Test Methods Uncertainty Statement Definitions and Methods

Walter Resch, Quality Engineer / Statistician, 3M, St. Paul, MN
and the PSTC Test Method Committee.

Introduction

Test methods provide a measured value of a specific property for a material or product. That single determination is a fixed value but it has random variability associated with it. If the test is repeated one or more times it will always produce a different result. If then a different operator at a different lab also performs the same test the results will have an even larger variability associated with them. This variability should be understood for all test methods that are used so that a real shift in results can be recognized and separated from just random noise. But what is a good way of measuring and then presenting this variability information? What information is important to know? What information could be misleading?

Many options are available and it becomes a matter of personal preference on how to analyze the data and what to present. ASTM suggests repeatability and reproducibility variances, standard deviations and/or coefficient of variation. Software packages supply prepared outputs for "Gage R&R's" that may include a myriad of results such as an ANOVA table with P values, variance components, ratios of the test variation to other variations and an assortment of graphs.

The PSTC Test Method subcommittee has been working on providing variability information for the PSTC test methods. A format has been developed by our members that seems to be the most universally useful and yet understandable to a non-statistician. The specific format may still be revised in the future as additional input is gathered.

Uncertainty Statement Methods

First some definitions:

Variable data: includes all real numbers, not restricted to discrete, categorical or attribute data (pass-fail, red-blue, 0-1-2-3…). The only type of data considered in this write-up is continuously variable data. However, if there are about 10 or more values of discrete (and ordered) data, they can be considered continuous enough so that the analysis will not be affected much.

Inter-laboratory study: a procedure or designed experiment for obtaining a precision statement for a test method, involving multiple laboratories, each generating replicate test results on multiple samples. This is the type of study discussed here. It is not the intent of these studies to estimate the bias or shift in results between laboratories as in a "Round Robin" study but to estimate the variability of a test method when done across a random selection of laboratories. The total variation from a study such as this is useful for comparing test values across different labs, companies or locations. For a specific company using the test method for quality control in manufacturing of a product, a more suitable measure of variation might be within a single lab but still encompassing multiple operators. The study plan and analysis would be the same as for the inter-lab study discussed here but instead of multiple labs there would just be multiple operators (testing in the same lab, on the same instrument). It is recommended that each company do a study such as this, in their own lab and get a better measurement of the error that they will experience.

Repeatability conditions: conditions where independent test results are obtained on identical test samples in the same laboratory by the same operator using the same equipment and method within short intervals of time.

Destructive testing: if testing a sample destroys the sample then the testing is destructive. Actual repeat testing is not possible in this case and "repeat" samples are taken in very close proximity to each other and assumed to be identical. Of course this is not strictly true and there will be some additional, but hopefully very small, variability added to the "repeatability" variation.

Reproducibility conditions: conditions where test results are obtained with the same method on identical test samples in different laboratories with different operators using equivalent but different equipment. It can also be useful to estimate a Reproducibility that refers only to different operators but within the same laboratory. Since Reproducibility can mean different things, the factors that are changed in measuring the Reproducibility need be defined.

Procedure for Conducting the Study

The general plan is to have multiple labs (or just multiple operators) test the same samples and test them repeatedly. The study is a balanced, two-factor experiment. The design is also considered crossed, not nested, because the "same" parts are tested at each laboratory.

Steps

Assemble 4-10 laboratories (the more the better), and 4-10 samples with varying properties (also referred to as parts). Each sample is to be measured 2-3 times by each laboratory. Include samples of test materials that will produce varying results and cover a range of interest. The results of this study will basically be valid within the range of values tested.

Create a data worksheet for the experiment such as in the example below. This helps with organizing and understanding the experimental plan.

Verify that the testing instruments are current with their calibration and that the operators are adequately trained on the test method.

Either circulate the samples from lab to lab or, if the test is destructive, provide a homogeneous quantity of each sample to each laboratory. The analysis assumes that each laboratory is testing the very same samples as the other laboratories. If the test destroys the samples then it is important to select individual test material that is known to be very homogeneous within a batch or roll so that each laboratory will be testing very similar material.

