Omair A. Khan

Understanding Statistics for Quality by Design

A TECHNICAL DOCUMENT SERIES

West Pharmaceutical Services · Exton, Pennsylvania

July 31, 2015

Contents

PART I PROCESS DESIGN AND PERFORMANCE

PART II PRODUCT DESIGN AND PERFORMANCE

Part I

Process Design and Performance

A Practical Guide to Utilizing C_p and C_{pk}

Omair A. Khan[1]

[1] West Pharmaceutical Services,
R&D Statistical Engineering Intern
omair@prettynumbe.rs

June 19, 2015

This document explains the use of process capability indices as a way to understand and improve manufacturing processes. It is intended to be an empirical and pragmatic approach to capability analysis without developing the underlying statistical theory. After reading this document, the reader will have a strong grasp on the motivation for process capability indices and what is needed to calculate them, focusing on C_p and C_{pk}.

Why quantify capability?

The ability to manufacture a product within a customer's specifications or tolerances is known as **capability**. **Statistical process control (SPC)** is a methodology for achieving process stability and improving capability through the reduction of variability. In any production process, a certain amount of natural variability will always exist (**chance or common causes of variation**). Occasionally, **assignable or special causes of variation** can be present in the output of a process, arising from improperly adjusted machines, operators, or defective raw material. This type of variability is usually large compared to natural variability and tends to suggest an unacceptable level of process performance. A process that is operating with only chance causes of variation is said to be **in statistical control**. Conversely, a process that is operating under the presence of assignable causes is said to be an **out-of-control process**.

Since process variation can never be totally eliminated, the control of this variation is the key to product quality. Maintaining a stable process average and systematically reducing process variation are the keys to achieving superior quality. If process variation is controlled, then a process becomes predictable. If predictability and consistency are achieved, then a description of the capability of the process to produce acceptable products is possible. A **process capability index** is a statistical measure of process capability. These indices can be used in the following ways:[2]

1. As a basis in the improvement process.
2. As an alarm clock.
3. As specifications for investments. By giving specifications for levels of process capability indices, expected to be reached by new machines, the purchasing process is facilitated.
4. As a certificate for customers. The supplier is able to attach the

[2] Deleryd M (1996)

result from the process capability studies conducted when the actual products were produced, with the delivery.

5. As a basis for new constructions. By knowing the capability of the production processes, the designer knows how to set reasonable specifications in order to make the product manufacturable.
6. For control of maintenance efforts. By continuously conducting process capability studies it is possible to see if some machines are gradually deteriorating.
7. As specifications for introducing new products.
8. For assessing the reasonableness of customer demands.
9. For motivation of co-workers.
10. For deciding priorities in the improvement process.
11. As a base for inspection activities.
12. As a receipt for improvements.
13. For formulating quality improvement programs.

Process capability indices in practice

SPC is primarily a method for monitoring process performance. Many engineers believe that C_{pk} can be used to quantify product quality. This is simply untrue. While C_{pk} can be used to calculate process fallout (Table 1), the decision to accept or reject a production lot of items must be made by **acceptance sampling**. Sampling plans can be derived using a variety of statistical techniques but are commonly chosen by consulting tables outlined in ANSI/ASQ Z1.4 (for attribute data) or ANSI/ASQ Z1.9 (for variables data).

C_{pk}	Sigma level	Yield	Fallout
0.33	1	68.27%	317311
0.67	2	95.45%	45500
1.00	3	99.73%	2700
1.33	4	99.99%	63
1.67	5	99.9999%	1
2.00	6	99.9999998%	0.002

Table 1: Relationship between C_{pk} and non-conforming items (measured in PPM).

As mentioned above, the main goal of capability analysis is to help reduce variability in the manufacturing process. Higher capability indices generally correspond to higher profits as they imply fewer non-conforming parts and better customer satisfaction. Table 2 contains commonly used minimum values for a variety of processes.

Situation	Minimum Capabilty
Existing Process	
Regular	1.33
Critical	1.50
New Process	
Regular	1.50
Critical	1.67
Six Sigma Process	2.00

Table 2: Recommended capability values for two-sided specifications.

Finally, it is important to understand that C_{pk} does not give us the whole picture. One of the disadvantages of C_{pk} is that it does not take into consideration the target or nominal specification. Figure 1 illustrates how the same C_{pk} value can describe two very different processes. For this reason, it is good practice not to base decisions solely on the numerical value of a statistic, but also to graphically visualize the data. Another way to address this difficulty is to use a process capability index that is a better indicator of centering, such as C_{pm} or C_{pkm}.

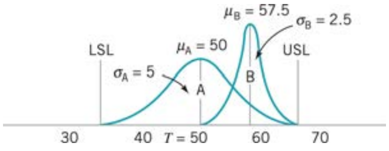


Figure 1: Two processes with $C_{pk} = 1.0$ (Montgomery, 2009).

