**Valid Statistical Techniques, v1.20**

Copyright 2024 by John N. Zorich Jr.; all rights reserved, with the following exception:
**Anyone may use all or part of this document in their own company's internal procedures.**

**To create your own copy, "select all", "copy", & "paste" into a blank Word document.**

**Some text boxes may need adjustment if they overlap when viewed in Word2010 or earlier.***Your compliments, criticisms, and suggestions regarding this document would be welcomed.*

For the most up-to-date version, write to: JOHNZORICH@YAHOO**.**COM or visit JOHNZORICH**.**COM

1. PURPOSE

The purpose of this procedure is to provide guidance on how to appropriately (*i.e.* validly) use some of the most common statistical techniques. In some cases, formulas are provided; however, **it is not the purpose of this procedure to provided instructions for how to perform the calculations for every statistical technique discussed here**; such instructions are easily obtained from published articles, reputable websites, or statistics textbooks.

1. SCOPE

This procedure applies to common statistical techniques that are used to analyze product sample statistics so as to determine the corresponding product population parameters, in order to draw conclusions about the acceptability of the population (*e.g.* a lot) or of the process that produced the population. Examples of such statistics include averages, percentages, and %-in-specification. **This procedure does not apply to statistical techniques used to analyze clinical trials data.**

1. GENERAL REFERENCES (other references are given in the body of the SOP)
2. *NIST/SEMATECH e-Handbook of Statistical Methods*
This book is freely available at: *http://www.itl.nist.gov/div898/handbook*

The National Institute of Science & Technology (NIST) is an agency of the United States Department of Commerce.

1. Wikipedia ( *https://en.wikipedia.org* ) contains articles on many of the topics covered here.
2. DEFINITIONS OF KEY TERMS (as used in this document)
3. *Product:* raw material, component, in-process or finished goods
4. *Process:* an activity that outputs a product
5. *Sample*: a representative subset of a population
6. *Population*: the entire data set or physical group from which the sample was taken; commonly, a population is a lot (or batch), but a population can also be regarded as the process from which a lot has been derived (*e.g.* during process validation, an entire pilot or validation lot can be considered a sample from the "population" process that produced it)
7. *Statistic*: a summary value (*e.g.* an average) that has been calculated based upon inspection or measurement of all the members of a sample; a sample *statistic* is used to estimate the corresponding *parameter*
8. *Parameter*: a summary value (*e.g.* an average) that has been calculated based upon inspection or measurement of all the members of a population
9. *Standard Error*: the standard deviation of the population of sample statistics (*e.g.* the standard error of the sample mean is the standard deviation of the population of means that is obtained by taking the mean of each of all possible samples of a given size from a given population)
10. *Excel:* the spreadsheet software program sold by Microsoft Corporation
11. LIST OF STATISTICAL TECHNIQUES IN THIS SOP
* Normality tests and normality transformations
* Confidence intervals and limits
* Coefficients of determination, and correlation coefficients
* Statistical power and tests of statistical significance
* Statistical process control (SPC)
* Process capability and confidence/reliability calculations
* AQL sampling plans
* Guard-banded QC specifications
1. FUNDAMENTAL VALIDITY CRITERIA FOR ALL STATISTICAL TECHNIQUES

Any statistical technique that is used within the scope of this procedure must meet the following criteria:

1. Techniques must be taken from textbooks or articles that have been published by reputable publishing houses, journals, or websites (in the case of websites, reputability is a judgement-call, the basis of which should be appropriately documented in quality system records, if it is not a general-recognized source); or taken from commercial statistical software programs.
2. Output of the technique should provide information about the relevant population parameter, at a chosen level of confidence; a technique that outputs only a sample statistic should not be used for *final* design verification, design or process or product validation, or product/lot disposition.
3. Output of the technique must be as accurate and exact as is practically possible; that is, pre-computer-era "approximation" techniques should not be used for *final* design verification, design or process or product validation, or product/lot disposition.
4. NORMALITY TESTING AND NORMALITY TRANSFORMATIONS

To assess whether a data-set is normally distributed, plot the data on a Normal Probability Plot (a.k.a. NPP), which can be created either by hand, by Excel, or by using programs such as Minitab or StatGraphics. The plot will appear to be a straight line *only* if the data is normally distributed; if the plot is not a straight line (with points above and below the line in an approximately random fashion), then the data is not normally distributed. Although use of an NPP plot in such a manner is somewhat subjective, it is the best of all normality evaluation methods, as explained in quotations below. There is no minimum sample size required; however, other popular tests of normality require a minimum of 5 (*e.g.* the Shapiro-Francia W' test) or 6 (*e.g.* the Anderson-Darling A2\* test).

*REFERENCES RELATED TO NORMALITY TESTS:*

* Gross, J., *A Normal Distribution Course*, 2004 by Peter Lang GmbH
* Shapiro, S. S. (1990), *How to Test Normality and Other Distributional Assumptions*, published by ASQC Press
* Thode, H. C. Jr. (2002), *Testing For Normality*, published by Marcel Dekker Inc.
* Tobias, P. A., and Trinidade, D. C., *Applied Reliability*, 3rd ed., 2012, published by Taylor & Francis Group
* Zylstra, R. R. (1994), "Normality Tests for Small Sample Sizes", *Quality Engineering*, 7(1), 45-58, published by ASQC Press

*QUOTATIONS REGARDING NORMALITY TESTS — To evaluate normality, NPP analysis is recommended rather than a "Normality Test" such as Anderson-Darling because such normality tests do not assess normality but rather assess only non-normality. The following quotes help explain:*

*"No single [normality test] statistic should be relied on to confirm normality. A good place to begin an analysis for normality... is with the use of a graphical methods such as the probability plot." (Zylstra, p. 53)*