Each lab should measure all samples in random order. Work to minimize day-to-day and other environmental sources of variation (noise). Follow the test method instructions carefully.

Select enough samples (or parts) so that (number of samples) X (number of laboratories) > 15. If this is not possible, choose the number of trials (repeats) as follows:
if S x L < 15, trials = 3
if S x L < 8, trials = 4 (but this is not preferred)
where S = # of Samples and L = # of Laboratories.

Example of a data sheet:

		Lab			
sample	repeat	A	B	C	D
1	1				
	2				
2	1				
	2				
3	1				
	2				
4	1				
	2				
5	1				
	2				

Down a data column, the samples 1 to 5 will have varying properties. Across a row, the data should come from the same sample, including repeat 1 and 2. If the testing is destructive, then the data across rows should come from samples with nearly identical properties.

Analysis
The suggested analysis basically uses standard deviations and metrics calculated from standard deviations to give the user information regarding the variability of the test method. Also important are the measurement ranges that were tested and how many data was gathered in the study. The listed items are categorized as “important” or “optional”.

Important: Means Tested
List the means or range of means of all the samples that were included in the study. The results are valid within this range. Outside this range the analysis would be considered extrapolated and we could not be certain that the variability remains constant.

Important: Verify Constant Variance
An ANOVA analysis makes a few assumptions and constant variance is the most important verifiable assumption of the analysis. Supply some evidence that the variance is constant over the whole range of means tested. One method is the Levene’s test supplied by software, which is a t-test comparing the absolute values of the deviations for each sample. Or alternatively, a simple rule of thumb is that the standard deviations should not vary by more than three times between the samples (if the number of measurements for each sample is equal). Or a plot of residuals versus fits can be shown to demonstrate that there is relatively uniform spread.

Important: Standard Deviations

Simply: List the standard deviations for repeatability, reproducibility and the total test method.

Details: The experiment is a crossed, balanced, two-factor design and the data are analyzed using ANOVA. The factors are considered to be random which means they are assumed to have a normal distribution with mean=0, and it is the factor variances that are estimated, not the factor effects. The factors are crossed, not nested. The ANOVA model includes the main effects of Parts (or samples), Laboratories, and the Part*Laboratory interaction.

Variances are calculated for each factor (Parts, Laboratory, Part*Laboratory interaction, and total) using expected mean squares by the software. The standard deviation for each component is then the square root of the Variances.

The "Repeatability" will be the "error" standard deviation. The "Reproducibility" will be the "Laboratory" standard deviation and will also include the interaction.

If the PSTC test method recommends that a specific number of tests are to be run and averaged, then a standard deviation for an average should also be displayed. For example the standard deviation for a sample of 3 would be equal to the individual standard deviation divided by the square root of 3.

The degrees of freedom for each component should also be provided to give the user an idea of how much data went into the estimates.

An example of an ANOVA table and components of variation calculations are in the appendix.

The user can then use these standard deviations to understand the typical variation in the test data or to compare to their specification limits or their process standard deviations to judge if this test method will be adequate for his or her purposes. See below.

Optional: Maximum Range Expected

Simply: the maximum range in test values that can be expected between multiple measurements made on the same sample.

Details: Reference ASTM standard C670 − 13 (page 4, table 1). The calculation is based on order statistics. The range is calculated from the standard deviations:

Number of Test Results	Multiplier of Standard Deviation
2	2.8
3	3.3
4	3.6
5	3.9

For example, if the standard deviation for an individual measurement was 2.5, then 95% of the time for a group of 3 measurements the maximum range between any two (of the 3) readings will not be greater

than 2.5 X 3.3 = 8.25. The appropriate standard deviation to use depends on the application. For example if comparing values across different labs then a total test standard deviation that includes lab-to-lab reproducibility as well as repeatability should be used.

This statistic gives the user a quick and easy understanding of the range of values that can be expected from a group of measurements.

