QUICK REFERENCE GUIDE

APPLICATION OF STATISTICS IN R&D AND MANUFACTURING

MODULE 1:

DISTRIBUTIONS, NORMALITY, CAPABILITY, PERFORMANCE

SUMMIT QUALITY CONSULTING

TABLE OF CONTENTS

DESCRIPTIVE STATISTICS

WHAT IT IS

Descriptive statistics characterize data, defining what the data is. Descriptive Statistics can be numerical or visual summaries of the information in a data set. The numerical summaries are typically measures of ***location*** and ***dispersion***. The visual summary is typically a bar chart, line graph, or a ***Histogram***.

WHEN TO APPLY IT

Descriptive statistics are ubiquitous, forming the foundation of all other statistics discussed in this manual.

WHAT TO KNOW

MEASURES OF LOCATION

Mean is the most common measure of central tendency, and is the arithmetic average of all measurements in a data set. The Mean of a sample is often expressed as x-bar ($\bar{x}$).

Median is the middle number, or center value of a data set when all the data are arranged in increasing order.

Mode is the value that occurs most frequently in a data set.

MEASURES OF DISPERSION

Range is the highest value minus the lowest value in a data set.

Standard Deviation is a measure of dispersion that indicates if values tend to be close to the mean (indicated by a low standard deviation value) or spread out away from the mean (high standard deviation value). Standard Deviation is expressed in the same units of measure as the mean and individual data points.
Standard Deviation of a sample can be expressed as sd or s or stdev, and is defined by the following equation:

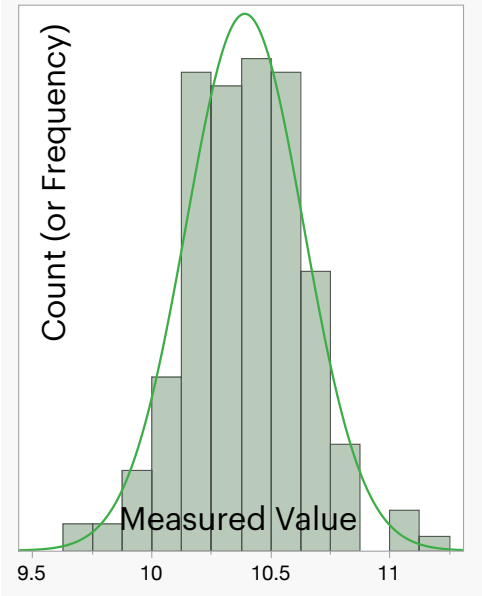
$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Where N = number of samples in the data set, x_i is an individual value, and $\bar{x}$ is the mean.

Variance is a standard deviation squared.

VISUAL DESCRIPTIVE STATISTICS: THE HISTOGRAM

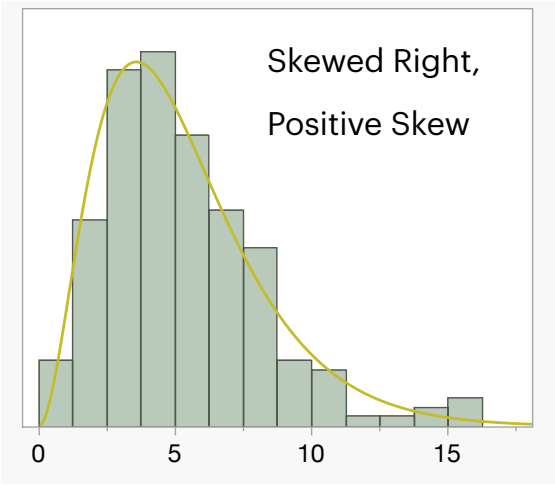
A ***Histogram*** is a plot of a frequency distribution, which displays the centrality and dispersion of the data. The measured value is on the x-axis and the frequency of that measure is on the y-axis.

