BURGE HUGHES WALSH

The Systems Thinking Tool Box

Dr Stuart Burge

".. bump, bump, bump, on the back of his head. It is, as far as he knows the only way of coming downstairs, but sometimes he feels that there really is another way, if only he could stop bumping for a moment and think of it."

Winnie the Pooh - A. A. Milne

Graphical Analysis

What is it and what does it do?

Graphical Analysis is about determining and understanding the nature of the variation displayed by a system through plotting collected data and looking for patterns. Graphical Analysis can be undertaken in two fundamental ways: as a ***frequency plot*** and as a ***time series plot*** as shown in Figure 1.

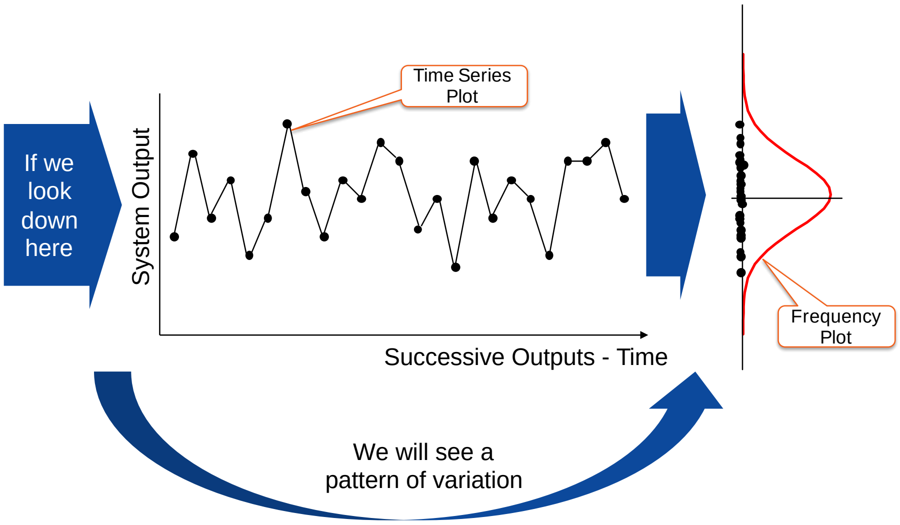


Figure 1: Graphical analysis through time series and frequency plots

Frequency Plots are used to identify and quantify the pattern of variation and have many forms including:

- Histograms
- Dot Plots
- Box Plots
- Stem and Leaf Plots.

Time Series Plots show pattern of variation over time and can signal unusual events and thereby identify the causes of variation that could lead to poor, or improved, system performance. There are several types that include:

- Time Series (run chart)
- Control Charts.

Why do it?

System parameters such as inputs, internal states and outputs will display variation. In particular, output variation, can affect the perception of the system performance. If we want to fully understand our system of interest, we have to comprehend this variation.

A key aspect of understanding a system is looking for the patterns of behaviour, but these are often not spotted because of variation. The natural variation in system can mask a pattern such that it is not recognised, or alternatively a pattern is recognised where none exists. In simple terms, variation can hide issues or opportunities.

It is also important to quantify the extent of the variation and whether this is acceptable. Figure 2 shows four potential variation scenarios of a system output that relate to whether the variation experienced is acceptable or not.

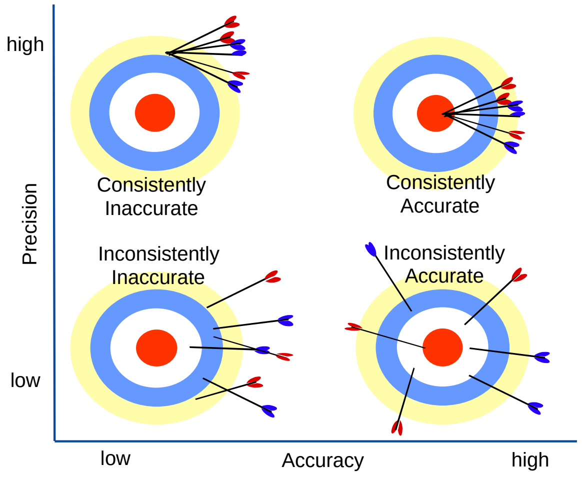


Figure 2: Variation Scenarios

The axes in Figure 2 comprise two critical measures in understanding variation:

- Accuracy: Whether we are hitting the target
- Precision: Whether we are consistent.

Clearly, the ideal situation is to be **consistently accurate** – to hit the desired target time after time. The remaining three represent differing degrees of disappointment for the customer or user of the output:

- Consistently Inaccurate
- Inconsistently Accurate
- Inconsistently Inaccurate.

The four scenarios of Figure 2 can also be visualised in terms of a Frequency Plot as shown in Figure 3. Also included in Figure 3 is the target and the allowable dispersion around that target as upper and lower limits. When the "tails" of the frequency plot lies outside the limits (one side or the other, or both) we are saying that the system performance on occasions will be unacceptable and result in customer complaints.

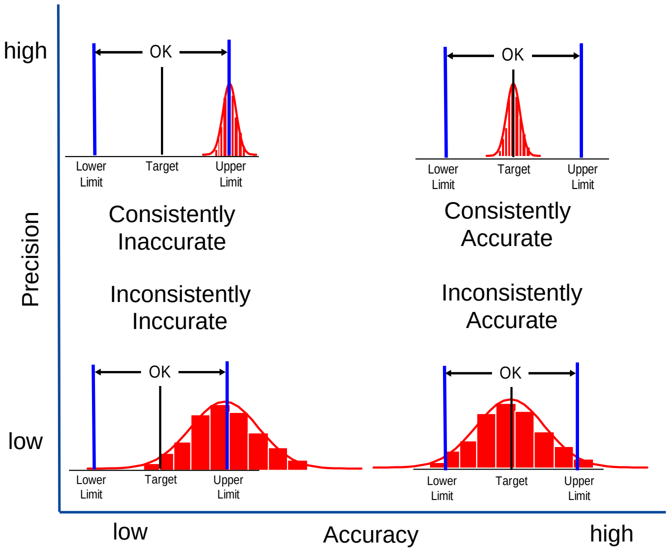


Figure 3: Frequency Plots of System Performance

The inclusion of upper and lower limits in figure provides us with a way of quantitatively assessing a system performance when variation is present.

Clearly, Figure 3 shows, as before, the ideal situation is to be consistently accurate. The next "best" situation is to be consistently inaccurate, although this could mean we are consistently upsetting the user/customer of the system output. The third best situation is to be inconsistently accurate and the last is to be inconsistently inaccurate. In Manufacturing Systems Engineering and Six Sigma various ***Process Capability*** indices are calculated to provide simple quantitative metrics to benchmark a system's output variation.

Being able to visualise and quantify the pattern of variation is therefore important in understanding system behaviour, especially that experienced by an end user/customer. However, Systems Thinking tells us that systems will also display events! Figure 4 shows a system where there is variation and an event.

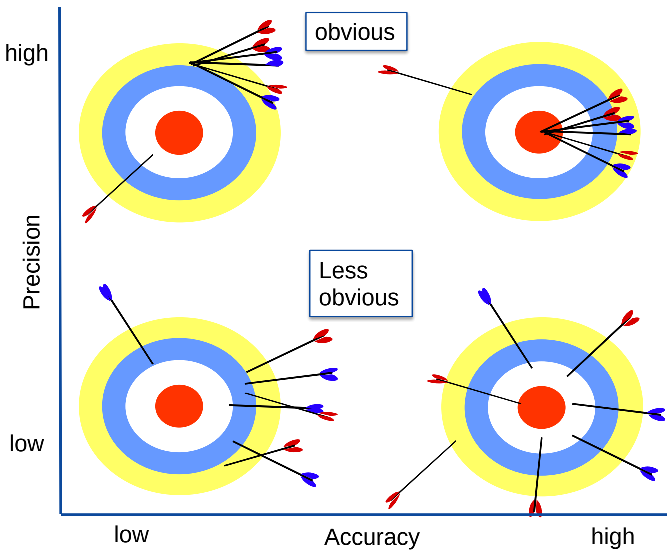
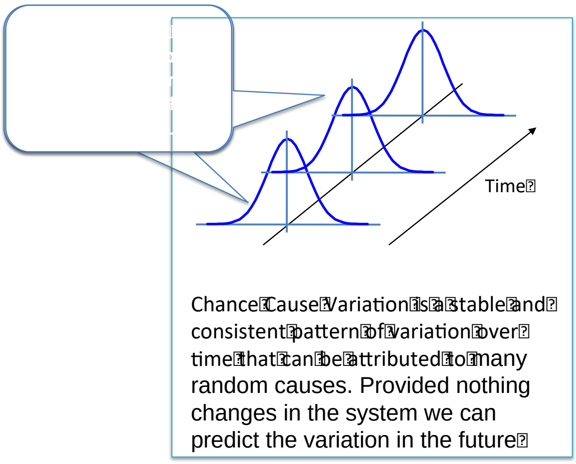


Figure 4: Variation and an Event

Such events often result in end user/customer dissatisfaction and are usually attributable to a single cause. It was Walter Shewhart who realised that while all variation is caused, it could be divided into assignable-cause and chance-cause variation.

Pure chance cause variation is that due to many sources. It is the variation that is inherent within the system that is present all the time. Indeed, chance cause variation is part of the system and is due to multiple causes such that we are not able to assign a single cause to the fluctuations experienced.

Pure chance cause variation when measured is random and if enough data is collected it will display a pattern that is stable over time. What this means is that although we cannot predict an individual outcome of the system output, that outcome will fit a pattern that is the same over time. Figure 5 shows this stable variation behaviour of the output of a system.



Chance Cause Variation is a stable and consistent pattern of variation over time that can be attributed to many random causes. Provided nothing changes in the system we can predict the variation in the future

Figure 5: Chance Cause Variation in the output of a System is stable over time

Assignable cause variation is due to a single cause that produces the variation event observed. We are able to say that the reason the system behaviour has changed; the present system output is different from the last is due to this cause. The cause can be:

- weather (season - time of day)
- a change in person
- equipment breakdown
- an accident
- unscheduled maintenance/repair
- etc

Such assignable causes of variation are either persistent, in which case and change is permanent, or transient. In the case of transient assignable causes of variation is that the system output returns to its previous behaviour. These two situations are shown in Figure 6.

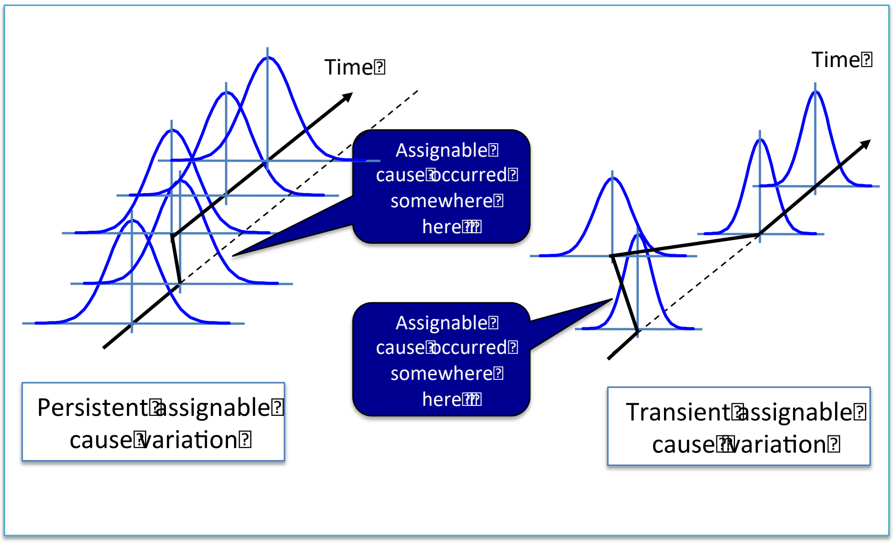

Time
Assignable cause occurred somewhere here
Persistent assignable cause variation
Time
Assignable cause occurred somewhere here
Transient assignable cause variation

Figure 6: Persistent and Transient Assignable Cause Variation

Figure 6 contains both good news and bad news, dependent on your point of view! Let’s look at the persistent assignable cause variation first. This single event has caused a shift of stable variation to the left in Figure 6. To understand whether this is good or bad news depends on context. If the assignable cause is a deliberate change to improve the performance, then there has been some success. Indeed, the way in which the “inaccurate” frequency plots of Figure 3 can be made “accurate” is by introducing a deliberate assignable cause (aka an improvement).