*"It is strongly recommended that...a test [of normality] should not replace a [normal] probability plot...[because] no single test statistic can give you as much information as a graphical display which shows the extent and type of departures from [linearity]...." After a normality test has failed to reject normality, "it is worth remembering that it was not proven that the measurements were normally distributed, it was only concluded that there was no evidence that one should not treat them as if they were normally distributed". (Shapiro, pp. 13, 21)*

*"When [on the basis of a normality test] we ...end up not rejecting [normality], we have not proved the hypothesis [of normality] to be correct....The statistical test gives us no such confidence... On the other hand, when we reject [normality], we are...saying that the sample evidence was so overwhelmingly against [normality]...that we have to reconsider our choice." (Tobias & Trindade, p. 59)*

To manually create an NPP, first arrange the data, as shown below for an n=12 sample:

|  |  |  |  |
| --- | --- | --- | --- |
| Raw Data (sorted) | Rank | Median Rank =(Rank − 0.3) / ( n + 0.4) | 1-sided Z-table value corresponding to Median Rank \*\* |
| 8.80 | 1 | 0.0565 | -1.585 |
| 9.30 | 2 | 0.1371 | -1.093 |
| 10.30 | 3 | 0.2177 | -0.780 |
| 10.80 | 4 | 0.2984 | -0.529 |
| 11.00 | 5 | 0.3790 | -0.308 |
| 12.40 | 6 | 0.4597 | -0.101 |
| 13.23 | 7 | 0.5403 | 0.101 |
| 13.34 | 8 | 0.6210 | 0.308 |
| 14.98 | 9 | 0.7016 | 0.529 |
| 17.20 | 10 | 0.7823 | 0.780 |
| 18.06 | 11 | 0.8629 | 1.093 |
| 20.80 | 12 | 0.9435 | 1.585 |

 \*\* Microsoft Excel function "NORM.S.INV" can be used to obtain the appropriate
Z-table value. For example, NORM.S.INV(0.9435) = 1.585

Then plot the Z-table value (the right-hand column above) versus the raw data (the left-hand column above), as shown in the example plot (below):



The plot above has a slight curve to it, as shown in the plot below of the same data:



If, as determined visually, the NPP has even a slight curve (as shown above), a transformation to normality must be sought that creates a straight-as-practically-possible NPP. Any subsequent statistical analysis is then used on the transformed values, not on the original raw data; and any numerical requirement (*e.g.* a QC specification) that is used in the statistical analysis must be transformed in the same way as the raw data. If a transformation is used, the final results of the statistical analysis (*e.g.* a confidence interval) should be reported in units of the original raw data (by reverse-transforming it), not in units of transformed values.

The following is a list of some commonly-used transformations to normality; if a transformation is needed, try them all and use the transformation that results in the straightest-looking NPP; only when it is *not* obvious which line is the straightest, choose the one with the largest *R****2*** value. When the range of raw data (and QC specifications) spans 1.00, it may be useful to try transformations after first adding 1.00 to each raw data value (and to the QC specs).

*WARNINGS:*

*The correlation coefficient (or its square, R****2****) should not be relied upon to indicate linearity; the correlation coefficient for the first plot above is 0.9706 ( = √0.9421 ), and yet the data is obviously curved, as shown in the second plot.*

*Similarly, an Anderson-Darling test of normality for the raw data in the plots above yields a p-value of 0.45, which "passes" by a wide margin (a p-value greater than 0.05 is considered passing); and therefore in order to most reliably determine normality, an NPP plot should be used rather than a "normality test" such as Anderson-Darling.*

|  |  |
| --- | --- |
| TRANSFORMATION | MS EXCEL EQUIVALENT |
| Inverse (X) | = 1 / X |
| Square root (X) | = SQRT ( X ) |
| Cube root (X) | = ( X ) ^ ( 1 / 3 ) |
| Quadratic (X) | = ( X ) ^ 2 |
| Cubic (X) | = ( X ) ^ 3 |
| Logarithm (X) | = LN ( X ) |
| Inverse Hyperbolic Sine (X) | = ASINH ( X ) |
| Inverse Hyperbolic Sine (Square Root (X)) | = ASINH ( SQRT ( X ) ) |
| Logit (X) [ used only when all X values are < 1.00 ] | = LN ( X / ( 1 − X) ) |

An "inverse" transformation was found to produce the best straight for the sample data used in the previous plot, as shown below:



1. **CONFIDENCE INTERVALS AND LIMITS**

Any statistic (*e.g.* a sample average) is an estimate of the true value for the population from which the sample was taken; that true value is called the population parameter. The classic way to determine how accurate is that estimate is to calculate the relevant "confidence interval". A confidence interval was described by its inventor as a range "in which we may assume are contained the [parameter] values of the estimated characters of the population"; a confidence level (*e.g.* 95%) can be assigned to such an assumption (J. Neyman, "On the Two Different Aspects of the Representative Method," *Journal of the Royal Statistical Society*, 97, no. 4 (1934): p. 562).

Two-sided confidence intervals are ranges that extend above and below the value of the sample statistic. One-sided confidence intervals extend from a value on one side of the statistic to a value *as far as possible* away from the statistic, on the opposite side of the statistic, as shown in the figure below.

negative
infinity

positive
infinity

value of
sample statistic

2-sided interval
(right-side edge = upper confidence limit;

left-side edge = lower confidence limit)

1-sided interval
(right-side edge = upper confidence limit)

1-sided interval
(left-side edge = lower confidence limit)

In some cases, a confidence interval has a natural limit; for example, the largest possible value in a confidence interval for a proportion is 100%, and the smallest possible value in a confidence interval for a weight is 0.000 gram.