How to calculate capability indices

While there are numerous process capability indices, the two that are most commonly used in industry are C_p and C_{pk}. These random variables are estimated with the following equations (note the use of the hat to denote the estimate):

$$\hat{C}_p = \frac{USL - LSL}{6\hat{\sigma}}$$
$$\hat{C}_{pk} = \min\left[\frac{USL - \hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu} - LSL}{3\hat{\sigma}}\right]$$

where USL and LSL are the upper and lower specification limits given by the customer, $\hat{\sigma}$ is the sample standard deviation (s), and $\hat{\mu}$ is the sample mean ($\bar{X}$). C_p estimates what the process is capable of producing if the process mean were to be centered between the specification limits. Many times, the mean is not exactly centered and C_p overestimates the process capability. C_{pk} more accurately quantifies capability in these cases and is generally used in place of C_p regardless of the location of the process mean. Note that $C_p = C_{pk}$ when the mean is actually centered between the specification limits.

It is recommended that standard deviation be estimated as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

rather than $\hat{\sigma} = \bar{R}/d_2$. This second equation is commonly used by Six Sigma practitioners but is less statistically tractable than the first equation.

Verifying assumptions

In order to use C_p and C_{pk} properly, three main assumptions must be verified:

1. The individual data must be normally distributed. Normality can be verified by visually inspecting a Q-Q plot or by using the Anderson-Darling or Shapiro-Wilk tests.
2. The individual data must be independent (a particular observation X_t cannot depend on a previous observation X_{t-1}). Independence can be assumed if a plot of the data against the order it was collected displays no obvious pattern. One can also use the Durbin-Watson test for autocorrelation.
3. The process must be under statistical control, which is verified using Shewart control charts. All data points (or subgroup averages) must fall in between the calculated control limits (not to be confused with customer determined specification limits).

Data that deviates from normality can sometimes be transformed to behave better. In practice, one should instead determine the cause for non-normality if the data is expected to be normal (e.g. dimensional data).

If any of these assumptions is not true, then the process capability indices have absolutely no interpretive value!

Determining how much to sample

Simulations have shown that one must have at least 30 samples in order to estimate C_p or C_{pk}. The exact number needed is dependent on the desired power and the type I error one is willing to tolerate. One must also take into account the length of an operator's shift and the type of manufacturing process to determine the frequency of sampling. **OC (operating characteristic) curves** are typically used for these types of calculations. However, it is usually easier to use the sampling plans tabulated in the aforementioned ANSI/ASQ standards.

Using confidence intervals

Because in practice we must estimate C_{pk} with $\hat{C}_{pk}$, the point estimate is subject to a certain degree of error. If we would like to ensure that our process has a C_{pk} of $c_k = 1.33$, for example, our measured $\hat{C}_{pk}$ must be higher. This value (the lower confidence bound) is a function of the desired C_{pk} (c_k), the sample size (n), and the probability of type I error one is willing to tolerate (usually $\alpha = 0.05$). Table 3 shows a few of these values for a variety of sample sizes.

c_k	**10**	**20**	**30**	**40**	**50**	**75**	**100**	**125**	**150**
1.30	2.29	1.87	1.73	1.66	1.61	1.55	1.51	1.48	1.47
1.40	2.45	2.01	1.86	1.78	1.73	1.66	1.62	1.59	1.58
1.50	2.62	2.14	1.99	1.90	1.85	1.78	1.73	1.71	1.69

Table 3: The minimum value of $\hat{C}_{pk}$ for which the process is considered capable (i.e. $C_{pk} \geq c_k$) 95% of the time. (Adapted from Chou *et al.*, 1990)

These values assume that the data were collected individually. When rational subgrouping is employed, the required minimum value of $\hat{C}_{pk}$ will be less than what is tabulated. The exact calculation is beyond the scope of this document, and the reader is referred to Scholz and Vangel (1998) for more details.

Final words: beware of statistical terrorism

This section is largely adapted from Kotz and Lovelace (1998).

C_p and C_{pk} can be extremely useful when used as part of a more comprehensive capability plan. However, these process capability indices have a high potential to be misused. The result is often an atmosphere of "statistical terrorism" within an organization. Burke *et al.* (1991) define statistical terrorism as "the use (or misuse) of valid statistical techniques along with threats and intimidation to achieve a business objective, even if the objective may be reasonable." Below are some examples of statistical terrorism that Burke *et al.* have outlined:

- *"Bandwagon" terrorism.* Customers require suppliers to commit to implementing SPC aggressively and may even demand a commitment to a deadline date for SPC implementation, after which proof of quality via control charts will be required with each shipment. The result? Vendors ignore the statistical methodology and focus on making attractive charts. The vendors simply won't send out-of-control charts to the customer, since they fear the material will be rejected. In this case, statistical terrorism causes the vendor to lie.

- *"Russian roulette" terrorism.* Vendors are contractually bound to a specific quality criteria, measured statistically with C_p or C_{pk} during a special qualification run. Because of the "random variability" of random variables, which include C_p or C_{pk}, sampling variability may result in a calculated value of C_p or C_{pk} below the specified minimum value, even if the process is truly capable. If only a single estimate of C_p or C_{pk} is required, and the process is exactly capable (say 1.33), there is a 50% chance that the estimate will be below the minimum value. Without including confidence limits, you are playing Russian roulette in terms of meeting their requirements.
- *"Tax audit" terrorism.* The use of standards by large customer companies forces vendors to estimate capability based on their guidelines, which may not be appropriate, for example, with non-normal data. The rigid standards deny the vendors the opportunity to understand their own processes and adjust estimation techniques to match them. The consequences of not meeting the standard may not be made clear, and these standards keep the improvement focus on the products, not the processes. The processes have to be improved in order to improve the products.
- *Other forms of terrorism.* These include "self-inflicted wound" terrorism, which results from extreme pressure that managers place upon their own employees to achieve some statistical goal. There is also "academy award" terrorism, which is the requirement that an organization compete for some renowned quality award, internal or external. Finally, there is the "one true statistician" terrorism, where an organization succumbs to the teachings of a specific individual to the exclusion of any other perspective.

