Using Exams for Teaching Concepts in Probability and Statistics

Andrew Gelman
Columbia University

Keywords: *calibration, correlation, instruction, prediction, regression*

We present several classroom demonstrations that have sparked student involvement in our introductory undergraduate courses in probability and statistics. The demonstrations involve both experimentation using exams and statistical analysis and adjustment of exam scores.

Our courses for undergraduates typically include two midterm exams and a final. Students are of course very interested in their exam scores; here, we present some tricks we have used to channel this interest into thinking about statistics.

Guessing Exam Scores

We include a question at the end of the first midterm asking the student to guess his or her total score on the other questions of the exam. As an incentive, the student receives 5 points extra credit if the guess is within 10 points of the actual score (which is on a scale of 0 to 125). When the students complete their exams, we keep track of the order in which they are handed in, so that we can later check to see if students who finish the exam early are more or less accurate in their self-assessments than the students who take the full hour. When grading the exams, we do not look at the guessed score until all the other questions are graded. We then record the guessed grade, actual grade, and order of finish for each student. We have three reasons for including the self-evaluation question. First, it forces the students to check their work before turning in the exam. Second, it teaches them that subjective predictions can have systematic bias (in this case, students tend to be overconfident about their scores). And third, the students' guesses provide us with data for a class discussion, as described below.

We thank Phillip Price, Seth Roberts, and Howard Wainer for helpful comments and the National Science Foundation for Young Investigator Award DMS-9457824.

Figure 1 displays the actual and guessed scores for each student in a class of 53, with students indicated by solid circles (women), empty circles (men), and ? for a student of unknown gender. (This student had an indeterminate name, was not known by the teaching assistants, and dropped the course after the exam.) The points are mostly below the 45° line, indicating that most students guessed too high. Perhaps surprisingly, men do not differ appreciably from women. The dotted line shows the linear regression of actual score on guessed score and displays the typical "regression to the mean" behavior. A class discussion should bring out the natural reasons for this effect. Figure 2 shows the difference between actual and guessed scores, plotted against the order of finish. Many of the first 20 or 25 students, who finished early, were highly overconfident, whereas the remaining students,

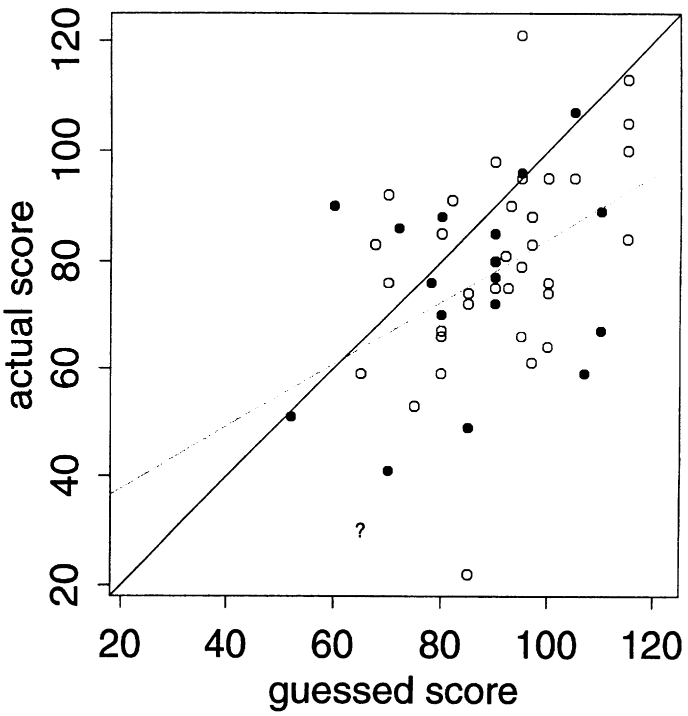

FIGURE 1. *Actual versus guessed midterm exam scores for a class of 53 students*

Note. Each symbol represents a student; empty circles are men, solid circles are women, and ? is of unknown gender. The 45° line represents perfect guessing, and the dotted line is the linear regression of actual score on guessed score. (The separate regression lines for men and women were similar.) Both men and women tended to perform worse than their guesses. That the slope of the regression line is less than 1 is an instance of the "regression effect": If a student's guessed score is x points higher than the mean guess, then his or her actual score is, on average, only about $0.6x$ higher than the mean score. A square scatterplot is used because the horizontal and vertical axes are on the same scale.