Optional: Sample Size Required to Detect a Difference between Two Samples

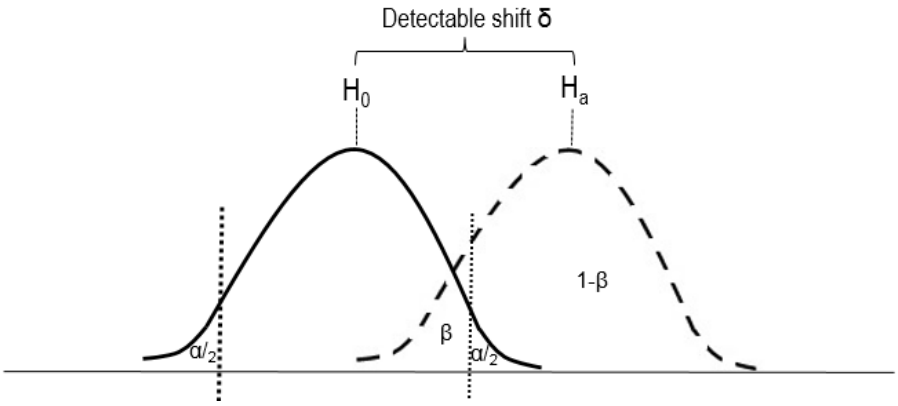
Simply: the sample size (number of repeated measurements on each sample) that is required to detect a specific difference in product performance between two samples.

Details: This is known as a power calculation for a two sample t-test. Under the assumption that there is a specified difference between two samples, this is the probability that the difference will be detected with various sample sizes or, conversely, what difference can be detected with a specific probability (such as 80% or 0.80 power) using various sample sizes. This is solved iteratively using the below equation. The appropriate standard deviation is the repeatability if the testing is all done by one operator on one instrument.

$$\delta = u_1 - u_2 = (t_{2(n-1),1-\alpha/2} + t_{2(n-1),1-\beta}) S_p * \text{sqrt}(2/n)$$

$\delta = u_1 - u_2$ = the actual difference of the samples.
n = number of repeated tests done for each sample
$1-\beta$ = power, typically 0.80, the probability of detecting δ under the assumption that it is there (β is the type 2 error; probability accepting the null hypothesis, H_0, when it is false)
α = type 1 error, usually 0.05 (probability of rejecting H_0 when it is true)
t = Student's t-distribution
S_p = pooled standard deviation

The method is shown graphically -



The minimum detectable shift is such that 80% (1-β) of the hypothesized distribution (H_a) is beyond the $\alpha/2$ level (vertical dotted lines) of the null hypothesis distribution (H_0), causing it to be rejected. The spread of the distributions are determined by the standard deviation and the sample size.

Optional: Specification Width Needed for Test Method Adequacy.

If the test is used to release product, this is a useful way to judge if the test method has the recommended amount of precision.

Simply: The minimum width of the specification that would be necessary for this test method to have adequate precision. If the user's specs are narrower then there could be an issue with excessive test method variability that should be addressed.

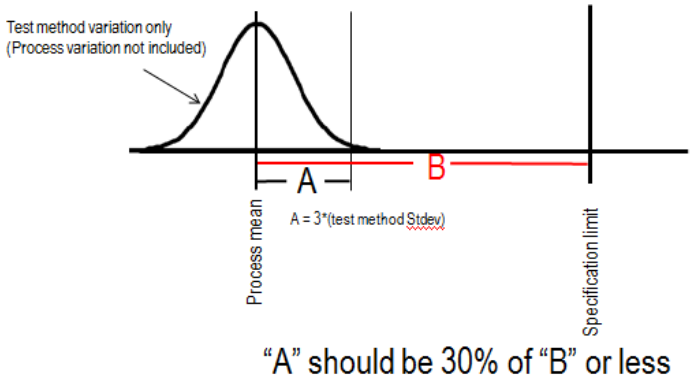
Details: The precision-to-tolerance ratio (P/T) can be used to assess the adequacy of a test method. This is a ratio of the total test method variability (precision, P = 6*StDev) to the width of the specification (tolerance, T = $Spec_{max}$ - $Spec_{min}$). The smaller this ratio is, the better. Fifty percent or 0.5 is considered a maximum for a destructible test. Thirty percent or less is preferable.

