Combining Statistical Evidence for Increased Power in Detecting Cheating

April 6, 2009

Dennis Maynes
Caveon Test Security

Abstract

A method is presented for combining statistically independent components of probability evidence for improving inferences concerning potential test fraud. The method is general and relies upon a minimal set of assumptions, so long as the probability density functions of the underlying statistics can be accurately approximated. This method has been used at Caveon with very good success and it is simple to apply.

Overview

In the science of forensics, additional items of evidence can strengthen or weaken the inferences that are made concerning the guilt or innocence of an individual. In the same way, inferences made in data forensics (the statistical analysis of test data for detecting test fraud) can be greatly enhanced when more than one piece of evidence is utilized. Because we make probability statements when stating the results of a forensic analysis, we achieve our goal of using multiple pieces of evidence when we probabilistically combine the data forensics statistics. This short paper presents a method for probabilistically combining data forensics evidence. While the context and content of the discussion is applied to test fraud, our method generalizes to all forms of statistical hypothesis testing, subject to the inference rules and assumptions that govern the methods.

Theory

Because we are seeking strong results, we begin with some rather stringent requirements. These are:

1. Each statistic must have a probability density or probability mass function that may be approximated with sufficient accuracy.
2. We may reasonably assume the statistics that we combine are stochastically independent when the null hypothesis (i.e., normal test taking) is true.
3. If the statistical distributions depend upon test-taking performance (i.e., theta or the raw score), conditional distributions may be derived or computed, thus removing the performance dependency.
4. The statistics that are computed must have relevance for the types of test fraud that are being investigated.

Many of the statistics in the literature fail to satisfy one or more of the above conditions. For example, probability density functions have not been published for most person-fit statistics. In the case of ℓ_0, for example, the statistical distribution functionally depends

upon theta, and the normal approximation (creating ℓ_z) is very poor due to large variances for the underlying probability mass functions. The above requirements have impelled Caveon to devise algorithms for appropriately approximating the probability density functions of the utilized statistics. It is beyond the scope of this paper to discuss the computational methods for approximating the probability density functions. In this paper, Caveon may disclose mathematics and methods, but it holds many of the computational algorithms as intellectual property.

The Null Hypothesis

In all data forensics analyses, the null hypothesis is "normal test taking." We define normal test taking as that property or behavior which is collectively established by the test-taking population. Our analyses proceed by searching for anomalous data or outliers from the model as defined by the body of data that has been collected.

We remark that some forms of test-taking (i.e., accommodations) are rare and decidedly not normal. We are not surprised when these data are detected as being anomalous by the forensics procedures. Thus, the analyst must always remember that an observed outlier may have a plausible explanation due to some unusual test-taking behavior or an assumption that has been violated. Examples of such behaviors are:

1. A teacher instructs students to "mark" items in the answer sheet when unsure about the answer and then return and change the answer later. This results in a spurious number of erasures.
2. An individual runs out of time on the test and selects the same answer choice (e.g., "C") for all remaining items. These item selections do not follow the test-taking model and their probabilities do not depend upon theta.
3. The form code for an answer sheet is coded incorrectly and the answer sheet is scored using the wrong answer key.

At Caveon, we have adopted Bock's nominal response model [Bock, 1972] for estimating response probabilities because it has great generality, encompassing the 1-parameter logistic, 2-parameter logistic, graded-response, and McMaster's partial credit models. We use regularization techniques to provide stable estimation under low sample size situations. The nominal response model is shown as Equation (1).

Equation (1)

$$p(x_{ij} = k \mid \theta_i) = \frac{e^{a_{jk}\theta_i + g_{jk}}}{\sum_m e^{a_{jm}\theta_i + g_{jm}}}$$

In Equation (1), the subscript "i" refers to the individual, the subscript "j" refers to the item, the subscript "k" refers to the selected response, and the subscript "m" is used to sum across all modeled responses. The variable "x" contains the subscript of the selected response, the variable "a" is the multiplier for the modeled response (acts in the role of the discrimination parameter), the variable "g" is the constant for the modeled response (acts in the role of the difficulty parameter), and the parameter θ is the individual's level of ability. We have found this model to provide very good estimates of response probabilities. In general, values of item parameters are important for computing

probabilities in data forensics applications, but they need not be interpreted and analyzed as is done with psychometric analyses. We also are not usually in a position to accept or reject items. Once the test results are "in," the test and its items are not changeable.

The distributions of many of the statistics used in data forensics analysis may be computed or approximated using Equation (1). When we use the nominal response model we also assume local independence—a standard assumption of Item Response Theory (IRT). The distributions of similarity (or answer-copying) statistics, aberrance (or person-fit) statistics, and conditional raw scores may be approximated using the probabilities from Equation (1). Distributions of erasure statistics (e.g., the number of wrong-to-right answer changes) may be modeled using binomial, trinomial, or logistic distributions depending upon the amount of information that is available. In each of these situations, a statistic is chosen (such as the number of identical answers or the number of answer changes) and the data are then evaluated for extremeness.

Directionality (The Alternative Hypothesis)

After one or more statistics have been selected and computed we may test them for extremeness. Without loss of generality, we can write the upper-tail probability of the statistic as is shown in Equation (2).

