Modeling for Starters

Daniel T. Kaplan

March 6, 2014

Does this sound familiar? "A parameter is to a population as a statistic is to a sample." [source](http://en.wikipedia.org/wiki/Statistical_parameter#Examples) In introductory statistics, students encounter descriptive statistics such as the proportion, mean, variance, and correlation coefficient. These descriptions then form the object of inferential statistics, what one can validly say about the population parameter based on sample statistics.

There's something important missing here: the *purpose* behind the statistical analysis. Knowing why you're doing something is essential to deciding how to do it. A key word here is "modeling," the process of making choices about how to describe a system. Or, as Starfield, Smith, and Blelock consisely phrased it 20-years ago, a model is a "representation for a given purpose." In building a model, you have to decide what features of the system are essential to your purpose and which can be omitted. Models are always approximations, but the quality of the model depends on how closely the approximation suits the purpose. As George Box famously said, "All models are wrong. Some models are useful."

The sorts of decision-making that guide modeling are not usually emphasized in introductory statistics. In a typical course, there are few choices to be made and they are usually related simply to the setting: Is the data categorical or quantitative? Is the description a proportion or a mean? Is one group being described, or the difference between two groups, or the differences among more than two groups?

It's understandable that things developed this way. Egon Pearson laid out many of the factors at work in his 1926 review of R.A. Fisher's "Statistical Methods for Research Workers" (*Science Progress*, **20**:733-737 available [here](http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/esp.htm).)

Mr. Fisher has undertaken the very difficult task of attempting to put before research workers ... without any special mathematical training, a summary covering a great range of methods ..... The book is chiefly concerned with the best methods of handling small samples .... After conscientiously working through the examples, the student should feel able to apply the methods to exactly similar problems.... [[ E. Pearson, (1926) "Review of Statistical Methods for Research Workers (R. A. Fisher)", *Science Progress*, **20**:733-734.]]

An introduction to statistics should take into account the many ways the statistical environment the has changed since 1926. In the era of "big data," the problems of "handling small samples" are more peripheral. Statistics is applied to much more diverse and complicated problems, not just experiments but the interpretation of observational data in the social sciences, medicine, commerce, and government. Whereas the experimentalist sets the conditions of the system her data are collected from, the analyst of observational data needs to account for the conditions **such as they may be**. At a start this requires identifying possible confounders, choosing which factors might be important and which not. In a word: modeling.

How to teach modeling to introductory students? Computation seems essential and, fortunately, excellent statistical software such as R is easily available to students. Beyond being free, R supports a notation that's well suited to modeling. We developed the mosaic package to build on this notation to make modeling accessible, even at an introductory level.

To illustrate, consider the NHANES dataset on human body shape, disease and mortality. In R, you can easily access both the dataset and documentation:

require(mosaic)  
help(NHANES)

## No documentation for 'NHANES' in specified packages and libraries:  
## you could try '??NHANES'

# Will need this later on, but doesn't seem to work  
options(na.rm=TRUE)

The NHANES dataset comes from the [National Health and Nutrition Examination Survey](http://www.cdc.gov/nchs/nhanes/nh3data.htm). The motivation for such a survey is to understand better the relationship between diet and health so that people can make helpful changes in their diets. As such, the presumed causal chain diet health is central. Being an observational study, NHANES-III is not ideally suited to drawing causal inference. But experiment is not feasible and data such as NHANES are the best we have to guide our conclusions.

You may have heard about the link between overweight and type-II diabetes. Originally named "adult-onset diabetes," it is now considered an "epidemic" in children. (See, for instance, this [news report](http://www.washingtonpost.com/politics/fighting-the-growing-type-2-diabetes-epidemic-in-children/2013/01/06/361c3b00-5854-11e2-9fa9-5fbdc9530eb9_story.html).)

What's the relation between diabetes and overweight? With almost 30,000 people represented, of whom more than 1500 have diabetes, NHANES and similar data may provide insight.

Body-mass index (BMI) is a common measure of overweight. BMI is a number: weight divided by height-squared (in kg and m, respectively). In terms of BMI, "normal" weight is defined as , which overweight is . BMI higher than 30 is considered obese; BMI lower than 20 is underweight.

An introductory student might be tempted to start by looking at differences in group means.

mean( BMI ~ diabetic, data=NHANES,   
 na.rm=TRUE ) # need to kill the na.rm part.

## diabetic nondiabetic   
## 30.9 24.5

A diabetic of average weight would be obese, an average non-diabetic is normal in weight.

The notation BMI ~ diabetic used by the mean() function in mosaic, clearly identifies the calculation being performed, the data on which the calculation will be based. The expression BMI ~ diabetic means "break BMI down by diabetes classification," or, "account for bmi by diabetes," or "model bmi by diabetes."

The same notation works for a t-test (use t.test( ) instead of mean( )) giving in this case a p-value on the difference of means .

A graphical depiction --- another sort of representation --- tells more of the story: use bwplot() in place of mean() or t.test().

Clearly, the diabetics as a group tend to have higher BMI than the non-diabetics. But there is a lot of overlap. It is only for very underweight BMI --- say below 15 --- that there is what might be called "protection" against diabetes. Medically, such a low BMI indicates malnutrion, an eating disorder, or other health problems.

One thing that's missing here is the role of age, which plays an important role in the relationship between BMI and diabetes. A scatter plot tells a lot:

xyplot( BMI ~ age, group=diabetic, data=NHANES )

You can see that BMI depends on age. Children have BMI less than 20; young and middle-aged adults have BMI > 20 and heading up toward much higher; in the elderly, BMI trends lower than for middle aged. A lot of the diabetes for people with BMI < 30 occurs in the elderly.