Burke *et al.* (1991) suggest that statistical terrorism may be countered in the same way as physical terrorism: by intelligence, speed, and strength. The vendor should be intimately familiar with what the customer needs in their products (intelligence). Statistical expertise should be developed in-house or be readily available from a qualified outside source, so that quick statistical analysis requests by the customer can be met accurately (speed). Finally, the strength of the quality program comes from knowledge, knowledge of your own processes and how to statistically analyze them in an accurate manner.

References and Recommended Reading

ANSI/ASQ Z1.4–2003 (R2013). Sampling Procedures and Tables for Inspection by Attributes.

ANSI/ASQ Z1.9–2003 (R2013). Sampling Procedures and Tables for Inspection by Variables for Percent Nonconforming.

Chou Y-M, Owen DB, and Borrego SA (1990). Lower confidence limits on process capability indices. *Journal of Quality Technology*, **22**(3), 223-229.

Deleryd M (1996). Process capability studies in theory and practice. *Licentiate thesis*, Lulea University of Technology, Luleå, Sweden.

Kotz S and Lovelace CR (1998). *Process Capability Indices in Theory and Practice.* Arnold, London.

Sholz F and Vangel M (1998). Tolerance Bounds and Cpk Confidence Bounds Under Batch Effects. In Kahle W, *et al.* (Eds.), *Advances in Stochastic Models for Reliability, Quality and Safety* (pp. 361-379). Birkhäuser, Boston.

Montgomery DC (2009). *Introduction to Statistical Quality Control*, 6th edn. John Wiley and Sons, New York.

Confidence Intervals for Process Capability Indices

Omair A. Khan[1]

[1] West Pharmaceutical Services,
R&D Statistical Engineering Intern
omair@prettynumbe.rs

July 3, 2015

This document introduces the use of confidence intervals for process capability indices. We begin with a basic description of estimation followed by various equations for calculating the interval for C_p and C_{pk}. The final section describes a custom developed web application to automate these calculations for West engineers.

Interval estimation

Process capability indices are most commonly reported as single point estimates. The **point estimates** of C_p and C_{pk} are calculated as

$$\hat{C}_p = \frac{USL - LSL}{6\hat{\sigma}}$$

and

$$\hat{C}_{pk} = \min\left[\frac{USL - \hat{\mu}}{3\hat{\sigma}}, \frac{\hat{\mu} - LSL}{3\hat{\sigma}}\right].$$

Due to the variability involved in sampling, this is not the most accurate method for quantifying capability. When a process is exactly capable, for example, there is a 50% chance that the estimate will be below the minimum value.[2]

A better estimate can be obtained by calculating the $100(1-\alpha)\%$ confidence interval of the process capability index. Here, α is the probability of type I error one is willing to tolerate. The most common choice for α is 0.05, resulting in a **95% confidence interval**. This is interpreted as follows: if repeated samples are taken and the 95% confidence interval is computed for each sample, 95% of the intervals will contain the true population parameter. Higher confidence levels correspond to wider intervals.

Note that it is not entirely correct to say that there is a 95% chance that the population parameter lies within the interval.

Confidence interval for C_p

Assuming normally distributed process data, $\hat{C}_p$ follows a chi-square distribution. A $100(1-\alpha)\%$ confidence interval for C_p is simple to calculate using Equation 1 (Kane, 1986):

$$\hat{C}_p\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}} \leq C_p \leq \hat{C}_p\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}} \tag{1}$$

[2] Khan OA (June 19, 2015). A Practical Guide to Utilizing C_p and C_{pk}. West Pharmaceutical Services Technical Document.

Confidence interval for C_{pk}

The construction of confidence intervals for C_{pk} is difficult to obtain because the distribution involves the joint distribution of two non-central t-distributed random variables. Several authors have proposed approximate confidence intervals based on various arguments. No single equation is considered best in practice (Kotz & Lovelace, 1998). Four equations are included in this section.

The interested reader is referred to Pearn and Lin (2004) for the exact derivation of the cumulative distribution function of $\hat{C}_{pk}$.

