Application of OCR "margin of error" to API Award Programs

The blood pressure parable reveals 'margin of error' folly
CA schools falling within their 'margin of error' have probability .97 and larger of positive true improvement--should we 'round' .03 up to 1.0?
Different award rules will give different results; Do we care about False Negatives?

September 9, 2002

By DAVID ROGOSA
Stanford University

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

In the first part of this series, I explained false positives and obtained the empirical estimates: 2% of schools and 1% of funds. Admittedly, the 2% false positive result is much less of a head-line grabber than 35%. In this second part, I build on that 2% result to explain why the OCRegister 1/3 or 35% figure is so off the mark and also why their "margin of error" construct–mean of 20 for the API and 26 for year-to-year API improvement--is so misunderstood and misapplied.

The Orange County Register stated results:

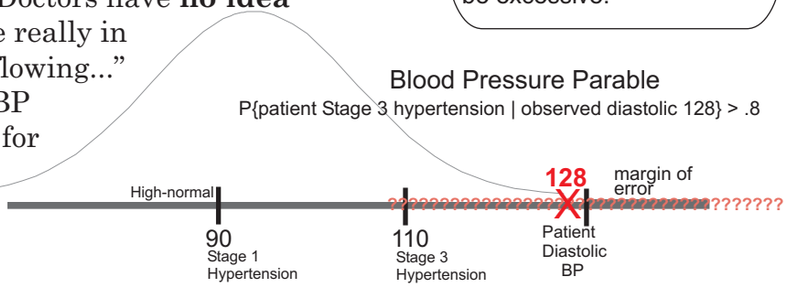
a. the 35% figure, "of the $67.3 million the state plans to give to schools as rewards this year, 35 percent is slated for schools where improvement had little or no statistical significance. (OCR 8/11)"
b. the one-third figure in their lead "the state has no idea whether one-third of the schools receiving cash awards actually earned the money (OCR 8/11) " or "about one-third of the award-winners gains were within the error margin making it impossible to tell if the school really improved" (OCR 8/16)

are obtained by the following sort of 'calculation' [see Note 1]. Count up all the schools receiving GPA awards for 1999-2000 and for 2000-2001. Tag a school with the designation "state has no idea if it really improved or earned the money" whenever the second year API score does not exceed the API target by at least 20 points (their mean margin of error). The proportion of schools (combining 1999-2000 and 2000-2001 award cycles) so tagged is .347. Hence the widely-cited 35% (OCR 8/11) [again, see Note 1, as for prose sanity the simplest version is given above and their actual calculation is one of the more complicated rules listed]. This OCRegister calculation has advantages of being simple to carry out and easy to explain; the only downside to their calculation is that it's horribly wrong.

One example used later is an Elementary School (CDS 19647336018253) which had year 1999 school API 616 and year 2000 API 647. This school received a $40,262 GPA award for that performance. But that school has a margin of error for improvement of 34.9 which exceeds the observed gain of 31 points. Does tagging this school make sense? What can we say about this school's true improvement? Is it really "impossible to tell"?

Blood Pressure Parable

As a lead-in to API data demonstrations, consider this artificial setting, using diastolic blood pressure (DBP) for hypertension diagnosis. The error of measurement in diastolic blood pressure is assumed to have standard deviation 10.2 to yield a OCRegister-style margin of error of 20 points. A diagnosis of Stage 3 hypertension, DPB 110 and above, will lead to drug intervention with beta-blockers, diuretics etc. Following the margin of error 'logic,' no DPB reading below 130 is *really* Stage 3 hypertension. The OCRegister headline would be: *Billions of Dollars wasted on diuretics and beta blockers,* and their lead would read: "Doctors have **no idea** whether millions of diagnosed patients are really in Stage 3 hypertension, yet the drugs keep flowing..." This parable considers a patient with a DBP reading of 128, within the margin of error for Stage 3 hypertension. Does the doctor really have *no idea* whether this patient



Related Documents

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

What's the Magnitude of False Positives in GPA 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

