Introduction to Biostatistics (4120)
Breheny

Lab #8

Lab #8 concentrates largely on the normal distribution. First we'll look at a couple simple examples of how to calculate individual probabilities in the normal setting. Then we'll delve into sampling distributions and the central limit theorem. You will each analyze your own personal data set (which was drawn at random from a larger data set) and enter your results into a spreadsheet on your TA's computer. We will then see sampling distributions in action when we look at the class's results.

If we have time, we will practice simulating random samples from the binomial distribution and see how the central limit theorem applies for it.

1 Normal Probability Exercises

A local farm discovers that the size of the eggs their chickens produce seems to follow a normal distribution with a mean of 2 oz. and a standard deviation of 0.33 oz.

- In the USA, a "jumbo" egg is categorized as an egg greater than 2.5 oz. What is the probability that a randomly chosen egg is jumbo?

- Suppose this farm produces 1000 eggs a week. How many jumbo eggs should they expect to produce each week?

- What is the probability that the farmers produce 100 jumbo eggs in a week?

Last year, the distribution of test scores on a particular exam (not in this class) was approximately normally distributed a mean of 79 and a standard deviation of 11 points.

- If the professor decides to use the 60/70/80/90 rule for grading, what proportion of students should she expect will fail on the exam this year?

- What proportion of students should she expect will get an A?

- If there are 35 students in the class, how many will get either a B or a C?

2 Sampling distributions and the central limit theorem

As part of the NHANES study, the triglyceride levels of 3026 adult women were measured. Triglyceride, the main constituent of both vegetable oil and animal fat, has been linked to atherosclerosis, heart disease, and stroke.

Download the data set `lipids.txt`. One can get a visual idea of whether or not a variable follows a normal distribution by typing:

```
nhanes <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lipids.txt")
attach(nhanes)
hist(TRG,freq=FALSE, breaks =15)

## To draw a normal curve:
x <- seq(0,max(TRG),len=101)
lines(x,dnorm(x,mean(TRG),sd(TRG)))
```

This creates a histogram of the triglyceride levels `trg`, and draws the normal curve which best fits the data over the top (don't worry about the details of drawing the normal curve in `R`). One can see from this picture that triglyceride levels do not follow a normal distribution very closely – they are highly skewed to the right.

Furthermore, using procedures and techniques that we have talked about in earlier labs, one can determine that the mean triglyceride level is 116.9 mg/dl, that the standard deviation is 67.9 mg/dl, and that 12.7% of individuals have triglyceride levels over 200 mg/dl, which the American Heart Association defines as having high levels of triglycerides.

However, suppose that you lack the resources to sample 3026 women, and instead obtain only a smaller sample. Let's simulate this scenario for when we can only get say 10, 20, or 30 people into our study. Each of you can create such a dataset uniquely by running the `sample()` function, which just picks out observations from the data like balls from an urn.

```
smallsample <- sample(TRG, 10)
medsample <- sample(TRG, 20)
largesample <- sample(TRG, 30)
```

Calculate the mean of your triglyceride levels in each of these groups, then enter that information into your TA's spreadsheet. The TA will use everyone's calculations to look at the sampling distribution of these statistics.

While he's working, use your knowledge of the central limit theorem to predict the mean and standard deviation of the class's sample means. What does this distribution look like? When everyone has entered their sample statistics into the spreadsheet, we can calculate the actual values from our class's sampling distribution and compare.

3 Binomial simulation: Making the computer flip coins for you

R has a function called `rbinom` for generating random variables from the binomial distribution – in other words, flipping a coin for you. You can even change the probability of heads to something other than 50%, which doesn't necessarily make sense if you're interested in coins, but makes a lot of sense if you're studying the survival of babies or the success of a therapy. So, for example:

```
> rbinom(1,size=10,prob=.5)
[1] 3
```

R just flipped a coin for me 10 times, and got 3 heads. Some of you (specifically, about 12% of you) will also get 3 heads, but the rest of you will get some other number when you submit the above code. Another example: let's say we're at a hospital in which 80% of babies born at 25 weeks gestation survive.

```
> rbinom(1,size=50,prob=.8)
[1] 42
```

R has just simulated the birth of 50 premature babies at our hospital – in my hospital, 42 survived. In most of your hospitals (specifically, about 69%), your babies won't do as well as the ones in my hospital did.

4 Repeating the experiment

As you may have guessed, R can do this over and over again – this is what that first number in `rbinom` is for. So, for example:

```
> rbinom(100,10,.5)
  [1] 4 5 3 5 4 5 3 4 7 7 3 6 6 7 7 2 5 3 7 4 5 4 3 6 5 8 5 4 8 5 7 3 6 6 8 6 5
 [38] 6 5 7 6 5 5 3 5 6 5 8 5 4 5 6 6 4 6 6 4 4 6 4 4 4 4 6 3 5 9 6 5 5 5 8 5 7
 [75] 6 4 3 6 4 6 5 5 7 7 6 4 4 3 9 6 7 5 7 3 6 6 6 4 6 3
```

As the binomial distribution would predict, we get lots of 4s, 5s, and 6s – very few 8s, 9s, and 2s. We can save the results of our random experiments:

```
> counts <- rbinom(100,10,.5)
> avgs <- counts/10
```

As we see, this gives us the ability to calculate things like the average number of heads in each "experiment" (each time we flipped a coin 10 times). Try typing `counts` and `avgs` to see what they look like. We can see the results that we alluded to during lecture:

```
> mean(counts)
[1] 5.1
> mean(avgs)
[1] 0.51
```

The average number of heads was right around $np = 10(.5) = 5$, while the average of the averages was right around $p = .5$. Let's look at standard deviation and standard error:

```
> sd(counts)
[1] 1.283146
> sd(avgs)
[1] 0.1283146
> sqrt(10*.5*.5)
[1] 1.581139
> sqrt(.5*.5/10)
[1] 0.1581139
```

Of course, your actual numbers for the first two will vary because they depend on random flips of the coin. Some of you will have numbers will be smaller than $\sqrt{np(1-p)}$ or $\sqrt{p(1-p)/n}$, for others, your numbers will be bigger. But they should be fairly close to what the binomial distribution predicts. Finally, let's look at a histogram of our results:

```
> hist(counts)
> hist(avgs)
```

My histogram looked vaguely normal, although there were certainly differences.

We can now see how the central limit theorem applies to the binomial distribution in a very hands-on way. For each of the following, what would you expect to happen? What happens when you try it out?

- As you increase the number of experiments (the number of times that you flip 10 coins), what happens to the mean of the counts? The standard error? The histogram?
- As you increase the number of times you flip the coin in each experiment, what happens to the mean of the counts? The standard error? The histogram?
- As you increase p to, say, .9, what happens to the averages? To the standard error? To the histogram?