Statistical Modeling in an Introductory Course

Daniel T. Kaplan
Macalester College, Saint Paul, MN 55105
kaplan@macalester.edu

May 6, 2005

We are teaching a novel introductory statistics course that introduces basic statistical concepts in the context of statistical modeling. We believe this helps to increase student motivation, since students can ask complicated, realistic questions of data. It also allows us to provide multivariate analysis techniques that are increasingly demanded in client departments such as economics and biology.

The course is built on some fundamental principles that make statistical modeling accessible to introductory students:

1. Whenever possible, use geometry rather than algebra for derivations. Almost all students have good geometry skills, knowing about angles and lengths and shadows. They all know the pythagorean theorem and understand the geometry of right triangles. All we need to do is get them to generalize these skills to N dimensions.

2. Use simulation and resampling to construct sampling distributions and confidence intervals. Who doesn't understand about shuffling a deck and dealing?

3. Use sophisticated computation routinely. What was a supercomputer 20 years ago is now available at the local mall for $500, and sophisticated software such as R is free.

The tools provided by an introductory statistics course are not up to the complex questions that students need to answer in fields such as economics, business, or biology. Introductory statistics courses build an impressive theoretical structure to answer a rather simple question: Are these two groups different? Answering this question is important, but it is not the only sort of question our students will need to address. They will need to deal with multiple explanatory variables.

We might use examples of Simpson's paradox to show the importance of considering more than one explanatory variable, but conventionally we don't give students the tools to deal with such situations: just a cautionary tale.

Challenging the Assumptions

At Macalester, we had a chance to challenge some of the assumptions about introductory statistics students.

First, we revised our introductory calculus course so that it covers functions of multiple variables, modeling concepts, and important ideas from linear algebra (projection, linear combinations, subspaces, etc.) This means that the students entering our new introductory statistics course all know about vectors, how to compute angles between vectors, multi-variable polynomial models (out to quadratic order), and so on.

Second, we make available to every student sophisticated computing capability. This isn't so hard with computer prices falling and computers widely available at home. With about one-twentieth of what a year's college education costs, one can nowadays buy a laptop computer and install on it free, but sophisticated statistical software such as R. We then teach every student the basics of using such software: plotting, defining variables, reading in data, defining a function.

With the computer in hand, we are then freed from the need to teach students algorithms. Instead, we can focus on the concepts behind the algorithms, so that students develop the understanding that they need to reason about the statistics rather than being forced to follow the calculations.

Example 1: Multivariate Modeling

Our course has a very heavy emphasis on modeling. Even simple descriptive statistics such as the mean are presented as a kind of model. We don't distinguish between simple regression and multiple regression: all of it is projection of the vector of the response variable onto the subspace spanned by the explanatory variable.

In a conventional introductory statistics course, one might teach simple linear regression by having students draw a line over a scatter plot and indicate the residuals. This is a good exercise for our students too, but we also work through the calculation in a vector-space formalism: plotting the response and explanatory variables (for $N = 2$), finding the orthogonal projection of the response variable onto the model subspace, finding the residuals. These easy geometrical calculations can then be matched up to the computer's calculations. It's also feasible to do graphical calculations with $N = 3$. Once a student sees how the projection works, and sees how the graphical solution can be computed using dot products, etc., it's easy for them to understand how the computer does things for large N.

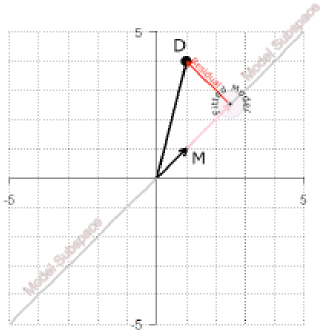

Figure 1. A graphical depiction of fitting the model $M = (1,1)$ to the data $D = (1,4)$. Fitting the model means finding the multiplier a to make the length of the residual vector $D - aM$ as small as possible. This particular model corresponds to finding the mean of the values in D. The fitted values, shown as a light-colored vector, are a projection of the data vector onto the model subspace. The residual vector connects the fitted values to the response values.

Example 2: Confidence Intervals

Confidence intervals play a justifiably important role in introductory statistics courses. They emphasize the variability in measurements and how to quantify them; they show why it's helpful to have large N. With the computer, it's natural to teach about sampling distributions by using simulation and resampling. No formulas are necessary. However, some central ideas — the width of the interval depends on the confidence level, the dependence on N — can still easily be expressed and mastered.

With the conceptual apparatus supplied by resampling, theoretical concepts such as power can readily be approached on a practical level. For example, suppose one has a data set that hints at the relationship between two variables, but without reaching statistical significance. If we take the data set as reflecting our alternative hypothesis, it is a straightforward matter to resample from that data set, varying the sample size as necessary to reach a desired p-value. At any given sample size, the number of trials in which the p-value is small enough indicates the power for that sample size.

Similarly, the sampling distribution under the null hypothesis can be generated by resampling an explanatory variable.

Example 3: ANOVA

Just about every student knows the pythagorean theorem and understands the geometry of right triangles. With that, and the notion of projections, it's straightforward to understand analysis of variance. The computer allows students to do the calculations on large data sets, but students can do the basic calculations themselves with a ruler.

Here is the ANOVA report, using R software:

```
> summary(aov( c(1,4) ~ c(1,1)-1))
          Df Sum Sq Mean Sq F value Pr(>F)
c(1, 1)    1   12.5    12.5  2.7778 0.3440
Residuals  1    4.5     4.5
```

The sum of squares can be estimated graphically just by measuring the length-squared of each vector. For example, the length of the fitted model is about 23/32in, and the length of the residual is 7/16in. Taking the square of the ratio of lengths gives us the F value.

ANOVA becomes the basic training ground for hypothesis testing. It is in many ways a more natural training ground than, say, the z-test because it has a simple geometrical interpretation. For example, in the graph below, I seek to know whether the model vector M is significantly associated with the response vector D. The null hypothesis is that M is unrelated to D, that is, that M might be equally likely to point in any direction relative to D. The p-value is the probability that a randomly pointing vector would be more closely aligned with D than M is. This p-value is easily computed as a ratio of angles.

Of course, the point isn't to replace standard regression and ANOVA calculations with graphics, but to develop an intuition about the calculations so that students can successfully use them to reason about complex multi-variable relationships in the systems they are interested in studying.

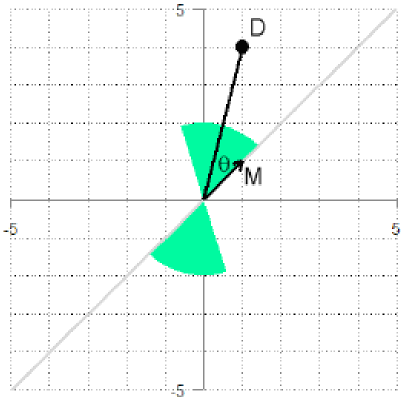

Figure 2. The p-value is the probability of a randomly pointing model vector falling closer to the response vector: that is, anywhere in the two cones shown. The fraction of the circle subtended by these cones is the p-value. In this case, since the model vector is $(1,1)$ we are effectively testing whether the mean of D is non-zero; the ratio of angles is exactly the same as p-value generated by a one-sample t-test or ANOVA. The angle θ is 31 degrees, giving a p-value of $4 \times 31/360 = 0.344$.