Findings

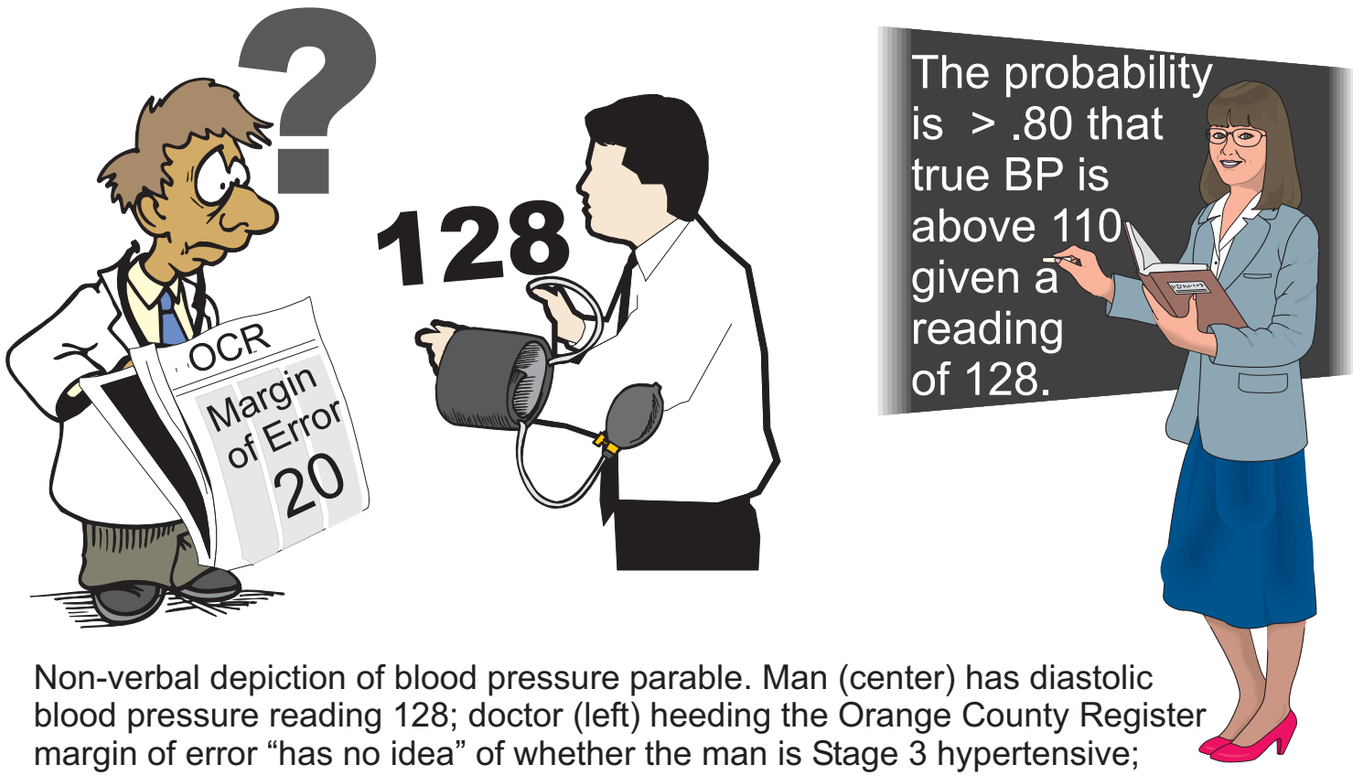
The margin of error, the main OCRegister tool for analysis of the API awards, is shown to have no value and to lead to preposterous results. A series of examples show CA schools with improvement in API falling within this margin of error (and therefore not real according to OCRegister) also having probability .98 and larger that true change is greater than 0.

This disconnect between the margin of error and the actual probabilities also provides an explanation for the inflated OCRegister result (~35%) for false positives. Application of the margin of error to the GPA awards is equivalent to rounding probabilities of magnitude .02 up to 1.0.

Alternative award rules implicit in the OCRegister analysis --e.g. requiring improvement of 20 points past the current API target--are used to illustrate the tradeoff between false positives and false negatives. The cost of reducing false positives can be excessive.

is properly diagnosed in Stage 3 hypertension? The relevant calculation is :
Probability{true DBP at least 110 given measurement of 128}. That probability is greater than .8 (depending on the details of the population distribution etc). As the patient, would odds of 4 to 1 that you are in Stage 3 hypertension provide useful information for therapy or does the 20 point margin of error make it "impossible to tell"? [Small reality check: the 10.2 value is set high for convenience, and real diagnosis is based on multiple BP measurements.]