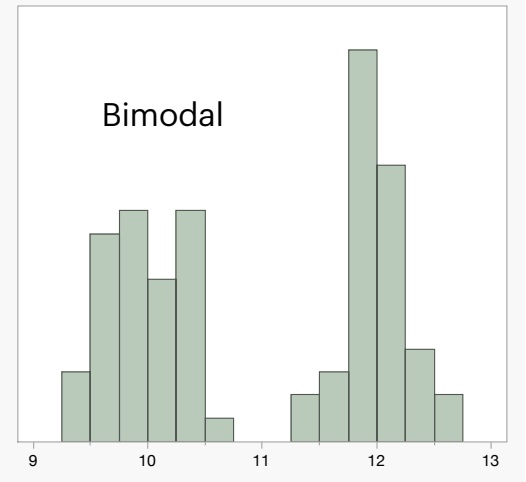


Data displayed in histograms are often symmetrical, as shown on the right. However, some data will distribute differently. Two considerations to be aware of are *skewness* and *bimodality*.

Skewness

Skewed data sets have a tail that disperses either to the left (negative skew) or the right (positive skew). Data can display skewness when the mean approaches an absolute, such as 0 or 100 as happens when plotting percentages. Skewness can also appear when efforts are made to minimize (positive skew) or maximize (negative skew) the measured value.





Bimodality

A distribution with two modes, or most frequently occurring value, is said to be bimodal. Bimodal histograms indicate that there is a factor influencing the data set that creates two distinct populations: two populations are lumped together in one data set. With bimodal distributions, the measures of location and dispersion, such as mean and standard deviation, are not meaningful until the data set is split into its distinct populations.

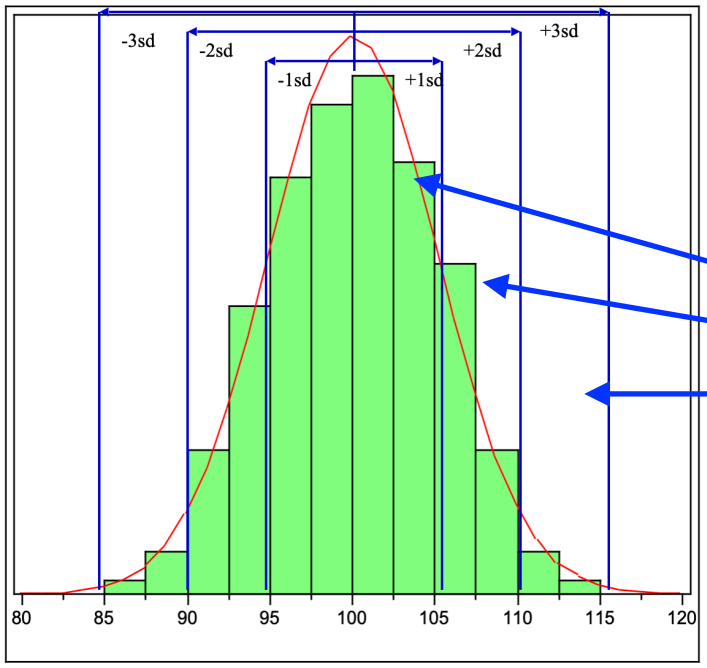
THE NORMAL DISTRIBUTION

WHAT IT IS

he Normal Distribution is a continuous, bell-shaped, symmetric distribution such that data frequency within the data set can be defined by two parameters: the mean and the standard deviation.

WHEN TO APPLY IT

The Normal Distribution is the basis for many other statistical analyses. It is a requirement for using analyses such as the Z score, capability analysis, t-tests, and other variables-data analyses.



The Normal Distribution

WHAT TO KNOW

The Normal Distribution is the simplest and most common of several ***Probability Distributions***. All you need to know is the mean and the standard deviation to understand probabilities of dada distribution.

UNIVERSAL PROBABILITIES

Every normal distribution has a universal probability of data points residing within n standard deviations of the mean. For example:

68.26% of the population will fall within **+/- 1** standard deviation of the mean.

95.46% of the population will fall within **+/- 2** standard deviations of the mean, and

99.73% of the population will fall within **+/- 3** standard deviations of the mean.

