Defensibility of Caveon's Statistics

April 25, 2006
Dennis Maynes
Caveon Test Security

Overview

This document reviews the standards and processes that are followed for producing a Data Forensics analysis. Caveon has followed the general approach that all statistics are based in probability distributions with a specification of the null distribution to facilitate a concise understanding and interpretation of the statistics. Caveon's statistical methodology follows standard statistical practice. Additionally, conservative test techniques are employed to prevent alpha inflation.

Statistics for Individual Tests

Data Forensics employs the following statistics that are computed from individual test results:

- Very similar test responses on different test records with a low probability of occurring by chance. Very similar test responses could indicate that two or more students did not independently answer the test questions.
- Multiple marks on answer sheets that generally occur through smudging the answer sheet or erasing and changing answers. Multiple marks (sometimes referred to as "erasures") are measured by the scanning software. An extreme number of multiple marks could indicate that a student's answer sheet was inappropriately modified.
- A larger score gain in a student's score than would seem reasonable using scores from prior years, or scores from other assessments. An extremely higher-than-expected score gain could indicate that the student's performance does not measure the student's actual knowledge.
- Aberrant or unusual response patterns that indicate the student's performance on the exam is inconsistent with demonstrated knowledge. Inconsistencies, such as students missing very easy questions, while answering the difficult questions correctly, may suggest that the student has had unfair access to portions of the test content.

Very similar test responses are measured by comparing pairs of test answer sheets and computing a statistic that measures agreement between them. The statistic is a multivariate statistic and consists of the number of identical correct and incorrect answers between two answer sheets. The distribution of the statistic is derived under an assumption that the tests are answered independently and follows a generalized trinomial distribution. The tail of the distribution of the test statistic is tested using an appropriate sub-ordering principle[1]. Probabilities are computed from this distribution and reported for

[1] Barnett V. and Lewis T. (1994), "Outliers in Statistical Data," 3rd Edition, p. 269-270.

each pair of tests. The statistic indicates whether the agreement or similarity between the answers is greater than would be expected by chance alone.

The similarity statistic explicitly takes into account student performance differences on the exam. This means that response probabilities for the different answer choices vary for every student depending on the knowledge and abilities of the student. It also means that tests with perfect scores are not detected as highly similar, since the students' performance indicates that both students in such a pair should be responding correctly. The strongest evidence that two tests are highly similar is provided by large number of incorrect identical responses since these are very low probability events.

Multiple marks are detected and counted during the answer sheet processing. The number of multiple marks is considered statistically excessive when the probability of that number of observed multiple marks is extremely low. Data Forensics uses several probability distributions to assess multiple mark counts. The best statistical results are obtained when probabilities of multiple marks are computed for each item that are conditional upon student performance on the exam. The correct small sample statistical distribution is used (i.e., binomial or trinomial) in assessing the evidence of multiple marks. The computed probabilities provide an objective assessment of chance observance of the number of multiples marks on each answer sheet.

A trinomial distribution is used when the scanning procedure categorizes the erasures by wrong-to-right, wrong-to-wrong, right-to-wrong and no erasure. If the scanning procedure only indicates whether an erasure was measured or not, then a binomial distribution is used.

High gain scores are measured by evaluating the change in a student's test performance relative to other or previous assessments. The presence of large numbers of high gain scores within a classroom or school may indicate the occurrence of a testing irregularity. Standard statistical methods based on regression models are used to assess gain scores. Robust regression procedures[2] are used as appropriate to protect against outlier influence in order to obtain an accurate assessment of outliers, under the hypothesis that the regression model is correct.

Probabilities from this model are computed using standardized residuals and standard regression assumptions. In general, one-tailed outlier tests are used for assessing whether the gain scores is unusually high or not.

Test aberrance is measured as a particular statistical inconsistency of the test response pattern, as compared to the Item Response model. Aberrance in a set of test responses occurs when the student's response pattern on some questions is inconsistent with demonstrated knowledge for other test questions on the exam. The simplest example of

[2] Rousseeuw, P.J. and Driessen, K. V. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," American Statistical Association, (http://www.amstat.org/publications/TECH/index.cfm?fuseaction=rousseeuwaug1999)

aberrance is when the student is able to answer difficult questions correctly, but is unable to answer easy questions correctly. In addition to testing irregularities, other atypical behaviors can contribute to aberrance. These other behaviors include fatigue, poor preparation, illness, running out of time, lack of motivation, guessing, differential test preparation (knowing some content well, but not knowing other content), and so forth.