A numerical result closer to that seen for California schools is obtained for a diagnosis of Stage 1 hypertension (90 or above) and a patient DPB reading of 108.. Again the reading is within a margin of error (20 points) of the target but Probability{true DBP at least 90 given measurement of 108} = .94. Are odds of 15 to 1 consistent with the OCRegister assertion that "it's impossible to tell"?



Non-verbal depiction of blood pressure parable. Man (center) has diastolic blood pressure reading 128; doctor (left) heeding the Orange County Register margin of error "has no idea" of whether the man is Stage 3 hypertensive; statistician (right) explains the probability is about .80 or more (depending on the details of the artificial scenario) that the patient's perfectly measured diastolic BP is above 110.

School Examples: Margin of Error vs Probability Calculations

Back to the API context where it is shown that the disconnect between the margin of error and any reasonable probability statement is considerably larger than that shown in the Blood Pressure Parable. Three examples, all telling the same story, follow below. Bottom line: application of the OCRegister margin of error to API scores produces nonsense.

Example 1. OCRegister Hypothetical (8/11, "API's error margin leaves a lot to chance")
Elementary School n=900

	2000	2001		
API	620	640	Growth target for 2001 would be 9	Margin of error
se(API)	8.32	8.32	assume it received GPA award	API 16.3
n	900	900		improvement 21.2

The OCRegister piece, by Ronald Campbell, poses the question "How do you know if the score really increased"? His answer is that the margin of error for improvement (here 21.2 points) indicates "how much the score could have risen by chance alone." His conclusion is "Bottom line: Chance, not improved academics, might be behind that 20 point increase." Anything "might" or "could" happen; in much of life we use probability (formally or informally) to quantify the "mights" and "coulds". Here are the relevant calculations, using the specificied hypothetical school and the actual California 2000-2001 Elementary school data.
Calculations:
P{true change ≤ 0| observed data} = .0189
P{true change ≤ 9| observed data} = .125

The probability exceeds .98 that the true change is positive for this school (and the odds are 7 to 1 that the true improvement exceeded the growth target). But according to the OCRegister this hypothetical school would be a school for which it's "impossible to tell if the school really improved"

(because the 20 point improvement does not exceed their 21.2 margin of error). Most importantly, to link with the follies in the previous part, this school would be counted in their 35% total; that is, such a school is counted as having probability 1 (i.e. certainty) of no real change. What this amounts to is rounding the .0189 probability up to 1.0 (in effect multiplying the actual probability by more than a factor of 50). Pile on enough "might haves" upon "could haves" and then count all of those as certainties, and it's not hard to obtain a dramatically inflated number such as the 1/3 or 35% that OCRegister trumpets. Rounding .02 up to 1.0 does not make for sound public policy for California, nor does it contribute positively to the frail numeracy of California's schoolchildren.

Example 2. CDS 19644516057616 (chosen for similarity to the OCRegister hypothetical school) This Middle School has 3 significant subgroups (SD Hisp Wht). Year-to-year improvement of 16 points is well within the margin of error for improvement of 19.1 points. Improvement is 10 points above the growth target.

	2000	2001		Margin of error
API	685	701	Growth target for 2000 was 6	API 14.7
se(API)	7.496	----	Received GPA award $44,404	improvement 19.1
n	900	1002		

Calculations:

P{true change ≤ 0 | observed data} = .0386

P{true change ≤ 6 | observed data} = .14

Again the probability that true (real) change is positive is very large, above .96, even though the school's improvement is less than the stated margin of error and, therefore, not real according to the OCRegister. The OCRegister calculation for the award programs counts this school as one for which “it’s impossible to tell if the school really improved (OCR 8/16)” and includes this school in the count to obtain their 35% figure, in effect taking the small probability .0386 and rounding it up to 1.0. Also note that the odds are better than 6 to 1 that this school's true change exceeded its growth target of 6.

As a small aside, to link with the content of part 1 of this series, for this school the probability of statistical variability alone producing an award: P{GPA award | no real improvement} is .077. To illustrate "saved by the subgroups" theme discussed in the first part and in the Plan&Preview document, note that this probability would be .226 if awards were based solely on the school API without the additional subgroup criteria.

Example 3. Elementary School 19647336018253 1999-2000

Year-to-year improvement of 31 points is 22 points above it's growth target. Margin of error for improvement is 34.9 points.

