal distribution, so we visualize it with a sequence of spikes. Let's represent a Bernouilli distribution where the probability of a 0 is 0.3 and the probability of getting a 1 is 0.7:

```
x <- c(0,1)
y <- c(0.3, 0.7)
plot(x, y, type = "h", las = 1, xlim = c(-1, 2),
      ylim = c(0, 1), lwd = 2, col = "darkblue",
       ylab = "p")
points(x, y, pch = 16, cex = 2, col = "red")
```

Now, let's toss a "fair" coin ten times. *What is the probability of obtaining 3 or fewer heads?* First, let's visualize the distribution. Here `n` is the number of flips; `p` is the probability of getting a heads, `x` is all the possible outcomes, and `pr` determines the probability of getting 0 to 10 heads out of 10 flips of a coin.

```
n = 10
p = 1/2
```

```
x = 0:10
pr = dbinom(x, size = n, prob = p)
plot(x, pr, type="h",xlim = c(-1, 11), ylim = c(0, 0.5),
      las = 1, lwd = 2, col = "blue",
       ylab = "Probability", xlab = "Number of heads")
points(x, pr, pch = 16, cex = 2, col = "dark red")
```

One approach is to realize that the probability of obtaining 3 or fewer heads is the sum of the individual probabilities: add up the probabilities of obtaining 0, 1, 2, or 3, heads.

```
sum(pr[1:4])
```

which is the same as

```
sum(dbinom(0:3,size = 10,prob = 0.5))
```

So what is the dbinom() function doing? It is calculating the binomial probability, using the equation, $P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X}$.

Let's write this equation in R:

```
p  <- 0.5
probX <- (factorial(n)/(factorial(x)*factorial(n-
x)))*p^x*(1-p)^(n-x)
```

Note that we used the same variables as above, so `n` and `x` are the same as above.

```
sum(probX[1:4])
```

How does this compare to the probability found using `dbinom()`?

Problem #2: Say you role a fair coin 20 times. What is the probability of obtaining 5 or fewer heads or more than 5 heads? Show how you would answer this question using the `dbinom` function and calculating it with the binomial equation.

Problem #3: Suppose there are 20 multiple-choice questions on a Stats quiz. Each question has four possible answers, only one of which is correct. Find the probability of answering 17 or more answers correctly if you attempt to answer them at random.

Poisson probabilities

Like the binomial distribution, the Poisson distribution is a discrete probability function. However, instead of modeling successes and failures, it models *counts* of outcomes. If lambda, λ, is the mean occurrence of a count, then the probability of having x occurrences is given as: $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$, where x = 0,1, 2, 3, …

Instead of using `dbinom()` to find the probability of an occurrence, we replace it

with `dpois()` for the Poisson distribution. For example, if the mean number of lightening strikes on top of Mt. Baldy is 3 per year, then the probability of the mountain only being struck once in 2014 is found by:

```
dpois(x = 1, lambda = 3)
```

Problem #4: If there are 4 butterflies feeding at a flower per hour on average, find the probability of having 9 or more butterflies feeding at the flower in a particular hour. Although the possible maximum count of butterflies *could* be much greater, let's limit the maximum number of butterflies to be 13. Do this by (1) writing out the Poisson formula in **R**, and (2) by using `dpois()`. Please also graph the probability distribution.

Problem #5: Suppose there are on average 20 species of birds in a forest. You are asked to sample the forest to evaluate the distribution of species richness, which you do by conducting 25 survey counts. Find the probability of counting exactly four species in a point count assuming that species richness is distributed randomly. What is the probability of counting 15 or more species? Please create a graph of the probability distribution for this example. (In the graph you can limit the maximum number of species observed to 40 (i.e. your x-axis limit is 40).