

Section 8

Statistical Techniques

Statistics are used in metrology to summarize experimental data, to provide the basis for assessing its quality, and to provide a basis for making probabilistic decisions in its use. The essential basic statistical information for describing a simple data set is:

The mean of the sample,	\bar{x}
The standard deviation of the sample,	s
The number of individuals in the sample,	n

If the set is a random sample of the population from which it was derived, if the measurement process is in statistical control, and if all of the observations are independent of one another, then s is an estimate of the population standard deviation, σ , and \bar{x} is an unbiased estimate of the mean, μ .

The population consists of all possible measurements that could have been made under the test conditions for a stable test sample. In this regard, the metrologist must be aware that any changes in the measurement system (known or unknown) could possibly result in significant changes in its operational characteristics, and, hence the values of the mean and standard deviation. Whenever there is doubt, statistical tests should be made to determine the significance of any apparent differences and whether statistics should be combined (pooled).

The following discussion reviews some useful statistical techniques for interpreting measurement data. In presenting this information, it is assumed that the reader is already familiar with basic statistical concepts. For a detailed discussion of the following techniques and others not presented here, it is recommended that the reader consult NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, January 2018 (or NBS Handbook 91 - Experimental Statistics, by Mary G. Natrella, (1963), reprinted 1966. Handbook 91 contains comprehensive statistical tables from which many of the tables contained in Section 9 of this publication were taken.

8.1 Estimation of Standard Deviation from a Series of Measurements on a Given Object

Given n measurements $x_1, x_2, x_3, \dots, x_n$

Mean:
$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}$$

Standard deviation estimate:
$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

The estimate, s , is based on $n-1$ degrees of freedom.

8.2 Estimation of Standard Deviation from the Differences of k Sets of Duplicate Measurements

Given k differences of duplicate measurements, $d_1, d_2, d_3, \dots, d_k$, a useful formula for estimating the standard deviation is:

$$s_d = \frac{\sum d_i^2}{2k} \text{ where } s_d \text{ is based on } k \text{ degrees of freedom.}$$

Note that $d_i = \bar{x}'_i - \bar{x}''_i$, for example.

The values d_1, d_2 etc., may be differences of duplicate measurements of the same sample (or object) at various times, or they may be the differences of duplicate measurements of several similar samples (or objects). Note: It is recommended to calculate the standard deviation for each set of duplicates and then to calculate a pooled standard deviation whenever possible.

8.3 Estimation of Standard Deviation from the Average Range of Several Sets of Measurements

The range, R , is defined as the difference between the largest and smallest values in a set of measurements.

Given $R_1, R_2, R_3, \dots, R_k$

Mean:
$$\bar{R} = \frac{(R_1 + R_2 + R_3 + \dots + R_k)}{k}$$

Standard deviation can be estimated by the formula:
$$s_R = \frac{\bar{R}}{d_2^*}$$

The value of d_2^* will depend on the number of sets of measurements used to calculate s_R , and on the number of measurements in each set, i.e., 2 for duplicates, 3 for triplicates, etc. Consult a table such as Table 9.1 for the appropriate value of d_2^* to use. The effective number of degrees of freedom for s_R is in the table. Note: It is recommended to calculate the standard deviation for each set of replicates and then to calculate a pooled standard deviation whenever possible.

8.4 Pooling Estimates of Standard Deviations

Estimates of the standard deviation obtained at several times may be combined (pooled) to obtain a better estimate based upon more degrees of freedom if F-tests demonstrate that variation is statistically similar. The following equation may be used for this purpose of pooling:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + \dots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \dots + (n_k - 1)}} \text{ where}$$

s_p will be based on $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \dots + (n_k - 1)$ degrees of freedom.

8.5 "Within" and "Between" Standard Deviation

See the NIST/SEMATECH e-Handbook of Statistical Methods, Section 2.3.3.3.4. Calculation of standard deviations for 1,1,1,1 design, for detailed descriptions of within and between standard deviations. <http://www.itl.nist.gov/div898/handbook/mpc/section3/mpc3334.htm>. See NISTIR 5672 for mass calibration applications.