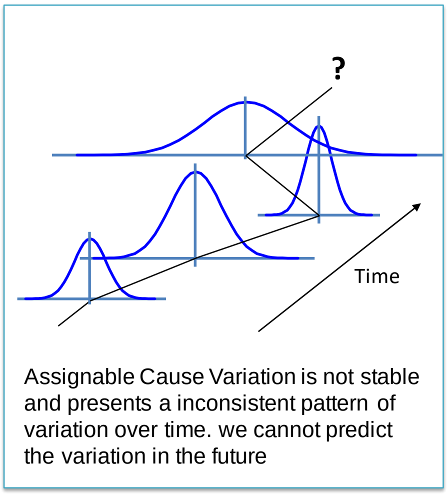
If in the case of the persistent assignable cause variation, the single event was not deliberate change, then something has changed in the system. Again, whether this is good or bad news depends upon the direction and magnitude of the change. If it is good news, then we must find out what the single cause was to ensure we can maintain it and also see if further improvement can be made. If the change is bad news, we must find out what the cause is to remove it. Note that is both situations we "must find out" what the cause was. If we monitor our system, then we should be able to quickly identify the cause. If we do not we are in a bit of a mess! In terms of what the cause might be, there are only two suspects:

- A change in one of the system inputs.
- A change in the internal workings of the system.

If we turn our attention to the transient cause variation, the observations are similar but subtly different. Because the change is not permanent we only get a glimpse of what might be. Again, it depends on whether the magnitude and direction of change are good or bad. If it is good, then we have briefly seen what is possible and if the cause can be identified and institutionalized we are in the position to improve the performance of our system. If it is bad, we live in fear of it happening again but because it is random, we will have no idea when it will occur. Again the only recourse is to find the root cause and in this instance take action to remove it. What we should never do is react to assignable cause variation. Indeed, reacting to assignable cause variation can make the situation worse – but that's what we do. Actually, by attempting to fix a problem we often introduce a new assignable cause, which leads to yet another fix – another assignable cause *ad nauseam*. This situation is the system archetype "Shifting the Burden" or "Fixes that Fail"[1] In such situations we end up in a mess as shown in Figure 7 where there is no predictability.

[1] System Archetypes System Archetypes describe common patterns of behaviour in systems that can provide diagnostic and prognostic insight to system behaviours. The common ones are:

- Shifting the burden (Fixes that Fail)
- Tragedy of the Commons
- Drift to low performance
- Escalation
- Success to the successful
- Addiction
- Rule Beating
- Seeking the wrong Goal



Assignable Cause Variation is not stable and presents a inconsistent pattern of variation over time. we cannot predict the variation in the future

Figure 7: Assignable Cause Variation in the output of a System is NOT stable over time

What is clear is the need to signal the occurrence of assignable cause variation in order to identify what the cause actually is. This activity often goes under the name of Root Cause Analysis.

Where and when to use it?

All systems will experience variation. We are often concerned with the outputs of the system and how they vary, but it is important to note that what comes out if a system often depends on what goes in and what goes on within it as shown in figure 8.

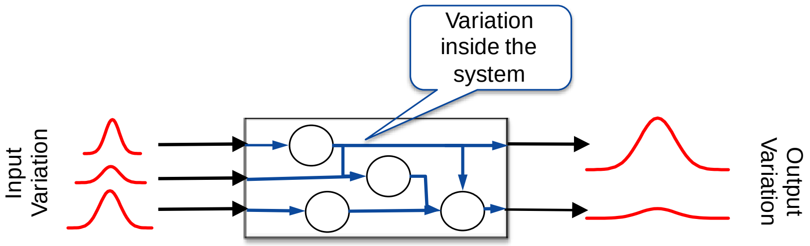


Figure 8: Variation in a System

Variation in the inputs and variation in how the system is operated, particularly human activity systems, will cause variation in the output. Graphical Analysis can be used to investigate the nature of the variation in all aspects of a system; its inputs, internal states and parameters and the outputs.

Graphical Analysis is particularly useful to help improve a system whose output is displaying variation sufficient to cause “end user/customer” dissatisfaction. In such situations, the graphical analysis can help to quantify the variation before and after any improvement activity to demonstrate in an evidence-based fashion that change has occurred. In other words, the improvement occurred. There is a link here with the common system archetype “shifting the burden” where changes are made but no evidence is collected to prove that the change worked.

Graphical analysis can also be used to monitor a system to signal the presence of assignable cause variation. This in turn will allow for the identification of and removal of the true root cause. It can also help us to manage a system correctly both chance and assignable cause variation can mask what is actually happening in a system leading to inappropriate and potentially disastrous management intervention.

Who does it?

An individual or team can undertake Graphical Analysis. It does require, however, the collection of suitable data from the system, and also access to a suitable software package. The software package I use is Minitab^{TM}.

How to do it?

Graphical Analysis requires us to collect data from our system of interest. This activity is its own right needs careful planning and execution. Poorly planned and/or executed data collection can itself introduce variation that is greater than the actual natural system variation under observation[2]. Common to both data collection and analysis (graphical or analytic) is an understanding of variation data.

In order to understand and quantify variation it is necessary to record a number of observations of a system parameter (often called by mathematicians a *variable*) which give rise to a set of data, known as **raw data**. How many observations we make is important and introduces two key terms. The collection of all possible observations is called the **population**. In many practical cases the population is either too large, or since the system is still running, to collect all observations. In such cases we take a small set or **sample** set of observations of the population. Provided this sample is representative of the population, then information about the population can be inferred from the sample. This is an important point as a poor sample can introduce patterns that do not exist. One of the most famous examples of bad sampling occurred during the 1936 presidential race when the *Literary Digest* sampled the American population and used this to predict that Alf Landon to win the election over Franklin D. Roosevelt. They got it terribly wrong because the *Literary Digest* selected their sample from phonebooks and automobile registrations. The sample was biased because at that time, people with phones and cars in 1936 were wealthier and therefore more likely to be Republicans and therefore not representative of the population.

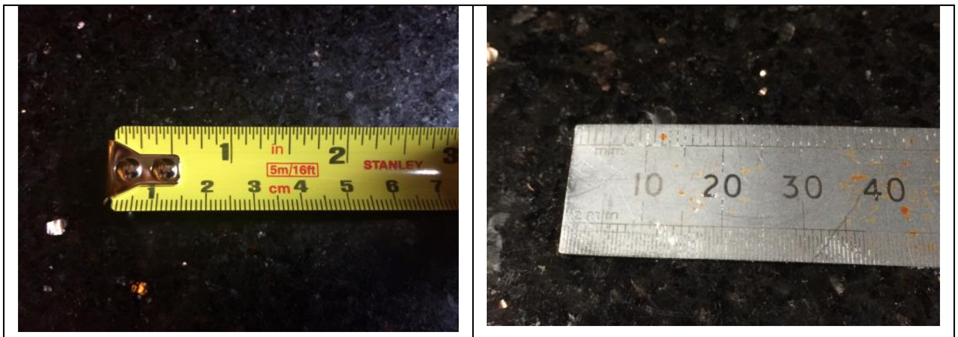
Assuming our data collection is okay, the numbers we collected can be categorised as comprising either discrete or continuous raw data.

[2] If you want to find out more about Data Collection and Measurement System Analysis visit BHW's website at www.burgehgheswalsh.co.uk.

Continuous or Variable data is measured on a ‘continuous scale’. Examples include: time, volumes, temperatures, weights, speeds etc. For example:

- *Weight of a person*
- *Room temperature*
- *Time of the winning runner in a marathon*

Continuous data are measured on a scale which is infinitely fine – our results are limited only by the resolution of our measurement system. For example, I have a tape measure, shown in Figure 9a, I use for DIY activities. Its resolution is 1mm. I can use it to measure lengths to a resolution of 1mm. I also own an “engineer’s rule” that is shown in Figure 9b – it can resolve down to 0.5mm.



(a) DIY Tape Measure

(b) Engineer’s Rule

Figure 9: A DIY Tape Measure and “Engineer’s Rule

Discrete or Attribute data is usually counted. It categorises things in some way such as:

- *meets requirements /does not*
- *present /not present,*
- *within standard/outside standard*
- *Red/yellow/green*

Discrete data can often be represented as a percentage or count data. For example:

- The percentage of calls lost in a call centre (A call can be taken or lost; it is a discrete classification).
- The number of lightning strikes in London in a given period of time is clearly discrete; one cannot have 1.345 strikes.

Sometimes large amounts of discrete data can be treated as continuous data, which is very useful for analysis purposes. For example, the number of calls lost per day in a call centre can be treated as continuous if there a large number of lost calls/day. There are several different types of discrete data:

Discrete proportion/percentage:

- Data where we can count the number of occurrences and also number of non-occurrences. For example, what percentage of tax bills were paid on time this month? What percentage of students obtained grade C or better?

Discrete count:

- The number of occurrences of a certain "event" or characteristic. For example, how many accidents did we have this month? How many employees were ill in December?

Discrete attribute:

- In this type of discrete data, the "Events" are placed into different category descriptions. For example, what are the different types of accidents? What are the illnesses?

Discrete ordinal:

- Uses a ranking scale to denote a range from better to worse or vice versa. This type of information can be very useful since it tells us about performance of one item against another. For example, How well do our customers rate us on a 1-5 scale?

Of the two, continuous data is the richest in terms of information content, but often is more difficult to collect than discrete data. There is also not a clear division between discrete and continuous data. For example, the number of telephone calls during an hour is discrete, but the number of telephone calls during a week could be treated as continuous. In the latter case, the discrete steps are so small in comparison to the timeframe of one week that they appear to be continuous. This feature can be very important since many statistical analyses are easier to apply for continuous data than discrete data.

Graphical Representations: Frequency Plots

The most common method of graphically representing a set of raw data is to form a histogram. This is done by dividing the range (the difference between the largest and smallest number in a data set) into a number of classes, and then count the number or frequency (how often) of variables in each class. These frequencies can be plotted as a bar chart.

For example, Table 1 shows data collected over a 40-month period to record the number of appointment "no shows" in a medical centre.

83	80	91	81	88	82	87	97	83	99
75	83	72	84	90	87	78	93	92	98
86	80	93	86	88	82	101	89	89	82
85	95	80	84	92	76	81	103	94	89

Table 1: Appointment "no shows" in a medical centre

From this table, the smallest number of no-shows is 72 whilst the largest is 102. This gives a range of 102 – 72 = 31. We now have a choice because it is up to us to decide on the classes. Here I use Millers[3] magic 7±2 as a starting point – can I divide up the range into 7 suitable classes? The answer is yes:

70 – 74, 75 – 79, 80 – 84, 85 – 89, 90 – 94, 95 – 99, 100 - 104

It is now a question of allocating the data in table to the appropriate class using a tally chart like that shown in table 2.

Class	Tally	Frequency
70 – 74	1	1
75 – 79	111	3
80 – 84	1111 1111 111 / /	13
85 – 89	1111 1111 / /	10
90 – 94	1111 11 /	7
95 – 99	1111	4
100 –104	11	2

Table 2: A Tally Chart

The frequency vs the class can now be plotted as the bar chart or histogram shown in Figure 10.

[3] George Miller in his 1956 seminal paper, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", recognised we all have a processing limit and when faced with complexity beyond this limit we make mistakes.