Caveon uses a particular aberrance statistic based on a likelihood ratio test which compares production of the test responses using two ability levels (i.e. values of θ) and the production of test responses using one ability level. Caveon has not derived a statistical distribution for this statistic. Therefore, following standard statistical methodology, a simulation is performed to estimate appropriate thresholds for determining whether a particular test is "extremely" aberrant. Simulation studies of this "bimodality ratio" indicate the statistic is very powerful in detecting cheating that may be exhibited through pre-knowledge which is a reasonable alternative hypothesis.

Statistics for Schools and Classrooms

The general statistical approach followed by Caveon Data Forensics for assessing the "unusualness" of the test result data from a particular school or classroom is based on assuming a binomial or hypergeometric statistical distribution. For these groups of tests the statistical distribution assumes that (1) the particular statistic computed for the individual test (e.g., similarity, aberrance, multiple marks, or high gains) is unrelated to student demographics and performance, and (2) the particular behavior being measured is assumed to be randomly distributed throughout the state. Using these distributions accepted statistical outlier detection methodology is used to assess the probability of observing the test results that have been collected. These probabilities provide an objective assessment of the "unusualness" of the test result data.

Protection against alpha-inflation

When large numbers of statistical hypothesis tests are conducted, by chance alone, many tests should fail. This is a normal result of using probability theory as the basis for accepting or rejecting a statistical hypothesis. Caveon uses the distribution of the maximum order statistic to control the alpha-level across all statistical tests.

This procedure employs a very conservative statistical method. The probability thresholds are so low so that when every school is examined (among all the schools in the state) the chance that *any* school or *even one* school out of all the schools is reported with a statistical anomaly (or as an outlier) is one chance in 100 (i.e., the simultaneous, experiment-wide Type I error is .01). This is very different than stating that the probability of a statistical inconsistency is one in 100, for in that case we would expect to see 1% of the schools reported with statistical inconsistencies by chance alone. Instead we desire to see only .01 statistical inconsistencies by chance alone. In order to satisfy this desire, the probability of observing a statistical inconsistency must be very small (typically less than one chance in 500,000). Conceptually this is the same as inspecting 100 bags containing 5,000 raisins each and only expecting to see one bad raisin in all 100 bags. Restating the concept in terms of this analysis, the threshold is set so that if this analysis were done every year for one hundred years we would expect to see *only one*

detected statistical inconsistency in all 100 years when no testing irregularities are present.

Control of alpha at the experiment-wide level is achieved using the extreme value distribution (i.e., distribution of the maximum order statistic). The typical nominal level of the extreme value distribution is set at .01. This level is chosen such that the maximum order statistic would only exceed the associated critical value 1 time in 100 when the null hypothesis is true.

Technically, the distribution function of the maximal order statistic[3] is used:

$$F_n(y_n) = [F(y_n)]^n$$

If the probability of the maximal order statistic is .01 or less of exceeding the observed statistic, the critical value can be expressed in terms of the original distribution function, as shown algebraically below:

$$P(y_n > c) = 1 - F_n(c) = [1 - F(c)]^n = 1 - \alpha = .99$$

$$\ln(1 - F(c)) = \frac{\ln(1-\alpha)}{n}$$

$$1 - F(c) = \exp\{\frac{\ln(1-\alpha)}{n}\} = \alpha_n$$

Therefore, if the probability of the observed value is less than α_n then the probability of the maximum order statistic exceeding the observed value is .01 or less.

Conclusion

Caveon Data Forensics models are based in fundamental statistical and probability principles. If the assumptions of the models are true, then the results of the analyses have an objective, mathematical interpretation that is consistent with accepted statistical practice.

[3] Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics, Fourth Edition. Macmillian Publishing Co. (1978), pp. 154-161, Section 4.6, "Distributions of Order Statistics."