Since only the user knows the applicable specifications, we do not calculate the P/T ratio but we calculate the width of the specification needed for this test to have an adequate P/T. If the user's specification is at least that wide then he or she is in good shape with this test. If not, then something should be done to reduce test variation such as doing multiple measurements and averaging the result.

For a one-sided specification the test variability is 3*StDev and the tolerance is the distance of the process mean to the specification limit. So in this case P = 3*StDev, and T = | mean – spec |. This calculation may also be useful in the case of the 2-sided specification where the P/T to the nearest specification is of interest.

The result should be calculated using a within-lab test standard deviation which includes repeatability and may also include reproducibility, but only the operator variability, not the lab-to-lab variability, as it is likely for a specific piece of production equipment that the testing will always be done on the same equipment in the same lab but by multiple operators. If this type of reproducibility variation is not available then just the repeatability standard deviation should be used. However, it is recommended that the user measure his or her within-lab reproducibility as well. It goes without saying that this assessment is only meaningful if the specification limits are meaningful and set according to what is really required of the process to make good product.

Example for a single-sided specification (or nearest specification)



Another measure of test method adequacy is to compare test method variation to the process variation either as a ratio of variances or standard deviations or as "Number of distinct categories." If a historical

process variation is known, this is another good way to assess test method adequacy. However, the process variance may not be well known, and this variation is likely different for different products. Statistical software provides a measure of process variation from the samples selected in the study but most of the time these are hand-picked samples (and are sometimes even different products) and so do not give a good estimate of the actual process variation. It is strongly advised to avoid any ratios that involve the sample-to-sample variation estimated in the study.

Summary

The important items to include in the statement of variability for test methods are:

Range of means – for establishing the valid range of the results;
Evidence of constant variance – for verifying an important assumption made by the analysis; and
Standard deviations – the fundamental statistic for variation.

There are many other possible types of analysis and graphs that could be employed and it is a matter of opinion on what are the most useful. The committee selected the following additional items because they seemed to be the most universally applicable as well as being relatively easy to understand:

Expected maximum range – to help visualize possible variation;
Sample size needed to detect a specific difference – helpful for planning a study; and
Required tolerance for test method adequacy – a measure of test method adequacy based on specification width.

Many types of analysis can be useful but, as was said before, sample variation generated from the study and comparisons of it to the test method variation should be avoided. But this is commonly done because many software packages are set up to do just that. If such a ratio is required, for example by an auto manufacturer, a known historical process standard deviation should be manually plugged into the software to calculate the appropriate ratios.

An example for PSTC-17 Shear Adhesion Failure Temperature (SAFT):

A test method variability study was done where 4 different labs tested 3 different products: Aluminum Foil Tape (acrylic adhesive), Duct Tape (rubber adhesive), and Masking Tape (rubber adhesive). Each product was tested 3 times by one operator at each lab.

(1) The variability study included products that produced means ranging from 71 to 106 ^{0}C for the three different products.

(2) Variance of the test was constant over the range of means tested (Levene's Test, p-value = 0.979). The data was showing signs of increasing variability at higher means. If data is well outside the range tested here then it is possible that the standard deviations may be different than shown.

(3) Test standard deviation:

Source of Variability	Degrees of Freedom	Standard Deviation ^{0}C (sample of 1)	Standard Deviation ^{0}C (sample of 3)	Description
Repeatability	30	2.75	1.59	variation from repeated tests
Reproducibility	4	7.25	4.19	variation from different labs
Overall	-	7.75	4.48	Total Test method variation

(4) The maximum range that can be expected from multiple measurements made on the same sample with this test method (95% of the time).