Equation (2)

$$p(s >= S \mid H_0) = \int_S^{\infty} f(s)ds$$

or

$$p(s >= S \mid H_0) = \sum_{k=S}^{\infty} p(k)$$

If a lower-tail test is desired then the direction of the integration or summation should be modified appropriately. It is nearly always the case that the alternative hypothesis is expressed in the form of a directional test because larger or smaller values of the statistic provide stronger evidence of potential test fraud. It is possible to perform two-tail tests with an appropriate modification to the integrations and summations of Equation (2). Rather than proceed forward with all of the combinations of upper-tail, lower-tail, two-tail and discrete or continuous probability functions, we rewrite Equation (2) using the distribution function and trust the reader in making the appropriate modification when these equations are implemented. This is shown in Equation (3).

Equation (3)

$$p(s >= S \mid H_0) = 1 - F(S)$$

We have found it very convenient to express our methods for combining statistical evidence probabilistically using Equation (3) because the random variable y, shown in Equation 4, has two important statistical properties.

Equation (4)

$$y = F(S)$$
$$y \sim U(0,1)$$
$$u = -2\ln y$$
$$u \sim \mathrm{X}_2^2$$

The distribution function transformation of Equation (4) yields a uniformly distributed random variable. The logarithmic transformation yields a Chi-Square distributed random variable with two degrees of freedom.

We acknowledge that when the random variables are discretely distributed (i.e., only having a finite number of values with spacing) Equation (4) does not hold strictly. The probability as represented by the distribution function (y in Equation (4)) is not uniformly distributed because the distribution function is not a continuous function; it is a step function. As a result, statistical granularity will add noise to the approximations that are used. This is not much different than the situation in current psychometric practice because most of the computed statistics and quantities are discretely distributed random variables. After recognizing this source of noise, we state that our methods follow the correct forms and we urge the practitioner to err on the side of being conservative when interpreting the analyses.

Testing a Single Outlier

Introductory texts in statistics usually introduce the concept of testing an outlier by computing the number of standard deviations the value differs from the mean. This is a reasonable didactic device, but it is disastrous when applied in practice [Barnett and Lewis, 1994][1]. The proper understanding is achieved when we realize that the observation will be rejected as an outlier only when it is the most extreme (i.e., maximum or minimum) value in the entire sample; and, when as an extreme value, the observation is anomalous.

The notion of maximum and minimum is relevant whether a single value is inspected or whether all values are inspected (such as occurs in data mining). Because we always inspect and suspect the most extreme value in the data set, the hypothesized population distribution is no longer appropriate for testing the extreme value. Instead, we must use the distribution function of the maximal (or minimal) order statistic. The distribution functions for these order statistics are found in Equation (5) [Hogg and Craig, 1978].

Equation (5)

$$F_n(y_n) = [F(y_n)]^n$$
$$F_1(y_1) = 1 - [1 - F(y_1)]^n$$

[1] Barnett and Lewis comment concerning the common practice of excluding observations based on three standard deviations: "This highlights two general defects… [1] [failing] to distinguish between population and sample variance… [2] erroneously based on the distributional behaviour of a random sample value rather than on that of an appropriate sample *extreme*."

We note that Equation (5) was derived under the assumption that all of the observations in the sample were independently and identically distributed. There are at least two additional observations that should be made in connection with this equation. First, the equation provides implicit protection against alpha inflation. The outlier test is performed simultaneously upon all observations. Second, the functional forms of the equations for the minimal and maximal order statistics are asymmetrically related.

We find it convenient to represent tail probabilities using logarithms. There are several reasons for this:

1. Extremely small probabilities require many leading zeros to print, while the associated logarithm counts of the number of leading zeros and is more conveniently displayed.
2. Manipulations involving Equation (5) require logarithmic transformations. Therefore, logarithms are a natural representation.
3. The logarithms of the tail probabilities are distributed with a Chi-Square distribution, assuming the data are representative samples from the estimated probability density functions.
4. We shall see that the logarithmic transformation provides for convenient and simple computations.

We emphasize from Equation (5) that the sample size is always relevant and that *n* is a parameter in the probability density function of the extreme statistics.

Testing a Multivariate Outlier

The idea of combining statistical evidence to increase detection power, of necessity, leads us into the area of multivariate statistics. We are interested in two kinds of multivariate outliers: (1) a single outlier from among a group of observations, and (2) an outlier in which all of the observations are extreme. The two general tests of hypothesis that we employ are:

H_0(any): None of the observations are extreme
H_1(any): At least one of the observations is extreme,

and

H_0(all): Not all of the observations are extreme
H_1(all): All of the observations are extreme.

We have found it convenient to express the alternative hypotheses using operations from propositional logic: "OR" (H_1(any)) and "AND" (H_1 (all)). Equation (5) provides us with the necessary theory. The "OR" hypothesis is tested using the upper tail of the maximum order statistic because we reject the null hypothesis if any observation is extreme. The "AND" hypothesis is tested using the upper tail of the minimum order statistic because we reject the null hypothesis only when all of the observations are extreme. Thus, the order statistics provide us with a foundation for testing the multivariate outliers.