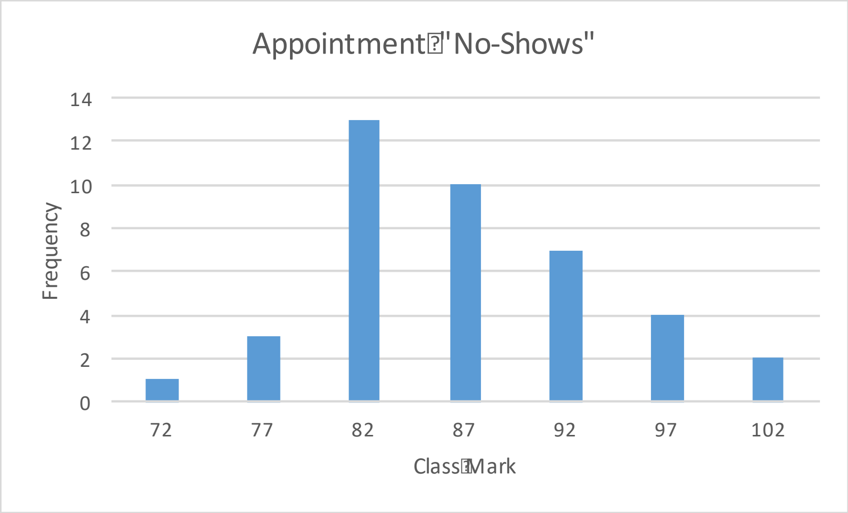


Figure 10: Histogram of Appointment "No Shows"

There are a few points to note about the last example.

- The data was discrete.
- The histogram is constructed about the mid-point of each class known as the class mark.
- The choice of class size or range is arbitrary, but too wide a class size and information is lost, and too small a class size makes comprehension difficult.
- The histogram is sometimes called a frequency distribution.
- Often the bars on the histograms are drawn 'touching'.

The previous example was for discrete data; we can do the same thing for continuous data. For example, in a manufacturing plant making internal combustion engines one of the operations is to "turn" the piston diameter in a lathe. Fifty-six consecutive pistons were measured and the data captured in a tally chart shown in Table 3.

Diameter (mm)	Tally	Frequency
55.0-55.09	1	1
55.10-55.19	11	2
55.20-55.29	1111	4
55.30-55.39	1111 11	7
55.40-55.49	1111 1111	10
55.50-55.59	1111 1111 11	12
55.60-55.69	1111 1111	9
55.70-55.79	1111 1	6
55.80-55.89	111	3
55.90-60.00	11	2

Table 3: Tally Chart for Piston Diameters

Before plotting the histogram there are a number of points that can be raised about continuous data. In this example the data is continuous, but classes appear not to be, with a ‘gap' of 0.01 between each class. The question arises which class does a piston 55.394mm belong to? The answer is that each class extends 0.005mm below and above the class limits as shown in Figure 11.

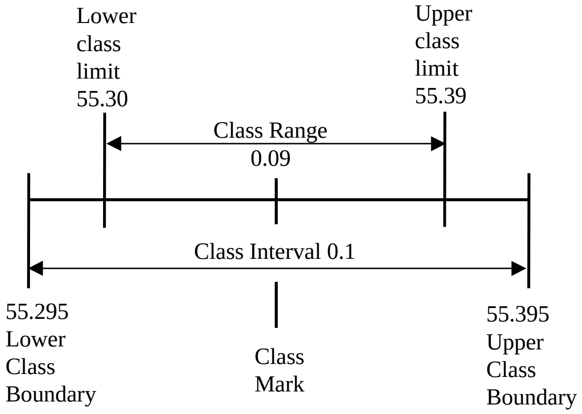


Figure 11: Class Boundaries for Continuous Data

Thus, the upper class boundary is the lower class boundary of the next class. Note that rounding up for halfway values is the most common. Figure 12 shows the final histogram.

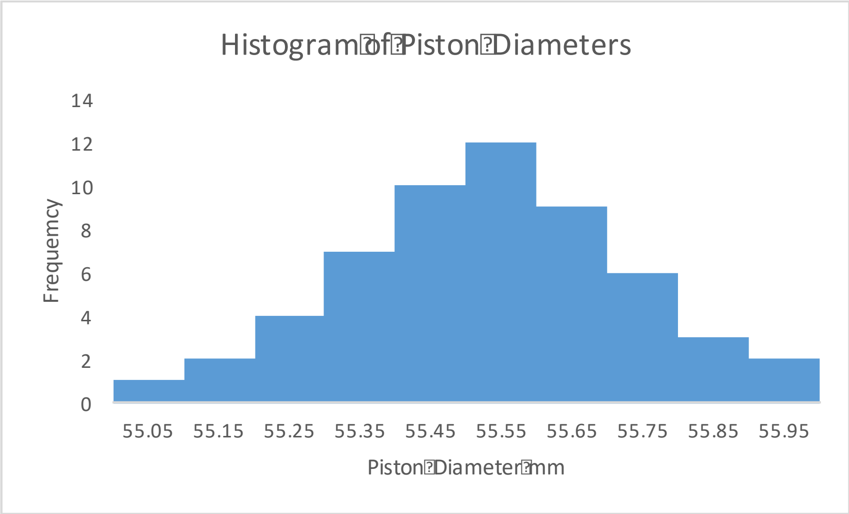
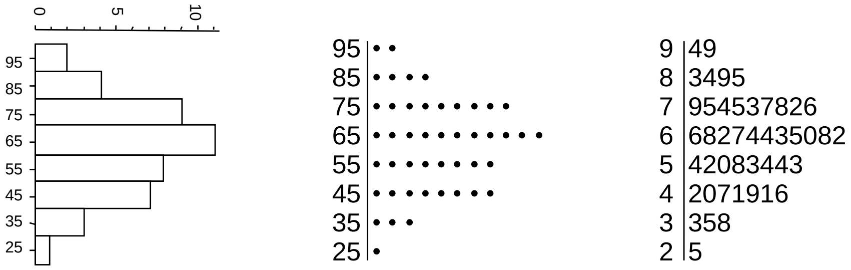


Figure 12: Histogram of Piston Diameters

Since the data is continuous, the block form of the histogram is not really representative, and it is common to draw a smooth line through all the class marks.

While the Histogram is perhaps the most common form of Frequency Plot, there are several other types that are useful. These are shown in figure 13.



Histogram

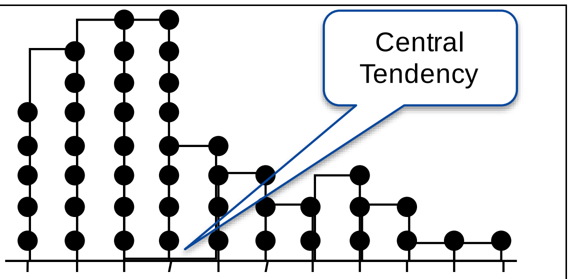
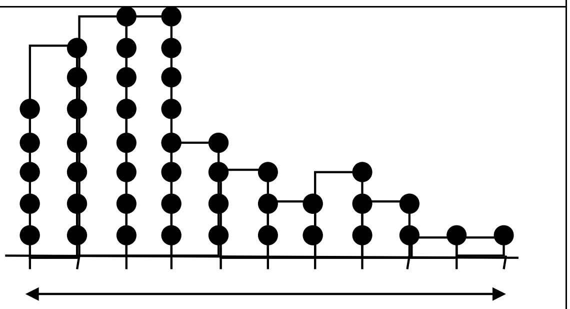
Dot Plot

Stem-and-Leaf Plot

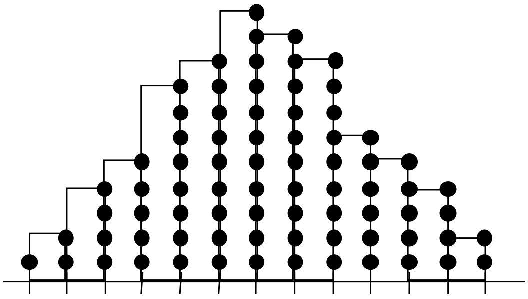
Figure 13: Alternatives forms of histograms.

Dot Plots can be particularly useful as you obtain a clear picture of the variation distribution but also through the dots a clearer picture of the quantities of data points in a class. This is taken further with Stem-and-Leaf plot. Here, not only can we see the variation distribution but the actual numerical values that make up the data set. Note, however, this type of frequency plot is less useful with continuous data that is to several decimal places.

Frequency Plots enable us to see the pattern of variation with respect to central tendency and spread.

Centre of the data, the point at which the data clusters usually we use the average. Here there are 3 possible options of the Mean, Median and Mode[4] We can also look at the shape of the distribution. Is it symmetric or asymmetric (one sided)? It is flat or peaky?	
The spread or dispersion of the data indicates the "amount" variation. If there are targets and limits it is possible to see and quantify the actual system performance against those limits. If the tails of the distribution hang outside the limits then it is an indication of poor performance.	 Spread or Dispersion

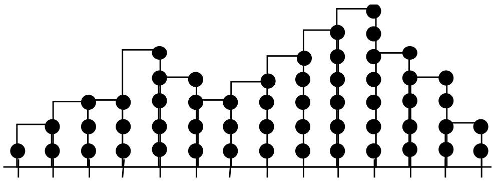
The following show the common situations that are experience when plotting data as histograms or dot plots.

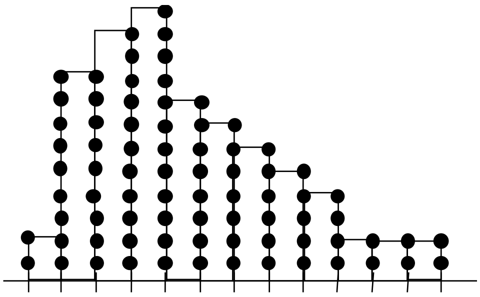
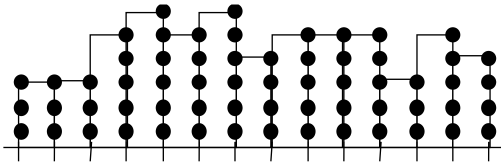
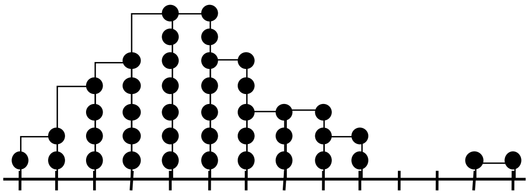
Interpretation: No indication of assignable causes and hence data, *may* come from a stable system. (This is no guarantee of the absence of assignable causes, they may appear on a control chart) **Implication:** A common situation that is often called a "normal" distribution which can be modelled mathematically. **Action:** If variation is too great will need to make changes to improve a stable system.	 Bell shaped. Symmetrical.

[4] There are three recognised and commonly used measures of central tendency

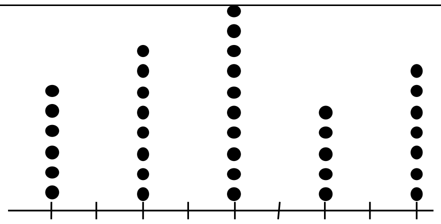
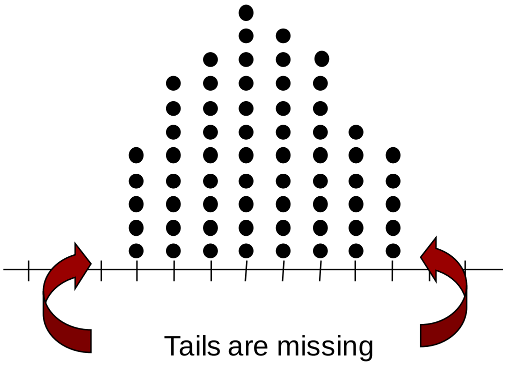
- Mean: the sum of the individual values divided by the number of values.
- Median: the central value when the values of the data set are ranked in order of magnitude. If there are an even number, then the median is taken as the mean of the central pair.
- Mode: The most frequently occurring value in a data set.

If the frequency histogram is symmetric, then the Mean, Median and Mode all take the same value. If, however, the frequency histogram is non-symmetric or skewed then the Mean, Median and Mode will take different values.

Interpretation: We have either mixed two data sources or, what we thought was one system is operating like two! **Implication:** Time plots and control charts may provide misleading information. **Action:** Need to separate the data to seek out causes for two humps. May require further data collection.	 Two humps. Bimodal.