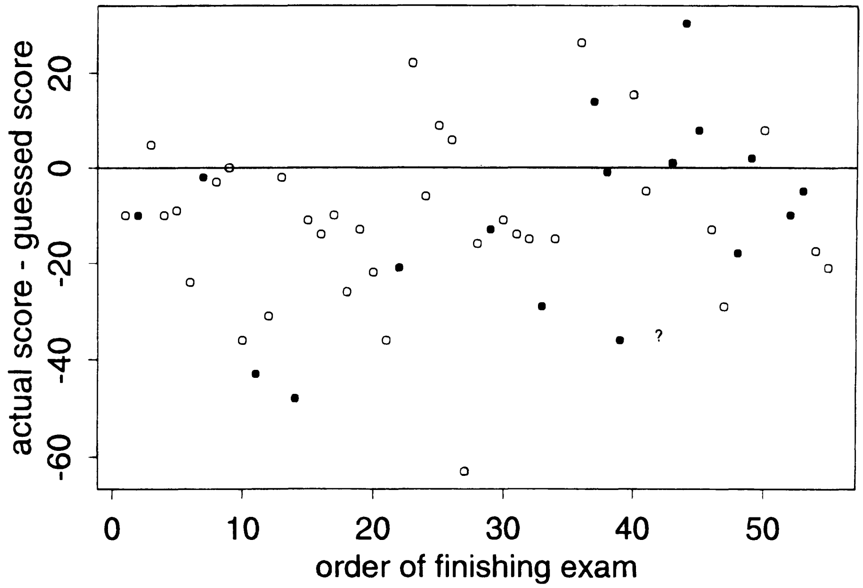

FIGURE 2. *Difference between actual and guessed midterm exam scores, plotted against the order of finishing the exam*

Note. The exact order is relevant only for the first 20 or 25 students, who finished early; the others all finished within 5 minutes of each other at the end of the class period. Each symbol represents a student; empty circles are men, solid circles are women, and ? is of unknown gender. The horizontal line represents perfect guessing. The students who finished early were highly overconfident, whereas the other students were less biased in their predictions.

who took basically the full hour to complete the exam, were close to unbiased in their predictions. Perhaps this suggests that students who finish early should take more time to check their results. (The students who finished early did, however, have higher-than-average scores on the exam.)

When teaching this course again, we varied the procedure by handing out Figures 1 and 2 a week before the midterm exam, discussing the overconfidence phenomenon, and warning them that the same question would appear on their exam. We were encouraged to find that, thus prepared, the students' guesses were less biased than those of the earlier class. Figure 3 displays the results for the unprepared class (indicated by dots, the same data as displayed in Figure 1) and the prepared class (indicated by asterisks).

Correlations and Regressions

A nice way to illustrate the regression effect is with a scatterplot of students' scores on the two midterm exams. The regression line of the second on the first typically has a slope less than 1; the students who score the highest on the first exam typically do worse on the second exam ("regression to the

