LA-UR-01-3666
Approved for public release;
distribution is unlimited.

Title: **Type I and II Error Associated with Verification and Confirmation Measurements**

Author(s): Brian G. Scott

Submitted to:

http://lib-www.lanl.gov/la-pubs/00796208.pdf

Type I and II Error Associated with Verification and Confirmation Measurements

Brian G. Scott
Los Alamos National Laboratory
Decision Applications Division
Probabilistic Risk and Hazard Analysis Group

Abstract

DOE M 474.1 states that "For Category I and II items, the acceptance/rejection criteria for verification measurements and, where possible, for confirmation measurements must be based on the standard deviation for the measurement method under operating conditions." Determination of the acceptance/rejection criteria for confirmation and verification measurements may involve null and alternative hypothesis testing. The specific values used for computing the appropriate test statistics are dependent on the desired type I and II errors. If hypothesis testing is applied to a verification measurement (such as verification measurement signal interpretation equals nuclear material book value), then the successful application of hypothesis testing and accurate determination of type I and type II error is dependent on the existing statistical assumptions. This paper discusses some common statistical assumptions that may be used in support of confirmation and verification measurements. The implications of these assumptions on type I and II error are briefly analyzed. Suggestions for validating the desired type I and II error for confirmation and verification measurements are provided. This paper, by providing an introduction to some common statistical assumptions and the associated implications on error, aids safeguards professionals in their ability to determine the validity on the statistical models that support their verification and confirmation measurement programs.

1. Introduction

Safeguards measurements can be of two categories: confirmatory and verification. Verification measurements assess the quantity of a material with the intent of comparing the measured value to the book value. Confirmatory measurements determine whether an attribute of the item is present. Measurements can provide confidence that a particular scenario did or did not occur.

Scenarios of diversion of an item are postulated in figure 1. The diversion scenario can either be a full or partial diversion of the item contents. Replacement of the item contents can either occur or not. As the below figure illustrates, different measurement types provide confidence that different postulated scenarios did not occur.

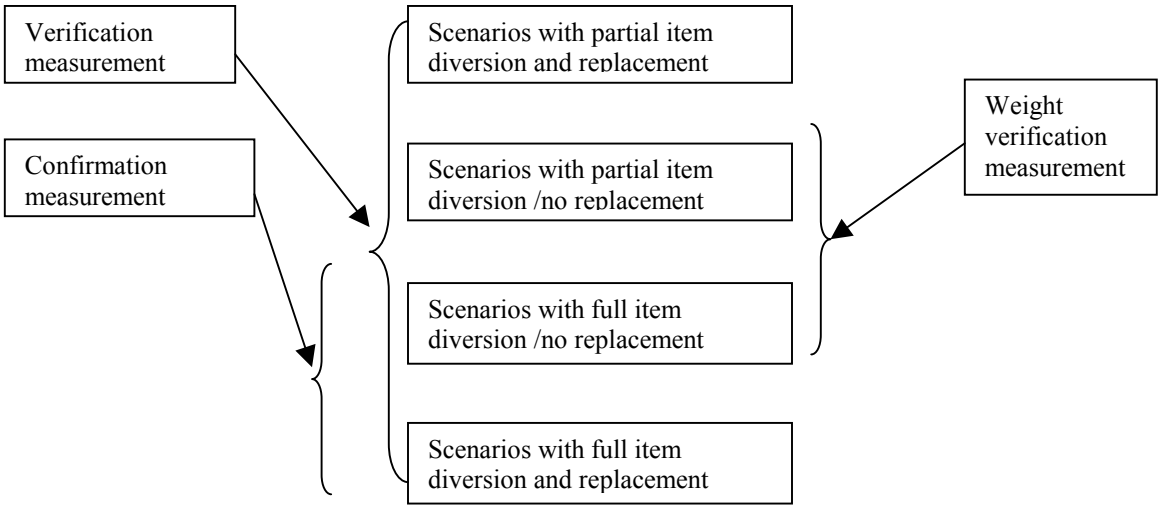

Figure 1. Measurements types and related diversion scenarios

The above figure illustrates that a sufficient verification measurement can provide confidence that any of the four diversion scenarios did not occur. From the diverter's point-of-view, scenarios with partial diversion can be considered more difficult than full diversion, and scenarios that include replacement can be considered more difficult than scenarios without replacement. If the most difficult scenario set (scenarios with partial item diversion and replacement), is not likely to occur (possibly due to the item configuration or the material control and protection afforded the item), then a combination of a confirmatory measurement and a weight verification will provide confidence that the other three postulated scenarios did not occur. For example, if a Pu rod is determined to not have plausible partial with replacement diversion scenario then a isotopic confirmatory measurement combined with a second confirmatory measurement (i.e., length or weight measurement) would be sufficient to provide confidence that a diversion did not occur.