Interpretation: Could be due to data collection errors or data is truly skewed. **Implication:** Time plots, and control charts may provide misleading information. **Action:** Data analysis techniques need to be used with caution. Can lead to false conclusions.	 Asymmetric - Skewed.
Interpretation: system may be "drifting" over time or, system could have a mix of several operating conditions or modes. **Action:** Use time plots, run charts or control charts to track over time. Look for possible stratifying factors such as time of day, people changes (shift patterns).	 Basically flat
Interpretation: Outlier data points are likely the result of measurement error or something unusual happening – it could be an assignable cause. **Action:** Confirm outliers are not measurement error. Consider as an assignable cause and take appropriate action.	 One or more outliers

Version 2.0

Interpretation: Measurement System does not have enough resolution or, readings are being rounded. **Action:** Check measurement device and data collection/recording procedure.	 Five or fewer distinct values
Interpretation: has been subject to some form of inspection and sort operation to remove the "tails". This could be because they are beyond the limits. **Implication:** Underlying pattern is a "normal" distribution but with a large spread. while the system is potential meeting its performance targets, there are many outputs very close to the limits and the end customer will experience large variation of performance. **Action**: Investigate potential output sorting.	

Version 2.0

An alternative to the various types of Frequency plots is a Box Plot. This is a simple pictorial representation of the raw data to highlight key features of the variation as shown in Figure 14.

Note that in Figure 14 distribution is not symmetrical because the mean and median have different values.

Box plots are particular useful at quickly comparing data sets of variation to see if there are differences. Figure 15 shows an example of two Box Plots where it is possible to see the differences between situation 1 and 2. In this case, we can see that the spread is greater in case 2 and also that case two has a higher median average. Care must be exercised, in that while there may appear to be differences they could be due to the sampling. To be more certain we need to perform a hypothesis test to tell us if data is statistically different – but that is another story.

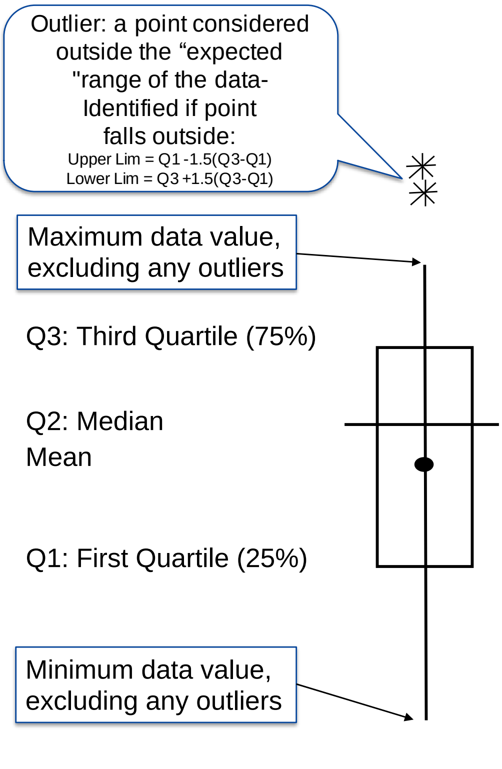

Outlier: a point considered outside the "expected "range of the data-
Identified if point falls outside:
Upper Lim = Q1 -1.5(Q3-Q1)
Lower Lim = Q3 +1.5(Q3-Q1)
Maximum data value, excluding any outliers
Q3: Third Quartile (75%)
Q2: Median
Mean
Q1: First Quartile (25%)
Minimum data value, excluding any outliers

Figure 14: The Box Plot

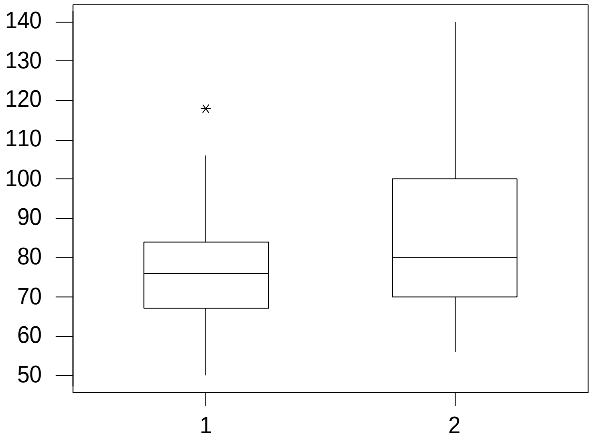


Figure 15: Box Plots used to compare two situations

Graphical Representations: Time Series Plots

Time Series Plots show pattern of variation over time and can signal unusual events and thereby identify the causes of variation that could lead to poor, or improved, system performance. There are several types that include:

- Time Series (run chart)
- Control Charts.

Time Series (Run Chart) Plots

A time series is a sequence of successive measurements of a system parameter made over a time interval. Typically, the measurements are made at regular time intervals, but they do not have to be. For example, you could collect data on the "dwell-time" of passenger trains at a railway station. Every time a passenger train stops at a station, the time from door open to doors shut could be collected. The trains do not arrive at regular intervals, but the data can still be plotted as a sequence of regularly sequenced observations.

A Time Series Plot or Run Chart is plot of the collected data in sequence order as shown in Figure 16. The observations whether regular or not are plotted on a regular scale and lines drawn to highlight the variation. It is common also to plot the median[5] as shown in Figure 16. The use of the median means 50% of the data points will be above the median line and 50% below. I chose this particular example because the frequency distribution is symmetric which will make the explanation of how a Run Chart works somewhat easier.

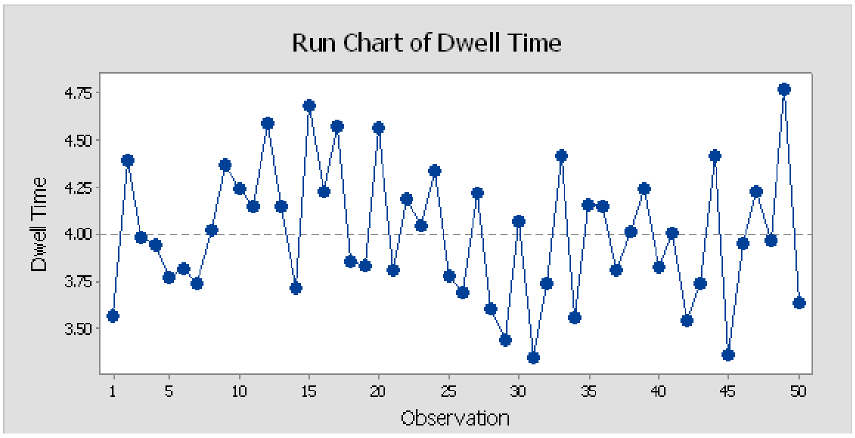


Figure 16: A Time Series Plot or Run Chart of Dwell Time

Figure 16 shows a stable system where the variation is entirely due to chance causes. I know this just by looking at the plot. It can be likened to tossing a coin for which there is a 50% chance of a head (or tail). The chance of getting two successive heads is actually 50% of 50% or 25%. The chance of three successive heads is 50% of 25% or 12.5% and so on. So, the chance of seven successive heads is 50% of 50% of 50% of 50% of 50% of 50% of 50%, which is 1/128 or 0.8%. Going back to the example above, having seven points above or below the median can happen but not very often. It is this idea that the pattern of variation is stable (a known unknown) that allows us to identify the presence of assignable cause variation. There are, using Run Charts, four signals of assignable cause variation:

1. Not counting points on the median, six or more consecutive data points above or below the median. This situation indicates a shift in the average meaning the accuracy has changed. Context is all important here because it could be good news or bad.

[5] The median is used in preference to the mean to take account of asymmetric frequency distribution.

2. Six or more consecutive data points that are either all increasing or decreasing in value. If two points are the same value, ignore one when counting.
3. A run is a consecutive series of data points above or below the median. Too many or too few runs (i.e. the median is crossed too many or too few times) indicates the presence of assignable causes.
4. An extreme data points that is clearly different from all others.

Clearly, Run Charts can be used to signal the presence of assignable cause variation, but they can also be used to provide evidence of “real” change.

Figure 17 shows weight data I collected on a daily basis (my weight first thing in the morning) using a set of digital home scales. It shows how my weight varied on a daily basis and the impact of effort on my part to lose weight. I know that everybody’s weight will vary from day to day due to the varying amount of energy expended, the varying food intake and our varying ability to “process’ the food (our metabolic rate varies). If I do nothing special my weight will vary. I decided to capture the extent of this variation as a “baseline” for my weight loss attempt. So each morning haven woken up and before doing anything else I weighed myself on the family digital scales. These were un-calibrated, so the weight reported might be out (the scales may not be accurate). However, I was interested in losing weight and while the scales may record 80kg when my true weight was 79.6kg didn’t matter. What did matter was that I took my weight in a consistent fashion to reduce the amount of measurement variation. I weighed myself every morning just after waking for a whole month (August 2015 to be precise).

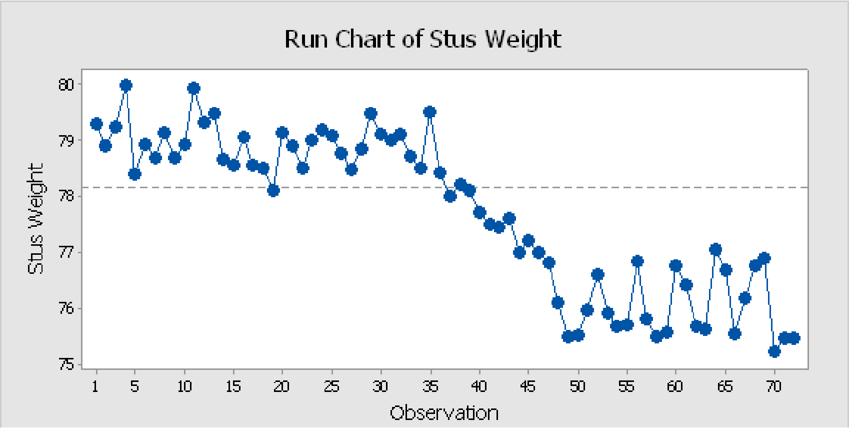


Figure 17: A Time Series Plot or Run Chart of my weight

There are several things to note about Figure 17:

- My weight varies daily.

- For the first 31 measurements (before the weight loss programme) show daily variation of approximately ±1kg. If I do nothing except follow my daily routine my weight will naturally vary by ±1kg (±4lbs). This is because although I may follow a routine each day is day is different. I walk different amounts, I go to the gym sometimes once a week sometimes 3 times a week, every meal is different, as is the time its eaten and so on. We all have a natural variation in our weight driven by our lifestyle choices!
- At day 32 the weight reduction programme began comprising a fundamental change in diet of significantly reduced carbohydrate intake. That is; no bread, or potatoes, rice or other starchy food; no sugars – sweets, puddings chocolate. This was done for 15 days, for which we can see a downward, but variable, trend. There is an interesting point at day 35 which was a “bad day” comprising missing the gym, having lunch and dinner out!
- Days 46 and onwards saw my returning to a more balanced diet that included some of the “carbs” denied (some because I kept up sugar based carbs – but had to go back to the odd sandwich!). Notice here that while there is variation it has returned to a horizontal pattern, albeit with an increased variation! This last point is interesting as its is saying that my weight has stabilised at a new and lower level, but the day to day variation is greater than the day to day variation in the first period.

There is no doubt from Figure 17 that something happened and I am at a lower weight which I am happy to report I am maintaining. Using test 1 above regarding points above and below the median, clearly shows there are assignable causes present – my diet. I need to be careful because I am basing my conclusions on data collected over the initial month period, is this a true representation of my natural variation. August is my normal holiday period and September is typically hectic. Are the changes really true? The answer is “I don't know” but there are tools that can indicate whether changes are causal of the effect observed – these tools are Control Charts.

Control Chart

Control charts are an incredibly powerful tool not only to understand the variation in a system but also, as the name suggests to help in the control of the system variation by signalling the presence of assignable cause variation. This leading to the identification of the root cause and thereby action to sustain good behaviour or action to eliminate unwanted variation. Because of their statistical nature control charts also provide a means of measuring whether any "real" change has taken place and monitor on- going performance. The principle of a control chart is relatively simple but does require a philosophical understanding of variation particularly around the concept of variation stability I talked about earlier. In this form of stability, we said that while it is not possible to predict an individual system measurement in a stable system the measurement will fall inside a pattern that is unchanging over time, proving the system itself doesn't change. It's about being able to accept the concept of the known unknowns. This can be proven in an analytic way through mathematics.