*Any sample size is valid* when calculating confidence intervals and limits. As can be seen in the figure below, the width of confidence intervals changes depending upon sample size. A narrow confidence interval based upon a large sample is no more valid than a wide confidence interval based upon a small sample size. They are both, *e.g.*, 95% confidence estimates for the population parameter.



* 1. **CONFIDENCE LIMITS FOR COUNT-DATA PROPORTIONS**

Proportions are ratios. For example, if there are 90 good parts in a sample of 100 parts, the proportion of good parts is 90/100. Similarly, if there are 10 defective parts in a sample of 100, the proportion of defective parts is 10/100. The numerator is always the count of the characteristic of interest (*e.g.* good, bad, green, red, etc.).

Proportions can be expressed as a decimal fraction, *e.g.* 0.10, or as a percentage, *e.g.* 10.0%. There are many published methods for calculating confidence intervals for a proportion; unfortunately, the intervals and limits produced by those various methods can be dramatically different, in part because some of the methods are rough approximations for the exact method. For example, the "Normal Approximation" Z-table method gives an incorrectly narrow interval with incorrectly placed upper and lower confidence limits. The method described below is called the "Exact Confidence Interval"; it is the oldest method and is generally considered the "gold standard" (L. D. Brown, *et. al.*, "Interval Estimation for a Binomial Proportion," Statistical Science, 16, no. 2 (2001), "comment" p. 117).

When using software (*e.g.* Minitab or StatGraphics) to calculate confidence intervals on a sample proportion, and if given a choice as to which method to use, always choose the "Exact Confidence Interval".

To calculate Exact confidence limits for proportions using Excel, use the following formulas:

Upper 2-sided limit = beta.inv ( 1 – (1 – C ) / 2 , k + 1 , n – k )

Lower 2-sided limit = beta.inv ( (1 – C ) / 2 , k , n – k + 1)

Upper 1-sided limit = beta.inv ( C, k + 1 , n – k )

Lower 1-sided limit = beta.inv ( 1 – C , k , n – k + 1)

where

C = confidence desired, as a decimal fraction (*e.g.* use 0.95 for 95% confidence)

n = sample size

k = count of the characteristic of interest (*e.g.* "defective part") in the sample

For example, if C = 0.95, k = 10, and n = 100, then the proportion is 10/100 = 10% and

 BETA.INV (1 – (1 – 0.95) / 2, 10 + 1, 100 – 10) = 0.176 = 17.6%

is the upper 2-sided 95% confidence limit.

*WARNING: Some of the other methods for calculating confidence intervals on proportions are very approximate (in many cases, accurate only to one or two digits). For example, the classic "Z" or "Poisson" table methods were developed in the pre-computer era, to simplify calculations that are nearly impossible by hand. In this 21st century, there is no reason to use such approximate methods, since the exact, accurate-to-many-digits method can be performed instantaneously by any computer capable of running a spreadsheet program such as MS Excel.*

* 1. **CONFIDENCE LIMITS FOR MEASUREMENT-DATA AVERAGES**

There are only two generally-accepted methods for calculating confidence intervals for sample means that are based upon measurement data; one method uses Z-tables and the other uses t-tables. Those 2 methods yield the same result (to many digits) when sample sizes are huge, but can differ greatly when sample sizes are not huge (when sample sizes are small, the Z-table method gives an incorrectly narrow interval with incorrectly placed upper and lower confidence limits, whereas the t-table method yields the correct interval and correct limits). Therefore, only the t-table method should be used.

*WARNING: The calculation of confidence limits described in this section are accurate only to the extent that the inputted averages and standard deviations were computed from raw data that is normally distributed or that has been transformed to normality. See discussion of normality, elsewhere in this document.*

To calculate confidence limits for averages using Excel, use the following formulas:

Upper 2-sided limit =

 SampleAverage + CONFIDENCE.T (1 − C, SampleStdev, n )

Lower 2-sided limit =

 SampleAverage − CONFIDENCE.T (1 − C, SampleStdev, n )

Upper 1-sided limit =

 SampleAverage + CONFIDENCE.T (2\*(1 − C), SampleStdev , n )

Lower 1-sided limit =

 SampleAverage − CONFIDENCE.T (2\*(1 − C), SampleStdev , n )

where

C = confidence desired, as a decimal fraction (e.g. use 0.95 for 95% confidence)

n = sample size

Stdev = standard deviation calculated using the "STDEV.S" Excel function

For example, if SampleAverage = 50, n =100, C = 0.95, and SampleStdev = 0.7, then

 50 + CONFIDENCE.T (1 − 0.95, 0.7, 100) = 50.139

is the upper 2-sided 95% confidence limit.

* 1. **CONFIDENCE LIMITS FOR MEASUREMENT-DATA STANDARD DEVIATIONS**

There are only two generally-accepted methods for calculating confidence intervals for sample standard deviations that are based upon measurement data; one method uses Z-tables and the other uses Chi-squared-tables. Those 2 methods yield the same result (to many digits) when sample sizes are huge, but can differ greatly when sample sizes are not huge (when sample sizes are small, the Z-table method gives an incorrectly narrow interval with incorrectly placed upper and lower confidence limits, whereas the Chi-squared method yields the correct interval and correct limits). Therefore, only the Chi-squared-table method should be used.

