What's the Magnitude of False Positives in GPA Award Programs?

Orange County Register indicates 1/3 or 35 percent; statistician estimates 2 percent of schools, 1 percent of award dollars
A good high school statistics student could/would have computed 2 or 3 percent

September 9, 2002

By DAVID ROGOSA
Stanford University

First part of a two-part discussion of the Orange County Register series on the California API

The Orange County Register analysis (8/11-16) of the California API awards is so riddled with statistical blunders and misstatements that credibility should not be given to their numerical assertions. On the other hand, the OCRegister should be applauded for taking on complex topics and for raising good questions. (Many of those questions are the ones I've been working on the past two years.) The issues regarding the accuracy of the API award programs are the ones discussed here.

To organize this statistical content, the best expository strategy I could come up with was to present the most obvious correct calculations in this first part and leave to the second part the dissection of the blunders in the OCRegister presentation. That is, obtain in this first part the correct answer on false positives of about 2% of schools, and use the contrast of 2% and the OCRegister 1/3 or 35% to motivate readers to consider in second part the explanations for the misinterpretations of their "margin of error" as applied to the API award programs.

Related Documents

Available from CDE
API Research Page
http://www.cde.ca.gov/psaa/apiresearch.htm

Application of OCR "margin of error" to API Award Programs
David Rogosa, Sept., 2002

Plan and Preview for API Accuracy Reports
David Rogosa, July 2002

Interpretive Notes for the API..., series
David Rogosa, 2000,2001,2002

In this first part, the fable of "The High School Intern and the API Dollars" is used to provide an initial plausibility check on the OCRegister claims, followed by more serious calculations that produce results remarkably close to the high school student calculation. Before considering those results, we need a little background on the statistical framework for studying the API awards.

Findings

The first part of this series provides an answer to the question posed by the OCRegister--What is the magnitude of false positives in GPA awards? Although the OCRegister claims more than 30%, the correct answer is 1.25% in 1999-2000 and closer to 3% in 2000-2001. Over the two award cycles the estimated number of false positives comprise 2% of the schools receiving GPA awards and about 1% of the funds.

In the second part of this series, the blunders and misunderstandings--i.e., their adoption of a “margin of error,”--that produce the inflated results are explained.

The false positives can be reduced further, but it’s a tradeoff with false negatives, as is shown by some school examples in the second part.

It also seems important to say up front that concern about the accuracy of the API award programs is not new nor has it been kept a secret. For example, Richard Rothstein, in his *New York Times* column January 24, 2001, http://www.nytimes.com/2001/01/24/national/24LESS.html “Flaws in Annual Testing” citing my work, covered false positives, false negatives, the saved-by-the-subgroups message and the effects of multiple subgroups on false negatives for the API awards.

Also at the beginning of July, I posted the “Plan and Preview” document but, perhaps because the numbers in that document were correct and the prose non-sensational, little attention ensued. The large, formal “Accuracy Reports” are still forthcoming, as a good bit of statistical research remains on how best to understand the properties of the award programs.

Medical Diagnostic Test Context

The statistical approach to the accuracy of award programs follows standard ideas from medical diagnostic and screening tests. The accuracy of the award programs is expressed in terms of false positive and false negative events, which are depicted in the chart on the following page (adapted from the exposition on the CDC web page). Commonly accepted medical tests have less than perfect accuracy. For example, prostate cancer screening (PSA) produces considerable false positives and in tuberculosis screening, false negatives (sending an infected patient into the general population) are of considerable concern. In the context of API awards, false positives describe events where

statistical variability alone (no real improvement) produces award eligibility. False Negatives describe events for which award status is denied due to statistical variability in the scores, despite a (specified) level of underlying ("real") improvement.

2x2 diagnostic accuracy table (see also CDC site http://www.cdc.gov/hiv/pubs/rt/sensitivity.htm)

	Good Real Improvement	NO Real Improvement
GPA Award	TRUE POSITIVE a	FALSE POSITIVE b
NO GPA Award	FALSE NEGATIVE c	TRUE NEGATIVE d

Note that P{no real improvement| award} = b/(a + b) , 1 - predictive value positive

This table serves to help define the terms in this discussion; application of the usual numerical tools for this setting are more complicated because individual schools differ widely in their accuracy properties.

In discussing accuracy of the award programs, a main factor is the subgroup criteria-- the descriptive phrases that I have used in prior discussions are “saved by the subgroups” and “herding cats”. The herding cats metaphor is that it’s unlikely that a set of cats will all move in the same direction (past the growth target) by accident, but a strong enough probe (real improvement) may persuade all the cats to move in unison. The number of significant subgroups is an important factor: having many subgroups in a school tends to make false positives less likely and make false negatives more likely (the more cats, the tougher to herd them). Furthermore, statistical variability in the school and subgroup scores makes growth targets far more formidable than these might appear because of the subgroup requirements (as each of the subgroups has larger uncertainty than the school index). To have high probability that all subgroup scores will meet the criteria requires underlying improvement that far exceeds (blows through) the seemingly modest growth target. Further discussion and numerical results are in Plan and Preview.

The High School Intern and the API $$$

The short version of this fable is expressed in the equation :

Smart High School Statistics Student + Publicly Available Information = correct answer

The setting is California, July 2002. A newspaper preparing a series on the API has a summer intern who has recently completed one of the fine (one semester) high school statistics courses. The intern is asked: “Do you think the finding that a third or more of the GPA award schools made no real improvement--P{no real improvement | GPA award} > .3-- is reasonable?”