8.6 Confidence Interval for the Mean

The estimation of the confidence interval for the mean of n measurements is one of the most frequently used statistical calculations. The formula used will depend on whether the population standard deviation, σ , is known or whether it is estimated on the basis of measurements of a sample(s) of the population. (See also section 8.16 to determine the uncertainty of the mean value for a calibration process.)

8.6.1 Using Population Standard Deviation, σ

Strictly speaking, σ , is never known for a measurement process. However, the formula for use in such a case is:

$$\bar{x} \pm \frac{z\sigma}{\sqrt{n}}$$

Table 1. Variables for equation in 8.6.1.

Variable	Description
\bar{x}	sample mean
s	known standard deviation
n	number of measurements of sample
z	standard normal variate, depending on the confidence level desired

For 95 % confidence $z = 1.960$; for 99.7 % confidence $z = 3.0$.

For other confidence levels, see Table 9.2

8.6.2 Using Estimate of Standard Deviation, s

In the usual situation, s is known, based on ν degrees of freedom and the formula for use is:

$$\bar{x} \pm \frac{ts}{\sqrt{n}}$$

Table 2. Variables for equation in 8.6.2.

Variable	Description
\bar{x}	sample mean
s	estimate of standard deviation
n	number of measurements on which the mean is based
t	Student's t value, based on the confidence level desired and the ν degrees of freedom associated with s (see Table 9.3).

Note that $t \rightarrow z$ as $n \rightarrow \infty$. For many practical purposes, the standard deviation may be considered as known when estimated by at least 30 degrees of freedom.

8.7 Confidence Interval for σ

The standard deviation, σ , is ordinarily not known but is, instead, an estimated value based on a limited number of measurements, using procedures such as have been described above. Such estimates may be pooled, as appropriate, to obtain better estimates. In any case, the uncertainty of the estimated value of the standard deviation may be of interest and can be expressed in the form of a confidence interval, computed as indicated below.

The interval is asymmetrical because the standard deviation is ordinarily underestimated when small numbers of measurements are involved due to the fact that large deviations occur infrequently in a limited measurement process. Indeed, it is the general experience of metrologists that a few measurements appear to be more precise than they really are.

The basic information required to compute the interval is an estimate of the standard deviation, s , and the number of degrees of freedom on which the estimate is based. The relationships to use are:

Lower limit	B_{LS}
Upper limit	B_{US}
Interval	B_{LS} to B_{US}

The values for B_L and B_U depend upon the confidence level and degrees of freedom associated with s . Values for use in calculating the confidence level are given in Table 9.7. A more extensive table (Table A-20) is available in NBS Handbook 91.

8.8 Statistical Tolerance Intervals

Statistical tolerance intervals define the bounds within which a percentage of the population is expected to lie with a given level of confidence. For example, one may wish to define the limits within which 95 % of measurements would be expected to lie with a 95 % confidence of being correct. The interval is symmetrical and is computed using the expression

$$\bar{x} \pm ks$$

where k depends on three things

Table 3. Variables used to select k .

Variable	Description
p	the proportion or percentage of the individual measurements to be included
γ	the confidence coefficient to be associated with the interval
n	the number of measurements on which the estimate, s , is based

Table 9.6 may be used to obtain values for k for frequently desired values of γ and p . A more extensive table is Table A-6 found in NBS Handbook 91.

8.9 Comparing Estimates of a Standard Deviation (F-Test)

The F-test may be used to decide whether there is sufficient reason to believe that two estimates of a standard deviation differ significantly. The ratio of the variances (square of the standard

deviation) is calculated and compared with tabulated values. Unless the computed ratio is larger than the tabulated value, there is no reason to believe that the observed standard deviations are significantly different.

The F-ratio is calculated using the generic equation:

$$F = \frac{s^2}{s^2}$$

The critical value of F depends on the significance level chosen for the decision (test) and the number of degrees of freedom associated with each standard deviation.