*WARNING: The calculation of confidence limits described in this section are accurate only to the extent that the standard deviations were computed from raw data that is normally distributed or that has been transformed to normality. See discussion of normality, elsewhere in this document.*

To calculate confidence limits for standard deviations using Excel, use the following formulas:

Upper 2-sided limit =

 ( SampleStdev ) \* SQRT ( (n − 1) / CHISQ.INV ( (1 – C ) / 2, n − 1) )

Lower 2-sided limit =

 ( SampleStdev ) \* SQRT ( (n − 1) / CHISQ.INV ( 1 – (1 – C ) / 2, n − 1) )

Upper 1-sided limit =

 ( SampleStdev ) \* SQRT ( (n − 1) / CHISQ.INV ( 1 – C, n − 1) )

Lower 1-sided limit =

 ( SampleStdev ) \* SQRT ( (n − 1) / CHISQ.INV ( C, n − 1) )

where

C = confidence desired, as a decimal fraction (*e.g.* use 0.95 for 95% confidence)

n = sample size

Stdev = standard deviation calculated using the "STDEV.S" Excel function

For example, if n =100, C = 0.95, and SampleStdev = 0.7, then

 0.7 \* SQRT ( ( 100 − 1 ) / CHISQ.INV ( ( 1 − 0.95 ) / 2 , 100 − 1 ) ) = 0.813

is the upper 2-sided 95% confidence limit.

1. COEFFICIENTS OF DETERMINATION AND CORRELATION COEFFICIENTS

The most commonly used symbol for the "coefficient of determination" is "*R2*" and for the "correlation coefficient" is "*r"*. Those coefficients are related to each other by the fact that the former is the square of the latter; *i.e. r = SquareRoot(* *R2* ). An *R2* can range in value from 0 to +1, whereas an *r* can range in value from −1 to +1. Both coefficients are outputs from linear regression analysis of a paired *X,Y* data-set. Both coefficients are indicators of the strength of the correlation between the *X,Y* values. There is no minimum sample size requirement to calculate these values, although for pairs = 2, both coefficients always equal 1.000 exactly.

Both coefficients have rigorous mathematical meaning at 0 and at the absolute value of 1. At a value of 0, they both mean that there is no correlation between the *X* and *Y* values; at an absolute value of 1, they both mean that there is perfect correlation. Values between those extreme have rigorous mathematical meaning for a coefficient of determination, whereas they have *no rigorous mathematical meaning for a correlation coefficient* (Healey, J. F. (1984), *Statistics: A Tool for Social Research*, Belmont: Wadsworth, p. 267).

Intermediary values of a coefficient of determination can be interpreted as a proportion on a linear scale, whereas intermediary values of a correlation coefficient cannot be interpreted as a proportion because they exist only on an ordinal scale, *i.e.* as indicators of relative position in a sequence of numbers (Selkirk, K. E. (1981), *Rediguide 32: Correlation and Regression*, Nottingham: Nottingham University, p. 17). Therefore, *if the goal of a linear regression analysis is to accurately determine the strength of correlation*, the signed coefficient of determination (symbolized by *SR2*) should be reported rather than the correlation coefficient (see "Reasons for Teaching and Using the Signed Coefficient of Determination Instead of the Correlation Coefficient", 2018 by John Zorich, in *MathAMATYC Educator,* 9(3):48-51). However, to check the "statistical significance" of an *SR2* value, calculate the square root of its absolute value (*i.e.* calculate the absolute value of the corresponding *r* value), and then use the classic test found in statistics books for the statistical significance of *r* (*i.e.* the correlation coefficient); the result of such a test applies to the original *SR2*value.

In order to indicate in a report whether the correlation is positive or negative (*i.e.* whether the linear regression line slopes upwards or downwards), a positive or negative sign is placed in front of the coefficient. The classic formula for *r* automatically includes such a sign; the following formula for the signed version of the coefficient of determination also includes such a sign:

Signed Coefficient of Determination = *SR2* = *r*3 / ( absolute value of *r* )

Software such as Minitab, StatGraphics, or Excel automatically calculate *R2* and *r*, but do not output a sign for *R2*; therefore, in reports, the sign must be added manually or calculated using the formula just given above.

Using that formula, the value of *SR2* varies from −1 to +1. The absolute values between 0 and 1 represent the proportion (*i.e.* percentage) of the variance in the *Y* data that can be explained by the variance in the *X* data, assuming that in fact the *X* and *Y* values are linearly correlated. That is,

*SR2* = % of *Y* variance that can be explained by its linear correlation with *X*

1 −*SR2* = % of *Y* variance that cannot be explained by its linear correlation with *X*;
this is the % of *Y* variance that is due to other "causes" such as random error, measurement error, poor measurement technique, etc.

*WARNING: The coefficient of determination is equivalent to the ratio of two variances (a variance is the square of a standard deviation). Variances can be added, subtracted, and/or proportioned in a mathematically rigorous manner. Therefore a coefficient of determination can be interpreted in a mathematically rigorous manner as a proportion. On the other hand, the correlation coefficient is equivalent to the ratio of two standard deviations. Standard deviations cannot be added, subtracted, and/or proportioned in a mathematically rigorous manner. Therefore a correlation coefficient cannot be interpreted in a mathematically rigorous manner as a proportion or percentage.
[Gravetter, F. J., and Wallnau, L. B. (2000), Statistics for the Behavioral Sciences, 5th ed., Belmont: Wadsworth, pp. 531 and 536]*

1. STATISTICAL POWER AND TESTS OF STATISTICAL SIGNIFICANCE

Tests of "statistical significance" include "Z tests", "t tests", "Chi-squared tests", ANOVA, F-tests, and so on [note: this section does not apply to "normality tests"]. The "alpha" value used to judge significance is an arbitrary choice, but alpha = 0.05 (i.e. 5%) is most commonly used these days. When not otherwise mandated by a requirement, alpha can be chosen based upon the desired level of confidence in a test's conclusion of "statistically significant" (%Confidence = 100% minus the %Alpha); therefore, confidence should be chosen based on Risk Management concerns, if applicable to the data being analyzed.