Heavlin (1988)

$$C_{pk} = \hat{C}_{pk} \pm Z_{1-\alpha/2}\sqrt{\frac{n-1}{9n(n-3)} + \frac{\hat{C}_{pk}^2}{2(n-3)}\left(1 + \frac{6}{n-1}\right)} \tag{2}$$

Bissell (1990)

$$C_{pk} = \hat{C}_{pk} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{9n} + \frac{\hat{C}_{pk}^2}{2(n-1)}} \tag{3}$$

Kushler-Hurley (1992)

$$C_{pk} = \hat{C}_{pk}\left(1 \pm \frac{Z_{1-\alpha/2}}{\sqrt{2(n-1)}}\right) \tag{4}$$

Minitab

This formula, used in Minitab 16 and 17, is unique in that it takes batch effects into consideration:

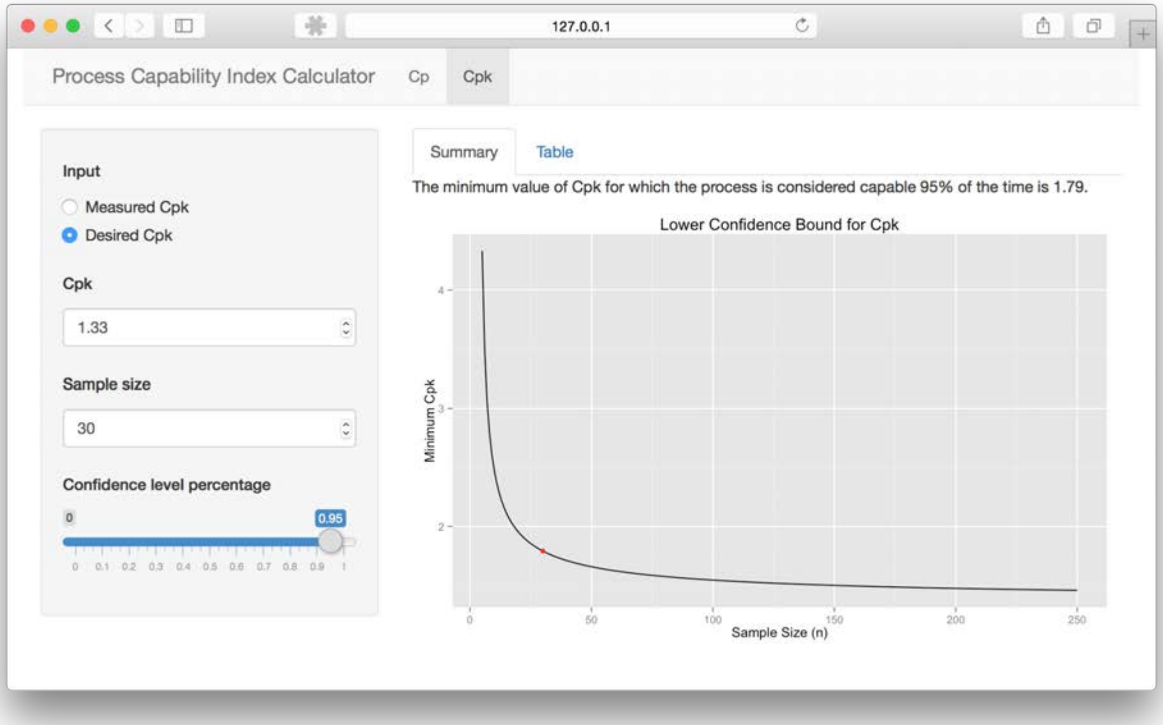
The source for this equation is not clear. It seems to be a modification of the formula proposed by Bissell (Equation 3). Minitab's technical support is trying to find a proper citation. This document will be updated when more information is available.

$$C_{pk} = \hat{C}_{pk} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{N + (m/2)^2} + \frac{\hat{C}_{pk}^2}{2\nu}} \tag{5}$$

Here, N is the total number of observations, m is the sigma tolerance value (6 by default), ν is the degrees of freedom (calculated as $\sum(n_i - 1)$ by default), and n_i is the subgroup size.

Web application for PCI interval estimation

An online tool for calculating the confidence interval of C_p and C_{pk} or its lower bound inverse is available at `https://westelastomer.shinyapps.io/pci_confidence`. A screenshot of the app is shown in Figure 1. The user can specify whether they want to input the measured PCI (output: $100(1-\alpha)\%$ confidence interval) or the desired PCI (output: lower confidence bound for which the process will be capable $100(1-\alpha)\%$ of the time). The second option also produces a plot of the inverse function and a searchable table of values.



Input

○ Measured Cpk

● Desired Cpk

Cpk: 1.33

Sample size: 30

Confidence level percentage: 0 — 0.95

The minimum value of Cpk for which the process is considered capable 95% of the time is 1.79.

Figure 1: Screenshot of web application.

The web application uses Equations 1 and 4 for the calculations. These were chosen because their lower bound inverses $f^{-1}(C_p)$ and $f^{-1}(C_{pk})$ are single-valued functions.

The application will load in any modern browser (including Internet Explorer 9). The tool was developed using Shiny, a web application framework for R. The source code is available on my GitHub page (`https://github.com/prettynumbers/cpk_app`) for forking and modification. Because of the usability limitations on `shinyapps.io` (25 active hours per month on the free account), I recommend that West host the application on their own server or upgrade to a paid account if it is found to be popular.

References and Recommended Reading

Bissell AF (1990). How reliable is your capability index? *Journal of Applied Statistics*, **39**, 331-340.

Chou Y-M, Owen DB, and Borrego SA (1990). Lower confidence limits on process capability indices. *Journal of Quality Technology*, **22**(3), 223-229.

