Lab 3: Distributions of Random Variables

In this lab we'll investigate the probability distribution that is most central to statistics: the normal distribution. If we are confident that our data are nearly normal, that opens the door to many powerful statistical methods. Here we'll use the graphical tools of R to assess the normality of our data and also learn how to generate random numbers from a normal distribution.

The Data

The data in this lab is concerned with heart disease diagnosis and was contributed to the UCI Machine Learning Repository (*http://archive.ics.uci.edu/ml/*) by the Cleveland Clinic Foundation. There are a total of 76 variables in the original dataset, but all publications utilizing this data have used a subset of 14 variables.

Note: The URL needed to import the data using **getURL** is *http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data*

```
require(RCurl)
heart <- getURL("http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.clevel
    ssl.verifypeer = FALSE)
heart <- read.table(text = heart, sep = ",", header = FALSE)
colnames(heart) <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
    "thalach", "exang", "oldpeak", "slope", "ca", "thal", "num")
```

Let's take a quick peek at the first few rows of the data.

```
head(heart)
```

Full descriptions of each variable can be found at *http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names*. Important variables for this lab will be *sex* ($male = 1$ and $female = 0$) and *chol* (cholestrol). The variable *num* adds up the number of major blood vessels with greater than 50% diameter narrowing. Lets look the range of major blood vessels that have $> 50\%$ diameter narrowing.

```
range(heart$num)
```

As you can see, subjects in this dataset have up to 4 major blood vessels that are of concern.

Since males and females tend to have different health patterns, let's look at them separately by subsetting the data into two groups:

```
mheart <- subset(heart, heart$sex == 1)

fheart <- subset(heart, heart$sex == 0)
```

> **Exercise 1** Make a histogram of men's cholestrol and a histogram of women's cholestrol. How would you compare the various aspects of the two distributions? Look at the means of each distribution, do they differ? By how much? On average, do males or females in this population have higher cholestrol?

The normal distribution

In your description of the distributions, did you use words like "bell-shaped" or "normal"? It's tempting to say so when faced with a unimodal symmetric distribution.

To see how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data. We'll be working with men's cholestrol, so let's store them as a separate object and then calculate some statistics that will be referenced later.

```
mcholmean <- mean(mheart$chol)

mcholsd <- sd(mheart$chol)
```

Next we make a density histogram to use as the backdrop and use the **lines** function to overlay a normal probability curve. The difference between a frequency histogram and a density histogram is that while in a frequency histogram the *heights* of the bars add up to the total number of observations, in a density histogram the *areas* of the bars add up to 1. The area of each bar can be calculated as simply (the height) $\times$ (the width of the bar). Using a density histogram allows us to properly overlay a normal distribution curve over the histogram since the curve is a normal probability density function. Frequency and density histograms both display the same exact shape; they only differ in their y-axis. You can verify this by comparing the frequency histogram you constructed earlier and the density histogram created by the commands below.

```
hist(mheart$chol, probability = TRUE)

x <- 126:353

y <- dnorm(x = x, mean = mcholmean, sd = mcholsd)

lines(x = x, y = y, col = "blue")
```

After plotting the density histogram with the first command, we create the x- and y-coordinates for the normal curve. We chose the x range as 126 to 353 in order to span the entire range of mheart$chol. To create y, we use **dnorm** to calculate the density of each of those x-values in a distribution that is normal with mean mcholmean and standard deviation mcholsd. The final command draws a curve on the existing plot (the density histogram) by connecting each of the points specified by x and y. The argument col simply sets the color for the line to be drawn. If we left it out, the line would be drawn in black.

Exercise 2 Based on the this plot, does it appear that the data follow a nearly normal distribution?