Source of Variability	Max Range ^{0}C of two test values* (2.8*Stdev)	Max Range ^{0}C of three test values* (3.3*Stdev)	Description
Repeatability	7.7	9.1	Measurements made on the same instrument with the same operator
Overall	21.7	25.6	Measurements made with different labs

* Individual test values (not averaged)

(5) Sample size required for a 2-sample t-test to detect a specific difference in tape performance with a power of 0.8 (80% probability). For example, if two rolls of tape are compared in the same lab (and same instrument), there is 80% probability of detecting a difference of 8.4 ^{0}C or greater, if three tests are done on each roll of tape.

Using Repeatability variability (StDev=2.75):

Sample Size	Power	Difference
3	0.8	8.4
5	0.8	5.6
10	0.8	3.6

The sample size is for each group.
2-Sample t-Test
Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05

(6) The precision to Tolerance ratio (P/T) is a ratio of the total test method variability (precision) to the width of the specification (tolerance). Fifty percent is considered a minimum for a destructible test. Thirty percent is preferable.

This test method is adequate if the associated specification range is at least as big as shown on this table:

	For a Double sided specification:	For a Single sided specification:
Precision to Tolerance ratio	Specification width	Distance of process mean from specification
for 50% P/T	33.0	16.5
for 30% P/T	55.0	27.5

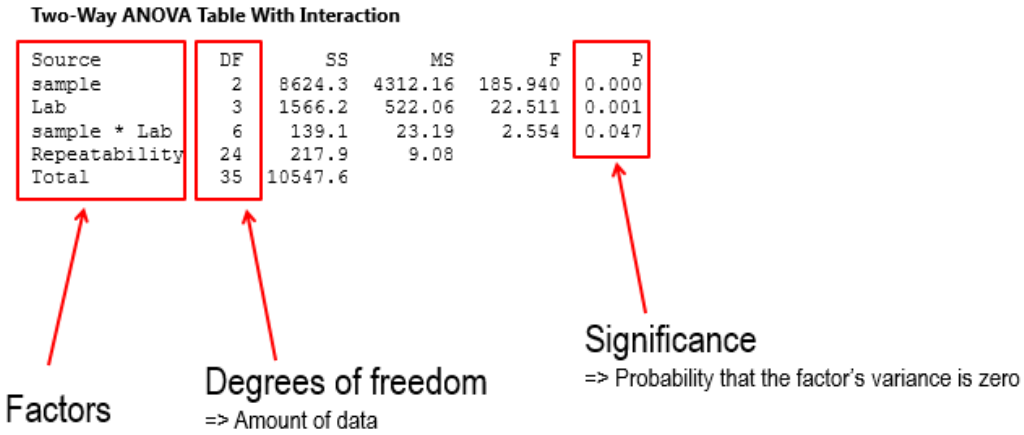
* uses multiplier of 6

* Individual test values (not averaged)