Items with TIDs (tamper indicating devices) do not require a verification measurement, while non-TID'd item may not require a verification measurement. While the above statement is afforded a regulatory basis, the associated technical rationale is that a partial diversion scenario (with or with out replacement) is not likely for a TID'd item. Occasionally, an item without a TID will be afforded sufficient material control and protection for one to conclude that some diversion scenarios are not likely. A TID is not the only way to reduce the likelihood of a particular diversion scenario. Various material and personnel control techniques can also afford equivalent protection. A verification on a non-TID'd item verifies that the amount of material currently within the container is the same as the value previously recorded. From the verification measurement, one (within statistical uncertainty) is able to conclude that the NM material has not been partially substituted with a different material.

Hypothesis Testing

As the above discussion indicates, confirmatory and verification measurements provide two different methods to detect diversion. While the decision criteria for both measurement methods can be modeled using the null hypothesis, different conclusions for the two methods may arise.

The null hypothesis test used to determine if a verification measurement value V_0 is statistically different from the book value B_0 can be formulated as

$$V_0 = B_0. \qquad (1)$$

The hypothesis testing procedures specify the values for the test statistic that lead to the rejection of the hypothesis. At Los Alamos National Laboratory (LANL), if V_0 is different from B_0 by more than two sigma (determined by historical information, tabulated as Precision and Accuracy Values or PAVs), then the verification measurement is determined to be statistically different from the book value measurement. The statistical difference is further investigated to determine whether the difference is a "false alarm" or a true difference that may indicated a diversion of material.

The null hypothesis test is predicted to have an error of the first kind α (false positives) 4.55% of the time (assuming two standard deviations). The probability of accepting the null hypothesis (false negative) when a true a difference exists is dependent on the true difference and the variance. For example, assuming that the book value possesses negligible variance, a true difference between B_0 and V_0 of two sigma will be detected 50% of the time and a true difference of three sigma will be detected about 84% of the time.

The null hypothesis may be constructed differently for a confirmation measurement that determines that absence or presence of an isotope. In this case the null hypothesis may be constructed as follows:

$$C_0 = 0 \qquad (1)$$

where C_0 represents the confirmatory measurement.

Verification measurements, while usually more sophisticated than confirmatory measurements, are relatively straightforward in correlating an instrument false positive (rejection of the null hypothesis, $B_0 \neq V_0$) to a diversion false positive. The conclusions drawn from false positive/false negative for a confirmatory instrument response may be different than those drawn from a verification measurement.

Confirmatory measurements: false negative of a full diversion

Assuming a normal distribution of background counts and an instrument false alarm setpoint at 2 sigma, the false alarm rate is 4.55%. therefore, the probability of not detecting a diversion is

$$P(\text{measurement false positive}) = P(\text{false negative of a diversion}) = 0.0455 \qquad (2)$$

In addition, the probability of a diversion false negative (instrument false positive) can increase due to a fluctuating a background. While a general statistical practice is to control the instrument false positive rate to an acceptable false alarm rate, consideration should be given to decreasing the instrument false alarm rate in areas susceptible to fluctuating background (such a small areas that require SNM movement during inventory).

Confirmatory measurements: false positive of a diversion

The probability of a false negative measurement (giving a false positive indication of diversion) is dependent on a target quantity providing a less than three times background measurement. The true false negative rate is dependent on the background, and the number of counts recorded by the instrument. The consequence of a false positive of a diversion is usually some immediate action, such as a re-measurement.