The most obvious purpose for exploring diabetes and BMI is risk avoidance. For an individual contemplating lifestyle changes, the relevant question is this: How does BMI affect the risk of diabetes?

But all of the calculations so far are looking at something different. They depict how BMI depends on diabetes, rather than on how diabetes depends on BMI.

A simple way to approach the relationship is to look at the probability of having diabetes for high and low BMI. Setting the dividing point at BMI 30:

tally( diabetic ~ BMI>30, data=NHANES, margins=TRUE )

## BMI > 30  
## diabetic TRUE FALSE  
## diabetic 0.1230 0.0347  
## nondiabetic 0.8770 0.9653  
## Total 1.0000 1.0000

The notation here, diab ~ BMI>30 can be read: account for diabetes by whether or not the BMI is greater than 30. This is reversed from the sort of question being addressed by the groupwise means or t-test: how BMI depends on diabetes.

tally( BMI>30 ~ diabetic, data=NHANES, margins=TRUE )

## diabetic  
## BMI > 30 diabetic nondiabetic  
## TRUE 0.478 0.190  
## FALSE 0.522 0.810  
## Total 1.000 1.000

It's well documented that most people, including physicians, do not distinguish well between the two probabilities and . [Gigerenzer] The choice of analysis by the t-test (or the box-and-whisker plot or the differences in group means), though seemingly simple, gets things the wrong way round.

For addressing the desired purpose, the representation should show the relationship of interest: risk of diabetes as a function of BMI and age. This can be expressed quantitatively as a *function* with risk of diabetes as the output, not the input.

Constructing a relevant model is straightforward

mod <- lm( diabetic=='diabetic' ~ BMI\*poly(age,2), data=NHANES )

You can read this as, "account for whether a person is diabetic by a combination of age and BMI."

As a computer statement, this is no more difficult than the t-test. What makes it seemingly difficult is the *interpretation* of the result. For instance, such models are often interpreted using model coefficients as sample statistics. R makes the calculation as simple as this:

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) -6.19e-02 -0.03825  
## BMI 2.83e-03 0.00382  
## poly(age, 2)1 -1.94e+01 -15.04136  
## poly(age, 2)2 -3.23e+00 1.12416  
## BMI:poly(age, 2)1 1.02e+00 1.18682  
## BMI:poly(age, 2)2 7.69e-04 0.17200

Coefficients, though, can be hard to interpret at face value. For instance, here the coefficients are negative for bmi and age --- suggesting that diabetes risk is lowered with higher BMI and age.

It's easier to see what the fitted model has to say by plotting it as a *function*: diabetes risk versus BMI for different ages. In mosaic, you have direct access to the fitted model function itself.

f <- makeFun(mod)  
# Simplify this statement for the article,   
# leaving out the labeling stuff  
plotFun( f(BMI, age=30) ~ BMI, BMI.lim=range(10,50),   
 ylab="Diabetes Risk", xlab="BMI",ylim=c(0,.8),  
 main="")  
plotFun( f(BMI, age=50) ~ BMI, add=TRUE, col='green')  
plotFun( f(BMI, age=70) ~ BMI, add=TRUE, col='red')

# Put some text in this plot to label the lines ????

It's comparatively easy to elaborate on the function by adding more components or changing the form of the fitted function. For instance, logistic functions are often used in modeling probabilities, and the dependence on age may not be simply linear.

mod2 <- glm( diabetic=='diabetic' ~ BMI\*poly(age,2), data=NHANES,   
 family='binomial' )

Even for such a seemingly complex model, the model functions make interpretation straightforward:

f2 <- makeFun(mod2)  
# Simplify this statement for the article,   
# leaving out the labeling stuff  
plotFun( f2(BMI, age=30) ~ BMI, BMI.lim=range(10,50),   
 ylab="Diabetes Risk", xlab="BMI", ylim=c(0,.8),  
 main="")  
plotFun( f2(BMI, age=50) ~ BMI, add=TRUE, col='green')  
plotFun( f2(BMI, age=70) ~ BMI, add=TRUE, col='red')

# Put some text in this plot to label the lines ????

Or, risk can be plotted as a function of *both* age and BMI:

plotFun( log(f2(bmi,age=age))~bmi&age,  
 bmi.lim=c(10,50), age.lim=c(0,70) )

The graph suggests that the risk of diabetes increases with BMI at any age.

You may ask, "Why that model and not some other? Why not include sex or other variables as a covariate? Why not split out the young and the elderly from the middle-aged?"

Our answer: Go and try it! Then you'll have to think about how to decide which of two models is better, a question which in part can be addressed through confidence intervals, etc.

For students, crafting models highlights the creative aspects of statistical reasoning. It's not just a matter of interpreting a p-value: you can examine many factors and build a model that addresses your actual purpose.

With software and comprehensible notation standing in for Fisher's "special mathematical training," students can be introduced to covariates and the process of judging whether a covariate contributes to the model in an important way. Indeed, without such capabilities, all one can do is preach caution about causation. But causation, even from observational data, is often central to the purpose of the statistical analysis.

We teach modeling in the introductory class. We focus on statistical concepts [FRESH] and use R/mosaic to empower students with an accessible and expressive way to apply those concepts to data. It's important, we think, to start modeling in the introductory course. The vast majority of statistics students take just the intro course, but everybody has to deal with confounding. If students don't learn this in their statistics course, where will they learn it?