Last lecture

I. Miscellaneous topics (some, but not all, of these are briefly mentioned in your text)

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:

\[ \hat{y}_{i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \]

Now the first subscript refers to the x-variable, and the second to the values for these x’s (and Y-hat and \( \epsilon \)) 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 some statistical software, get an answer and pat yourself on the back thinking you figured out how to do multiple regression.

B) Logistic regression.

1) This is also briefly discussed in your text (at least in the 3rd edition).

   - suppose you have a discrete (or binomial) dependent variable.

   - For example (using the example in your text), suppose you are trying to predict the probability that a esophageal cancer has spread to the lymph nodes:

      - your response (y) is either 0 (not spread) or 1 (spread)

      - your predictor (x) is the size of the tumor.

   - Problem - how do you do this?? You have a continuous variable that can take on a (presumably) infinite number of values, and a discrete variable that can only take on two values (so, in this case, is binomial).

   - You need to take the continuous variable and “map” it into [0,1] (where 0 is dead, and 1 is living).

   - You use something called a “link function”:

     \[
     \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}
     \]

     - here \( \pi(x) \) is the probability of y being “successful”, or as your book puts it, \( \Pr\{Y = 1\} \).

     - \( \pi(x) \) can go from 0 to 1 (just like a probability)

     - we could also reverse this and write it in terms of the regression equation (or in terms of \( \pi(x) \)):

       \[
       g(x) = \ln\left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x
       \]
- you need to “estimate” \( b_0 \) and \( b_1 \) using fairly advanced math (though a lot of software can do this). There are no simple formulas like in linear regression.

- Now you can plot this (see [fig. 12.41, p. 585], [fig. 12.8.10, p. 541]) \{12.8.7, p. 572\}. This curve gives you the probability that the cancer has spread, based on the size of the tumor.

- This can also be used to “classify” things:

  - e.g., if find the \( x \) for which the probability is more than 1/2

  - if you had to pick, then anything with an \( x \)-value bigger than this would be classified as “spread” and anything with a \( x \)-value smaller than this would be classified as “not spread”

  - there are ways to make this classification more accurate (we'll discuss this again below).

- One can also do “multiple” logistic regression.

C) 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)

D) 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:

$$T^2 = (\bar{x}_1 - \bar{x}_2)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\bar{x}_1 - \bar{x}_2) > c^2$$

Note that all the stuff in bold represents vectors or matrices. In particular, $\bar{x}_1$ is a vector containing the averages for the first population, $\bar{x}_2$ for the second population, and $S$ represents a variance-covariance matrix.

This is why we don’t cover multivariate stuff in a basic statistics course.

2) But hopefully you can see that these are quite powerful. You will quite probably come across 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?). Some of this (e.g., discriminant analysis) can be used in a similar way to logistic regression.
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.

vii) factor analysis. This is a favorite technique of psychologists because you can do almost anything you want with it. You take the original variables, transform them into something you think you like, and then do statistics on these transformed variables (note that the transformations I'm talking about are NOT the same as transformations on a single variable to make it “normal”, or to control for variance). A lot of statisticians (except those in psychology) don't like factor analysis.

vi) numerous others.

E) Computer techniques:

There's a huge “subdiscipline” in statistics known as computational statistics. It uses computer intensive techniques to discover statistical relationships.

Here's an example:

1) Bootstrapping.

Suppose you calculate the mean of a small sample. You want to know something about the distribution of your mean.

- If you have a large sample, you can assume that the distribution of the sample mean is approximately normal.

- But it's a small sample, so what do you do?

- Resample. The basic idea is this:

  Suppose you have the following data:

  12, 54, 76, 2, 6

  You “pretend” this is your population, and take a sample of five (usually) from this sample. You might get:

  54, 76, 2, 2, 54 (note the duplicate 2's and 54's)

  Now you calculate the mean for this “sub-sample”

  Do it again. In fact, do it many, many times, each time calculating the mean of your “sub-sample”. 
Once you have 100 or so means, you can plot these and get a histogram or distribution - this “bootstrap” distribution will estimate the actual distribution of your sample mean.

- You can also use this technique to get confidence intervals, calculate variances, and other things.

2) Other computer intensive techniques include:

- Jackknifing: leaving out part of your data to get a better idea of what is really going on (think about it - your data describes YOUR sample, not another sample; jackknifing tries to correct this bias)

- Monte Carlo methods: using simulation via random numbers to solve problems not otherwise solve-able.

  - for example, determining the significance level of a test statistic if we don't know the distribution.

  - (If you're interested, it's done (approximately) by getting a large number of values for our test statistic from the data, and then calculating how large it needs to be to include, say, 95% of our total number of “test statistics”).

- both bootstrapping and Jackknifing could be considered special cases of Monte Carlo methods.

- also used for many, many disciplines other than statistics:

  - simulating which path salmon will likely take.

  - simulating the value of π (a bit silly to do it this way, but quite possible).

  - used in engineering, physics, biology, chemistry, economics, etc.

- Many, many others.

F) There are many, many other techniques which we can’t even talk about. Some of you have come across some of these (things like the Cox hazard model, or the Mantel-Haenszel test, risk analysis, survival analysis, etc.)