The statistical methodology of testing multivariate outliers for these two hypotheses depends upon two assumptions:

1. The estimated probability density function is a very good approximation of the actual population density function. There are issues regarding sampling variation, as noted by Barnett and Lewis. Thus, this assumption does not hold strictly in practice, but we have found as long as we have stable approximations the results will be reasonable. If this assumption holds, we can transform the statistic into a uniformly distributed random variable (i.e., by using the distribution function of the statistic and Equation (5)).
2. The statistics are independently distributed when the null hypothesis is true. If the assumption of independence does not hold, Equation (5) may not provide a proper approximation for the distributions of the maximal and minimal order statistics. Thus, we need to realize if and in what manner the assumption of independence is violated in order to understand how our inferences may be affected.

Even though we have referred to the "OR" and "AND" operations of propositional logic, we stress that this method of combining statistical quantities does not lead to a system where the associative and distributive properties of propositional logic hold. Thus, the testing of a set of hypotheses depends upon the order in which the hypotheses are formulated and tested.

Computation of Extreme Probabilities

We have remarked that it is convenient to use logarithms of tail probabilities for representing extreme probabilities. At Caveon, we have adopted the term "*index*" to denote a probability that has been converted to a base-ten logarithm. The relation between the index and the probability is expressed in Equation (6).

Equation (6)

$$p = 10^{-index}$$

An index value of 6 means one chance in one million of natural occurrence, assuming the null hypothesis is true. An index value of 6.5 means the probability is equal to one chance in $10^{6.5}$, or one chance in 3,162,277. We never believe that the computed probabilities are precise, but we do presume they are reasonable estimates of the actual probabilities.

The probability for the upper tail of the maximum order statistic may be approximated by Equation (7) when the largest index value in the group of statistics being tested is large (e.g., when the index is greater than 4.0).

Equation (7)

$$
\begin{aligned}
p(Y >= y_n) &= 1 - F_n(y_n) \\
&= 1 - [F(y_n)]^n \\
x &= 10^{-I_n} \\
Log_{10}[p(Y >= y_n)] &= Log_{10}(1 - (1-x)^n) \\
(1-x)^n &= \sum_{k=0}^{n} \binom{n}{k} 1^{n-k} (-x)^k = \left(1 - \frac{n}{1}x + \frac{n(n-1)}{2!}x^2 - \frac{n(n-1)(n-2)}{3!}x^3 + \ldots\right) \\
Log_{10}[p(Y >= y_n)] &= Log_{10}\left(\frac{n}{1}x - \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 - \frac{n(n-1)(n-2)(n-3)}{4!}x^4 + \ldots\right) \\
&= Log_{10}\left(nx\left[1 - \frac{(n-1)}{2!}x^2 + \frac{(n-1)(n-2)}{3!}x^3 - \frac{(n-1)(n-2)(n-3)}{4!}x^4 + \ldots\right]\right) \\
&= -I_n + Log_{10}(n) + Log_{10}\left(1 - \frac{(n-1)}{2!}x + \frac{(n-1)(n-2)}{3!}x^2 - \frac{(n-1)(n-2)(n-3)}{4!}x^3 + \ldots\right) \\
&\approx -I_n + Log_{10}(n) \quad (\text{x} << .0001;\ 4\ \text{digits of accuracy}) \\
I_{\max} &\approx I_n - Log_{10}(n)
\end{aligned}
$$

The series expansion in Equation (7) is the binomial expansion and one reference that provides this expansion is "Handbook of Mathematical Functions" [Abramowitz and Stegun, 1970]. If x is very small and if this computation is the final result--not to be used in subsequent computations, the last term can be safely ignored.

In Equation (7), Y is the random variable representing the maximum order statistic and y_n is the observed value of the maximum order statistic. The function $F_n(y)$ is the distribution function of the maximum order statistic and the function F(y) is the distribution function of the population from which the maximum order statistic is drawn[2]. The value I_n is the index value associated with the maximum order statistic and it will be the largest observed index value. The value I_{max} is the index value for the desired probability of the multivariate observation under the null hypothesis.

Using Equation (5), the probability for the upper tail of the minimum order statistic is computable from the smallest index value. This is shown in Equation (8).

[2] After inverting the probability the distribution function F() is the distribution function of a uniformly distributed random variable, x, and is equal to x.

Equation (8)

$$
\begin{aligned}
p(Y >= y_1) &= 1 - F_1(y_1) \\
&= 1 - \left\{ - [1 - F(y_1)]^n \right\} \\
&= [1 - F(y_1)]^n \\
Log_{10}[p(Y >= y_1)] &= Log_{10}\left(\left\{ - (1 - 10^{-I_1}) \right\}^n \right) \\
&= nLog_{10}(10^{-I_1}) \\
&= -nI_1 \\
I_{\min} &= nI_1
\end{aligned}
$$

In Equation (8), Y is the random variable representing the minimum order statistic and y_1 is the observed value of the minimum order statistic. The function $F_1(y)$ is the distribution function of the minimum order statistic and the function F(y) is the distribution function of the population from which the minimum order statistic is drawn[3]. The value I_1 is the index value associated with the minimum order statistic and it will be the smallest observed index value. The value I_{min} is the index value for the desired probability of the multivariate observation under the null hypothesis.

Applications and Examples

This section applies the presented theory to three examples and illustrates the power of the methods.