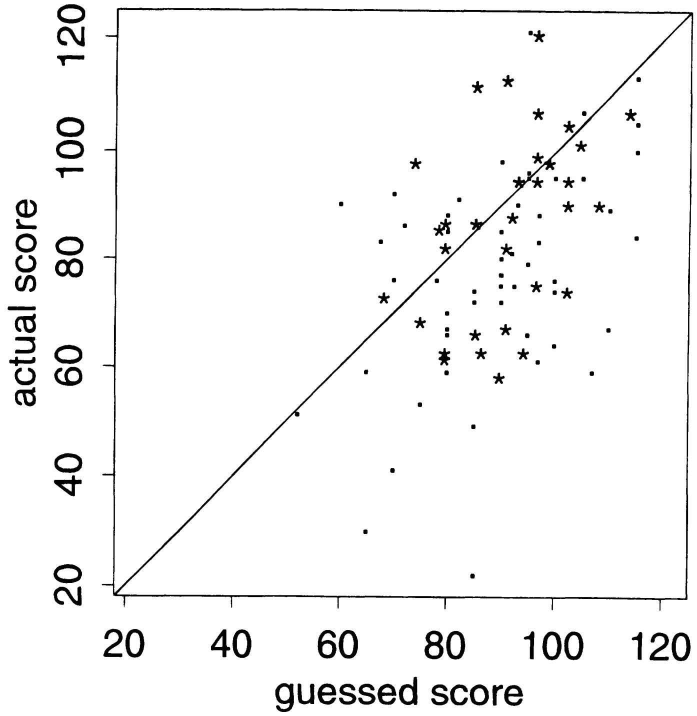

FIGURE 3. *Actual versus guessed midterm exam scores for students in two terms of introductory statistics classes*

Note. The dots represent students in the first term; the asterisks represent students in the second term, who were shown the data from the first term a week before the exam. The students in the second term gave predictions that were less biased. A square scatterplot is used because the horizontal and vertical axes are on the same scale.

mean"); and so forth. Many students are more interested in this example than in the traditional regression example of parents' and children's heights. Students commonly see exam scores represented as univariate distributions (for example, mean, median, and standard deviation of scores, or stem-and-leaf plots) but the bivariate display stimulates new thoughts.

One year, to make a different point, we recorded for each student the score on the final exam and the number of pages used by the student in the blue book to write the exam solutions. The two variables are negatively correlated. Since then, we have used these data to illustrate Simpson's paradox and the distinction between correlation and causation. A naive interpretation of the negative correlation between pages written and exam scores would suggest that students could raise their scores (on average) by writing less. But this is not so. For any given student, it would only help to write more. This is similar to the high scores of the students who finish early on exams (see the previous section): Students who require the entire class period to finish their exams have lower scores, on average, than those who finish early, but, for any given student, staying on and working through the entire class period can only increase his or her score.

These examples are a natural lead-in to other discussions of correlation and causation. For example, people who own BMWs have bigger bank balances, on average, than people who own VWs, but this does not mean that if you sell your VW for a BMW you will have more money in the bank. For another example, baseball players with higher batting averages receive higher salaries, on average; does this mean that if a professional baseball player raises his batting average he will likely get a higher salary? Well, yes . . . but why is that? Obviously the correlation alone is not enough to convince us.

Randomizing the Order of Exam Questions

Without the knowledge of the students in the class, we prepare two versions of the second midterm examination, identical in all respects except that the order of the questions is reversed. We prepare equal numbers of the two versions and mix them randomly before handing out one to each student for the exam. In grading, we are careful not to be influenced by the order of the questions. (If necessary, blindness can be achieved by having the students put their names on each sheet of the exam and then tearing the exams apart before grading, but we find grading of probability and statistics questions to be objective enough that such a formal procedure seems unnecessary.) We record the grades achieved by the two groups of students.

After returning the graded exams to the students, we reveal that there were two forms of the exam and present the aggregate results; for example, the average score was 65 for Exam A and 71 for Exam B. Should we adjust the scores of the Exam A students upward (and the Exam B students downward) to reflect that Exam A seems more difficult, in retrospect? A student who took Exam B objects, noting that the two exams had identical questions—just the order was different. But the order could have an effect, right? What if the two forms had been randomly given to 1,000 students and this difference had been observed—would it be "real" then? The goal here is to get the Exam A students and Exam B students all fired up and holding opposite positions.