Evaluating the normal distribution

Eyeballing the shape of the histogram is one way to determine if the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve. An alternative approach involves constructing a normal probability plot, also called a normal Q-Q plot for "quantile-quantile".

```
qqnorm(mheart$chol)

qqline(mheart$chol)
```

A data set that is nearly normal will result in a probability plot where the points closely follow the line. Any

deviations from normality leads to deviations of these points from the line. The plot for male cholestrol shows points that tend to follow the line but with some errant points towards the tails. We're left with the same problem that we encountered with the histogram above: how close is close enough?

A useful way to address this question is to rephrase it as: what do probability plots look like for data that I *know* came from a normal distribution? We can answer this by simulating data from a normal distribution using **rnorm**.

```
sim_norm <- rnorm(n = length(mheart$chol), mean = mcholmean, sd = mcholsd)
```

The first argument indicates how many numbers you'd like to generate, which we specify to be the same number of cholestrol records in the mheart data set using the **length** function. The last two arguments determine the mean and standard deviation of the normal distribution from which the simulated sample will be generated. We can take a look at the shape of our simulated data set, sim_norm, as well as its normal probability plot.

> **Exercise 3** Make a normal probability plot of sim_norm. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

Even better than comparing the original plot to a single plot generated from a normal distribution is to compare it to many more plots using the following function. Make sure to install the *StMoSim* package before trying to run the following code.

```
require(StMoSim)
qqnormSim(mheart$chol)
```

> **Exercise 4** Does the normal probability plot for mheart$chol look similar to the plots created for the simulated data? That is, do plots provide evidence that the male cholestrol records are nearly normal?

> **Exercise 5** Using the same technique, determine whether or not female cholestrol records appear to come from a normal distribution.

Normal probabilities

Okay, so now you have a slew of tools to judge whether or not a variable is normally distributed. Why should we care?

It turns out that statisticians know a lot about the normal distribution. Once we decide that a random variable is approximately normal, we can answer all sorts of questions about that variable related to probability. Take, for example, the question of, "What is the probability that a randomly chosen adult male has a cholestrol of over 260?"

If we assume that male cholestrol is normally distributed (a very close approximation is also okay), we can find this probability by calculating a Z score and consulting a Z table (also called a normal probability table). In R, this is done in one step with the function **pnorm**.

```
1 - pnorm(q = 260, mean = mcholmean, sd = mcholsd)
```

Note that the function **pnorm** gives the area under the normal curve below a given value, q, with a given mean and standard deviation. Since we're interested in the probability that someone has a cholestrol reading greater than 260, we have to take one minus that probability.

Assuming a normal distribution has allowed us to calculate a theoretical probability. If we want to calculate the probability empirically, we simply need to determine how many observations fall above 260 then divide this number by the total sample size.

```
sum(mheart$chol > 260)/length(mheart$chol)
```

Although the probabilities are not exactly the same, they are reasonably close. The closer that your distribution is to being normal, the more accurate the theoretical probabilities will be.

> **Exercise 6** Write out two probability questions that you would like to answer; one regarding male cholestrol and one regarding female cholestrol. Calculate the those probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which variable, male cholestrol or female cholestrol, had a closer agreement between the two methods?

On Your Own

1. Subset the data into two groups: Individuals with 1 or more blood vessels with $> 50\%$ diameter narrowing, and individuals with 0. Look at the mean cholestrol levels for each group. Is there a difference? Which group on average has higher cholestrol? Does this suprise you?
2. What is the range of *num*? What is the maxmum number of blood vessels with $> 50\%$ diameter narrowing? What is the mean cholestrol level for this group? Compare to the groups above.
3. You now should have three subsets of the data based on the variable *num*. Using the methods you learned above, determine which (if any) are approximately normally distributed.
4. What happens to the histograms and normal probability plots as the sample size gets smaller? Is it easier or more difficult to judge if the data is normally distributed with a smaller sample size or a larger one?
5. What is the probability that a randomly chosen adult with more than 2 blood vessels with $> 50\%$ diameter narrowing has a cholestrol of above 300? What about for an adult with 4 narrowing blood vessels?
6. Find the above probabilites empirically. Which probabilites are in closer agreement? What does this tell you?