Case Study 1: Cross-form Answer Copying

Story: A professor decided to administer two forms of the final exam with the items scrambled on one of the forms. Upon grading the tests, the professor noticed that one student received a very low score. The professor suspected cheating had occurred, so she graded that particular test with the answer key for the other form. The student received a much higher score (but lower than customary performance) using the answer key for the wrong form. When approached and asked for an explanation, the student asserted that no mistake had been made. The student alleged that the answer sheet was marked properly and the calculated score was correct. Caveon was asked to provide statistical evidence for or against these assertions.

Overview of Analysis: Because two forms were used, a cross-form answer-copying analysis was performed. The responses for each test were compared to every other test, using item order and not item identifier to align the answers (Caveon was not provided with the unscrambling arrangement of the items). An extreme pair of similar test instances was found, with a test instance being administered for each form. The pair of extremely similar tests was reanalyzed assuming that a form coding error had been made (We were not told if the answers were marked on the form or on a scan sheet). The pair

[3] See Footnote 2.

of tests was still extremely similar. These analyses were followed by evaluating test-taking performance under each scenario.

Cross-Form Analysis: We will refer to the test takers by their numbers within the data set, #32 and #121. The statistic used for the cross-form analysis is the number of matching answers. The distribution of this statistic can be computed using Equation (1) and assuming the two test papers were answered independently. For example, the probability of a matching answer for a particular question is shown in Equation (9).

Equation (9)

$$p(x_{32,j} = x_{121,j}) = \sum_k p_{form1}(x_{32,j} = k \mid \theta_{32}) p_{form2}(x_{121,j} = k \mid \theta_{121})$$

The number of matches (our data forensics statistic) follows a generalized binomial distribution and the probability mass function for this distribution is computable using standard recurrence equations [Tucker, 1980; Feller, 1968].

For the pair of extremely similar tests, a summary of performance on the 72 item test is provided in Table 1:

Table 1: Comparison of Scores – Case Study 1

	#32-Form1	**#121-Form2**
Score on Given Form	16	52
Score on Other Form	45	9
Expected Score on Other Form	21.8	48.1
Std. Dev. of Other Score	3.0	3.7
Index Value of Other Score	13.1	0.0

In Table 1, we see that Test Taker #32 scored very low (16 of 72) on the administered form and higher (45 of 72) on the alternative form, just as the professor noted. Likewise Test Taker #121 scored at a reasonable level (52 of 72) on the administered form and very poorly (9 of 72) on the alternative form. Because Form 2 was a scrambled version of Form 1, only 7 key values were present in the same sequential positions between the two forms. This count is about 2.25 standard deviations lower than we would expect with a random shuffling of the answer key. A simulation of a randomly generated answer key (with 5 choices) on 72 items repeated 100,000 times gave an average of 14.8 key values in common with a standard deviation of 3.4.

The expected values and standard deviations assume that the performance on the administered form is correct, but that the answer key from the other form is the correct answer key. We will use this statistic as we combine evidence below. The score distribution is computed by directly expanding the generalized binomial probabilities as provided by the nominal response model.

Table 2, below, summarizes the probability evidence of matching. The distribution of the same-form matching statistic (in this case Form2) is derived using the same methodology

as that used for Equation (9). A derivation of this statistic and some discussion of its computation may be found in van Der Linden and Sotaridona [2006].

Table 2: Count of Matching Responses – Case Study 1

	Cross-Form	Form2
Matching Responses	65	65
Expected Number of Matches	25.3	40.4
Standard Deviation of Matches	3.7	3.8
Index Value of Match	25.8	11.0

From Table 2, we see that the estimated probability these two individuals would have agreed upon 65 answers (aligned by position, but not form code) or greater is one in $10^{25.8}$. This is very extreme and it is expected. We also note that if there had been a form coding or answer key error, the probability that these two individuals would have agreed on 65 answers or greater is one in 10^{11}. This is still a very extreme probability. We note that the number of matches is approximately 6.3 standard deviations (using a continuity correction factor) above the expected number of matches.

The last component of statistical evidence that we computed was a differential score statistic. We computed the score difference for each test between the 65 items for which the test takers had the same response and the remaining seven items where the responses differed, assuming that Form2 was the correct form for both tests. These data are shown in Table 3.

Table 3: Score Difference Analysis – Case Study 1

	#32	#121
Score on Form	45	52
Score on 7 Non-matching Items	0	7
Score on 65 Matching Items	45	45
Observed Difference	-45	-38
Expected Difference	-34.4	-39.9
Std. Dev. Of Difference	2.06	1.72
Index Value of Difference	5.34	0.0
Proportion correct (7)	0.0	1.00
Proportion correct (65)	0.703	0.703

We note that Test Taker #32 did not answer any of the seven items correctly where the two test takers disagreed (assuming that Form 2 was the correct form), while Test Taker #121 answered all seven items correctly. The distribution of the difference statistic in Table 3 is derived from the joint probability distribution of the two sub-scores (computed using the nominal response model) and then conditioning upon the total score. It is beyond the scope of this paper to present this derivation.