If the desired output of a significance test is "statistically significant", then any sample size used is valid because a statistically significant result is literally mathematically equivalent to the Null Hypothesis value being outside the confidence interval for the sample statistic — a larger sample size would result in a narrower confidence interval and therefore an even *more* statistically significant result (see elsewhere in this document for a discussion of confidence intervals). Such tests can be performed by hand, by Excel, or by major statistical software programs (*e.g.* Minitab or StatGraphics).

If the desired output of a significance test is "statistically *non*-significant", and if the "non-significant" conclusion will be used as the basis for claiming conformance to a formal regulatory, product, or design requirement, then a valid statistical rationale is needed to justify the sample size used. Classically, that rationale is based on a calculation of statistical Power (see definition below) vs. a *specific* numerical value for the Alternate Hypothesis (see definition below). The numerical value of power is an arbitrary choice; values of 0.8 or 0.9 are commonly used these days, but the value should be chosen based on Risk Management concerns, if risk is applicable to the data being analyzed. It is recommended that statistical software programs be used to perform the calculations; if performed "by hand" the calculations should be in conformance with textbooks on the subject (see recommendations below).

*WARNINGS:*

*Virtually any test of statistical significance can be caused to output a statistically significant result simply by having a large enough sample size. Similarly, most tests of significance using real-life data and real-life specifications can be caused to output a statistically non-significant result simply by having a small enough sample size.*

*When using a test of significance to compare, for example, the means of samples of two different medical products, a non-significant result does not mean that the means are equivalent; in fact, even if not "statistically" significant, the difference in the means could be so large that one of the medical products may represent a danger to the user and/or patient. Similarly, even if "statistically" significant, the difference in the means could be so small that neither product is superior in any practical sense.*

* **Statistical power** is the probability of obtaining the observed sample statistic (*e.g.* a mean) or one that is even further away from the Null Hypothesis, assuming that the Alternate Hypothesis is the true parameter for the population from which the sample was taken; power is the level of confidence in a test's "statistically non-significant" conclusion (for example, if power = 0.90, then there is 100 x 0.90 = 90% confidence in the "non-significant" conclusion; that is, power is the confidence in the claim that the Alternate Hypothesis is *not* the population parameter).
* **Alternate hypothesis (a.k.a. "alternative hypothesis" or "critical difference")** is (for the purposes of power calculations) a numerical value, not simply a "greater than...", "less than..", or "not equal..." statement; this hypothesis represents a value that, *if it were the parameter* for the population from which the sample was taken, would signal that harm is not only possible but actual; such harm could be to the end-user or patient, or could instead be harm to the manufacturing company (*e.g.* an increased scrap rate); depending on the software program used or the "by hand" formulas used, the alternate hypothesis can also be expressed as the *difference* between the null hypothesis value and the "if it were the parameter" value mentioned above.

*REFERENCES ON POWER CALCULATIONS*

* *How Many Subjects? Statistical Power in Research
[1987, by H. C. Kraemer; published by Sage Publications]*
* *Statistical Power Analysis for the Behavioral Sciences
[2nd ed., 1988, by J. Cohen; published by Lawrence Erlbaum Associates]*
* *Statistical Power Analysis
[1998, by K. R. Murphy and B. Myors; published by Lawrence Erlbaum Associates]*
1. STATISTICAL PROCESS CONTROL

Statistical Process Control (a.k.a. SPC) is a tool for improving the quality of the output of a process; from an SPC point of view, quality is improved by reducing variation in that output. SPC's most important contribution to such improvement is to signal when the output of a process is so variable that an investigation may be able to "easily" discover the cause of that variation; a proper response to that discovery will help reduce process variation in the future. If no such investigations and responses are occurring in a timely manner, then it is not valid to claim that an SPC *program* is in place.

The upper and lower control-chart limits (*e.g.* for sample averages) are *literally* the upper and lower 99.7% confidence limits on the most recently calculated process average (*e.g.* the average of averages, or the average of ranges). Like any confidence interval, the distance between the limits widens as sample size decreases and narrows as sample size increases; thus *any sample size is valid* (see discussion of confidence intervals, elsewhere in this document).

SPC control charts indicate when the process being charted is "out of control", which is the signal to perform an investigation. There are many possible rules for determining out-of-control; it is recommended that only a few rules be used, so that the likelihood of a random-chance false signal is kept to a reasonable minimum. If only the following three rules are used, the combined chance of a false signal is about 1%:

* one or more plotted points found to be outside the control limits
* nine plotted points in a row, where either all are above the midline of the chart, or all are below the midline of the chart (where the midline is the most recently calculated process average, *e.g.* the average of averages, or the average of ranges)
* nine plotted points in a row, where either the 2nd through 9th points are each numerically larger than the previous point, or the 2nd through 9th points are each numerically smaller than the previous point.

Technically speaking, a control chart could be created using only two samples; however, from a practical point of view, it is best not to start SPC charting until there is adequate history or assurance that the process will remain unchanged for the foreseeable future (for example, no planned changes in production equipment, processes, types of raw materials, etc.).

When first starting an SPC chart, if it is found that the historical plotted data shows "out of control" signals, there is no requirement to perform investigations; investigation should be performed on subsequent data (not prior data) that signals out-of-control (reference: Wheeler).

Control charts can be created on paper, on Excel spreadsheets, or on specialized software (*e.g.* major statistical software programs such as Minitab and StatGraphics have SPC modules built into them).

SPC is typically used on production processes; however, it can be used on other processes, such as customer complaint management; in such a use, an out-of-control signal means not only that an investigation is warranted but that notification to regulatory bodies (*e.g.* the FDA and/or an EU Competent Authority) may be required. SPC is sometimes used to monitor "validated processes" in lieu of QC; in such a use, an out-of-control signal means not only that an investigation is warranted but that re-validation should be considered.