Heavlin WD (1988). Statistical properties of capability indices. *Technical Report 320*, Technical Library, Advanced Micro Devices, Inc., Sunnyvale, CA.

Johnson N and Kotz S (1970). *Continuous Univariate Distributions*, vol 2. 1st edn. John Wiley and Sons, New York. p. 224.

Kane VE (1986). Process capability indices. *Journal of Quality Technology* **18**(1), 41-52.

Kotz S and Lovelace CR (1998). *Process Capability Indices in Theory and Practice.* Arnold, London.

Kushler RH and Hurley P (1992). Confidence bounds for capability indices. *Journal of Quality Technology*, **24**, 188-196.

Montgomery DC (2009). *Introduction to Statistical Quality Control*, 6th edn. John Wiley and Sons, New York.

Nagata Y and Nagahata H (1994). Approximation formulas for the confidence intervals of process capability indices. *Okayama Economic Review*, **25**, 301-314.

Pearn WL and Lin PC (2004). Testing process performance based on capability index C_{pk} with critical values. *Computers & Industrial Engineering*, **47**, 351-369.

Part II

Product Design and Performance

Statistical Procedures for Inter-Rater Reliability

Omair A. Khan[1]

July 17, 2015

This document describes various statistical procedures for measuring inter-rater reliability. These methods quantify the homogeneity in ratings and can be used to show how well two methods of measurement agree.

ANOVA gauge R&R

This method for determining the capability of a measurement system utilizes a designed factorial experiment. The data from this experiment is analyzed using the **random effects model analysis of variance (ANOVA)**. If there are a randomly selected parts and b randomly selected operators, and each operator measures every part n times, then the measurements (i = part, j = operator, k = measurement) can be represented by the model

$$y_{ijk} = \mu + P_i + O_j + (PO)_{ij} + \epsilon_{ijk} \quad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases} \tag{1}$$

With this model, the variance of any observation is $V(y_{ijk}) = \sigma_P^2 + \sigma_O^2 + \sigma_{PO}^2 + \sigma^2 = \sigma_P^2 + \sigma_{\text{Gauge}}^2$. The gauge variability can be decomposed into the **repeatability** variance component (σ^2) and the gauge **reproducibility** ($\sigma_O^2 + \sigma_{PO}^2$). It is common to compare the estimate of gauge capability to the width of the specifications or the tolerance band for the part that is being measured. This is called the **precision-to-tolerance (P/T) ratio**:

The experiment can easily be extended to study different measurement systems by adding an M_ℓ term and its two-way and three-way interactions with P_i and O_j.

$$P/T = \frac{k\hat{\sigma}_{\text{Gauge}}}{USL - LSL} \tag{2}$$

In Equation 2, popular choices for the constant k are $k = 5.15$ and $k = 6$. The value $k = 5.15$ corresponds to the limiting value of the number of standard deviations between bounds of a 95% tolerance interval that contains at least 99% of a normal population, and $k = 6$ corresponds to the number of standard deviations between the usual natural tolerance limits of a normal population. Values of the estimated ratio P/T of 0.1 or less often are taken to imply adequate gauge capability (Montgomery, 2009).

[1] West Pharmaceutical Services,
R&D Statistical Engineering Intern
omair@prettynumbe.rs

Concordance correlation coefficient

The concordance correlation coefficient (r_c) for measuring **agreement** between continuous, normally-distributed variables X and Y is calculated as follows for an n-length data set:

$$r_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\overline{x} - \overline{y})^2} \tag{3}$$

Equation 3 is an estimate of the population concordance correlation coefficient:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

Just like the familiar Pearson correlation coefficient, a value of $r_c = +1$ corresponds to perfect agreement, a value of $r_c = -1$ corresponds to perfect negative agreement, and a value of $r_c = 0$ corresponds to no agreement.

McBride (2005) suggests the following descriptive scale for values of the concordance correlation coefficient:

Value of r_c	**Strength of agreement**
< 0.90	Poor
0.90 - 0.95	Moderate
0.95 - 0.99	Substantial
> 0.99	Almost perfect

Cohen's kappa

Cohen's kappa statistic (κ) is a measure of **agreement** between categorical variables X and Y. It is unique in that it takes into consideration agreement by chance. Kappa can be used to compare the ability of different raters to classify parts or defects into one of several groups. It can also be used to assess the agreement between alternative methods of categorical assessment when new techniques are under study.