Hypothesis 1: We assume that Test Taker #32 is telling the truth. Therefore, the probability of the observed number of agreed answers with Test Taker #121 is less than one in 10^{25} if the tests were answered independently (and if the nominal response model holds). We reject the null hypothesis and state that the tests were not taken independently. However, for didactic purposes, we form the composite hypothesis to state: *The tests were taken independently **and** when the item responses were scored in comparison to the "other answer key" the scores were consistent with expectations*. Using our rules of inference, we compute the probabilities of this hypothesis for each test taker as follows:

$P(H_0|\text{Test Taker \#32}) = 1-[1-\min(1-10^{-25.8},1-10^{-13.1})]^2 = 10^{-26.2}$
$P(H_0|\text{Test Taker \#121})=1-[1-\min(1-10^{-25.8},1-10^{-0.0})]^2 = 1$

Therefore, we reject the composite null hypothesis for Test Taker #32 and do not reject the composite null hypothesis for Test Taker #121. By combining the statistical evidence we find no inconsistency with Test Taker #121's performance. In this instance, we are led to believe that Test Taker #121 was not involved in the behavior that created the extremely similar tests. We also note that the extreme statistic for Test Taker #32 became slightly more extreme.

Hypothesis 2: We assume that an incongruous mistake was made and that Test Taker #32 was actually administered Form #2. Under this scenario, the probability of the number of agreed answers with Test Taker #121 is less than one in 10^{11} if the tests are answered independently (and if the nominal response model holds). We reject the null hypothesis and state that the tests were not taken independently. We now form the composite hypothesis to state: *The tests were taken independently **and** test scores on the "non-matching" items were consistent with test scores on the "matching" items*. Using our rules of inference, we compute the probabilities of this hypothesis for each test taker as follows:

$P(H_0|\text{Test Taker \#32}) = 1-[1-\min(1-10^{-11.0},1-10^{-5.34})]^2 = 10^{-10.7}$
$P(H_0|\text{Test Taker \#121})=1-[1-\min(1-10^{-11.0},1-10^{-0.0})]^2 = 1$

Therefore, we reject the composite null hypothesis for Test Taker #32 and do not reject the composite null hypothesis for Test Taker #121. By combining the statistical evidence we find no inconsistency with Test Taker #121's performance. Again, we are led to believe that Test Taker #121 was not involved in the behavior that created the extremely similar tests. Even though this result is similar to that found when we tested hypothesis #1, this second test appears to rule out any complicity on the part of Test Taker #121. We reason as follows: Test Taker #121 may have assumed that Test Taker #32 was given the same form, but the performance for Test Taker #121 did not change for the non-matching items, while it did change for Test Taker #32. If Test Taker #121 had been involved, we would have expected to see no change for Test Taker #32. Finally, we note that the extreme statistic for Test Taker #32 became slightly less extreme under the composite hypothesis.

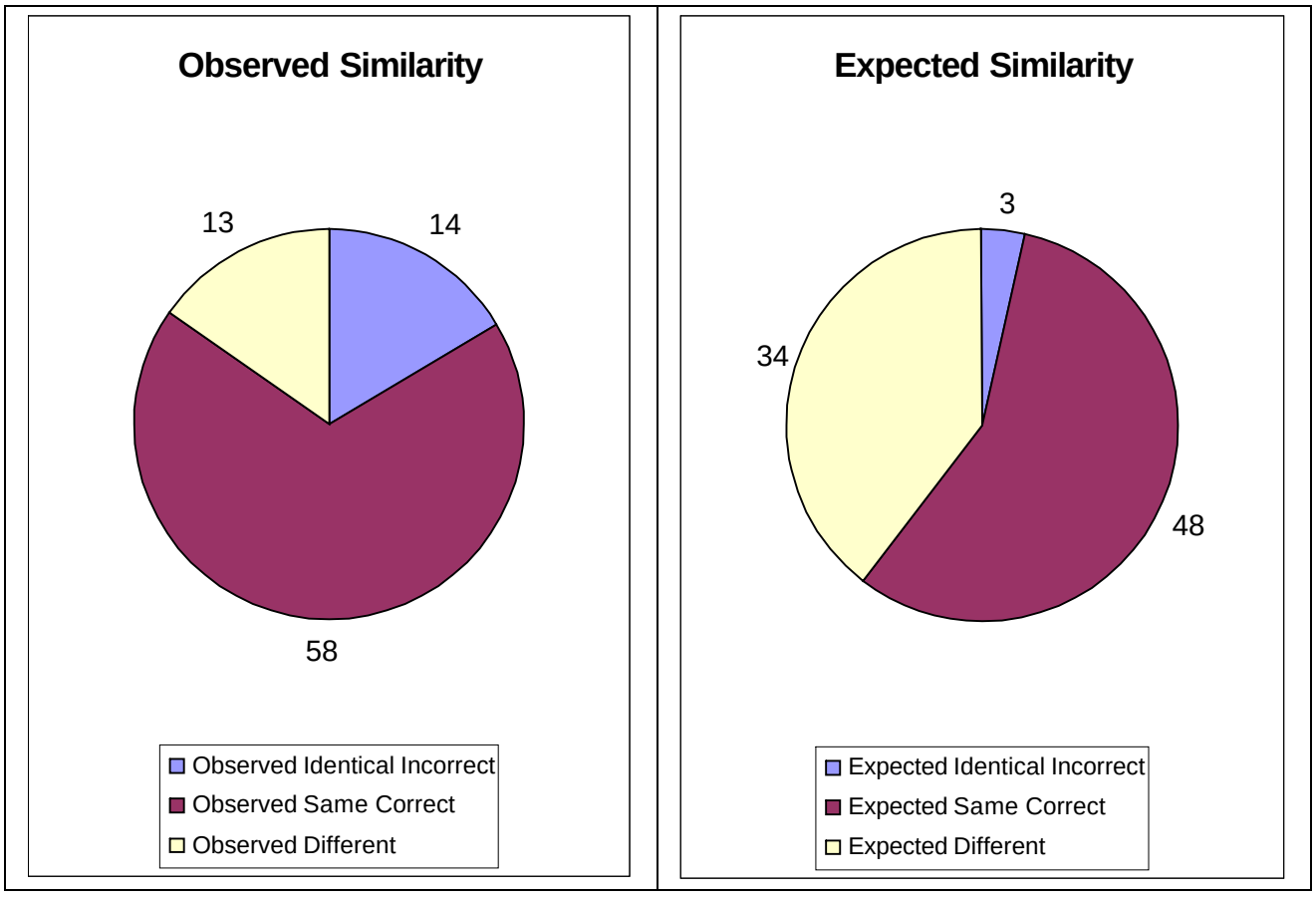
Case Study 2: Answer Copying as an Explanation for Erasures