± 1sd = 68.26% of the data
± 2sd = 95.46% of the data
± 3sd = 99.73% of the data

This is true for all Normal Distributions, regardless of the mean or sd.

TABLE OF PROBABILITIES

Refer to the following table for further information regarding probabilities of the Normal Distribution (refer to the section on Capability Analysis for the definition of Cpk):

SD Limits (+/-)	% of Population Within Limits	PPM Defective (Outside Limits)	Cpk
0.6745	50.00	500,000	0
1.00	68.27	317,300	0.33
2.00	95.45	45,500	0.67
2.36	98.00	20,000	0.79
3.00	99.73	2,700	1.00
3.29	99.90	1,000	1.10
3.54	99.96	400	1.18
3.89	99.99	100	1.30
4.00	99.9937	63	1.33
4.26	99.9980	20	1.42
4.42	99.9990	10	1.47
4.5	99.9997	3.4	1.50
5.00	99.99994	0.6	1.67
6.00	99.9999998	0.002	2.00

TEST FOR NORMALITY

WHAT IT IS

The Test for Normality is a statistical test to determine if the data in a sample are distributed according to the probabilities of the Normal Distribution.

WHEN TO APPLY IT

Use a Test for Normality to confirm a sample is from the Normal Distribution, including:

- Whenever you want to apply the probabilities of the Normal to your data analysis.
- Whenever another statistical test relies on the assumption that the data are distributed normally, such as a Z score, t-test, or Capability Analysis

WHAT TO KNOW

To test for normality, use a normality test in a statistical software package. Commonly accepted tests are the Shapiro-Wilk test or the Anderson-Darling test.

•Tests for normality require a sampling size between about 20 and 100 to be accurate, with an optimal sample size of about 30. If sample sizes are less than 20 or greater than 100, the normality test loses its power to determine if the probabilities of the normal distribution apply to the sample or not.

•If the sample size is less than 20, generate more data.

•If the sample size is greater than 100, select a random sample of 30 values from the data set to test for normality.

PROCEDURE:

- The Null Hypothesis is that the data are from the Normal Distribution. Small p-values reject the Null Hypothesis.
- From the software output, identify the p-value of the Normality Test. When the p-value is less than 0.05 we have a high degree of confidence that the data are **not** distributed normally (i.e. we consider the distribution to be non-normal).
- When the p-value is greater than 0.05 we cannot conclude that the distribution is non-normal, so we consider the distribution to be normal.

NON-NORMAL DATA

WHAT IT IS

Non-Normal data is a sample that does not conform to the probabilistic parameters of the Normal Distribution. When the Test for Normality fails (p-value is less than 0.05), then we consider the data non-normal.

SPECIAL CAUSES

Special Causes are sources of variation that are not always present, and only affect a portion the process output. They are often intermittent and unpredictable. Special Causes can result in a process output that is unstable, incapable, and unpredictable.

COMMON CAUSES

Common Causes are sources of variation that are always present and consistently act on the process. Examples include raw material lot variation, machine setting variation, variation within a machine from rep to rep, ect. Common Cause variation results in a process that is stable, repeatable, and therefore predictable over time.

WHAT TO KNOW

When data are non-normal, we cannot use certain statistical analyses that rely on normality. Follow one or more of the four approaches below for dealing with non-normal data:

1. Investigate Special Causes

Special Causes may be the source of non-normal data. First, investigate the data for special causes. IF a special cause is identified and remedied (you have assurance that the special cause will not recur), the special cause data may be removed from the sample and the normality test may be run again.