The High School statistics student makes the following presentation to the newspaper’s reporters: "In my class we learned about about false positives and false negatives, like the chart on the CDC website. To get the information I needed for the API, I did the following:

a. I went to Rogosa's "Plan and Preview" report on the CDE site. The information I get there is P{award | no improvement} for two examples, a typical elementary school with a .1 probability, and a typical high school with a .01 probability. So the best I can do is reason that middle schools will be in between elementary and high, and since there are more elementary schools, that probability might average out to .07 or .08.
b. But I'm not done because that's not the probability I was asked about. From my statistics course, I know a little about conditional probability
   P{no improvement | award} = P{award | no improvement}*P{no improvement}/P{award} .
   So I need a bit more information.
c. From newspapers or CDE site I see that the GPA award rate for 1999-2000 was just about 2/3.
d. From Rogosa's Year 2000 Interpretive Notes on the CDE site, I can get the observed distribution of year-to-year change in the API, and I calculate that proportion of schools with observed improvement less than or equal to 0 (which overestimates no true improvement) is approximately .1.

So now I can plug into my conditional probability formula and get a guesstimate for the 1999-2000 GPA awards P{no real improvement | GPA award} ≈ .07*.1/.67 = .01 . For 2000-2001 awards, less awards were given, and from Rogosa's Year 2001 growth Interpretive Notes I see that at least twice as many schools showed no improvement compared to 1999-2000. Combine those factors, and for 2000-2001 I get that P{no real improvement | GPA award} is at least .03. Average them out to an overall .02, and 1/50 is a whole lot less than 1/3.”

The Statistician's Probability Calculations

The statistician works considerably harder than the high school student, for essentially the same result. Calculate for each school P{true improvement $\leq$ 0 | data} for all GPA award schools. Do this (empirical Bayes) calculation separately by school type (elementary, middle, high) and by award cycle (1999-2000, 2000-2001), because improvement differs over school type and between year-to-year comparisons (one could further stratify these analyses by low and high scoring schools etc but that seems to be overkill). The aggregate results of this collection of six analyses are shown in the table below.

False Positive Results

Award Cycle	School type: Elementary	Middle	High
1999-2000	.0098 35.0	.0199 12.67	.0277 8.95
2000-2001	.0296 74.53	.0304 14.21	.0448 7.94

Each cell contains
the average probability of no improvement for GPA award schools
the expected number of schools having no real improvement and given award

The expected number of schools in each cell is simply the sum of the probabilities of all the schools (or the mean probability times the number of schools). The 1.25% result for 1999-2000 award cycle is obtained from (35 + 12.67 + 8.95)/4545 = 0.01246 and the 3% result for 2000-2001 is obtained from (74.53 + 14.21 + 7.94)/3167 = 0.03053. The cumulative 2% of schools is (35 + 12.67 + 8.95 + 74.53 + 14.21 + 7.94)/(4545 + 3167) = 0.01988. The total funds associated with the false positive estimates are closer to 1% of the award monies because of two factors: GPA awards in the 1999-2000 cycle were about twice as large as the awards in the 2000-2001 cycle, and because these funds are per student, the small schools which tend to have the higher false positive probabilities receive less total funds.

In the full technical "Accuracy Reports" these type of analyses will be greatly expanded. The tools will be forms of the statistical methodology of "False Discovery Rates," methods commonly used and under continuing development in genomics for GeneChip array data.

False Negatives and the Fallacy of Small School Advantage

The tradeoff between false positives and false negatives is the important policy decision in the formulation of an award or sanction system. False negative events are given mention in the OCRegister series, but since no numerical claims are made I'll leave a full analysis to the Accuracy Reports. The tradeoff between false positives and false negatives is revisited at the end of part 2, where alternative award rules accommodating the OCRegister "margin of error" are shown to reduce false positives (which aren't really the problem) at the cost of greatly increasing false negatives.

One important issue relating to false negatives is the claim that "small schools [have] an advantage in winning awards" (e.g. OCR 8/11). The fallacy lies in the neglect of false negatives. A small school having made no real improvement has statistical variability as its friend, in that a false positive result may occur more often than for a large school. But a small school that has made substantial real improvement (which so far has been the more likely event) has statistical uncertainty as its foe, in that a false negative result may occur more often than for a large school. An imperfect illustration using the examples from Plan and Preview (also part of the chart in part 2) compares the elementary school example (n=350) versus the high school example (n=1115); for true improvement 29 points the false negative probability P{no award | strong real improvement} is more than twice as large for the smaller school, .13 versus .29. This comparison understates the difference in false negatives between the two school sizes because the high school has 4 subgroups; a cleaner comparison would be formed from two elementary schools of similar subgroup configuration, but I didn't want to drag additional examples and numbers into these short pieces.

A personal note. I vastly prefer to help reporters understand these state testing data, rather than argue with their stories. I've put out a series of reports (the Interpretive Notes) as part of the effort to be helpful. The "Plan and Preview" document I put out on this site July 4 weekend was my best attempt to warn off the OCRegister from the direction their stories were taking. False positive and false negative probabilities for exemplar schools were presented, and in the 5 pages of text there were repeated warnings that you can't go from the API standard error to the properties of the award programs, plus plots demonstrating that fallacy. Regrettably, they ran through all those stop signs.

Please consider part 2: Application of OCR "margin of error" to API Award Programs and look forward to the full Accuracy Reports at the end of the year