Story: Scanning software routinely analyzes and counts erasures on bubble sheets in the public schools. One of the bubble sheets found with an extreme number of wrong-to-right erasures intrigued us. We performed a deeper analysis in order to understand the behavior associated with these erasures. The data forensics results revealed that the test taker with a large number of erasures also had an extremely similar test with another test taker. For convenience in describing the tests, we will refer to the two test takers with the extremely similar tests as BL and AS.

Overview of Analysis: Only one test form was administered. Therefore, we used our most powerful answer-copying statistic, the M4 Similarity statistic. The statistic counts the number of identical correct and identical incorrect answers separately and then combines the bivariate data by computing the tail probability of the "third-tail" in the trinomial distribution. As additional explainers of the detected oddness, we computed Guttman's G-statistic and the difference score statistic between the items where the two test takers agreed and those where they disagreed.

The M4 Similarity statistic is depicted in Figure 1, below.

Figure 1: Observed and Expected Agreement – Case Study 2



Two pie charts are shown in Figure 1, above. The pie chart on the left shows the observed number of identical answers between the two tests. The pie chart on the right shows the expected number of identical answers between the two tests, conditioned upon theta for each test.

The index value for the observed amount of similarity, assuming independent test taking, was 9.25. The test for AS was observed with no erasures on any of the 85 items. The test for BL was observed with 21 wrong-to-right erasures and 12 any-to-wrong erasures for a total of 33 erasures (or multiple marks) on the 85 items.

We have augmented this analysis with a modified version of Guttman's G-statistic and with a score difference statistic. These data are summarized in Table 4.

Table 4: Summary of Supporting Statistics – Case Study 2

	BL	**AS**
Guttman's G Statistic	17.28	11.31
Expected value of G	11.95	9.91
Std. Dev. of G	.99	.73
Index of G Statistic	5.29	1.11
Raw Score	58	71
Score on 72 Matching Items	58	58
Score on 13 Non-matching Items	0	13
Observed Difference	58	45
Expected Difference	40.66	49.64
Std. Dev. Of Difference	4.06	3.15
Index Value of Difference	3.83	0.0
Proportion correct (72)	0.81	0.81
Proportion correct (13)	0.00	1.00

As will be seen, testing the composite hypothesis will provide additional evidence for this case.

Hypothesis 1: Given the extreme index of 9.25 on the similarity, we have rejected the hypothesis that the tests were taken independently. We can expand the null hypothesis to state: *The tests were taken independently **and** no advantage in test score was gained for any items that were answered the same.*

$P(H_0|BL) = 1-[1-\min(1-10^{-9.25},1-10^{-3.83})]^2 = 10^{-7.66}$
$P(H_0|AS)=1-[1-\min(1-10^{-9.25},1-10^{-0.0})]^2 = 1$

We used the index value of the observed score difference as the additional piece of statistical evidence. We see that we would reject the composite hypothesis for Test Taker BL but we would not reject the composite hypothesis for Test Taker AS.

Hypothesis 2: Looking at the data, we naturally wonder about "aberrance" and consistent test taking. We have implemented a modified statistic based on Guttman's G statistic. This statistic is the sum of the ranks (or, in our case the sum of the scaled ranks) for the responses that were selected and conditioned upon the total test score. The expanded composite null hypothesis is: *The tests were taken independently **and** no advantage in test score was gained for any items that were answered the same **and** the test responses were given consistently as indicated by probabilities of selection.*

$P(H_0|BL) = 1-[1-\min(1-10^{-9.25},1-10^{-3.83},1-10^{-5.29})]^3 = 10^{-11.49}$
$P(H_0|AS)=1-[1-\min(1-10^{-9.25},1-10^{-0.0},1-10^{-1.11})]^3 = 1$

We used the index value of the observed score difference and the index value of the G statistic as the additional pieces of statistical evidence. We see that we would reject the composite hypothesis for Test Taker BL but we would not reject the composite hypothesis for Test Taker AS.