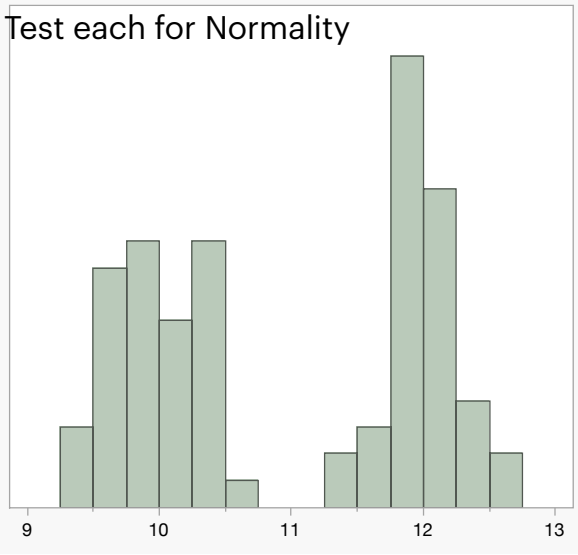
2. Split the Data

Sometimes the source of non-normality is a dependence on a particular contained in the data set, such as a particular machine, operator, or timeframe. This often looks like a bimodal or trimodal data set. In this case, the data sample may be split and each of the new samples evaluated individually for normality. Turn one data set into multiple and determine normality in each new set.

Find the cause,

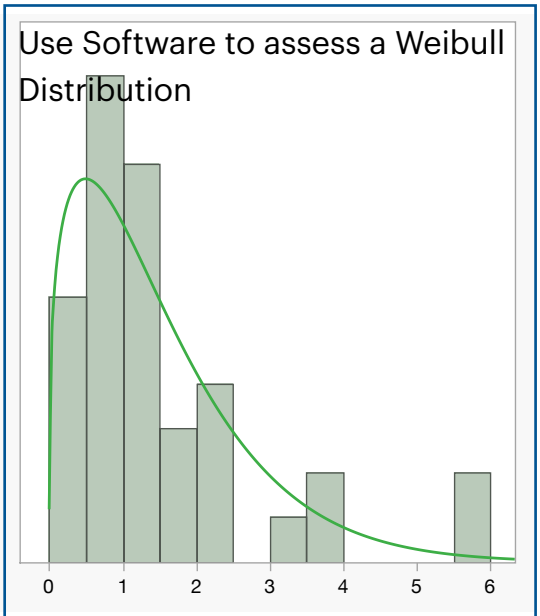
Split the data into two sets

Test each for Normality



3. Transformation or Best Fit

It may be possible to identify a transformation of the data that results in normally distributed values. A ***transformation*** is essentially a mathematical computation applied to each individual data point, and then the transformed data can be assessed for normality. The most common transformation is to take the logarithm of the values. If specification limits are relevant to the data analysis, the same transformation must be applied to any specification limits to render correct conclusions in an analysis.



Statistical Software is capable of completing many of the statistical analyses described in this document with standard probability distributions other than the Normal Distribution, such as a Johnson SI, exponential, or logarithmic. It is acceptable to use the software to complete analyses with these distributions.

- Ensure the sample size is between 20 and 100.
- Use the software to identify the distribution that best fits the data.
- Use the p-value to determine if the data adequately fit the distribution with a high degree of confidence.
- Complete the analysis with the software using the best fit distribution.

4. Convert to Attribute Data

If none of the above approaches are appropriate for your data set, then convert the data to an attribute (pass/fail) and complete the analysis with one of the attribute analysis techniques described in this Guide.

Z SCORE

WHAT IT IS

Z, or Z Score, is a term that describes how many standard deviations from the mean an individual data point resides in a normal distribution. The equation for Z is:

$$(X - \mu) / \sigma = Z$$

WHEN TO APPLY IT

Z is commonly used to determine probabilities from a normal distribution.

The Z tables are used to determine the proportion of the population that resides below a particular Z score.

WHAT TO KNOW

Z Score and the Z tables assume the data are from the Normal Distribution. Test for Normality prior to determining Z score.