Three cases using the F-test are considered here.

- 8.9.1 Evaluation of two samples to determine whether variability is statistically the same.

$$F = \frac{s_L^2}{s_s^2}$$

In this case, the larger of the two standard deviations is placed in the numerator (sL) and the smaller of the two in the denominator (ss).

Table 9.4 contains critical values for F at the 95 % level of confidence. The tabulated values of F are not expected to be exceeded with 95 % confidence on the basis of chance alone. As an example, if both the numerator and the denominator values for s were each based on 9 degrees of freedom, an F value no larger than 4.03 is expected with 95 % confidence, due to the uncertainties of the s values, themselves. Table A-5 of NBS Handbook 91 [19] contains values for F for other confidence levels.

This F-test is useful for comparing the precision of methods, equipment, laboratories, or metrologists, for example. An inspection of Table 9.4 shows that when either of the values of s is based on a small number of degrees of freedom, the F value is large. Consequently, the significance of decisions based on small changes in precision can be supported statistically only by a relatively large number of measurements. If such changes are suspected, but the data requirement is difficult to meet, the decision may need to be made on the basis of information about the measurement process itself.

This F-test is also useful for deciding whether estimates of the standard deviation made at various times differ significantly. Such questions need to be answered when deciding on whether to revise control limits of a control chart, for example.

- 8.9.2 Evaluation of observed within-process standard deviation compared to the accepted/pooled within process standard deviation to determine whether the process has degraded.

$$F = \frac{s_{w \text{ Observed}}^2}{s_{w \text{ Accepted}}^2}$$

In this case, the observed standard deviation of a measurement process is compared to the accepted standard deviation of the process that has been gathered and pooled over time. Table 9.12 provides the statistical limits at 95 % confidence for this statistic. If the F-test demonstrates that the observed standard deviation agrees with the accepted value to the extent that the values are statistically similar, the process is considered to remain stable.

This F-test is useful for integrating a measurement assurance check on the process into the measurement procedure before proceeding to report calibration results. This F-test is integrated into SOP 5 (Section 3.4), SOP 28, and weighing designs as presented in NBS Technical Note 952, NISTIR 5672, and the catalog of designs shown in the NIST/SEMATECH e-Handbook of Statistical Methods.

8.9.3 Evaluation of standard deviations from replicate measurement results.

When the laboratory performs replicate measurements as a part of each procedure, it may be valuable to evaluate the standard deviation of the data pool for a measurement process. The within-process standard deviation is calculated for each set of replicate measurements in a data pool.

Once all of within process standard deviations, s_w , are calculated, they are analyzed using the F-max test. The F-max test uses the equation:

$$F = \frac{s_{Max}^2}{s_{Min}^2}$$

where s signifies the standard deviation of each data set being compared, the subscript “Max” refers to the larger (maximum) standard deviation in the data pool and the subscript “Min” refers to the smallest standard deviation in the data pool. Equation scenario 1 in Section 8.9.1 is identical to the one provided here, but as presented in that case, it is used when comparing two samples rather than sets of replicate data. Table 9.13 is used to evaluate the level of significance for the F-max test, taking into consideration the number of replicate samples. When the ratio of the variances is close to 1, further analysis is not required.

This version of the F-test is useful when comparing many results from calibrations that use replicate measurements to determine a mean value, or several processes with similar equipment and standard deviations, especially in situations where check standards are not practical or feasible. For example, large prover volume calibrations of differing volumes that have similar within-process standard deviations, large mass weight carts where replicate measurements are made per SOP 33, or even multiple balances within a laboratory that are used for the same procedures. See Section 8.13 for more information on evaluation and use of the F-max test.