* uses repeatability stddev

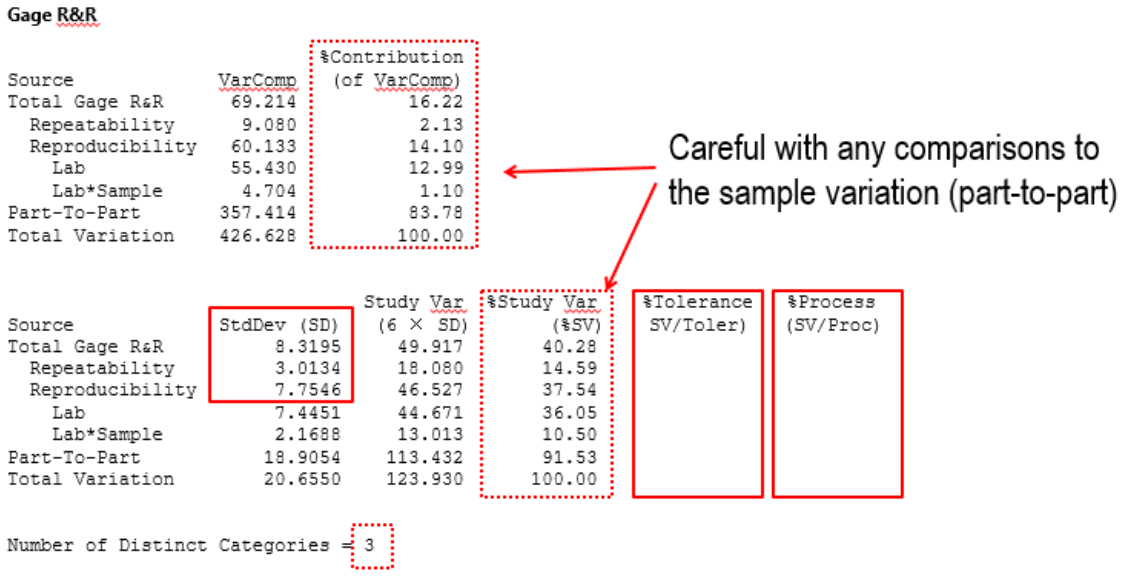
Appendix ----------

Example of ANOVA output:



The factors are listed under *Source*. The two factors are *sample* and *lab*. The interaction of the two is also included. The *repeatability* is then the "left-over" error.
The degrees of freedom are given under *DF* and indicate the amount of data. For example *Lab DF* = (number of labs) – 1 and *Total DF* = (total data points) – 1.
The *P* values in this random effects model give the probability that a factor's variance is zero. If the value is less than 0.05 then it is said that there is evidence that the factor's variance is statistically significant (nonzero).

Example of Components of Variance output:



Components of variance output may be a part of a *Gage R&R* function in software. Or it can be run using ANOVA and setting the factors to *random*.

VarComp are the variances. *Repeatability* is the error variance. *Reproducibility* is the variance of Lab and the interaction. *Total Gage* is the sum of *repeatability* and *reproducibility*. *Part-to-Part* is the variance from the samples. The sum of all variances is the *total variation*.
%*Contribution* is the percent of the total variance for each of the factors. This adds up to100% since the variances are additive. These two columns are useful when comparing test method factors to each other like *repeatability* to *reproducibility* – not to *part-to-part* or the *total* since these are calculated from the selection of the specific parts or samples which are rarely representative of the true variation.
The lower table gives the *standard deviations* which are the most important result from the output. *Standard deviation* is in the same units as the response. *Study Var (6XSD)* includes 99.7% of the probable values. If any of the given ratios or *Number of Distinct Categories* are supplied to Automotive manufactures, they allow the use of 5.15 (which covers 99%) instead of 6.
%Study Var is the percent of each factor's *standard deviation* in the study. These do not add to 100%. Because of that and also because the *part-to-part* variability comes in to play again, this column should be avoided. If specification limits were supplied, the %*Tolerance* column compares the variability to the specification width. If the specification limits are meaningful, this can be useful and this concept is used for providing the "Specification width needed for test method adequacy" in the paper. Also, if a process standard deviation is known, the last column, %*Process,* is useful for comparing the test method variability to the process variability. *Number of Distinct Categories* gives the number of categories the test method can distinguish between, within the part-to-part variability. This would be useful if the part-to-part variability gave a good estimate of the real process variability. But with hand-chosen parts, they rarely do.

References

1. ASTM "Standard Practice for Preparing Precision and Bias Statements for Test Methods for Construction Materials", Designation: C670 − 13
2. ASTM "Standard Practice for Use of the Terms Precision and Bias in ASTM Test Methods," Designation: E177 − 14
3. Leon Harter, "Order Statistics and Their Use in Testing and Estimation," Vol. 1, Aerospace Research Laboratories, United States Air Force.

Acknowledgments

The PSTC Test Methods Committee:

Mark Byrne, CHAIR	Shurtape
Kiran Malhotra, TM Advisor	Adchem Corp
Gary Avalon	Chemsultants
Laura Donkus	Dow Chemical
Nestor Hansen	Cray Valley
Michael Johnson	Intertape Polymer Group
Walt Resch	3M