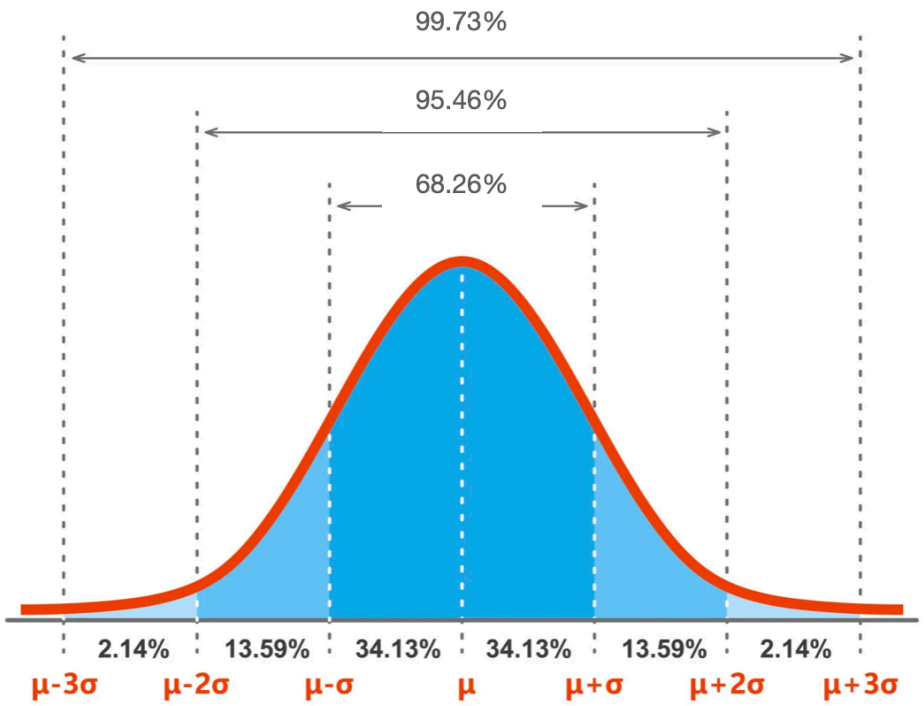
In the equation:

$$(X - \mu) / \sigma = Z$$

X is the data point being analyzed,
μ is the population mean, and
σ is the population standard deviation

To graphically represent the Z distribution, transform all data points to Z and plot them in a histogram. The mean will be 0 and the sd will be 1.

The Z tables are look-up tables that tell the proportion of the population below a particular Z score.
The Y-axis is the Z score to the first decimal place.
The X-axis is the second decimal place of the Z score.
For example, if we wish to know what proportion of the population is below -1.23 Z, we look to the row -1.2 (Y-axis) and the column .03 (X-axis) and find the proportion is 0.1093. In other words, 10.93% of the population of a normal distribution is less than -1.23 standard deviations from the mean.



Z TABLES

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

CAPABILITY ANALYSIS

WHAT IT IS

Capability Analysis is a set of tools that measures a process distribution against the process specifications. Process Capability measures will tell us if we can manufacture a product to specifications, as well as predict the proportion of units that will be out of specification.

WHEN TO APPLY IT

We use Capability Analysis:

- when establishing a process, to confirm that it will consistently meet specifications.
- when preparing or validating process changes, to confirm that the process changes will not adversely affect defect rates.
- during failure investigations to determine if the source of failures is from within lot variation, between lot variation, or from lack of centeredness.

WHAT TO KNOW

Each time we run a process, the process output will vary. The accumulation of many outputs from many runs of a process will produce a distribution of results; variation around a mean.

Process Capability measures assume a normal distribution. Before applying these measures, ensure the data are distributed normally.

We will discuss four measures of Process Capability: C_p, C_{pk}, P_p, and P_{pk}

The combination of these four indices will help us to understand the process centeredness, variation, and proportions in and out of specification.

DEFINITIONS

Short-term standard deviation:
A measure of standard deviation that is typically bound to one lot, one machine, one day, or all of these factors. Also known as “within” or “within lot sd”.

Long-term standard deviation:
A measure of standard deviation that includes measurements across multiple runs, days, operators, machines, etc. It includes all factors that contribute to variability in the process output over the long-term. Also known as “total sd” or “between lot sd”.

Stability:
The characteristic of having a predictable, small amount of variation over time.

Capability:
The ability to produce product that consistently meets specifications.