I want to keep this simple because the concept is - however you will have to trust me. I am also going to provide in a local appendix the mathematical view. Figure 5 and Figure 7 show respectively stable and unstable patterns of variation. In the unstable situation we cannot predict what the next output is. In the stable situation we can predict the range in which it will fall and the central value it will tend towards. A really important point here is that all systems WILL display variation: it cannot be avoided but it can be controlled. Unstable systems can be made stable by identifying the assignable cause and taking the appropriate action. Stable systems can also be changed. If the stable variation is too big, it is possible to investigate the chance causes to find the dominant ones (with pure chance cause variation the causes are manifold and complex (through interactions) but typically from the potential multitude a few will dominate - often cited here is the Pareto principle or 80:29 rule - 80% of the observed variation is due to 20% of the causes).

Measuring variation is not straightforward. We need knowledge of both the accuracy and precision. Accuracy is concerned with central tendency. In other words, when we collect data from a system, of a system output, parameter, or input, it will group around a central value. This central tendency can be measured by the mean (As we saw with the Run Chart in some circumstances the median can be used).

Measuring the spread or dispersion in trickier. Ideally, just like we use the mean to measure central tendency, we need an average variation around that mean. Figure 18 shows the idea where to get the “average” variation around the mean we first determine the deviations from the mean, which in Figure 18 are labelled as d_1, d_2, d_3, etc .

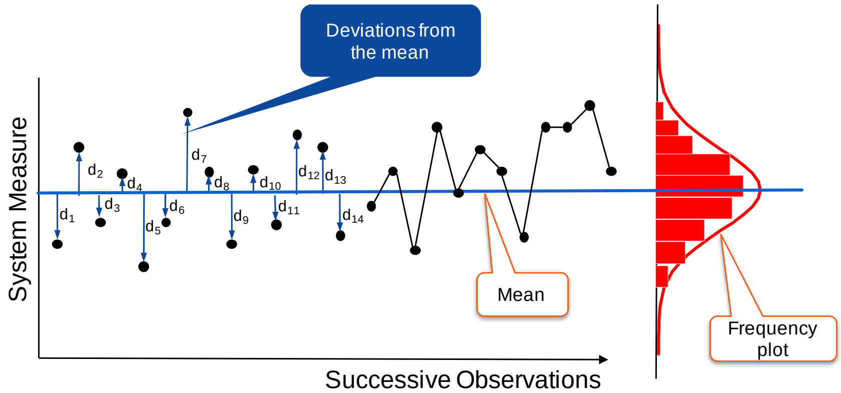


Figure 18: The deviations of the observed data from the mean

To determine the average deviations from the mean, all we have to do is add up the deviations and divide by the numbers of observations. What appears to be a simple idea becomes hard in practice because when we sum all the deviations from the mean it always adds up to zero. It is due to the way the mean is calculated and the deviation “below” the mean will exactly cancel those “above” the mean. Mathematicians overcome this by squaring the deviations first and then adding the squares together before dividing by the number of data points.

This particular measure is called the ***Variance*** and is equal to

> Sum of the squares of the deviations from the mean divided by the number of observations.

Mathematicians will describe this using an equation which if you interested I have put in the appendix. Unfortunately, the story is quite complete as the variance has the wrong units! Returning to my weight example shown in Figure 17. My weight was measured in units of kilograms (kg). If I add all the data and divide by the number of observations (72 in total) I get my mean weight of 77.7kg. My variance will be:

> Sum of the squares of the deviations from 77.7 divided by 72.

Which is 1.98kg^2. There is something very subtle about my answer here as the units are not kg but kg squared or kg^2. One of the issues with using the variance is that the units are wrong. My weight measurements are in kg, the mean is in kg, but the variance is in kg^2. The final trick is to take the square root of the variance to arrive at what is called the ***Standard Deviation.*** Standard Deviation is effectively the "average deviation from the mean". For my weight the standard deviation is 1.41kg.

Standard Deviation is usually represented by the Greek letter sigma σ[6]. It is a horrible thing, but a highly useful and important thing in quantifying variation. It is particularly useful in measuring precision.

In summary, we know we want our systems to be both accurate and precise: to be consistently on target. To measure this, we use the mean (accuracy) and standard deviation (precision).

The last building block in understanding Control Charts is the ***Probability Distribution*** in particular the ***Normal Distribution.*** A Probability Distribution is a mathematical model of that relates the values of a variable with the probability of observing the values. There are many different types of distribution that have been developed to model real world situations but the most common is the Normal or Gaussian[7] distribution.

The Normal Distribution

The normal distribution occurs very frequently in practice. It has a mean of μ and a standard deviation of σ and has the characteristic "bell" shape as shown in Figure 19.

[6] You may have come across the improvement approach known as 6-sigma or 6σ. The origin of the name comes from the concept of standard deviation because the approach is concerned about improving system performance by reducing the output variation to almost unmeasurable amounts. 6-sigma is a systems approach to reducing variation and in its pure form is highly effective at reducing chance cause variation. It has over the years been misused (used as a one sized fits all improvement approach), become watered-down (lean-sigma for example) and worst of all misunderstood.

[7] The origin of the Normal Distribution is often attributed to Johann Carl Friedrich Gauss, however, there is evidence to suggest it was in fact discovered by Abraham de Moivre.

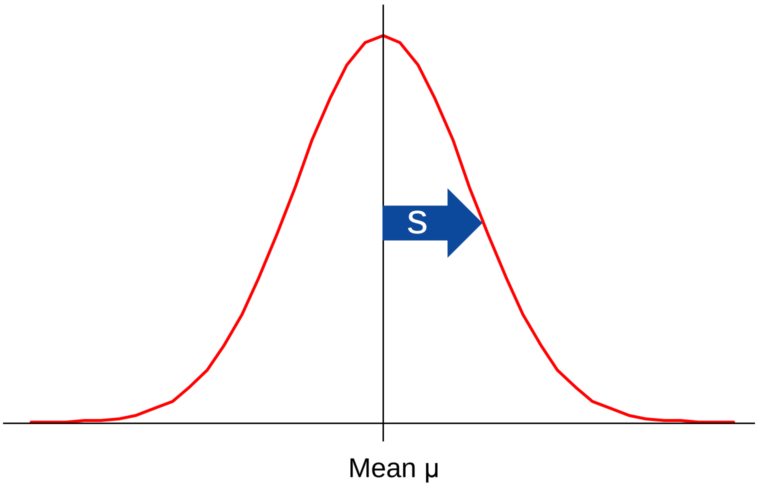


Figure 19: The Normal Distribution

Probability Distributions are scaled so the area under the curve is 1.0. the same is true for the Normal Distribution A very important feature of the Normal distribution is that the area under the curve is equal to the probability or chance of being in that region as shown in Figure 20.

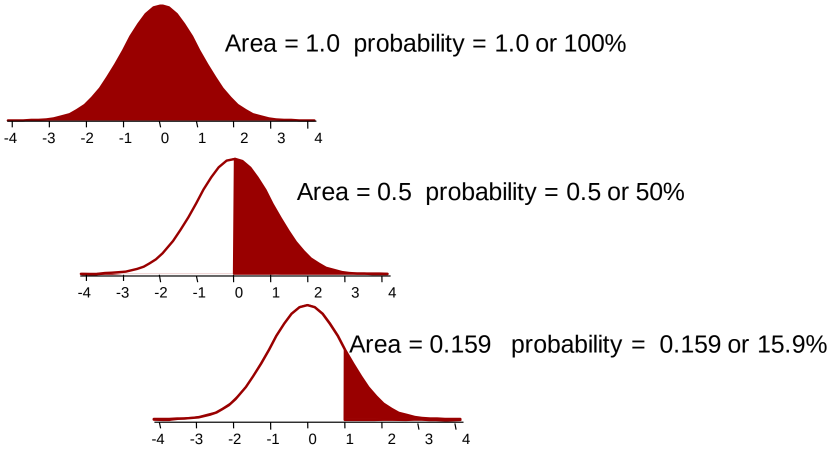


Figure 20: Probability Distributions, Area and Chance

This property is extremely useful since if some system parameter or characteristic is known to be “Normally” distributed, predictions can be made about the quantities that lie within certain ranges. In Figure 20, the scale on the horizontal axis is the number of standard deviations. Figure 21 shows the probability of lying in between bands of standard deviations.

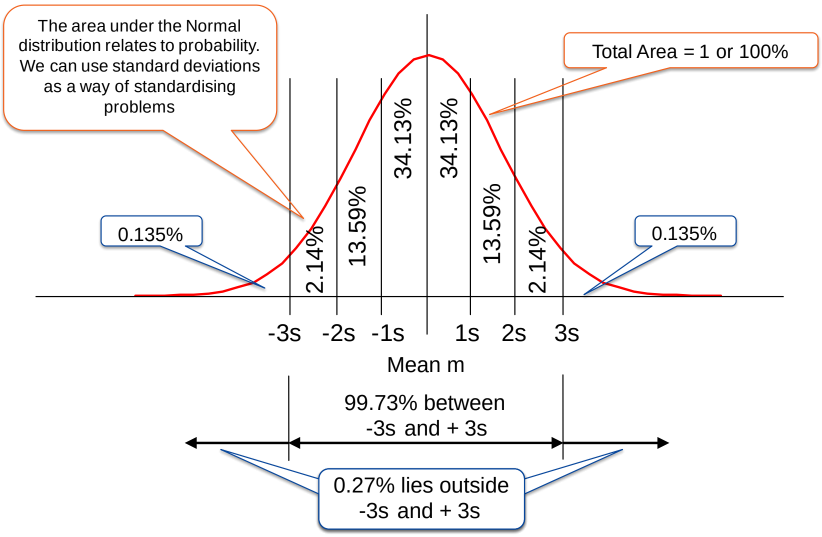

The area under the Normal distribution relates to probability. We can use standard deviations as a way of standardising problems
Total Area = 1 or 100%
0.135%
0.135%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
-3s
-2s
-1s
1s
2s
3s
Mean m
99.73% between -3s and + 3s
0.27% lies outside -3s and + 3s

Figure 21: Area, probability and standard deviation

We now have all the jigsaw pieces in place to explain the principle of Control Chart.

A Control chart is a time series plot on which is added the mean and Upper and Lower Control Limit lines as shown in figure 22.

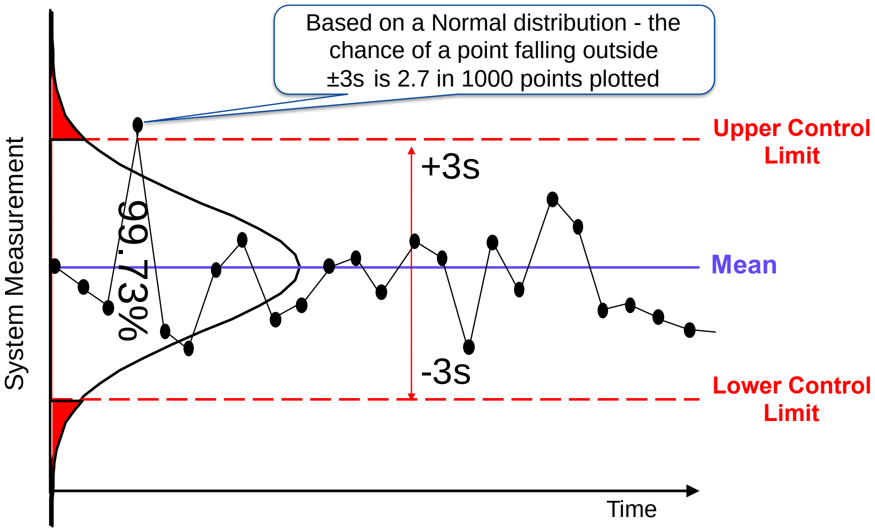


Figure 22: The Basic format of the Control Chart

Based on a Normal Distribution the control limits are drawn at ± 3 Standard Deviations. If the System is stable as defined by Figure 5, the chance of a point falling outside the upper or lower limit is less than three time in every 1000 points plotted. Quite rare! However, the presence of an assignable cause of variation will cause the distribution to shift location (mean) or change spread (standard deviation), in which case the change of points falling outside the Control Limits becomes significantly larger as shown in Figure 23.

