Last lecture

I. Miscellaneous topics (A & B are briefly mentioned in your text, C is not)

A) Multiple regression.

1) No, you won't learn how to do this. But here's what it is:

Suppose you have more than one x. Why would you have more than one x?

Example: you measure plant growth.

First x:	amount of light.
Second x:	amount of fertilizer.
Third x:	temperature.

Do you think you could get better estimates of plant growth if you used all three variables instead of just one?

Let's see what it would look like:

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Here the subscripts for x do not refer to specific values of x, they refer to the first, second or third x variable (i.e., light, fertilizer and temperature). You really need to start using two subscripts as in:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon$$

Now the first subscript refers to the x-variable, and the second to the values for these x's (and Y-hat and ε) for a particular i.

Aren't you glad you don't have to learn how to do this?

2) A couple of comments about multiple regression.

a) You really need to learn/use matrix (or linear) algebra in order to use and understand multiple regression.

b) Sometimes you may have too many x's. In this case you might try to figure out which x's are more important than others. You may have heard of words such as "stepwise", "forward", or "backward". They all refer to different ways of getting rid of excess x's.

c) Multiple regression is a very useful tool, and there's a good chance you'll come across this sometime.

d) Most of the assumptions you learned still apply in more or less the same way. You can even do residual plots, though they're just a bit more complicated since you're dealing with several x's.

e) If you find you need something like this, TALK TO A STATISTICIAN. Whatever you do, DO NOT just plug stuff into Minitab, get an answer and pat yourself on the back thinking you figured out how to do multiple regression.

B) ANCOVA

1) You might be interested to know if the relationship between height and weight is different form men and women. How would you do this?

a) you could measure a bunch of men and women at the same height and see if their weights are different.

you need to "control" for height - if you use different heights, then you don't know if the weight difference is due to sex or height!

b) but note the obvious - using just one height is very restrictive (you'd have to find men and women all the same height. Hopefully you're not surprised to learn that there's a better way:

2) ANCOVA can detect differences between groups when some of the variables are interfering with what want to discover. [Illustrate height/weight/sex example]

You are basically "adjusting" or "controlling" for height so that you can get at the difference in weights between men and women.

(Notice the different y-intercepts)

C) Multivariate designs.

1) What are these? Well, suppose you had two (or several) populations. Now you take a series of measurements on each (sort of like we did for our starlings). How would you analyze the results?

a) You don't do a t-test (or ANOVA if you have several groups) for each variable. Why not? Because doing multiple t-tests louses up your error rate. It's the same reason you do an ANOVA instead of multiple t-tests when you have more than two groups. There are also much more powerful ways of doing this.

b) Example: (illustrate on board)

c) A sample equation from multivariate statistics:

To reject a null hypothesis of two equal mean vectors, a test statistic, T-square (also known as Hotteling's T) is used as follows:

Reject H_0 if:

$$T^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \left[\left(\frac{\mathbf{1}}{\boldsymbol{n}_1} + \frac{\mathbf{1}}{\boldsymbol{n}_2} \right) \boldsymbol{S}_{pooled} \right]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) > \boldsymbol{c}^2$$

note that all the stuff in bold (except n_1 and n_2 represents vectors or matrices. In particular, x_1 is a vector containing the averages for the first population, x_2 for the second population, and S represents a variance-covariance matrix.

This is why we don't cover multivariate stuff in an introductory class.

2) But hopefully you can see that these are quite powerful. If you go on in biology you may wind up having to use techniques like this sometime.

3) Multivariate techniques include:

i) multivariate correlation: getting correlations between "groups" of variables.

ii) multivariate ANOVA's and t-tests: our example above.

iii) multivariate regression: you have several y's that you're trying to predict (think of having a multivariate and multiple regression!)

iv) classification techniques: measuring a bunch of variables on a number of different specimens to see if you can then classify these correctly. Very useful in taxonomy. In more advanced versions also useful in military applications (what makes an enemy tank an enemy tank?)

v) principal components: you measure 25 variables on two sets of head lice (e.g., body length, bristle length, body width, eye diameter, etc. etc.). Isn't 25 variables a bit much? Principal components let's you reduce the number of variables without losing too much information.

vi) numerous others.

D) There are many, many other techniques which we can't even talk about.