C_P, PROCESS POTENTIAL SHORT

C_p is a ratio of the spread of the distribution to the specification width. **C_p uses short-term standard deviation**, and is defined by the formula:

$$C_p = \frac{USL - LSL}{6sd_{short}}$$

Where:
USL is Upper Spec Limit
LSL is Lower Spec Limit

Interpreting *Cp*

- Typically, we consider a process with a C_p >1 to be a process that has good potential to meet specifications.
- In a normal distribution, when C_p = 1, then 99.73% of the data is expected to fall within specification, assuming the process is centered within the specifications and the process is stable over the long term.
- The greater the C_p, the greater the process' potential to produce units within specification.
- It is considered a best practice to design and control a process to have C_p > 1.5 to allow for process variation over time.

C_{PK}, PROCESS CAPABILITY SHORT

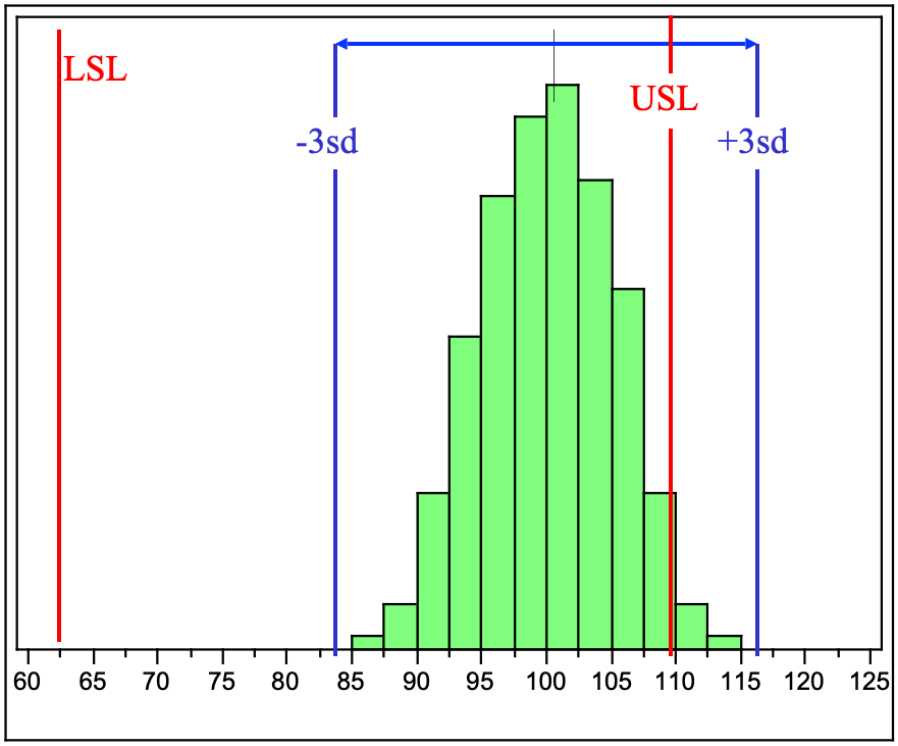
C_{pk} is a ratio of the spread of the distribution to the specification width, and includes an indication of *centeredness* of the distribution (Note that C_p does not indicate centeredness). Like C_p, the C_{pk} calculation uses short-term standard deviation. For C_{pk}, two computations are made, one against each specification, and then the lower of the two results is used as the C_{pk} index. Also, the C_{pk} index can be calculated with a one-sided specification, using only the one computation that is relevant.

$$C_{pk} = \text{smaller of} \left(\frac{USL - \bar{x}}{3sd_{short}} \; Or \; \frac{\bar{x} - LSL}{3sd_{short}} \right)$$

Where:
USL is Upper Spec Limit
LSL is Lower Spec Limit
$\bar{x}$ is the mean of the data

Interpreting *C_{pk}*

- C_{pk} is a more valuable measure than C_p because it takes into account centeredness. Process are rarely perfectly centered.
- Like the C_p measure, a good rule of manufacturing is to design and control processes to have a C_{pk} index of greater than 1.33, and the greater the C_{pk}, the more capable the process is to produce units within specification.
- When C_{pk} = 1, then 99.87% of the units are expected to fall within specification. It is considered a good practice to have C_{pk} greater than 1.33, to allow for long-term standard deviation and still have a high proportion of product within specification.