Kappa is calculated from the observed and expected frequencies on the diagonal of a square contingency table. Suppose that there are n parts on which X and Y are measured, and suppose that there are g distinct categorical outcomes for both X and Y. Let f_{ij} denote the frequency of the number of parts with the i^{th} categorical response for variable X and the j^{th} categorical response for variable Y. Then the frequencies can be arranged in the following $g \times g$ table:

	$\mathbf{Y=1}$	$\mathbf{Y=2}$	$\cdots$	$\mathbf{Y=g}$
$\mathbf{X=1}$	f_{11}	f_{12}	$\cdots$	f_{1g}
$\mathbf{X=2}$	f_{21}	f_{22}	$\cdots$	f_{2g}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{X=g}$	f_{g1}	f_{g2}	$\cdots$	f_{gg}

The observed proportional agreement between X and Y is defined using the diagonal values as:

$$p(a) = \frac{1}{n}\sum_{i=1}^{g} f_{ii} \tag{5}$$

and the expected agreement by chance is:

$$p(e) = \frac{1}{n^2} \sum_{i=1}^{g} f_{i+} f_{+i} \tag{6}$$

where f_{i+} is the total for the i^{th} row and f_{+i} is the total for the i^{th} column. The kappa statistic is:

$$\kappa = \frac{p(a) - p(e)}{1 - p(e)} \tag{7}$$

Cohen's kappa is generally between 0 and 1, however negative values are possible when there is less than chance agreement. For ordinal data and partial scoring, it is possible to use a weighted form of kappa (Cohen, 1968). When there are more than two categorical variables being compared, one can use Fleiss's kappa.

Viera & Garrett (2005) suggest the following descriptive scale for values of Cohen's kappa statistic:

Value of κ	**Strength of agreement**
< 0	Less than chance
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 0.99	Almost perfect

Bland-Altman plot

The Bland-Altman plot (also known as the Tukey mean-difference plot) provides a quick, graphical method to determine if two raters or test methods agree. Figure 1 shows an example of this plot:

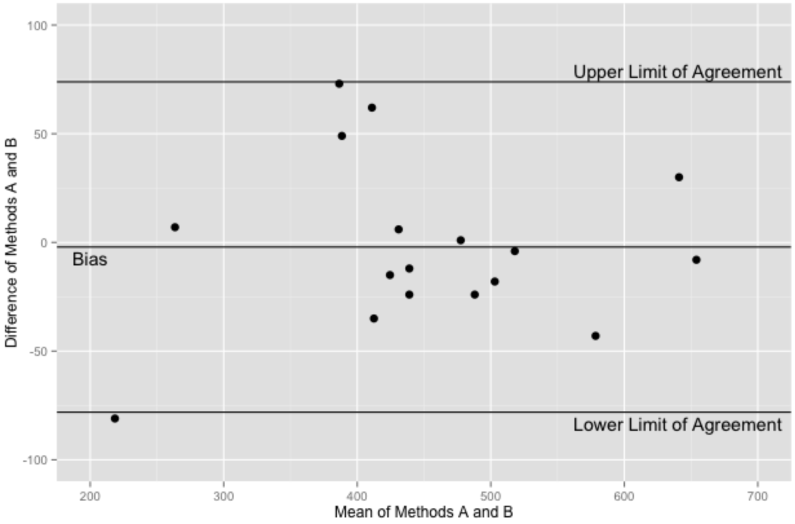


Figure 1: Bland-Altman plot. One observation in the lower-left corner is outside the reference interval.

Each test method is performed on n paired samples. The mean of the two tests is plotted on the horizontal axis and the difference is plotted on the vertical axis, i.e. $S(x, y) = \left(\frac{S_1+S_2}{2}, S_1 - S_2\right)$. The **bias** of the two methods is the mean of these differences ($\overline{S}_y$). A reference interval known as the **limits of agreement** is often calculated as

$\overline{S}_y \pm 1.96s$. However, this equation is not valid for smaller sample sizes. The most accurate formula which can be used in any case is (Hayes & Krippendorff, 2007):

$$\overline{S}_y \pm t_{0.05,n-1} s \sqrt{1 + \frac{1}{n}} \tag{8}$$

The limits of agreement provide insight into how much random variation may be influencing the ratings or test methods. If the measurements tend to agree, the differences between the two sets of observations will be near zero. If one rater or method is usually higher or lower than the other by a consistent amount, the bias will be different from zero.

References and Recommended Reading

Cohen J (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**(4), 213-220.

Hayes AF & Krippendorff K (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77-89.

McBride GB (2005) A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. NIWA Client Report: HAM2005-062.

Montgomery DC (2009). *Introduction to Statistical Quality Control*, 6th edn. John Wiley and Sons, New York.

Viera AJ & Garrett JM (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Journal of Family Medicine*, **35**(5), 360-363.

Statistical Procedures for Testing Equivalence

Omair A. Khan[1]

[1] West Pharmaceutical Services,
R&D Statistical Engineering Intern
omair@prettynumbe.rs

July 30, 2015

This document covers procedures for testing the equality of two or more means including t-tests, one-way ANOVA, and post-hoc procedures.

How can we ensure that a certain product produced by multiple manufacturing plants is actually the same? This question was the motivation behind this final technical document of the series. Once we can confirm that there is strong agreement between different measurement systems at different sites,[2] we can then perform statistical tests to verify equal product quality attributes. The paper describes a hypothesis testing approach to comparing product measurements. If it is found that two or more sets of products that should be the same are actually not equivalent, a closer inspection of the manufacturing processes is warranted.

Testing the equality of two means

The most common method for testing $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ is the t-test. The observations in each group must follow a normal distribution. The statistic is calculated differently for equal and unequal sample sizes and variances. The t-statistic is then compared to the value of $t_{\text{critical}} = t_{1-\alpha,\nu}$ (found in a table) to make a decision:

$$\begin{cases} \text{if } t < t_{\text{critical}} \text{ then do not reject } H_0 \\ \text{if } t \geq t_{\text{critical}} \text{ then reject } H_0 \end{cases}$$

The Mann-Whitney U test (Wilcoxon rank-sum test) is the analogous non-parametric test for testing whether two samples come from the same population. It does not require that the samples be normally distributed.