Hypothesis 3: Presuming that the answer-copying and advantage gained was the result of erasing, we seek to include these data into the analysis. Table 5 summarizes the erasure counts for Test Taker BL and their association to the items that matched.

Table 5: Association between Erased and Matching Answers – Case Study 2

	Wrong-to-right Erasure	**Anything-to-wrong Erasure**	**Total**
Matching Item	21	6	27
Non-matching Item	0	6	6
Total	21	12	33

Using Fisher's exact test, we can test the statistical independence between erased and matching items. We perform a one-sided probability computation[4], resulting in an estimated probability value of 0.0008 or an index value of 3.08.

We extend Hypothesis 2 by including this additional piece of evidence.

$P(H_0|BL) = 1-[1-\min(1-10^{-9.25},1-10^{-3.83},1-10^{-5.29},1-10^{-3.08})]^4 = 10^{-12.32}$

In summary, we conclude that Test Taker BL gained an advantage by looking at the answers on the test paper of AS and changing answers in a way that is inconsistent with BL's actual knowledge. The additional components of statistical evidence have allowed us to strengthen the evidence of this assertion and to suggest that Test Taker AS be eliminated as a suspect in the security breach of the exam. The final computed probability is 1,000 times smaller than the initial extreme probability computed using the M4 Similarity statistic.

[4] This computation is performed using the Hypergeometric Distribution.

Case Study 3: Detection of a Cram School

Story: Data forensics analysis is essentially a data-mining operation, and actual knowledge of security infractions is usually confirmed after anomalous results are detected. A detected site with rapidly answered tests and high pass rates left us puzzled. After-the-fact investigation confirmed that the detected site was a cram school where the exam content was being disclosed.

Overview of Analysis: We routinely compute several detection statistics when analyzing a set of data. These statistics are computed for each individual test instance. Group data are detected as being extreme or anomalous when the upper tails of the statistical distribution for the group are fatter or heavier than expected (as compared to the entire population). We use a liberal threshold for placing a test instance in the tail of the distribution and refer to these as "marginal" tests. While not perfect, this approach works quite well in distinguishing extreme group-based behaviors as opposed to extreme individual behaviors (which are detected using extreme statistics as demonstrated in Case Studies 1 and 2). In the current situation, the computed statistics were:

1. Guttman's G Statistic – The sum of the ranks of responses ordered by probability
2. Latency Aberrance – A multivariate-based statistic using three dimensions of response latency: fast response, widely varying response, and inconsistent response as predicted given response rate, item complexity, and estimated ability.
3. Fast-High Statistic – A combination statistic which is constructed in a similar matter to the statistics in Case Studies 1 and 2 for detecting high-scoring tests completed in unusually short test sessions.
4. Rapid Response Statistic – A counting statistic which detects with large numbers of items answered very quickly—quickness is determined using empirically derived thresholds.
5. M4 Similarity – A statistic which compares each test with every other test in order to detect collusion and answer copying on the exam.
6. Volatile Scores – A gain score statistic that analyzes test-retest results and attempts to find retests with larger gains than predicted using a regression model.
7. Perfect Tests – This counting statistic indicates whether the maximum score was achieved on a test.
8. Identical Tests – This is an extreme example of a similar test. Identical tests are not always detected by the M4 Similarity statistic under high-scoring conditions when nearly all items are answered correctly.
9. Retake Violations – This counting statistic indicates whether a retest attempt is in violation of the examination policy.

We compare the rates of these statistics with baseline or population rates and for most of these statistics we compare pass rates within the groups between the marginal tests and the remaining tests.

A large amount of data was computed for a large number of test sites. Rather than show all of these data, the data for the site of interest are summarized in Table 6.

Table 6: Summary Statistics for Case Study 3

Statistic	Baseline (all tests)	Site of Interest
Number of Tests	1152	78
Pass Rate	0.69	0.86
Pass Index	0.00	3.51
Response Aberrance Rate	0.08	0.13
Pass Rate With Response Aberrance	0.37	0.70
Pass Rate Without Response Aberrance	0.72	0.88
Response Aberrance Difference Index	0.00	0.01
Response Aberrance Rate Index		1.10
Latency Aberrance Rate	0.22	0.35
Pass Rate With Latency Aberrance	0.76	0.85
Pass Rate Without Latency Aberrance	0.67	0.86
Latency Aberrance Difference Index	2.31	0.16
Latency Aberrance Rate Index		2.14
Fast-High Rate	0.04	0.08
Fast-High Rate Index		0.97
Rapid Response Rate	0.17	0.51
Pass Rate With Rapid Response	0.73	0.98
Pass Rate Without Rapid Response	0.68	0.74
Rapid Response Difference Index	1.08	2.59
Rapid Response Rate Index		12.45
M4 Similarity Rate	0.04	0.03
Pass Rate With M4 Similarity	0.74	1.00
Pass Rate Without M4 Similarity	0.69	0.86
M4 Similarity Difference Index	0.57	0.13
M4 Similarity Rate Index		0.10
Volatile Scores Rate	0.09	0.00
Pass Rate With Volatile Scores	1.00	0.00
Pass Rate Without Volatile Scores	0.66	0.00
Volatile Scores Difference Index	2.17	0.00
Volatile Scores Rate Index		0.00
Identical Rate	0.00	0.00
Identical Rate Index		0.00
Perfect Rate	0.01	0.00
Perfect Rate Index		0.00
Retake Violations Rate	0.04	0.00
Pass Rate With Retake Violations	0.43	0.00
Pass Rate Without Retake Violations	0.68	0.00
Retake Violations Difference Index	0.02	0.00
Retake Violations Rate Index		0.00