8.10 Comparing a Set of Measurements with a Given Value

The question may arise as to whether a measured value agrees or significantly disagrees with a stated value for the measured object. The evaluation can be based on whether or not the confidence interval for the measured value encompasses the stated value. The confidence interval is calculated using the expression:

$$\bar{x} \pm \frac{ts}{\sqrt{n}}$$

as previously described in Section 8.6. In using this expression, n represents the number of measurements used to calculate the mean, \bar{x} , and t depends on the degrees of freedom, ν , associated with s and the confidence level needed when making the decision. Note that one may use historical data for estimating s , such as a control chart for example, in which case ν will represent the degrees of freedom associated with establishment of the control limits and may be considerably larger than $n-1$.

8.11 Comparing Two Sets of Measurements with Regard to Their Means (Two-sample t-test with 1. Equal variation and 2. Unequal variation)

This discussion is concerned with deciding whether the means of two measured values, A and B, are in agreement. The data sets used for this purpose may consist of the following:

$$\begin{array}{cc} \bar{x}_A & \bar{x}_B \\ s_A & s_B \\ n_A & n_B \end{array}$$

The first question to be resolved is whether s_A and s_B can be considered to be different estimates of the same standard deviation or whether they do, indeed, differ. An F-test may be used for this purpose. However, it will be recalled that this is not sensitive to small real differences, so the decision may need to be based on physical considerations, such as the known stability of the measurement process, for example.

8.11.1 Case I – Equal variation

Confirming (or assuming) that s_A and s_B are not significantly different, they are pooled, as already described (but repeated here for convenience) and used to calculate a confidence interval for the difference of the means. If this is larger than the observed difference, there is no reason to believe that the means differ. The steps to follow when making the calculation described above are:

Step 1. Choose α , the level of significance for the test.

Step 2. Calculate the pooled estimate of the standard deviation, s_p :

$$s_p = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}}$$

s_p will be estimated with $n_A + n_B - 2$ degrees of freedom.

Step 3. Calculate the respective variances of the means:

$$v_A = \frac{s_A^2}{n_A} \text{ and } v_B = \frac{s_B^2}{n_B}$$

Step 4. Calculate the uncertainty of $|X_A - X_B| = \Delta$:

$$U_{\Delta} = t\sqrt{(V_A + V_B)}$$

using a value for t based on $\frac{\alpha}{2}$ and $\nu = n_A + n_B - 2$.

Step 5. Compare U_{Δ} with Δ .

If $U_{\Delta} \geq \Delta$, there is no reason to believe that Δ is significant at the level of confidence chosen.

8.11.2 Case II – Unequal variation

Confirming (or assuming) that s_A and s_B are significantly different, their individual values are used to calculate U_{Δ} as outlined below.

Step 1. Choose α , the level of significance for the test.

$$v_A = \frac{s_A^2}{n_A} \text{ and } v_B = \frac{s_B^2}{n_B}$$

Step 2. Calculate the respective variances of the means:

Step 3. Calculate the uncertainty of $|X_A - X_B| = \Delta$:

$$U_{\Delta} = t\sqrt{(v_A + v_B)}$$

using a value for t based on $\frac{\alpha}{2}$ and f , the effective number of degrees of freedom calculated as described in Step 4.

Step 4. Calculate f , the effective number of degrees of freedom using the equation below or the Welch-Satterthwaite formula given in NIST Technical Note 1297 or the Guide to the Expression of Uncertainty in Measurement:

$$f = \left(\frac{\left(\frac{v_A^2}{f_A} + \frac{v_B^2}{f_B} \right)}{\frac{v_A^2}{f_A} + \frac{v_B^2}{f_B}} \right)$$

Step 5. Compare U_{Δ} with Δ .

If $U_{\Delta} \geq \Delta$, there is no reason to believe that Δ is significant at the level of confidence chosen.

One goal of these assessments is to determine if the difference between two mean values is less than the combined variation of the processes, keeping in mind that it does not evaluate the full uncertainty from a calibration process.

$$\text{Is } |\bar{x}_A - \bar{x}_B| < t \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} ?$$

Note: Another common way to view this test and equation is as follows (as presented, this equation may be used for both Case I and Case II).

$$\frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \leq t_{table}$$

8.12 Use of Random Numbers

Conducting operations in random sequences can avert problems of bias that might stem from a particular order of operations. For example, in the measurement of a series of items, it might be difficult to determine whether systematic trends in the measured values were due to differences in the items or to measurement system drift unless the items were measured in random order.