Equal sample sizes, equal variances

The t-statistic is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2}/\sqrt{n}} \tag{1}$$

where $s_{X_1X_2} = \sqrt{s^2_{X_1} + s^2_{X_2}}$ and the t-statistic has $\nu = 2n - 2$ degrees of freedom.

Unequal sample sizes, equal variances

The t-statistic is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

[2] Khan OA (July 17, 2015). Statistical Procedures for Inter-Rater Reliability. West Pharmaceutical Services Technical Document.

where $s_{X_1X_2} = \sqrt{\frac{(n_1-1)s_{X_1}^2+(n_2-1)s_{X_2}^2}{n_1+n_2-2}}$ and the t-statistic has $\nu = n_1 + n_2 - 2$ degrees of freedom

Equal or unequal sample sizes, unequal variances

Welch's *t*-test is an adaptation of Student's t-test for unequal variances. The t-statistic is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

and the degrees of freedom are calculated as:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}} \tag{4}$$

Here, $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are the degrees of freedom associated with the two variance estimates.

Paired samples

When the same set of samples is used in both groups, we can do the **paired *t*-test** to get more power. The t-statistic is calculated as:

$$t = \frac{\overline{X}_D}{s_D/\sqrt{n}}. \tag{5}$$

For this equation, the differences between all pairs must be calculated. The average ($\overline{X}_D$) and standard deviation (s_D) of those differences are used in the equation. The degrees of freedom for the hypothesis test are calculated as $\nu = n - 1$.

Two One-Sided Tests (TOST)

In some cases, it is acceptable to conclude equivalence if the difference of the two means falls between an upper and lower bound. The null hypotheses for non-equivalence are:

Minitab 17 has the functionality to do TOST under the menu heading "Equivalence Tests."

$$H_{0,1} : \mu_1 - \mu_2 \leq \delta_L \quad \text{and} \quad H_{0,2} : \mu_1 - \mu_2 \geq \delta_U$$

and the alternative hypothesis of equivalence is:

$$H_1 : \delta_L < \mu_1 - \mu_2 < \delta_U$$

If we can assume that the two groups of normally-distributed values have the same variance, the calculation of the two one-sided test statistics uses the following equations:

$$t_L = \frac{(\overline{X}_2 - \overline{X}_1) - \delta_L}{SE} \quad (6)$$

$$t_U = \frac{(\overline{X}_2 - \overline{X}_1) - \delta_U}{SE} \quad (7)$$

where the standard error is:

$$SE = \sqrt{\frac{\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (8)$$

The critical value $t_{\text{critical}} = t_{1-\alpha, n_1+n_2-2}$ is used to make a two-part decision:

$$\begin{cases} \text{if } t_L < t_{\text{critical}} \text{ and } t_U > t_{\text{critical}} \text{ then do not reject } H_0 \\ \text{if } t_L \geq t_{\text{critical}} \text{ or } t_U \leq t_{\text{critical}} \text{ then reject } H_0 \end{cases}$$

Testing the equality of three or more means

When we are interested in testing the equality of more than two means, we can perform a **one-way analysis of variance (ANOVA)**. In the special case of two groups, the F-test used in the ANOVA is equivalent to the t-test (since $F = t^2$). The hypothesis we are testing is:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{vs.} \quad H_1 : \text{at least one mean is different}$$

The theory and procedure of the ANOVA are beyond the scope of this document. The reader is encouraged to look at any introductory statistics book for a discussion on this versatile test. It is relatively robust to small deviations from normality, however the assumption of homoscedasticity (equal variance) must be satisfied. The ANOVA cannot be used to determine which means are different if the null hypothesis is rejected. Post-hoc testing procedures are therefore necessary and are described in the next section.

The Kruskal-Wallis one-way analysis of variance is the analogous nonparametric test for testing whether three or more samples come from the same population. It does not require that the samples be normally distributed.

Multiple comparisons

Sometimes we are interested in considering a set of statistical inferences simultaneously. It is not acceptable to sequentially perform these tests without alteration as the probability of incorrectly rejecting the null hypothesis (type I error) increases exponentially. For example, if we are interested in making 10 pairwise comparisons between $k = 5$ groups and try to do a series of t-tests with individual 95% confidence levels, our overall confidence level falls to $(95\%)^{10} = 59.9\%$!

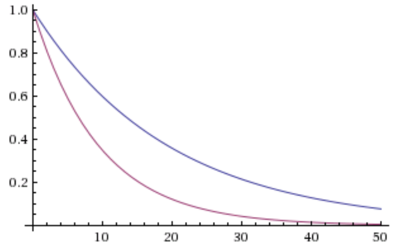


Figure 1: The overall confidence level of a set of simultaneous inferences. The blue and pink lines are for 95% and 90% confidence levels respectively.