Confirmatory measurements: false positive of a partial diversion

If partial diversion, without replacement is added to the postulated full diversion scenarios, then a confirmatory isotopic and weight measurement can provide confidence that a diversion did not occur. Frequently these, "double confirmatory" measurements are performed when a verification measurement can not be performed (such as when the material is in a glovebox or the material is "non-amenable to measurement"). In this case, two null hypotheses would be constructed: the confirmatory null hypothesis discussed above and the weight null hypothesis where the measured weight equals the book value. If the isotopic measurement is performed through a glove box, the measurement may susceptible to diversion false negatives for cases where the background radiation is not sufficiently taken into account.

Statistical Assumptions

The precision values from the PAV tables are determined by calculating the variance of a set of historical data generated by the same measurement technique. The accuracy values provided in the PAV tables are determined by calculating the mean difference of a set of verification measurement values from the book value. If the verification technique is different than the measurement technique used for the book value, then a precision and an accuracy value can be assigned to the null hypothesis test. If the book and verification techniques are the same then only a precision value is assigned the null hypothesis test.

The above is relatively straightforward; however, the application and subsequent conclusions that can be inferred from statistical tests is sensitive to the assumptions made and the actual model applied. Some of these considerations are described below.

The formula used (and described in the 2000 LANL MC&A Plan) to determine if a verification measurement fails to validate the original measurement is

$$|X_1 - X_2| > |\beta_1 X_1| + |\beta_2 X_2| + 2\sqrt{\{(\sigma_1 X_1)^2 + (\sigma_2 X_2)^2\}} \qquad (3)$$

where subscript 1 represents the book-associated values and subscript 2 represents the verification-associated values, X_1 is the book value and X_2 is the verification value, and β and σ represent bias and precision respectively.

Since the bias depends on a comparison between two different measurements, and, the book value is usually a value that is considered a "gold standard" or a value without comparison" equation 2 can be written as

$$|X_1 - X_2| > |\beta_{2/1} X_2| + 2\sqrt{\{(\sigma_1 X_1)^2 + (\sigma_2 X_2)^2\}} \qquad (4)$$

where $\beta_{2/1}$ is the relative accuracy between measure method 1 and measurement method 2.

Some of the assumptions used for the above hypothesis testing are:

- Variance of the book and verification measurements are known (not estimated).
- Any bias is a fixed (not random).
- The distribution of the measurement is normal (not skewed, no large tails etc.).

The validity of the above assumptions is dependent on the measurements performed.

Assumption 1: Variance is known

Assumption 1 is frequently an approximation of the true environment. Occasionally, for well-characterized, well-controlled measurements, measurement error is sufficiently controlled such that the theoretical Poisson (counting statistics) error becomes the dominant error and can be used. For almost all other cases, measurement associated variances are calculated. Calculated variances imply that the *t* distribution should be used; however, for large enough sample sizes, the t distribution approximates to the normal distribution. For n=5 and assuming a normal distribution with a two sigma failure decision level, a 10% false alarm rate would be predicted, rather than a 4.55% distribution predicted if the variance was known.

The consequences of the assumption of a normal distribution, when a *t* distribution would be more appropriate, would be an increase of the false alarm rate for that particulate sample set. A model review concerning assumption 1, would focus on the number of data points used to determine the random error component. If the number of points is less than 20, the expectation for false alarms should increase or a *t* statistic may be applied.

Assumption 2: Fixed Bias

Assumption 2 assumes the bias to be fixed, not random. Fixed systematic errors (fixed bias) traditionally linearly add to the total error, while precision error (standard deviation) is added in quadature. Bias listed in the PAV tables is calculated at LANL by a mean estimate of the ratio between the assayed value and the book value. The bias may possess a random component. Since, measurement bias may be due in part to sample matrix, this systematic random component may be due to the imperfect information known about a set of measurement items.

If the bias is assumed to be fixed and estimated about the mean,, but actually possesses a random component as described above, then entire sets of samples with similar matrices could be under- or over-estimated. The measurement safeguards significance of this over- or under-estimation is dependent on the number of samples and the ratio of the true bias over the estimated bias.

In order to bound the error due to the random component of bias, a statistical model review may concentrate on review the distribution of the bias elements that comprise the bias estimate. The variance of this bias estimate should be known. In addition, any deviations from normality should be noted. The inclusion of a random component term for systematic error may be considered.