Use of tables of random numbers is a convenient means for randomizing measurement operations. The operations, test objects, and other matters requiring randomization may be assigned serial numbers. The order of selection is then determined by use of a random number table, as described below. When the number of operations or test items is less than 100, a table such as Table 9.11, reproduced from NIST Handbook 91 [19], may be used conveniently. One may start from any arbitrarily selected position in the table and proceed from it in any pre-determined arbitrary manner. If the first number encountered is not that of one of any item, ignore it and proceed until a valid match is encountered. This becomes the first item in the sequence. Continuing in the same manner, items are selected in the sequence in which their serial numbers are encountered ignoring the repetition of previously identified items. The procedure is continued until all items have been randomly selected.

As an example, select 10 specimens (numbered 01 to 10) in random order. Start from a randomly selected place, say column 2, row 5 of Table 9.11. Proceed from this point along the table as one would read a book. The starting number is 14, which is not usable. The first useful number encountered is 08, the next 03, and so on. Using the procedure described above, the following random order was found:

Specimen No.									
08	03	09	05	06	02	07	10	04	01
Order									
1	2	3	4	5	6	7	8	9	10

8.13 Estimating a Pooled Standard Deviation for Multiple Sets of Replicate Measurements

When uncertainty estimates require the use of process variability and degrees of freedom from replicate measurements where check standards are not feasible, and where reported measurement results are mean values from replicate measurements, the following approaches may be applied.

Step 1. First calculate the within process standard deviation for each set of replicate measurements in a data pool.

Step 2. Once all of the within process standard deviations, s_w , are calculated, analyze them using the F-max test (Section 8.9, Case 3) to evaluate whether the levels of variability from replicate measurements are sufficiently similar. Table 9.13 is used to evaluate the level of significance for the F-max test, taking into consideration the number of replicate samples. When the ratio of the variances is close to 1, further analysis is not required.

Step 3. If the standard deviations pass the F-max test, signifying that the samples in the data pool represent the same population of measurements, then following 8.4 above, calculate the pooled standard deviation of all measurements in the pool of values.

When a new s_w value is obtained from additional replicate measurements, assess it with another F-test. If it is in statistical agreement with the other s_w values, add it to the data pool and combine with the pooled s_w values.

Using this method, within process standard deviations of measurement results that have unequal nominal values (and different measurement results) can be combined based on common parameters of the measurement system. These common parameters are typically a combination of factors like the resolution of the device where the measurement data originates, i.e., scale plate resolution or balance resolution, and similarity of the process, i.e., multiple deliveries from a volumetric standard or replicate weighings of a weight cart.

The pooled standard deviation is then treated as the process standard deviation, s_p , and combined in quadrature with other uncertainty components deemed significant to the measurement result per the measurement SOP and SOP 29.

$$U_{(k=?)} = k\sqrt{(u_s^2 + s_p^2 + u_1^2 + \dots + u_x^2)}$$

The degrees of freedom associated with the estimated process standard deviation are combined

based on $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \dots + (n_k - 1)$ for each set of replicates in the data pool as shown in Section 8.4. Pooled degrees of freedom associated with this estimate of process variability are then used with other estimates of degrees of freedom to calculate effective degrees of freedom for the uncertainty using the usual methods.

The benefit of this process is to provide a better estimate for the standard deviation of a measurement process over time, and increase the degrees of freedom to reduce the coverage factor, k , required for the uncertainty calculations. Monitoring the standard

deviations of replicate processes on a standard deviation chart also allows visual assessment of the process, which helps to determine the stability of the variation over time.