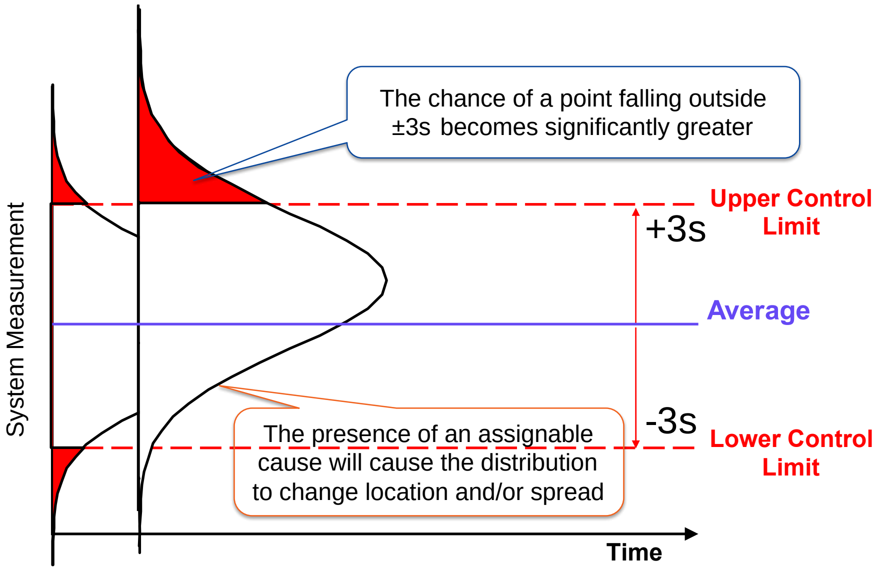

The chance of a point falling outside ±3s becomes significantly greater
Upper Control Limit
+3s
Average
-3s
Lower Control Limit
The presence of an assignable cause will cause the distribution to change location and/or spread
System Measurement
Time

Figure 23: The Effect of Assignable Cause Variation

Hence, when we construct Control Charts and find pints outside the Control Limits, there are two possible reasons:

- Pure chance (approximately 3 in 1000).
- Assignable cause variation.

We assume that all points outside the Control Limits are due to assignable causes. This means that sometimes we could be drawing the wrong conclusions!

Figure 24 shows a Control Chart for journey times from my home to my office collected over several months. I do not go to the office every day and although the intervals between successive points are equal, they are not some could be one day apart, some could be seven days apart. This collected data has been used to calculate the mean at 35.36 minutes. It is also used to calculate the standard deviation at 1.25 minutes. With this data the control limits can be calculated as:

Upper Control Limit = 35.36 + 3x1.25 = 39.11

Lower Control Limit = 35.36 – 3x1.25 = 31.61

In practice, I use the software package Minitab^{TM}, as given the raw data it will perform all the necessary calculations and construct the plot shown in Figure 24. You will notice it has three extra lines and red square with the number 1 nearby. This is signalling the presence of something unusual signifying that this particular point does not fit with the remainder of the data.

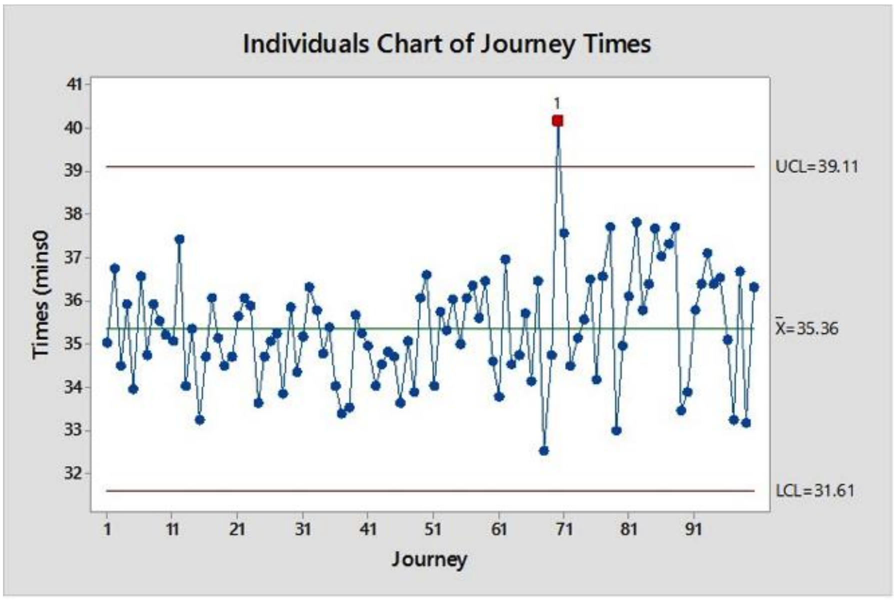


Figure 24: An Individuals Control Chart for my Journey Times to my Office

Indeed, this point is telling me that something happened on that day that was different from the other days. This could be true or it could be one of those few 3 in 1000 collected data points that will be greater than ± 3 standard deviations. In this case it was in fact a broken down car blocking a lane of the motorway.

The particular Control Chart shown in Figure 24 is called an Individuals Chart, because individual values are plotted. For this particular type of Control Chart (there are several others which I will introduce later), it is important that the data is Normally distributed. Actually, having plotted the Control Limits on Figure 24 using all the data points, we should recalculate them without the assignable cause data point. The inclusion of the assignable cause data point results in slightly wider Control Limits that could mask further unstable behaviour. It is also important to have enough data to provide a good estimate of the mean and standard deviation. The recommendation is to have at least 20 data points.

The Normal distribution is common, but there are others that also occur frequently, particularly the “skewed” distribution. An example is shown in Figure 25 which shows the histogram for a system with output behaviour that is naturally skewed. Skewed variation is not symmetric and has a “long tail” one side or the other. The concept of a Control Chart is based upon a Normal Distribution which is symmetric. Figure 26 shows an Individuals Chart for the skewed data. You can see that there are three points signalled as being potential assignable cause variation.

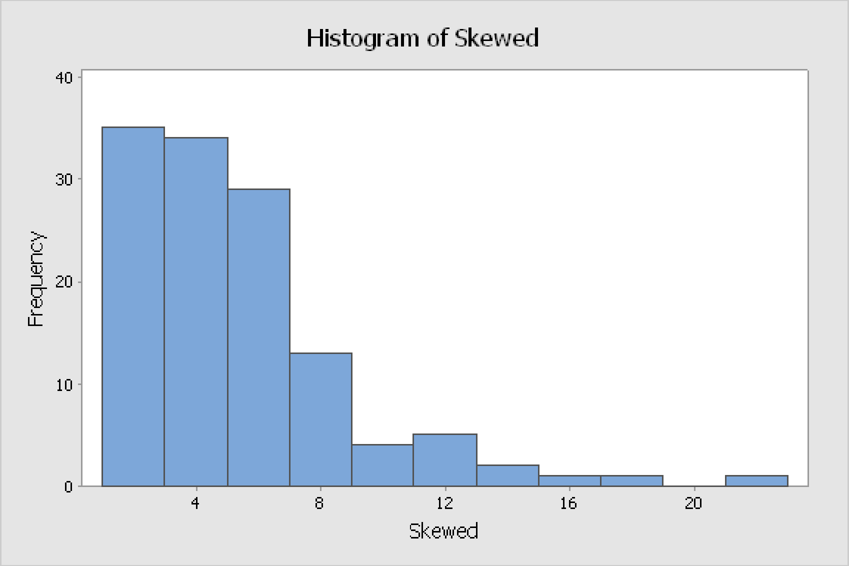


Figure 25: Frequency Plot of the "Skewed" Data Set showing it to be clearly non-Normal.

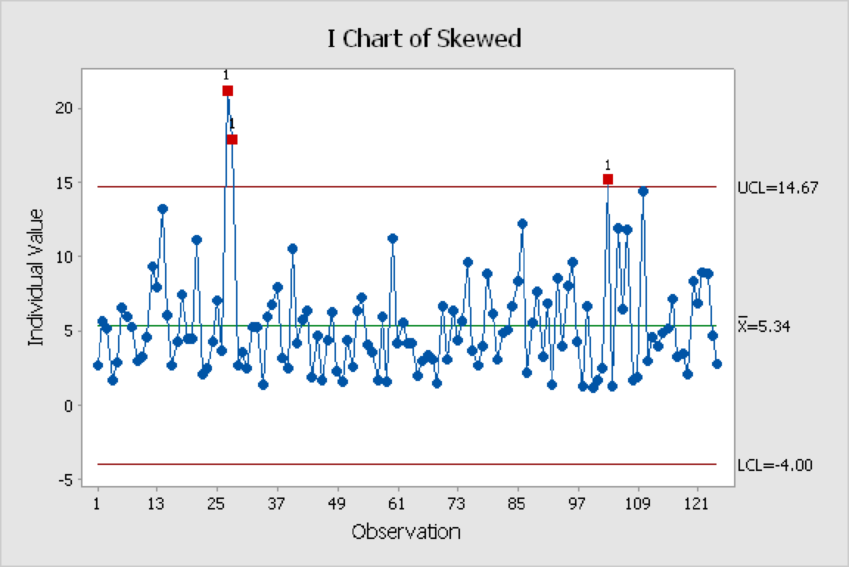


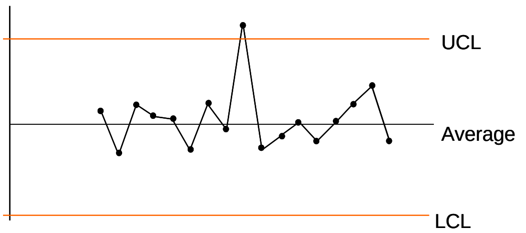
Figure 26: An Individuals Chart for the Data Set "Skewed"

Care has to be to be exercised however. Interpreting whether or not a set of collected data is Normally distributed or not from frequency plot can be difficult. There are mathematical tests that can be performed on the raw data to check for "normality". While not difficult to perform if you have the right software, it is perhaps an advanced topic. In extreme cases, like that shown in figure 25 and 26, you can tell from the Individuals chart that something is amiss. You will notice that there are no data points between 0 and the Lower Control Limit set at -4.0. There should be! This band of nothing it also indicative of a non-normal distribution.

Points falling outside the Control Limits are not the only signal for the presence of assignable cause variation. Picking up on the theme start with the Run Chart, there are other patterns that can be tested for. Table 4 shows the eight possible tests for assignable cause variation.

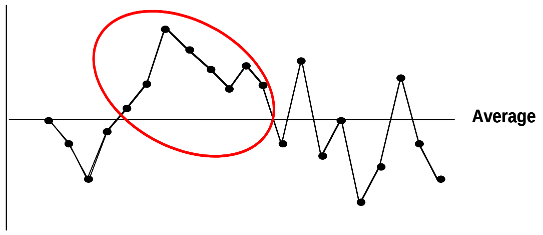
TEST 1: Point outside Control Limits:

Indicates there is something different about this point.



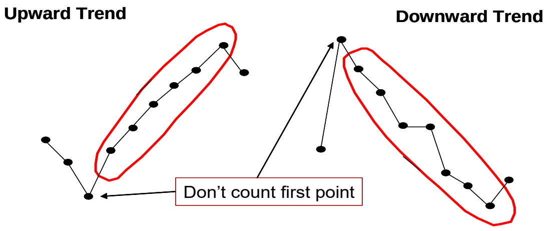
TEST 2: Points Above or Below Average:

8 or 9 or more points in a row above or below the average indicates a system shift (8 is more sensitive – 9 is more conclusive).



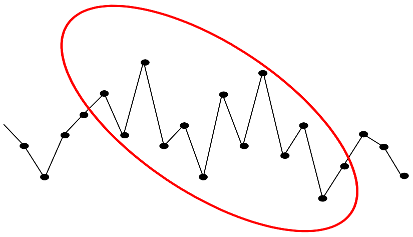
TEST 3: Treads Up or Down:

6 or more points in a row increasing or decreasing indicates a trend in the in the system.