*RECOMMENDED REFERENCES ON SPC:*

* *Understanding Statistical Process Control*, by D. J. Wheeler (any edition)
* *Advanced Topics in Statistical Process Control,* by D. J. Wheeler (any edition)
1. PROCESS CAPABILITY, AND CONFIDENCE/RELIABILITY CALCULATIONS

Each of the methods described here outputs a conclusion as to

* what percent of the population is "in specification" (here, "%reliable" means the same as "% in-specification")
* how confident we are in that %-in-specification conclusion.

Such conclusions are called "*Confidence/Reliability statements*". All of them, as described here, are *literally* the lower 1-sided confidence limits on the %-in-specification found in or indicated by the sample, and therefore *any sample size is valid* (see discussion of confidence intervals, elsewhere in this document).

Whenever possible, the confidence level and reliability level target for such statements should be based upon information derived from the company's "risk management" program if the calculated %reliability will be used to prove conformance to a product, regulatory, or design requirement. For example, if an FMEA (which is one possible type of risk-management document) states that risk to the customer is acceptable as long as a given component is at least 98% reliable at a confidence level of 95%, then the QC or validation criterion should be that the component must have at least 98% reliability at 95% confidence. If no risk-management document mentions the given component, then it is valid to assume that there is no risk to the customer and therefore the confidence and reliability levels can instead be chosen based upon business risks (*e.g.* cost or lead-time issues).

* 1. STATEMENTS BASED UPON COUNT DATA

In effect, the confidence/reliability statement for a sample %-in-specification (*i.e.* for count data) is simply the lower 1-sided confidence limit on the %-in-specification observed in the sample (see discussion of "count data" confidence intervals, elsewhere in this document). For convenience, the relevant formula is given again, here, modified for ease of use in this application:

Reliability ( = % in-spec ) = beta.inv ( 1 – C , n – F , F + 1 )

where...

C = Confidence desired (expressed as a decimal fraction)

n = sample size ( lot size is assumed to be infinite)

F = # of failures seen in the sample

For example, if sample size = 46, failures = 1, desired confidence = 0.95, then...

BETA.INV ( 1 − 0.95, 46 − 1, 1 + 1 ) = 0.90 = 90%

is the % reliability that can be claimed at that confidence level.

* 1. STATEMENTS BASED UPON MEASUREMENT DATA
		1. NORMALLY DISTRIBUTED DATA (or data transformed to normality)
		The most commonly used method is referred to as "Normal K-Tables" (a.k.a. Normal Tolerance Factor Tables, or Statistical Tolerance Factor Tables). These tables are similar to Z-tables in the sense that the body of the table is a list of Z-scores (in K-tables, they are called "Observed K's" or simply "K" or "k" values); however, the row and column headings hold different information than does a Z-table. The row headings in a K-table typically indicate the size of the sample that generated the mean and standard deviation that were used to calculate the observed K value; the major column headings typically indicate various levels of %Confidence; the minor column headings typically provide the lower 1-sided confidence limit on the %-in-specification statistic ("statistic" here equals the %-in-specification what would have been calculated had a Z-table been use rather than a K-table). See example table and calculation, below.



Example calculations for %reliability vs. a 1-sided QC specification:

Sample size = 7

Sample mean = 139.72

Sample standard deviation = 10

1-sided QC specification = 100

Observed K = (139.72 – 100) / 10 = 3.972

%reliability (= %-in-specification) (at 90% confidence) = 0.99 = 99%
which was determined by using the 1-sided-K table shown above.

For 2-sided QC specifications, calculate the Observed K using only whichever side of the QC specification is *closer* to the sample mean; and then use a *2-sided*-K table to determine the %-in-specification.

If an Observed K falls between values in the K-table, a %-in-specification can be determined by interpolation of a *4th-order polynomial curve*, as shown below for a 1-sided specification (this example curve was created using Excel).



*REFERENCES:*

* *ANSI/ASQ Z1.9, Sampling Procedures and Tables for Inspection by Variables*
* *Juran's Quality Handbook, Table V ("One-Sided and Two-Sided...Factors k")*

*CORRECTION FOR OFF-CENTER SAMPLE AVERAGES USED WITH 2-SIDED K-TABLES*

*If the sample average is off-center vs. the QC specifications (i.e. not exactly midway between the upper and lower QC specification limits), a more accurate %-in-specification can be calculated:*

* *Using a 2-sided K-table, calculate the %-in-specification as if the upper QC specification were "closest to the sample mean".*
* *Using a 2-sided K-table, calculate the %-in-specification as if the lower QC specification were "closest to the sample mean".*
* *Add those two values and then multiply by 50%. The resulting value is a more accurate %-in-specification.*
	+ 1. DATA FROM UNKNOWN DISTRIBUTIONS

			1. 1-Sided QC specifications

If *all* of a sample's values are within specification, and the shape of distribution of the population from which the sample was taken is *unknown*, then (using an Excel equation)...

%-in-specification = ( 1 − C ) ^ ( 1 / n )

where

C = confidence as a decimal (*e.g.* use 0.95 for 95%)

n = sample size

For example, if n =100, C = 0.95, then...

 ( 1 − 0.95 ) ^ ( 1 / 100 ) = 0.9705 = 97.05%

is the %-in-specification that can be claimed at 0.95 = 95% confidence.

Formula reference: *Engineering Statistics*, by A. H. Bowker & G. J.
 Lieberman (2nd ed., 1972 by Prentice-Hall) page 311

* + - 1. 2-Sided QC specifications

If *all* of a sample's values are within specification, and the shape of distribution of the population from which the sample was taken is *unknown*, then (using Excel equations)...