In Table 6, rows that pertain to each statistic have been alternately shaded in gray to help the reader find relevant table entries. The index values of interest for the detected cram-school test site have been highlighted in gold. An explanation of each statistic was provided immediately preceding Table 6. Our analysis method splits the data into "marginal" tests and the remaining tests. The rates of marginal tests are reported for the

baseline and the site of interest on the lines where the row heading ends in the word "Rate." As an example, the line 'M4 Similarity Rate' provides the proportion of marginally similar tests for the baseline and for the site of interest. The pass rates for the marginal tests are compared with the remaining tests and these have the prefix labels "Pass Rate With" and "Pass Rate Without," in the Table. There are two kinds of statistical tests being performed in this analysis. We compare the pass rates for the marginal tests and for the remaining tests within the group using an upper-tail test and these comparisons have the suffix label "Difference Index." We compare the observed rate of marginal tests in the group baseline and these comparisons have the suffix label "Rate Index."

Hypothesis 1: We ask the question whether *any* of the statistical observations in Table 6 has an extreme index value. Thus, our statistical hypothesis is stated: *All rates are not larger than the population rates **and** all pass rate differences for marginal test instances for each statistic are not larger than the pass rate for the remaining tests.*

$P(H_0|\text{Site of Interest}) = 1-[\max(1-10^{-0.01},1-10^{-1.10},\ldots,1-10^{-0.0})]^{15} = 10^{-11.27}$

We note that the only extreme value in Table 6 is the index value of 12.45 associated with the Rapid Response Rate Index. This site did have a very high pass rate of 86% compared with the baseline rate of 69%. Given that there were 78 test instances administered at this site, the one-sided test of significance on the pass rates yielded a probability value of 0.0003. We know that pass rates vary from site to site and we should not expect them to all be the same. However, given the high pass rate and the high rate of rapidly answered tests, we thought these data were significant when we first analyzed them. At that time, we made a few remarks shown below.

> Since this behavior [i.e., high pass rates and rapidly answered tests] seems to be limited to a single site and predominately one exam [title], it could indicate local exposure of examination content.
>
> The difference between the pass rate for "rapid responders" and the pass rate for the "remaining" test takers was quite large (98% versus 74%). We estimate that 10 test takers[5] in this group passed the test illicitly. Twenty four test takers finished [the exam] in less than 20 minutes with a 100% pass rate (4 of these were finished in less than 10 minutes).
>
> [Maynes, Parr, Mulkey 2008]

After investigation, the above observations were confirmed. The site was engaged in improperly coaching and disclosing the exam content.

[5] We estimate this by assuming that the rapid response distribution is appropriately stabilized for performance and being performance-neutral the test takers with rapid responses should have the same pass rate as the remaining test takers: 10 = (78 tests) x (.51282 RR rate) x (.975 pass rate for RR - .73684 pass rate for remaining test takers).

Because several index values were used and compared, the extreme value of 12.45 was lowered slightly to the reported value of 11.27. This procedure may remind the reader of the Bonferroni adjustment to protect against alpha inflation. And, indeed the Bonferroni adjustment returns exactly the same result. That adjustment multiplies the extreme upper-tail probability value by the number of elements examined, and is:

Equation (10)

$$Bonferroni = -\log_{10}\left(15 \times 10^{-12.45}\right) = 12.45 - \log_{10}(15) = 11.27$$

The Bonferroni procedure works because the power series function for values extremely close to one is practically linear. And, it can be approximated very accurately with just one linear term (see Equation (7)).

Conclusions

A method has been presented for combining data forensics evidence statistically. This method requires carefully creating uni-directional arguments and properly modeling the distributions of the underlying statistics. The method provides principled guidance for examining and testing propositions concerning test fraud in the data. At Caveon, we have used this and similar methods effectively for strengthening the security of exams.

References

Abramowitz, M. and Stegun, I. (1970). Handbook of Mathematical Functions, Ninth Printing, Dover Publications, Inc. New York, New York.

Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data, Third Edition, John Wiley & Sons, Ltd. West Sussex, England, pp. 30-31.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *46*, 443-459.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Volume I, Third Edition, John Wiley and Sons, Inc., New York, New York. pp. 264-270.

Hogg, R. V. and Craig, A. T. (1978). Introduction to Mathematical Statistics, Fourth Edition. MacMillian Publishing Co. New York, New York, pp. 154-161.

Maynes, D. D., Parr, R., and Mulkey, J. (2008). Unpublished data forensics report, dated August 13, 2008. [For reasons of confidentiality a complete citation is not permitted.]

Tucker, A. (1980). Applied Combinatorics, John Wiley and Sons, Inc., New York, New York. pp. 76-94.

van der Linden W. J. and Sotaridona L. (2006). "Detecting Answer Copying When the Regular Response Process Follows a Known Response Model", Journal of Educational and Behavioral Statistics, Fall 2006, Vol. 31, No. 3, pp. 283–304