Assumption 3: The distribution of the measurement is normal

Frequently, the assumption of normality is not reviewed. Statistical discussions concerning normality frequently reference the Central Limit Theorem or large-sample theory in support of Gaussian distribution assumptions. Referencing the Central Limit Theorem is a acceptable analytical method when independent errors of approximately the same order are added. Briefly, the central limit theorem can be paraphrased as follows: the averages of samples from any distribution will approach a normal distribution as the number of samples in the average increase. From a practical point-of-view, at least three assumptions are necessary to apply the central limit theorem for safeguards measurements: (1) the errors from the distribution are additive-no strong multiplicative effect exists between errors, (2) the sample average set comprises of more than one error, and (3) the errors are more-or-less evenly weighted.

While many observed phenomena are in accordance with the normal distribution, exceptions from normality, such as many environmental-based distributions that approximate lognormal distributions, exist. Lognormal distributions can imply a multiplicative effect between errors, or that one error component has overwhelmed the other error components. Well-characterized safeguards measurements can probably sufficiently characterized by assuming the errors are normally distributed. If, however, the measurements error could be significantly changed by a single factor (such as self-shielding, alpha-n reaction, sample matrix, etc) then further evaluation of the error distributions could be warranted.

Statistical Model Validation

All statistically based models should be validated. Global validation tests have the advantage of speed, perceived cost effectiveness and providing a “big picture” assessment. More exact validation tests such a inter-laboratory cross checks and specific statistical testing will probably be more expensive, labor-intensive and more narrowly focused.

Since a global test, such as false alarm testing can appear attractive, a brief discussion concerning the application of false alarm validation follows.

Depending on the statistical model used, the results of a false alarm verification may provide vague information. If the statistical model assumes the components of error to be random and fixed, the false alarms generated would be the result of the random component.

A false alarm is defined as the incorrect rejection of the null hypothesis. If a false alarm occurred due to random error, then upon re-measurement with the same measurement method, the item will likely satisfy the statistical condition. If the error is systematic, then re-measurement of the item may not be sufficient to satisfy the statistical condition and the null hypothesis. Occasionally, a book value will not equal the verification measurement for a certain method and matrix, no matter how many times the item is re-measured on the same instrument. LANL-based observations of repeated observations on the same instrument indicate that "true random errors" (i.e. re-measurement on the same instrument) do not cause the majority false alarms.

If the null hypothesis has been rejected for an item measurement, the item is usually re-measured by a different measurement method. If the null hypothesis is accepted after the re-measurement by a different method, it may appear plausible that the initial item measurement should be categorized as a "false alarm. However, false alarms are only modeled based on the random component; therefore, it is possible to generate an acceptable false alarm rate that has no relationship to the statistical model described.

If no random bias exists (only fixed systematic error is present), model validation using false alarms should still be approached with caution, especially for large measurement data sets. Large measurement sets, can be comprised of sub-sample sets that are well understood and those that are poorly understood. Experience indicates that the initial error provided by subject matter-experts is usually an underestimation of the true error, while later-estimated error can be an overestimate for the majority of samples (due to skewed distributions). It is possible to obtain a reasonable false alarm rate, but have a distribution of that rate unacceptable. For example, consider a sample set that is 50% overestimated in error and 50% underestimated in error. The overestimated section provides no false alarms and the underestimated provides a false alarm rate of 10%. The average false alarm rate for the entire sample set would be an acceptable 5%, while the two sub-sample sets probably possess unacceptable false alarm rates.

Conclusion

Statistical assumptions and practices tied to a MC&A Program should be reviewed periodically. Practices such as controlling the instrument type I error to 95% can lead to higher than desired non-detection of diversion for confirmatory measurements. In addition, if the "go-no-go" nature of confirmatory measurements leads to neglecting background fluctuations, the diversion non-detection can also increase.

Statistical models applied to relatively large, diverse safeguards-related measurement data set possess inherent challenges. The major challenges can include the understanding of the model assumptions, the application of cost-effective model validation techniques and the application of programmatic broad-scope solutions to a diverse data set. Periodic

review of the statistical models applied, the data set used to generate errors and the adequacy of validation testing appear to be a fundamental part of a safeguards statistical quality control program.

References

[1] See Jaech, John L., "Statistical Analysis of Measurement Errors" 1985 John Wiley and Sons, Inc and Jaech, John L "Statistical Methods in Nuclear Material Control", 1973 National Technical Information Center, for appropriate disclaimers associated with large-sample theory assumptions.