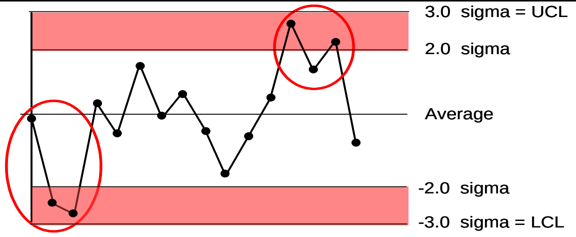
TEST 4: Alternating Up and Down:

14 or more points in a row alternating up and down indicates data collection Problems, or over-adjustment of system.



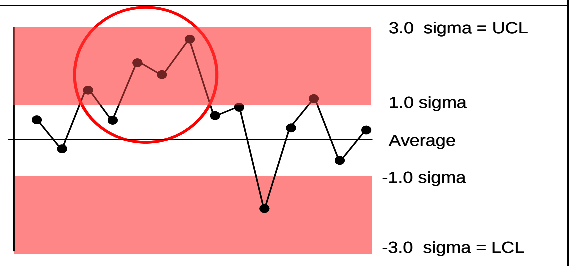
TEST 5: 2 out of 3 beyond 2 Sigma:

2 out of 3 consecutive points beyond 2 standard deviations (on the same side) indicates a shift in the system.



TEST 6: 4 out of 5 beyond 1 Sigma:

4 out of 5 consecutive points beyond 1 standard deviation (on the same side) indicates a shift in the system.



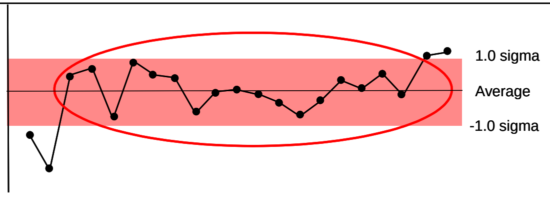
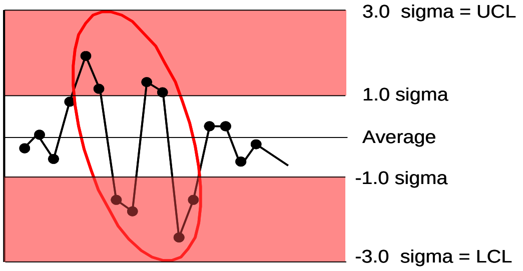
TEST 7: Middle 1/3 15 consecutive points within 1 standard deviation (above or below the average).	
TEST 8: Outer 2/3: 8 consecutive points beyond 1 standard deviation (either side) indicates: • Overcompensation of system • Multiple sources of variation – poor data collection.	

Table 4: The 8 test for signalling the presence of Assignable Cause Variation

Let’s have a look at some of these tests in action. Figure 17 shows a Run Chart for my weight collected over a number of consecutive days. Figure 27 shows the same data plotting as an Individuals Chart. In Figure 27, all of the data has been used to calculate the mean and standard deviation and thence the Upper and Lower Control Limits.

The first striking point about Figure 27 is the large number of assignable cause signals. There are assignable causes present in Figure 27, but not as many as is being signalled. I know because I deliberately introduced them.

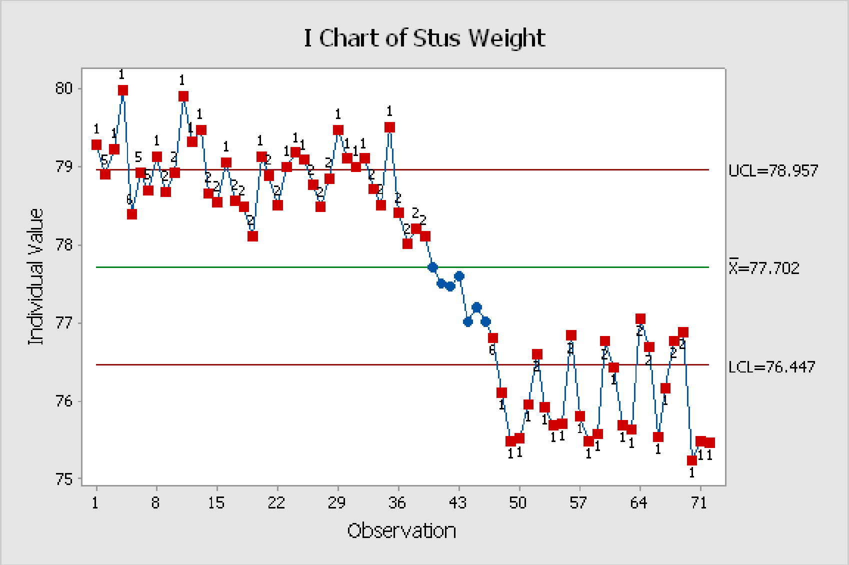


Figure 27: My Weight Data Plotted as an Individuals Chart

The first 31 points correspond to me “baselining” my natural day-to-day variation. While my weight will vary day-to-day it should be a stable predictable pattern of variation. Following the baselining period, I then introduced the changes in my diet for 15 days. The last observations are me adjusting back to a normal diet. Hence, there are three distinct phases with two potential assignable causes: Me introducing the diet and me reverting back to my normal diet. In such situations we need to construct three sets of Control Limits for each phase of my diet. This can be seen in Figure 28.

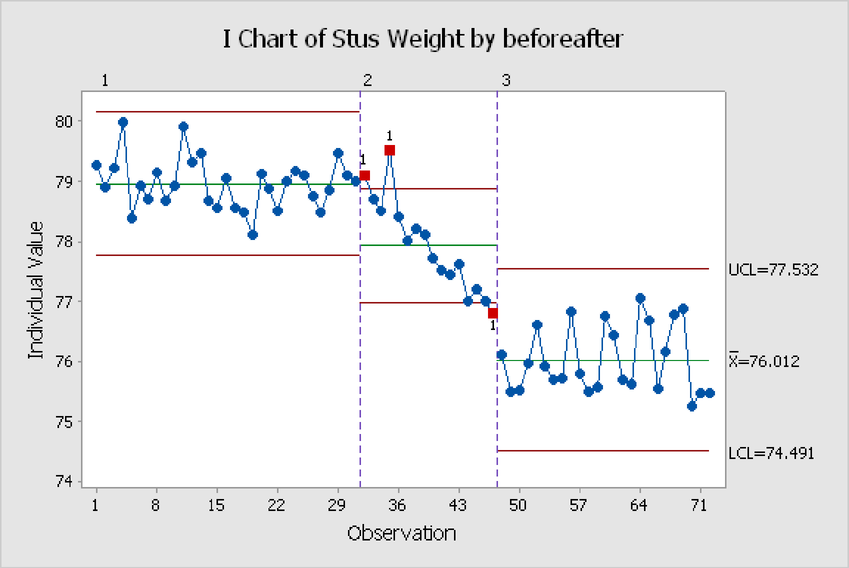


Figure 28: An Individual’s Chart for My Weight during a diet.

We can see from Figure 28 that during the “baselining” period of one month (31 days) my weight did vary day-to-day, but it was stable. The mean and Control Limits were calculated using just the 31 data points.

The next period on the plot is during the actual diet period. Here only the 15 data points collected during the diet period were used to calculate the second set of mean and Control Limits. Note there are assignable cause signals here: which I would expect. The plot even picks out a “bad day” during the diet on day 3 (observation 34) which resulted in a significant weight increase on day 4 (observation 35).

The final period following the 15 days of dieting represents the third stage for which control limits are calculated. Although the day-to-day variation is actually greater than that of the first period, it is stable. What is perhaps most important is that my diet has worked. Because there are two stable periods with different means and Control Limits a change has occurred. The new mean is lower by just under 3Kg. The fact following the diet my weight, although varying day-to-day, shows a stable pattern means I have returned to maintainable weight – but lower.

Given any system the Individuals chart can tell you so much about what is going on. We often start by measuring the system outputs. Typically, we discover, especially for Human Activity Systems, the system is not in control and subject to the vagaries of "management interference". Figure 29 shows a "classic" situation where the change is data from one point to another appears to be so large that action is deemed necessary. In Figure 29 we see the monthly training costs for a large organization. The training Manager notices that the last December figure is some £106,100. The previous month was some £94,300. An increase of over 12%.

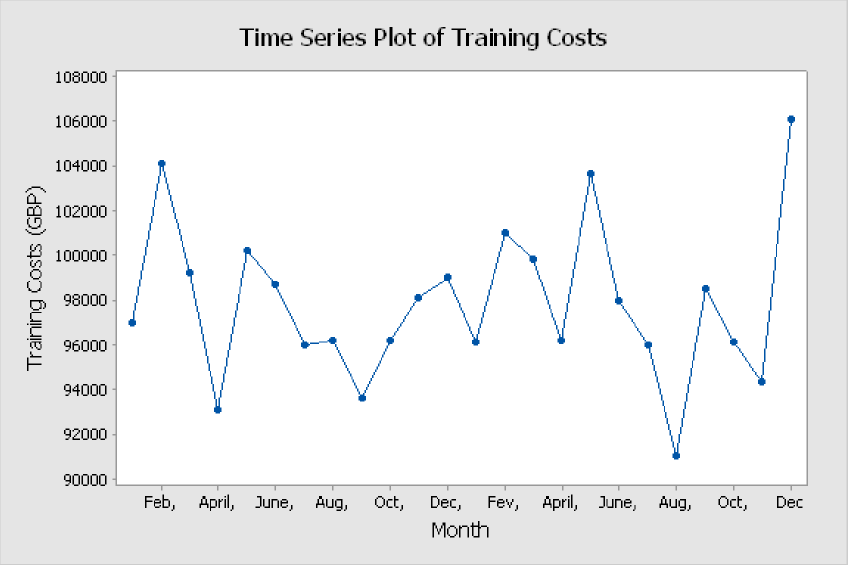


Figure 29: Time Series plot of Training Costs over a two-year period.

My experience is that faced with such data, many managers will initiate an investigation that will result in the introduction of changes (at some cost). Figure 30 shows the same data plotted on an Individuals Chart, from which it can be clearly seen that the variation is stable – no action is needed whatsoever.

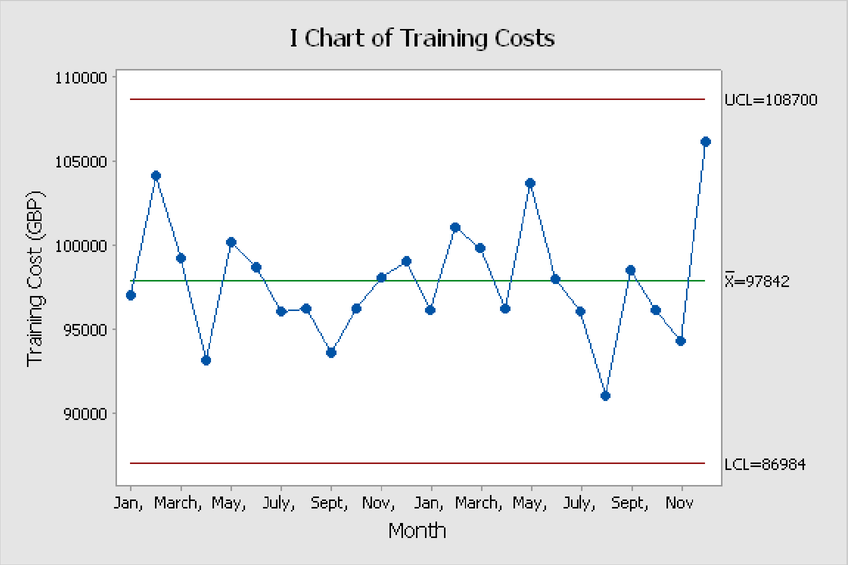


Figure 30: Individuals Chart for the Monthly Training Costs

In fact, Figure 30 shows that the monthly training costs could be as £108,700 before there is any need to be concerned.

Let's look at another example. This concerns factory that manufactures die cast model cars. The company uses water from a nearby stream to cool the moulding equipment. They are allowed to recycle the water back into the stream as long as the impurities from the dies does not exceed 50mg/1000 litres. A sample of the cooling water is taken and analysed each day in the company laboratory. The results plotted on the Individuals Chart shown in Figure 31.

