Last statistics lecture

I. Miscellaneous topics

A) Categorical data analysis (see chapters 22 and 23)

A large group of tests that we did not discuss (no one needed to use this).

You've probably heard of the χ^2 goodness of fit test.

It's often used to figure out if the distribution of proportions you're interested in follows some specified hypothesis.

For example, you observe 240 brown mice and 97 black mice - do these follow a 3:1 Mendelian ratio?

Note that this may be one of the few instances where you're sometimes more interested in H_0 than H_1:

H_0: mice follow 3:1 ratio $\qquad$ H_1: mice do not follow 3:1 ratio

The math is very easy:

$$\chi^{2*} = \frac{\sum_{i=1}^{c} (O_i - E_i)^2}{E_i}$$

Then compare this to χ^2_{table} with c-1 d.f. (c = number of categories).

Another large group of test includes contingency tables. Again, you usually use the χ^2 distribution:

Safety equipment in use	Injury		
	Fatal	Non-fatal	Total
None	1,601	165,527	167,128
Seat belt	510	412,368	412,878

And now you analyze this data to see if seatbelts save lives.

You calculate χ^{2*} the same way as above but now c = number of cells, and

d.f. = (number of rows - 1)(number of columns - 1)

You calculate expected values by doing

$$\text{Expected value} = \frac{(\text{Row total}) \times (\text{Column total})}{(\text{Grand total})}$$

Contingency tables are superficially very simple, but are used by a lot of people even for rather advanced studies.

From contingency tables one can quickly get into odds ratios and relative risk.

For example, the relative risk for the above table

The proportion of fatalities wearing seat belts is:

$$510/412{,}878 = \quad \hat{p}_2 \quad = .001235$$

The proportion of fatalities for not wearing a seat belt is:

$$1{,}601/165{,}128 = \quad \hat{p}_1 \quad = .00958$$

And relative risk is .009672/.001235 = 7.83.

In other words, the risk of dying is 7.8 times higher if one is not wearing a seatbelt than if one is wearing a seatbelt.

This makes it sound pretty good for wearing a seatbelt!

Odds ratios are similar, and more versatile then Relative Risk (the Odds Ratio for the seatbelt example is 7.822, close but not quite the same).

Odds ratios and relative risk are used a *lot*, particularly in medical studies.

B) Multiple regression (see chapter 20).

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 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.

Your text tries to present this, but really can't without resorting to some "primitive" matrix algebra.

If you start using this a lot, it's easier to just learn the necessary matrix algebra

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 (you usually do fitted values vs. residuals instead).

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.

(multiple regression is really simple in R... be careful!)

C) Logistic regression (see chapter 24, section 18).

Suppose you have a discrete (or binomial) dependent variable.

For example, 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 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.

The math involves estimation methods (look up Newton-Raphson to get an idea).

Now you can plot this (see fig. 24.2, p. 580). This curve gives you the probability that the cancer has spread, based on the size of the tumor.

Logistic regression can also be used to "classify" things:

For example, 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).

Obviously, one can also do "multiple" logistic regression.

D) ANCOVA (see chapter 18)

Your book claims it isn't doing ANCOVA (see 12.10), which seems odd. Perhaps the author defines ANCOVA a bit differently?

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)

E) Other multivariate designs (not in your text).

We already discussed MANOVA, but there are others. Let's summarize some of them:

Multivariate correlation: getting correlations between "groups" of variables.

For example, we have length, width, and height of seeds on one side, and we compare that to volume and weight.

Multivariate ANOVA's and t-tests: our example above.

Yes, we already discussed these, but you should realize that a lot of the "designs" we discussed (e.g., nested, two way, blocked), can all be extended to multivariate ANOVA.

Multivariate regression

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

Classification techniques

You measure 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 these (e.g., discriminant analysis) can be used in a similar way to logistic regression.

Principal components

Usually a variable reduction technique:

For example, 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.

Two ways of using Principal components:

Identify the most important variables, then discard the rest

This is easy to interpret, but wastes data (some information is in the "less" important variables).

Combine the original variables into "new" variables, that contain almost all the information. Do your analyses on these "principal components".

This is the better way to use principal components since you preserve most of the data, but it's much more difficult to interpret.

What do your "new" variables actually mean? They're a "linear" combination of the original variables, but that usually doesn't help in interpretation.

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 the stranger forms of factor analysis.

There are some forms of factor that are actually pretty good - in fact PCA is a special type of factor analysis.

F) 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

You then calculate how large this test statistic needs to be to include, say, 95% of the 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, mathematics, etc.

Many, many others.

G) There are many, many other techniques which we simply don't have time to talk about. Things like the Cox proportional hazards model, the Mantel-Haenszel test, various risk or survival analyses (Cox is an example), other non-parametric techniques (rank regression, etc.) and so on.