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 know or assume that our data are normally distributed, 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

This week we'll be working with measurements of body dimensions.

```
download.file("http://www.openintro.org/stat/data/bdims.RData",
destfile = "bdims.RDat")
load("bdims.RData")
```

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

```
head(bdims)
```

You'll see that for every observation we have 25 measurements, many of which are either diameters or girths. A key to the variable names can be found online[1], but we'll be focusing on just three columns to get started: weight in kg (`wgt`), height in cm (`hgt`), and sex(1 indicates male, 0 indicates female).

This dataset contains observations on 507 healthy young men and women. Since genders tend to have different body dimensions, it will be handy to create two additional datasets - one with only men and another with only women.

```
mdims <- subset(bdims, bdims$sex == 1)
fdims <- subset(bdims, bdims$sex == 0)
```

> **Exercise 1** Make a histogram of men's height and histogram of women's height. How would you describe their distributions?

The normal distribution

In your description of the distribution 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 if the data follow a normal distribution. This normal curve should have

[1] This dataset comes from a paper entitled *Exploring Relationships in Body Dimensions* by Heinz et. al. For more information on the dataset, visit *http://www.amstat.org/publications/jse/datasets/body.txt*

the same mean and standard deviation as the data. Working with the women's heights, let's calculate those statistics first and store them as `fhgtmean` and `fhgtsd` in order to be able to reference them later.

```
fhgtmean <- mean(fdims$hgt)
fhgtsd <- sd(fdims$hgt)
```

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 relative frequency histogram and a density histogram is that while in a relative frequency histogram the *heights* of the bars add up to 1, 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 × 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, relative frequency, and density histograms all display the same exact shape, they only differ in their y-axis. You can verify this by scrolling between the frequency histogram you constructed earlier and the density histogram you just plotted.

```
hist(fdims$hgt, probability = TRUE)
lines(x = 140:190, y = dnorm(x = 140:190, mean = fhgtmean, sd = fhgtsd),
    col = "blue")
```

The above code plots a line over an existing plot (the density histogram) with x coordinates between 140 and 190 and y coordinates that follow a normal distribution with mean `fhgtmean` and standard deviation `fhgtsd`. We chose the x range as 140 to 190 in order to span the entire range of `fhgt`. The last argument `col` simply sets the color for the line to be drawn. If we left the last argument out the line would be drawn in black.

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

Evaluating the normal distribution

Eyeballing the shape of the histogram is one way to determine if the data appear to be 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.

```
qqnorm(fdims$hgt)
qqline(fdims$hgt)
```

A dataset that is perfectly normal will result in a probability plot where the points follow the line. Any deviations from normality lead to deviations of these points from the line. The plot for female heights 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: how close are the probability plots of datasets that I *know* came from a normal distribution? We can answer this by simulating data from a normal distribution using `rnorm()`.

```
simnorm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
```

The first argument specifies how many numbers you'd like to generate. The last two arguments determine the mean and standard deviation of the normal distribution from which they'll be generated. We can take a look at the shape of our simulated dataset, `simnorm`, as well as its normal probability plot.

> **Exercise 3** Make a normal probability plot of `simnorm`. Do all of the points fall on the line? How does this plot compare to the probaiblity 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 eight plots, which we can do using the following function (it may be helpful to click the "zoom" button in the plot window).

```
qqnormsim(fdims$hgt)
```

> **Exercise 4** Does the normal probability plot for your data look similar to the plots simulated from the normal distribution? That is, is it likely that your sample of female heights comes from a normal distribution?

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

Normal probabilities

Ok, so now you have a slew of tools to judge whether or not a variable is normally distributed. It's about time you ask the question: why should I 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 young adult female is greater than 6 feet (about 182 cm) tall?"

If we assume that female height is normally distributed, we can find this probability by calculating a z-score and consulting a z-table. In R, this is done in one step with the function `pnorm()`.

```
1 - pnorm(q = 182, mean = fhgtmean, sd = fhgtsd)
```

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 interesting in the probability that someone is greater than 6 feet tall, 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 182 and divide this number by the total sample size.

```
sum(fdims$hgt > 182) / length(fdims$hgt)
```

You'll see that although the probabilities aren't the exact same, they are reasonably similar. This indicates that, at least in this particular part of the distribution, female heights is behaving as if it is normally distributed.

This is the take-home lesson for assessing normality. While you always have the option of assuming a normal distribution, your probabilites only become accurate when the distribution is close to being normal.

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

On Your Own

Create a vector called `math5` representing the scores of 5^{th} graders on the standardized math test.

1. Make a density histogram of the scores of 5^{th} graders on the math test and overlay a normal distribution curve on the histogram. Describe the distribution of the scores of 5^{th} graders on the math test.

2. Make a normal probability plot of these scores and evaluate if the data follow a normal distribution.

3. Determine if the data follow the 68-95-99.7% rule.

4. Calculate the probability that a randomly chosen 5^{th} grader scores between 1,500 and 2,500 on the math test. Calculate this probability empirically and also using a normal distribution with the same mean and standard deviation as the data.

5. Make a normal probability plot of scores from the geometry (`testID == "geometry"`) test as well as a normal probability plot of scores from the science 4 (`testID == "sci 4"`) test. Based on these normal probability plots determine if these distributions are symmetric, right skewed or left skewed. You can use a histogram to confirm your findings.

6. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere, e.g. lecture, discussion section, previous labs, or homework problems? Be specific in your answer.