%-in-specification = ( Y − 1 ) / (Y + 1 )

where

Y = ( n − 0.5 ) / ( CHISQ.INV( C, 4 ) / 4 )

C = confidence as a decimal (*e.g.* use 0.95 for 95%)

n = sample size

For example, if n =100, C = 0.95, then...

 Y = ( 100 − 0.5 ) / ( CHISQ.INV(0.95,4)/4) = 41.9489

 ( 41.9489 − 1 ) / ( 41.9489 + 1 ) = 0.9534 = 95.34%

is the %-in-specification that can be claimed at 0.95 = 95% confidence.

Formula reference: *Engineering Statistics*, by A. H. Bowker & G. J.
 Lieberman (2nd ed., 1972 by Prentice-Hall) page 311

* + 1. PROCESS CAPABILITY INDICES

To calculate %-in-specification, do *not* use process capability indices (= PCI's, namely: Cp, Cpk, Pp, or Ppk). Instead, calculate %-in-specification directly using the other methods in this section for "confidence/reliability" determinations.

Unless mandated by a government agency or corporate SOP, a PCI that has been calculated per formulas found in SPC or other textbooks should not be reported because it is just a statistic without any associated confidence level. Instead, report the lower 1-sided confidence limit on that PCI, as calculated here next (using an Excel equation)...

1-sided lower confidence limit for PCI =

PCI − NORM.S.INV( C ) \* SQRT( ( 1 / ( 9\*n ) ) + ( PCI^2 ) / (2\*n − 2 ) )

where

C = confidence as a decimal (e.g., use 0.95 for 95%)

n = sample size (for Pp and Ppk, this is the number of data points used in the calculation of the standard deviation that was used to calculate the PCI; but for Cp and Cpk, this is the result of multiplying the number of samples used to calculate the current sample-averages control-chart limits, times the size of each of those samples).

For example, if Cpk (without confidence) = 1.812, and C = 0.95, n = 100 (*e.g.* 20 averages, each created with a sample size of 5, with 20 x 5 = 100), then...

 1.812 − NORM.S.INV(0.95)\*SQRT((1/(9\*100) + (1.812^2) / (2\*100 − 2)) = 1.593

is the Cpk that can be claimed at 0.95 = 95% confidence.

Formula reference: *NIST Handbook of Statistical Methods*, section 6.1.6, found at:

 www.itl.nist.gov/div898/handbook/pmc/section1/pmc16.htm

* + 1. ELECTRONIC EQUIPMENT: MTTF, MTBF, & RELIABILITY

A Mean Time to Failure (MTTF, for single-use-only equipment) or Mean Time Between Failure (MTBF, for repairable equipment) that has been calculated per formulas found in reliability textbooks should not be reported (unless required by regulations or corporate SOP) because they are just statistics without any associated confidence level. Instead, report their lower 1-sided confidence limit, as calculated here next (using Excel equations).

Lower 1-sided confidence limit for either an MTTF or MTBF:

(for Type-I studies = carried out for a pre-determined amount of time or cycles):

= ( 2 \* T ) / CHISQ.INV( C, 2\*F + 2 )

(for Type II studies = carried out until a pre-determined # of failures occurs):

= ( 2 \* T ) / CHISQ.INV( C, 2\*F )

where

T = total time (or cycles) that equipment was working during the study
 (this T is sum of time or cycles of all equipment units studied)

C = confidence as a decimal (*e.g.* use 0.95 for 95%)

F = total number of equipment failure incidents observed during T time.

For example, if C = 0.95, F = 5, T = 1000, then...

 ( 2 \* 1000) / CHISQ.INV( 0.95, 2\*5 + 2) = 95.12

is what can be claimed at 95% confidence for a Type-I study, and...

 ( 2 \* 1000) / CHISQ.INV( 0.95, 2\*5) = 109.25

is what can be claimed at 95% confidence for a Type-II study.

To calculate a *reliability* value for a population of equipment (*i.e.* a lot) *vs.* a defined QC specification (*e.g.* 8 hours or cycles), use the following equation:

Reliability (at confidence level) = e ^ ( − TQC / MTconf )

where the "confidence level" is whatever was used to calculate MTconf, and

 e = base of the natural logarithm ( = 2.718282...)

 TQC = QC specification (in same units as used to calculate MTconf)

MTconf = lower 1-sided confidence limit for MTTF or MTBF (calculated as described above)

For example, if Type I or II, 95% confidence MTTF = 95.12, and TQC = 8, then...

 2.718282 ^ ( − 8 / 95.12 ) = 0.9193 = 91.93%

is the %-in-specification that can be claimed for the lot, at 0.95 = 95% confidence.

Formula reference: *Reliability Statistics* by R. A. Dovich
 (1990 by ASQC Quality Press) chapters 4 and 8

1. AQL SAMPLING PLANS

In order to choose a statistically valid AQL sampling plan for a given product specification, start by determining whether or not a failure of that specification could lead to death or injury to the patient or end-user of the finished product. Such a determination is to be made in reference to the point in the production/inspection process where the AQL sampling plan would be used; this could be at Receiving (*i.e.* incoming) Inspection, In-process Inspection, or Finished Goods QC. Such a determination must be made by reference to relevant "Risk Management" documents (*e.g.* FMEA or relevant SOP); if such documents indicate that failure to meet specification at this point in the production/inspection process is a risk that is controlled *by means of sampling plans*, then the controls described here apply. However, *if Risk Management documents say nothing about the given specification at this point in the production/inspection process, then the controls described here do not apply*; in such a case, any AQL level and any sampling plan is valid, because there is no patient or user risk.