How can we address the question of whether the observed difference is due to the exams or just because, say, the better students happened to take Exam B? We can consider this as an experiment designed to measure the difference in exam difficulties and use the standard methods to obtain an estimate and standard error. Is the difference statistically significant? Should we adjust the scores, and, if so, by how much? To round out the discussion, we ask, What if the exams differed not just in their ordering, but in the questions themselves? How would/should our statistical methods change? This is of course a subtle question with no easy answers. Students have also raised ethical questions about basing grades on different forms of the exam.

In practice, the quality of the discussion is highly influenced by the observed difference between groups, which cannot be predicted ahead of time. If the true difference between the exams is approximately zero and the standard deviation of exam scores is 20, then with a class of 40 students the observed difference in means has an expected value of zero and a standard deviation of 6.3. An adjustment of 5 to 10 points is enough for students to care about, but anything less than 5 points might not spark much interest.

We return to this example at the end of the course when covering multiple regression. There are two basic explanations for the differences in average scores between the two exams: different difficulty levels of the two exam forms and different quality of students taking the exams. Randomization balances out the latter factor on average, but only on average. We ask the class, How would we be able to tell if better students were taking Exam B? What other information do we have about their abilities? That's right—their scores on the first exam! A difference in difficulty between the two exams should appear as a nonzero regression coefficient on the variable *exam type*, in a regression of exam score, after controlling for score on the first exam. Of course, the above comparison of means is equivalent to this regression, but *without* controlling for the first exam score. Including the control variable should improve our estimate of the relative difficulties of the exams.

Probabilistic Answers to True-False Questions

In our course on decision theory, we introduce the Brier (1950) score for evaluating probabilistic forecasts of binary outcomes. If a forecaster assigns the probability p to an event, the forecaster's Brier score is defined as $1 - (1 - p)^2$ if the event occurs, or $1 - p^2$ if the outcome does not occur. This scoring system is designed to give an advantage to forecasters who are calibrated (given that the forecast probability is p, the event should actually occur with frequency p) and precise (p should be as close as possible to 0 or 1, while remaining calibrated). If a forecaster has a subjective probability π that an event will occur, the expected Brier score will be maximized by setting $p = \pi$; that is, it is a "proper" scoring rule (see Dawid, 1986, for more on this topic).

We cover the Brier score extensively in class, using examples such as weather forecasting (the original motivation for the method). But we really bring the subject to life by including on the midterm exam several true-false questions, for which each student is asked to give a subjective probability p that the correct answer is "True." Their score for each question is 5 times the Brier score. We have found that students tend to be overconfident in their answers, frequently assigning probabilities of 0 or 1 (indicating certainty that the answer is "False" or "True," respectively) but being wrong. They have not internalized the mathematics of the Brier score. For example, suppose you think that the correct answer to a question is "True," but you are not *completely* sure. If you write 0.8, you will receive 4.8 points (out of a possible

5) if you are correct and 1.8 points if you are wrong. Even a blind guess of 0.5 nets you a certain 3.75 points. Students have a greater appreciation of calibration of forecasts after losing exam points from overconfident guessing. There are many other possible methods of adjusting exam scores to take into account students' uncertainties in their answers; see, for example, Coombs, Milholland, and Womer (1956) and DeFinetti (1965).

Conclusions

Exams are an important teaching tool for two obvious reasons: (a) the students' direct experience in working out the problems during the exam, and (b) the learning that occurs while studying and preparing before the exam. This article discusses some ways in which we have used exams to teach statistical concepts in a third way, as direct experience, by harnessing students' interest in their grades. We have had success using these techniques to involve the students in class discussions, and we believe there is the potential for much more work in this area. However, these demonstrations must be conducted with care: Students take their grades seriously, and it is important to make it clear that their exam grades are not manipulated in an arbitrary or random manner.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1–3.

Coombs, C. H., Milholland, J. E., & Womer, J. F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 15*, 337–352.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 7, pp. 210–218). New York: Wiley.

DeFinetti, B. (1965). Methods for discriminating level of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 18*, 87–123.

Author

ANDREW GELMAN is Associate Professor, Department of Statistics, Columbia University, New York, NY 10027. He specializes in statistics.

Received October 25, 1996
Accepted October 28, 1996