8.14 Calculating Effective Degrees of Freedom for Uncertainty Calculations

The degrees of freedom used to calculate the uncertainty expansion factor, k , can be the degrees of freedom associated with the pooled standard deviation, or more correctly the effective degrees of freedom, v_{eff} , calculated using the Welch-Satterthwaite equation from the Guide to the Expression of Uncertainty in Measurement (JCGM 100:2008, G.4.1):

$$v_{eff} = \frac{u_c^4}{\sum_{i=1}^N \frac{u_i^4(y)}{v_i}}$$

For this application, we can rewrite this equation:

$$v_{eff} = \frac{u_c^4}{\frac{u_s^4}{df_{u_s}} + \frac{s_p^4}{df_{s_p}} + \frac{u_1^4}{df_{u_1}} + \dots + \frac{u_x^4}{df_{u_x}}}$$

Table 4. Variables for equations in 8.14.

Variable	Description
u_c	combined standard uncertainty
u_s	standard uncertainty attributed to the standards used
u_1	first additional uncertainty component
u_x	x^{th} additional uncertainty component
s_p	standard uncertainty attributed to the process obtained by using the standard deviation of a check standard or the pooled standard deviation of replicate measurements
df	degrees of freedom associated with the indicated standard uncertainty

The Welch-Satterthwaite equation weights the impact of each uncertainty component based on the significance of the component to the overall uncertainty and considers the associated degrees of freedom of each component in the calculation. The components having a more significant uncertainty contribution also contribute more to the effective degrees of freedom calculation.

8.15 Comparing an Observed Value of a Check Standard to the Mean Value from a Control Chart Using the t-Statistic

A t-test may be incorporated to check the observed value of the check standard against the mean value using the following equation and a 95 % confidence level. All values must be entered in the control chart, even if failing this statistic, to ensure the variability obtained for the process is not unduly reduced over time. The observed value of the check standard is compared to the accepted mean value of the check standard and divided by the standard deviation for the check standard observations over time.

$$t = \frac{(S_c - \bar{S}_c)}{s_p}$$

This equation monitors the output of a measurement process compared to the stable value over time. A calculated t-value less than two is within the warning limits of the process. A calculated t-value between two and three represents a value between the warning limits and control/action limits. A calculated t-value exceeding three represents a value outside of the control/action limits and suitable action must be taken. Calculated values of the t-statistic may also be monitored over time to determine the presence of drift in the measurement standards.

This t-test is routinely incorporated into mass measurements, weighing designs, and the NIST Mass Code as standard practice.

8.16 Comparing the Mean Value of a Check Standard to a Reference Value

The Normalized Error, E_n , may be used to compare the mean value of a check standard, S_c to the calibrated reference value, $S_{c(cal)}$ taking care to ensure adequate metrological traceability for the reference value. The uncertainty of the reference value is taken from the calibration certificate and the uncertainty of the mean value is determined using the following equation, using the standard deviation of the process from the control chart, s_p , and where n is the number of relevant data points; other components are the uncertainty for the standards, u_s , and any other critical components to be considered, u_o based on the SOP in use. The coverage factor, k , is determined based on the desired level of confidence and the associated effective degrees of freedom.

$$u_c = \sqrt{\frac{s_p^2}{n} + u_s^2 + u_o^2}$$

$$U_{\bar{S}_c} = u_c \times k$$

E_n is calculated using the following equation:

$$E_n = \frac{|\bar{S}_c - S_{c(cal)}|}{\sqrt{U_{\bar{S}_c}^2 + U_{S_{c(cal)}}^2}}$$

The E_n value must be less than one to pass. If the E_n is greater than one, corrective action is required. In the usual method for propagating expanded uncertainties, standard uncertainties will be combined at the $k = 1$ level, using the root sum square method, and then multiplied by the effective coverage factor. This version of the equation for normalized error is commonly used for proficiency testing.

Note: the reference value may exceed acceptable limits even when a measurement process appears statistically in control. When tolerances are large compared to process variability, reference value offsets may simply be noted. A control chart should be reviewed prior to performing this evaluation to ensure that the measurement process appears stable, otherwise, corrective action is needed prior to attempting this assessment.