For the sampling plan being used, find the relevant OC curve (see example below) in the sampling plan's documentation; the curve must be for the AQL level *and* lot size being used (curves can differ greatly, depending on lot size). Using that OC curve, determine the %defective that corresponds to a 5% pass-rate (see example below, where 4% is the determined %defective). Compare that %defective to the %defective that Risk Management documents indicate is an acceptable failure rate. If that %defective is equal to or smaller than the failure rate that is acceptable by Risk Management, then the current AQL level is acceptable. If that %defective is larger than the acceptable failure rate, then the current AQL level is *un*acceptable because it is not in compliance with Risk Management requirements.

If the current AQL level is thereby found to be unacceptable, then determine an acceptable AQL level by identifying a smaller AQL level that has an OC curve whose 5% pass-rate %defective is *not* larger than the failure rate that is acceptable by Risk Management. A new statistically valid sampling plan is the combination of that smaller AQL, its corresponding sample size and pass/fail requirements (per the sampling plan's documentation), and a fixed range of lot sizes.





1. GUARD-BANDED QC SPECIFICATIONS (BASED UPON METROLOGY DATA)

*BACKGROUND INFORMATON:*

*Any measurement is the output not simply of a measurement instrument but rather a "measurement system", which includes the processes, equipment, and personnel involved in a measurement. That includes, for example, the process of calibration, the "gold standards" used during calibration, the people who are typically available to make a given measurement, and the various equipment available for making that measurement.*

*Product conformance to product design specifications cannot be assured if the variation in the measurement system is so large that random chance can cause an unacceptably high proportion of non-conforming product to be classified as conforming. If measurement system variation could cause a non-conforming product to be incorrectly classified as conforming, and when such a non-conforming product could cause death or injury to an end-user or patient, and when there is no other relevant "down stream" inspection/test process prior to release for shipment to the customer, then the relevant pass/fail specifications (e.g. QC specs) must be no larger than "guard banded" Design specifications, as described below.*

Determine the standard deviation of the various causes of measurement variability (see A, B, C... list, below); not all of those values may be available, and so in this next step, use only what values are available or that can be reasonably estimated. Square each of the standard deviations (thus creating variances), sum the variances, and take the square root of the sum. Multiplying that square root by an appropriate factor will create a "measurement uncertainty interval". In order to guard against having measurement system variation cause border-line non-conforming product units to be incorrectly classified as conforming, ensure that the QC specification interval is not larger than the Design specification interval minus the "measurement uncertainty interval", as shown in the figure and example calculations below.



The Measurement Uncertainty Interval is calculated as follows:

99% uncertainty interval = 5.15 x SquareRoot( A2 + B2 + C2 + D2 + E2 + X2 )

(use 3.92 instead of 5.15, if a 95% interval is desired instead of a 99% one).

where

A = standard deviation of the accuracy of the reference calibrator (a.k.a. gold standard) used by the calibration company to calibrate the instrument used to make the measurement for which a guard-banded QC specification is sought; unless otherwise instructed by the calibration company, use the following value:

A = (1 / 4 ) x (length of reference calibrator's calibration tolerance interval)

B = standard deviation of the accuracy of the instrument used to make the measurement for which a guard-banded QC specification is sought; use the following value:

B = ( 1 / 4 ) x (length of the measurement instrument's calibration tolerance interval)

C = standard deviation of the variation due to measurements being taken by different people; this value can be derived from a Gage R+R study.

D = standard deviation of the variation due to imprecision of the measurement equipment; this value can be derived from a Gage R+R study.

E = standard deviation of the variation in the "as found" (pre-calibration adjustment) accuracy of the measurement equipment from one calibration date to the next (a.k.a. "drift" or "instability").

X = standard deviation of any other known source of measurement variation.

METROLOGY REFERENCES:

* EA-4/02 (1999): "Expressions of Uncertainty in Measurement in Calibration", section 3.3
* "Simplified Method for Assessing Uncertainties in a Commercial Production Environment (found on May 3, 2004, on the Agilent Technologies', "Metrology Forum: Basics" website, at Agilent.com).
* *Measurement System Analysis*, by AIAG (Auto Industry Action Group)
* *Evaluating the Measurement Process and Using Imperfect Data*, by D. Wheeler (note: Wheeler takes a much different approach than does AIAG).

CONTINUED ON THE NEXT PAGE

1. REVISION HISTORY

|  |  |  |
| --- | --- | --- |
| **REVISION** | **DESCRIPTION of changes since last version** | **DATE** |
| 1.01 | Initial release of first version of this SOP | Aug 16, 2017 |
| 1.02 | Added instructions regarding copy/paste and regarding compatibility issues with MSWord 2010 | Aug 17, 2017 |
| 1.03 | Changed "Word 2010" to "Word 2011" in the instruction regarding compatibility issues. | Aug 31, 2017 |
| 1.04 | Corrected the formula for the signed version of the Coefficient of Determination (in the denominator, replaced "r" with "absolute value of r") | July 19, 2018 |
| 1.10 | Removed some unnecessary parentheses from formulas in section "8.a".Copied the formula for a "lower 1-sided confidence limit on a proportion" (in section "8.a"), rearranged it, and then put it into section "12.a".Miscellaneous minor wording changes. | August 7, 2019 |
| 1.11 | Miscellaneous minor wording / grammar changes. | Aug. 19, 2020 |
| 1.12 | Miscellaneous minor wording changes. | Sept. 3, 2021 |
| 1.20 | Corrected all 3 plots in the Normality section; the Y-axes on each plot were labeled as "Z-table Values" but actually they had mistakenly been constructed using Median Rank values. | March 20, 2024 |
|  |  |  |
|  |  |  |
|  |  |  |