	1999	2000		Margin of error
API	616	647	Growth target for 2000 was 9	API 27.8, 25.9
se(API)	14.21	13.2	Received GPA award $40,262	improvement 34.9
n	349	355		

Calculations:

P{true change ≤ 0 | observed data} < .01

P{true change ≤ 9 | observed data} < .04

Thus the probability exceeds 99% that the true improvement is greater than 0, and the odds are better than 25 to 1 that the true improvement exceeded the year 2000 growth target. Yet according to the OCRegister, because the observed improvement of 31 points is less than the margin of error for improvement of 34.9, we are to have "no idea" whether this school actually improved. True, .992 is not 1.0, but the OCRegister is horribly wrong to round this probability down to zero (or more exactly to round .01 up to 1.0 in their counting of “impossible to tell” schools). To take a current event analogy, I believe that if one were told that the probability was above .96 that an equity would beat it's target, many would clamor for the ticker symbol.

One additional aggregate comment from the calculations in part 1. As an example consider the set of the 1999-2000 elementary school GPA award winners. For these 4545 schools over 75% had P{true change > 0 | observed data} above .99, and over 90% had P{true change > 0 | observed data} above .97. Yet the OCRegister applies their margin of error to assert "no idea" of whether many of the schools really improved.

Hopefully these examples will expunge this margin of error from any discourse about the API awards or any other assessment program. Understanding the accuracy (statistical properties) of all aspects of an assessment program is essential; botching this important effort of understanding accuracy must be avoided.

Alternative Award Rules

The OCRegister claimed that in 35% of schools statistical uncertainty is too great to justify an award; although that headline is seen to be far off the mark, it can serve one useful purpose as a segue to the constructive exercise of considering the properties of alternative award rules. For example, suppose a more stringent GPA award rule withheld awards from those schools that are currently eligible but do not exceed their API targets by at least 20 points. That's in the spirit of the OCRegister analysis. Another alternative award rule would apply that 20 point surcharge also to the subgroups (since the subgroups have far more statistical variability than the school scores, it is surprising that OCRegister didn't apply their "margin of error" to those scores as well in its determination of who really improved or earned an award). The table on the following page augments the results in the July Plan and Preview document by adding OCRegister-style rules.

The purpose of these calculations, the details of which are described in the Plan and Preview document, are to illustrate the tradeoffs between false positives and false negatives for various award rules. The two schools shown are the schools used in Plan and Preview. The first two columns of the table are the current GPA rule and the rule used for AB1114 award eligibility (the doubled growth targets, a possible alternative for the GPA awards). The third and fourth columns are the results of OCRegister-style award rules. The short summary is that false positives in GPA awards can be lowered further by more stringent rules, but the cost is a large increase in false negatives (i.e. lower probability of award for a school that really improved). If false positives aren't much of a problem (which is the case if most schools are making good improvement), then reducing those further is not the best policy option. The colored bands in the table show the false positive row (light green), 1 - false negative for "moderate" real improvement (orange), and 1 - false negative for "stronger" improvement (red).

Additional asides

The standard error of the API is so very easy to compute to good approximation; therefore it is very hard to take seriously OCRegister complaints that information has been suppressed. The high school student knows that the standard error of the mean is the standard deviation of the scores divided by the square root of the number of students. Any school official having the student test data could compute such with a spreadsheet program on a Palm Pilot. But the point also is that this standard error doesn't tell you much about the properties of the award programs.

Decile accuracy. The quote from a deputy superintendent "I wouldn't buy a house based on the API" (OCR 8/11) fits in here because the state decile rank is a number that has real estate interest, and the accuracy of the state decile rank does have a relation to the standard error of the API. The quoted statement is correct for a long list of reasons, but that list does not include concerns about the accuracy of the API. In Plan and Preview I show the accuracy of the reported decile rank in terms of the decile accuracy hit-rate = 1 - P{ statistical uncertainty in API score moves the school out of its assigned decile}. Decile accuracy is reasonably good, but the similar schools rank is seen to be far less accurate.