C_p = 1.33

C_{pk} = 0.55

In the picture above:

- The C_p is 1.33, indicating that the spread of the distribution (*variation*) is sufficiently smaller than the specification width. The process has the *potential* to consistently meet specifications.
- However, the C_{pk} is 0.55, indicating that there is a significant portion of the process that is outside of specification. The process is not *capable* of consistently meeting specifications.
- If we only had looked at the Cp measure, we would have missed that the process is off-center.
- The combination of the two numbers Cp and Cpk indicate where the problem resides, within a short term timeframe: the process variation is good, but the process is off-center.

P_P, PROCESS POTENTIAL LONG

The index P_p is similar to C_p, but with one important difference: P_p uses ***long-term standard deviation***, also called sd total, instead of short-term.

With this in mind, if we use the C_p to make inferences about the process in the long term, then we are assuming that the process is *stable* over the long-term. This is often a poor assumption. When we use P_p, we are more able to correctly judge future potential of the process.

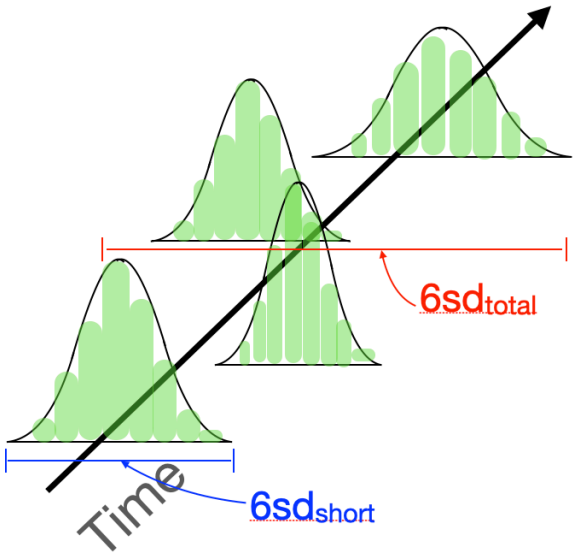
The equation for P_p is:

$$P_p = \frac{USL - LSL}{6sd_{Total}}$$

Interpreting P_p

- P_p is interpreted similarly to C_p in that it can judge the potential of a process to meet specifications.
- Like C_p, it does not consider centeredness, which is why we use the term *potential* (and not *capable*).
- With P_p, like C_p, the higher the number, the greater the potential to meet specifications. It is considered best practice to have P_p, greater than 1.33.

Short Term vs Long Term SD



The picture to the left depicts the difference between short-term and long-term sd.

The various histograms represent short-term production batches: for example one lot or one day.

Short-term sd is confined within one batch.

Long-term sd measures and includes the variation between all batches over a long time.

Long-term sd is conceptually greater than short-term. However, it is possible, on occasion, for calculations to render a greater sd value for short-term sd compared to long-term. This can happen when the short-term sd is based on a small data set or one that is greater than the typical short-term distribution. (You picked the fattest distribution to measure short-term!)

P_{PK}, PROCESS PERFORMANCE

The index P_{pk} takes into account centeredness, and uses long-term standard deviation. Because of these parameters in the calculation, it is the most powerful metric for judging if a process can meet specifications, when used as a stand-alone metric.

The equation for P_{pk} is:

$$P_{pk} = \text{smaller of} \left(\frac{USL - \bar{x}}{3sd_{Total}} \; Or \; \frac{\bar{x} - LSL}{3sd_{Total}} \right)$$

Interpreting P_{pk}

- It is considered a best practice to design and control a process to have a $P_{pk} > 1$.
- When $P_{pk} = 1$, then 99.87% of units are expected to fall within specifications in the long term.
- It is considered best practice to have P_{pk} greater than 1.