To overcome this issue, multiple testing correction methods must be used. This section covers the most commonly used procedures. Except for the Bonferroni correction, all other methods are post-hoc analyses and should only be run if the ANOVA procedure indicates that the means are not equal. In calculating the test statistic for these methods, the MSE is the mean squared error from the ANOVA output.

Bonferroni correction

This is the simplest method for multiple comparisons. It does not require an ANOVA to be run prior to performing the tests. The procedure involves a series of t-tests performed at an adjusted confidence level of $100(1-\alpha^*)\%$ where $\alpha^* = {}^{\alpha}/_{k}$. While the individual pairwise tests are performed at a higher confidence level, the overall confidence level is still approximately $100(1-\alpha)\%$.

Tukey-Kramer (Tukey's HSD) test

When doing all pairwise comparisons, this method is considered the best available for unequal sample sizes. When samples sizes are equal and confidence intervals are not needed Tukey's test is slightly less powerful than the Bonferroni correction, but the loss in power is very small unless the groups are large. We are interested in testing the hypothesis:

$$H_0 : \mu_i = \mu_j \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_j$$

The test statistic is calculated as:

$$q_{\text{obs}} = \frac{\overline{X}_i - \overline{X}_j}{SE} \tag{9}$$

where $SE = \sqrt{\left(\frac{MSE}{2}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$ is the standard error. The critical value $q_{\text{critical}} = q_{\alpha,k,N-k}$ can be found in a table of values and is used to make a decision for each pairwise comparison:

$$\begin{cases} \text{if } |q_{\text{obs}}| < q_{\text{critical}} \text{ then do not reject } H_0 \\ \text{if } |q_{\text{obs}}| \geq q_{\text{critical}} \text{ then reject } H_0 \end{cases}$$

Dunnett's test

When we are only interested in comparing k treatments against a control (for a total of $k+1$ groups), Dunnett's test is the preferred post-hoc analysis. Observations are allocated as n for each treatment and $n_{\text{control}} = n\sqrt{k}$ for the control group. We are interested in testing the hypothesis:

$$H_0 : \mu_i = \mu_{\text{control}} \quad \text{vs.} \quad H_1 : \mu_i \neq \mu_{\text{control}}$$

For example, if we choose a total sample size of $N = 60$ with $k = 4$ treatments, then each treatment should have $n = N/(k + \sqrt{k}) = 60/(4 + \sqrt{4}) = 10$ observations and the control should have $n_{\text{control}} = n\sqrt{k} = 10\sqrt{4} = 20$ observations.

The test statistic is calculated as:

$$q_{\text{obs}} = \frac{\overline{X}_{\text{control}} - \overline{X}_i}{SE} \tag{10}$$

where $SE = \sqrt{MSE\left(\frac{1}{n_{\text{control}}} + \frac{1}{n_i}\right)}$ is the standard error. The critical value $q_{\text{critical}} = q_{\alpha,k+1,N-k+1}$ can be found in a table of values and is used to make a decision for each pairwise comparison with the control:

This is not the same q_{critical} as the one for Tukey's HSD test.

$$\begin{cases} \text{if } |q_{\text{obs}}| < q_{\text{critical}} \text{ then do not reject } H_0 \\ \text{if } |q_{\text{obs}}| \geq q_{\text{critical}} \text{ then reject } H_0 \end{cases}$$

Scheffé's test

This is the most flexible multiple testing procedure as it allows for comparing any number of possible contrasts. If only pairwise comparisons are to be made, the Tukey-Kramer method will result in a narrower confidence limit, which is preferable. In the general case when many or all contrasts might be of interest, Scheffé's test tends to give narrower confidence limits and is therefore the recommended method.

Some of the possible null hypotheses for Scheffé's test include:

$$H_0 : \mu_i = \mu_j$$

$$H_0 : \mu_3 - \frac{\mu_1 + \mu_2}{2} = 0$$

$$H_0 : \frac{\mu_3 + \mu_4 + \mu_5}{3} - \frac{\mu_1 + \mu_2}{2} = 0$$

For an arbitrary contrast $C = \sum_{i=1}^{k} c_i \mu_i$ where $\sum_{i=1}^{k} c_i = 0$, the test statistic is calculated as:

$$S_{\text{obs}} = \frac{\left|\sum c_i \overline{X}_i\right|}{SE} \tag{11}$$

where $SE = \sqrt{MSE\left(\sum \frac{c_i^2}{n_i}\right)}$ is the standard error. The critical value $S_{\text{critical}} = \sqrt{(k-1)\, F_{\alpha,k-1,N-k}}$ is calculated and used to make a decision:

$$\begin{cases} \text{if } S_{\text{obs}} < S_{\text{critical}} \text{ then do not reject } H_0 \\ \text{if } S_{\text{obs}} \geq S_{\text{critical}} \text{ then reject } H_0 \end{cases}$$

References and Recommended Reading

DeGroot MH and Schervish MJ (2011). *Probability and Statistics*, 4th edn. Addison-Wesley, Boston.

Hogg RV, Tanis E, and Zimmermann D (2014). *Probability and Statistical Inference*, 9th edn. Pearson.

Lehmann EL and Romano JP (2005). *Testing Statistical Hypotheses*, 3rd edn. Springer, New York.