NOTE 1. As all the OCRegister-style calculations are seen to be statistically unsound, the exact details don't much matter. But for completeness, here's a little detail on my attempts to recreate their 1/3 or 35% numbers.
Version 1. If the criteria is set to be 20 points (their mean margin of error) above the API GPA award target for that year's improvement then the proportion passing the criteria (99-2k, 2k-01 cycles) is (3263 + 1785)/(4526 + 3200) = .653.
Version 2. Almost equivalently, if the calculation is done school-by-school, using the criteria 1.96*(standard error for that school's API score) above the API target for that year's improvement then the proportion passing the criteria is (3230 + 1726)/(4526 + 3200) = .641.
Version 3. Furthermore, if the criteria is set in terms of improvement rather than moving past the API school target for GPA award, then a OCRegister-style criteria would be 1.3*1.96*(standard error for that school's API score) above the previous year API, and the proportion passing the criteria is (3357 + 1895)/(4526 + 3200) = .68.

I believe Version 3 is actually a calculation they report. The point to make here is that most any version of the calculation produces numbers reasonably consistent with the OCRegister results.

Note that my analysis files used for these calculations do not include the "small schools" (school-type S) included in the 2000-2001 cycle but not the 1999-2000 award cycle. Those small schools, accounting for about 158 GPA awards in 2000-2001, I think of as a separate issue that may well need to be revisited as small schools are the most technically equivocal category in terms of accuracy (e.g., concerns about False Positives and False Negatives).

Comparing Award Rules: Probabilities of Award Eligibility:

Elementary School Example CDS 19643376011951

n= 350, CA Rank = 5, Sim Rank = 6, API = 613, s.e.(API) = 13.7, Sig Subgroups: SD, Hispanic, White

Incrementation (real improvement)	API	GPA PrAPI&Subgr>Targ1	AB1114 PrAPI&Subgr>Targ2	OCRegister MOE PrAPI-20& Subgr>Targ1	PrAPI-20& Subgr-20>Targ1	
P0	610	0.0655	0.0198	0.0080	0.0028	
Base (I0)	**613**	**0.1002**	**0.0354**	**0.0169**	**0.0036**	False Positive
P1	615	0.1275	0.0513	0.0234	0.0054	
I1	621	0.2446	0.1125	0.0597	0.0196	
P2	624	0.3111	0.1576	0.0849	0.0309	
I2	630	0.4590	0.2787	0.1857	0.0774	
P3	**634**	**0.5321**	**0.3553**	**0.2602**	**0.1180**	1 - False Negative
I3	640	0.6515	0.4832	0.3995	0.1963	
P4	642	0.7136	0.5540	0.4766	0.2588	
I4	**647**	**0.7927**	**0.6609**	**0.5992**	**0.3639**	1 - False Negative
P5	651	0.8639	0.7543	0.7105	0.4752	
I5	658	0.9299	0.8657	0.8566	0.6345	
P6	661	0.9564	0.9097	0.9017	0.7275	
I6	668	0.9832	0.9625	0.9665	0.8647	

High School Example CDS 15635291530708

n= 1115, CA Rank = 5, Sim Rank = 3, API = 609, s.e.(API) = 7.8, Sig Subgroups: SD, African-American, Hispanic, White

Incrementation (real improvement)	API	GPA PrAPI&Subgr>Targ1	AB1114 PrAPI&Subgr>Targ2	OCRegister MOE PrAPI-20& Subgr>Targ1	PrAPI-20& Subgr-20>Targ1	
P0	605	0.0015	0.0000	0.0000	0.0000	
Base (I0)	**609**	**0.0097**	**0.0002**	**0.0002**	**0.0000**	False Positive
P1	613	0.0307	0.0015	0.0002	0.0000	
I1	618	0.1457	0.0165	0.0025	0.0000	
P2	622	0.2700	0.0450	0.0145	0.0002	
I2	626	0.4480	0.1352	0.0525	0.0052	
P3	**629**	**0.5737**	**0.2122**	**0.1047**	**0.0150**	1 - False Negative
I3	634	0.7207	0.3857	0.2432	0.0512	
P4	638	0.8717	0.6002	0.4515	0.1532	
I4	**644**	**0.9555**	**0.8335**	**0.7725**	**0.3917**	1 - False Negative
P5	648	0.9792	0.9180	0.8825	0.5647	
I5	653	0.9935	0.9635	0.9690	0.7792	
P6	655	0.9932	0.9750	0.9830	0.8152	
I6	662	0.9982	0.9925	0.9987	0.9405	