APPLYING THE FOUR MEASURES IN COMBINATION

Note that sd_{total} will typically be greater than sd_{short}. Therefore, C_{pk} will be smaller than C_p and P_{pk} will be smaller than P_p.

Likewise, C_p will always be larger than P_p and C_{pk} will always be larger than P_{pk} because of the assumption or non-assumption of centeredness. In summary:

$$P_{pk} < C_{pk} < C_p \quad \text{and} \quad P_{pk} < P_p < C_p$$

The four indices can be used together to make inferences about process stability and process capability.

If	And…	Label	Source	Action
Ppk < 1	Pp > 1.33	Centeredness	Centeredness	Adjust process controls to center the process. Consider DOE and component specifications.
Ppk < 1	Cpk > 1.33	Unstable	Between run variation	Investigate causes of lot-to-lot variation: raw material lots, process settings, operator or shift biases
Cp < 1.33	Cpk = Cp	Incapable	Within run variation (but centered)	Reduce sources of variation: Consider DOE, process controls, component specifications
Cp < 1.33	Pp = Cp	Incapable	Within run variation (but stable)	Reduce sources of variation: Consider DOE, process controls, component specifications
Cp < 1.33	Cpk < 1 And ≠ Cp	Incapable, Centeredness	Within run variation and not centered	Reduce sources of variation, and center process: Consider DOE, process controls, component specifications
Cp < 1.33	Pp ≠ Cp	Incapable, Unstable	Within run and between run variation	Reduce sources of short and total variation: Consider DOE, process controls, component specifications

OUTLIER ANALYSIS

WHAT IT IS

Outlier Analysis is a method for identifying aberrant data, known as outliers. This section will also suggest best practice for how to deal with identified outliers.

WHEN TO APPLY IT

We use Outlier Analysis whenever we suspect that a datum point might be outside of the probabilities defined by the normal distribution.

Note that the techniques described here assume that data are distributed normally.

WHAT TO KNOW

Aberrant data can bias the summary statistics and conclusions of an analysis. An aberrant data point is also called an **outlier**.

The magnitude of the influence of such aberrant data depends on the sample size and the number of outliers.

Outliers can be caused by transcription error, misreading a measurement, equipment failure, or procedural error.

There are many reliable tests identify outliers. This manual will describe two: **Dixon's Test,** which we use for small sample sizes, and the **5Z test**, which we use for large sample sizes.

DIXON'S TEST:

- Dixon's test is a simple calculation that identifies outliers with 95% probability.
- Use this method when the sample size is small, **between 5 and 20.**
- To declare an outlier, compare the suspected aberrant value(s) to the "normal majority"; the other data points in the sample.

DIXON'S TEST PROCEDURE:

1. Arrange the data in increasing order. Call the ordered data x1, x2, x3, x4, …, xN
2. Calculate the ratio (denoted *r*) as follows:
   - For an outlier suspected at the high end: $r = (Xn - Xn\text{-}1)/(Xn - X1)$
   - For an outlier suspected at the low end: $r = (X2 - X1)/(Xn - X1)$
3. Compare the value r to the critical value in the table below for the sample size in the data set. If the r value is greater than the critical value, then the point can be considered an outlier.

DIXON'S TABLE OF CRITICAL VALUES:

Sample Size	Critical Value (95%)	Sample Size	Critical Value (95%)
5	0.6423	13	0.3615
6	0.5624	14	0.3496
7	0.5077	15	0.3389
8	0.4673	16	0.3293
9	0.4363	17	0.3208
10	0.4122	18	0.3135
11	0.3922	19	0.3068
12	0.3755	20	0.3005

5Z TEST:

The probability of a data point from a Normal Distribution being greater than 5sd from the mean is < 0.0001%.

This method can be used when the estimates of mean and standard deviation are robust, such as when the data set has 20 or greater data points, and when the distribution is approximately normal.

5Z TEST PROCEDURE:

1. Confirm the data set has 20 or greater data points
2. Test for Normality. Confirm the data are not non-normal.
3. Calculate the *Z* score of the suspected outlier. If $Z < -5$ or $Z > 5$, then the point can be considered an outlier.