Introduction and objectives

Course objectives: first we want to take a look at what the objectives are in this course.

1) To become literate in the use, applications, and mis-applications of statistics.

For example, when are statistics appropriate?

Example 1: Determining the effect of a medication. We give two groups of people medication for blood pressure. One group gets a placebo, the other the actual medication.

We notice a drop in average blood pressure in the group getting the medication:

	placebo	medicine
	165	153
	136	141
	143	131
	155	160
average:	**149.5**	**146.25**

Is the observed drop in blood pressure caused by the medication or by *random* events?

Statistics lets us answer this question using mathematics. It does this by calculating the *probability* of getting the observed result. If the probability is really low, then we say the result is *not* due to random events. Notice that we'll have to learn something about probability to answer this question.

Example 2: Figuring out if men and women have different heights. We know the answer to this one, but notice that many women are taller than individual men. So what do we mean when we say men are taller than women?

We use statistics to tell us how to evaluate this and figure out exactly what is meant by men being taller than women. Incidentally, notice that for many other organisms, females are bigger than males.

Example 3: Determining how an ectothermic (cold-blooded) animal (e.g. a snake or lizard) responds to heat. What happens if we increase the temperature in a cage with a snake? Being ectothermic, a snake will become more active if the temperature is increased. But by how much? For instance, if we increase the temperature by 1°C, how much more active does the snake become?

Again, we can evaluate this using statistics.

Example 4: Some non-biological examples (statistics are used in *many* non-biological applications):

Who will win the next presidential election?

Will more people buy brand x or brand y?

What will the effect be of a 0.25% increase in the interest rate?

And many, many more.

When are statistics mis-applied (mis-used)?

Unfortunately statistics are used to lie and mislead all the time.

Example 1: *9 out of 10 doctors...* . Many statements like this are based on doctors who work for or are paid by drug companies (yes!). The result is not unbiased.

Example 2: *Biased questions:*

Note the leading nature of the following questions (based on a real survey!):

Knowing that forests are being destroyed at record rates by destructive and unnecessary human activities, do you support a moratorium on logging?

This is then followed by a poll that shows most Americans support forest preservation. Even causes that we might like are guilty of *lying* with statistics.

Example 3: *Politics (admittedly non-biological)*

For particularly egregious examples of lying with statistics take a look at some of the political web pages (Democratic *or* Republican - both are guilty).

Many years ago (1954), Huff wrote a book entitled *How to lie with Statistics.* It's sad that most of the techniques in his book are still used every day. The book is actually quite good, but the drawings are very dated.

2) To be able to use and apply some basic statistical concepts.

This is actually most of the course. We will learn many basic statistical techniques (confidence intervals, t-tests, contingency tables, regression, etc.) that will be useful for a career in biology, medicine or related fields. We can't possible learn everything (see below), but we should get a good basic knowledge of statistics.

3) To ask meaningful questions of a statistician and learn your limitations.

Ask meaningful questions of a statistician

For example, statisticians hate to hear the following: ”I have this data that I can’t figure out - can you help me?”

You should know how to ask better questions. Such as: ”I collected a bunch of weight measurements on two populations of rats - can you help me figure out if there is a weight difference in these populations?”

Realize your limitations

Suppose you measure the length of 17 male turtles and 23 female turtles. At the end of the semester, you should be able to figure out if there is a difference in length between male female turtles.

But now suppose you measure both length and height of these turtles. You want to compare male and female turtles using both variables *at the same time*. You will *not* learn how to do this in this class (or at least, you won’t learn a good way).

Other, more specific examples: ANOVA, ANCOVA, multiple regression, survival analysis, or computer simulations are some of the techniques we won’t learn.

So here’s the important thing:

Given that you wont be able to analyze everything, you should be able to figure out when you need to talk to a statistician (and do it in such a way so that you can communicate your ideas effectively).

Some basic definitions and examples: Let’s start by looking at a few basic ideas.

statistic:

A value calculated or derived from the data. Some examples might be the mean, median, standard deviation, or simply one of the data points. Note that this also includes non-numerical values (a person’s eye color or blood type can be a statistic).

noise/error:

In a sense, this is one of the main reasons we need to use statistics. Data are variable. If you give the same medication to two different people, they will most likely react differently (sometimes only a little differently). In other words, there is variation in the things we’re interested in, and we need to sort out what is variation and what is really the result of some *experiment*.

Example 1: Three different people measure the same mouse from nose to tail. Will everyone get exactly the same measurement? Why not? Has the mouse changed size (of course not)? Different people measure things differently - the result is noise.

Example 2: Height in people has been shown to depend on diet. A good diet will let you grow taller. So we give exactly the same diet to two people from birth (probably impossibly, and possibly unethical). If the two people are unrelated, will they grow up to the same height?

Of course not. Genetics has a lot to say about how tall people get. Since we're interested in diet, not genetics, the noise in this example is coming from genetics.

Somehow we need to account for the differences in height caused by genetics (the noise) if we want to examine the effects of diet on height.

Let's do a couple of real examples that illustrate some of the issues with statistics and noise.

Example 3: *Anthrax in livestock.*

In order to convince people that he had developed a vaccine for anthrax, Pasteur performed a famous experiment. He took 25 sheep, one goat and several cows and vaccinated them with his vaccine. After a month, he took this group of animals and another group of similar animals and exposed both groups to anthrax. Let's worry just about the sheep:

	Not vaccinated	Vaccinated
Died	25	0
Lived	0	25
Total	25	25

Is there any noise here? Since there's a pretty good sample size it appears that this is pretty straight forward. You get vaccinated, you live, you don't get vaccinated, you die. Yes, we could do a statistical analysis here, but the outcome is fairly obvious without that.

Example 4: *Malaria and sickle cell anemia*

This example is made up, but is loosely based on actual numbers from a study in Mali in 2012. People who are carriers for sickle cell turn out to be less susceptible to malaria. Here are some data based on infants and how many are malaria free 8 months after exposure:

	Sickle cell carrier	Normal
Infected with malaria	18	30
Not infected with malaria	22	10
Total	40	40

This time it's more difficult to figure out what's happening. Does sickle cell protect from malaria? Just because you're a carrier for sickle cell does not give you 100% immunity. The problem is noise. This time we definitely need to employ statistics to answer the question.

We will learn how to analyze data like these later, but in the meantime we can get a glimpse of how to proceed by calculating the proportions of people with malaria in both groups. For carriers we calculate $\hat{p}_1 = 18/40 = 0.45$ and for normal people (non-carriers), we calculate $\hat{p}_2 = 30/40 = 0.75$.

This at least lets us see that the proportions of infected people really are different. Is it a meaningful (or significant) difference? That's the part that will have to wait until we learn how to do categorical data analysis.