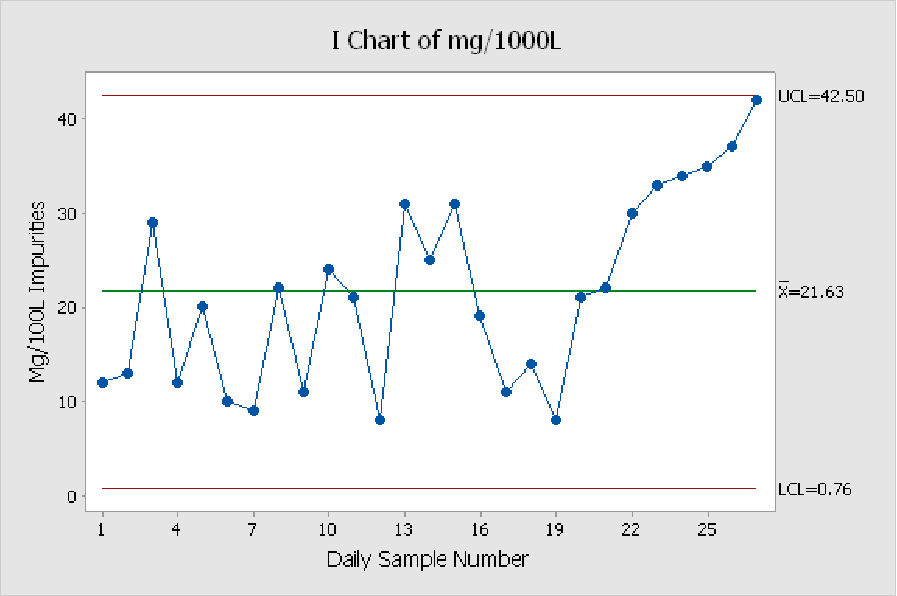


Figure 31: Individual Chart of Cooling Water Impurities

Figure 31 demonstrates the importance of applying all the tests for assignable causes. The situation shown in Figure 31 is not stable and assignable causes are present. In this particular case it is test 3 that signals the assignable cause with 6 consecutive points rising. Figure 32 shows this with the test activated in the software package Minitab^{TM}.

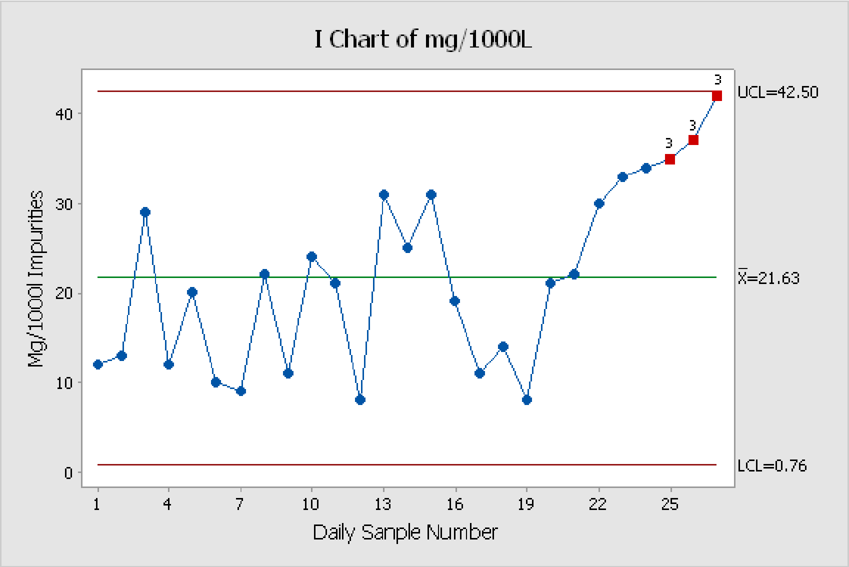


Figure 32: Individual Chart of Cooling Water Impurities with all test active

Note is Figure 32, none of the samples has violated the 50mg/1000litre level, but unless action is taken it will soon occur. Perhaps more important is the advanced warning the Control Chart provides and also an indication of when the assignable cause happened. Clearly something occurred at the discast facility after day 19.

In explaining the concept and principles of Control Charts I have focused on one type; the Individuals Chart. There are many others worthy of note. It is common when plotting Individuals Charts to also plot the "Moving Range". The Moving Range is the difference between the current data point and the last data point. It is used to visualise how the dispersion or spread of the data is changing over time.

Figure 33 shows a I-MR Chart (Individuals-Moving Range) for the water impurities data. The individuals element is the same as figure 32. The lower plot is the Moving Range which starts at sample 2 and is the difference between sample 2 and sample 1 (in this case it is 13 – 12 = 1Mg/1000L). The next point is the difference between sample 3 and sample 2 (19 – 13 = 16Mg/1000L).

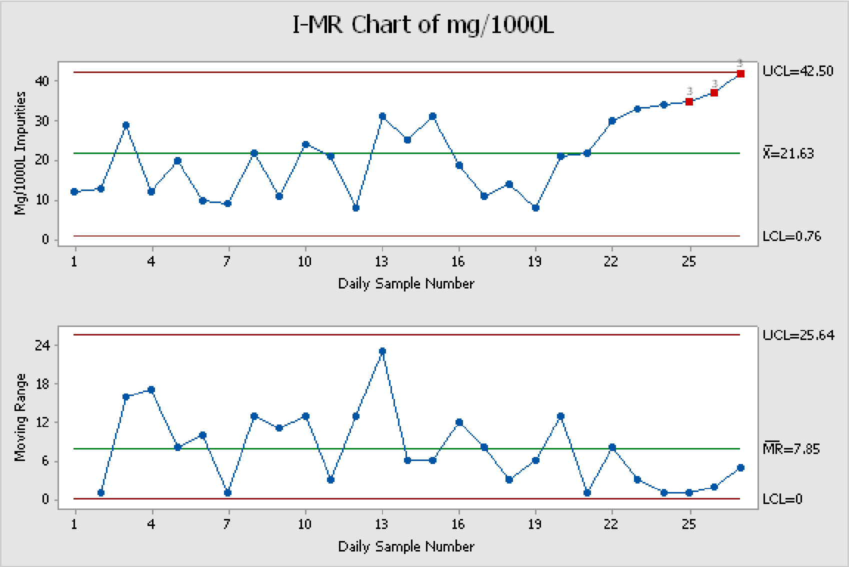


Figure 33: An I-MR Control Chart for Water Impurities

Individuals Charts are really good if the data collected from the system is Normally distributed. In some situations, this will not be the case. If we can model the variation with a known distribution, it is possible to either transform the data (make appear Normal) or use the known distribution to determine suitable control limits. Alternatively, an Xbar – R chart is used.

With an Xbar–R Chart the data is collected in small samples, typically a small sample of 5, rather than indivual values. For example, in the case of the water impurities rather than one daily sample, a set of 5 could be taken daily. How and when we take the samples is important as they are used to estimate the chance casuse varaiation. Hence taking all 5 water samples at the same time will not allow any estimate of the likely daily variation. Equally, speading them out over the day can allow asignable cause variation to be included as the chance cause variation. System Context is all important here and it is necessary to have a sound undersatnding of the system through the use of Systems Thinking tools.

From the small samples two sample properties are calculated:

- The sample mean or Xbar.
- The sample range or R.

The sample mean, Xbar, is used to monitor the accuracy of the system, while the sample range, R, is used to monitor the precision of the system. Here we are back to the arguments of Figure 2 and Figure 3 and our desire to be consistently accurate.

Figure 34 shows an Xbar – R Control Chart.

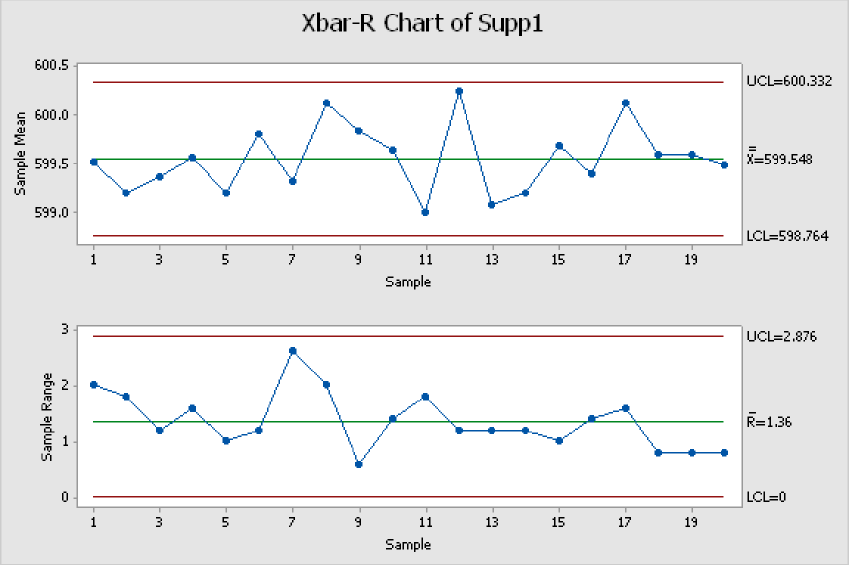


Figure 34: An Xbar – R Control Chart

Figure 34 shows the system to be stable in both mean and spread. It is another question, however, that the mean and range are what is desired. Xbar – R Control Charts can detect shifts in mean and spread as shown in figure 35.

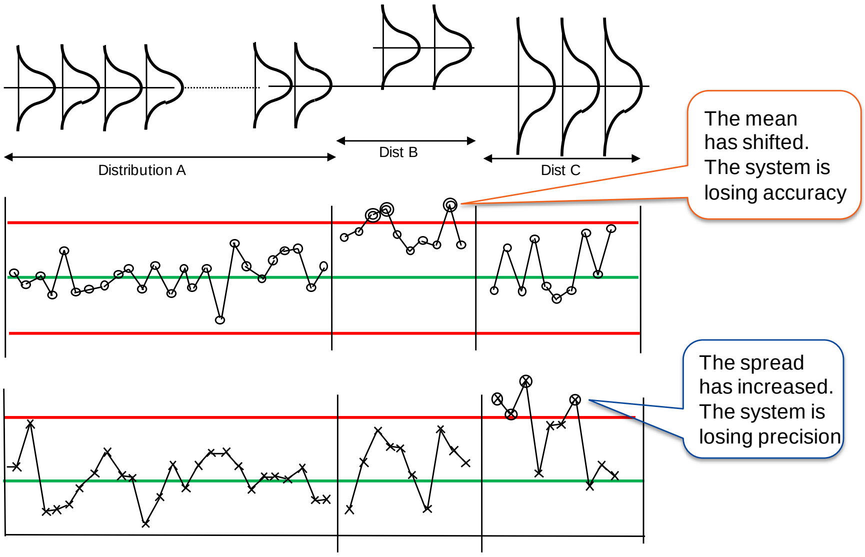

Distribution A
Dist B
Dist C
The mean has shifted. The system is losing accuracy
The spread has increased. The system is losing precision

Figure 35: The Xbar – R Control Chart can detect shifts in the mean of variation and changes in the spread of the variation.

The big advantage of the Xbar – R Control Chart is because we are looking at sample characteristics the underlying system variation distribution does not have to be Normal. Here we are relying on what is known as the *Central Limit Theorem*[8].

[8] The Central Limit Theorem states that irrespective of the distribution of the measured system quantity, if this is sampled, the means of samples will be distributed Normally.

Both I-MR and Xbar-R Control Charts are for continuous data. Some system data is likely to be discrete. For example:

- The number or proportion of incorrectly completed application forms.
- The number of errors on an application form.
- The number of paint blemishes on a new motorcar.
- The number of reported safety incidents per month.

In each of these examples we are able to count the occurrences of the event described. Discrete data Control Charts work on the same principle as continuous data Control Charts in that the Control Limits are set as:

Control Limits = average ± 3 standard deviations

How the standard deviation is calculated is not based on the Normal Distribution but either on the Binomial or Poisson Distribution. Therefore, discrete Control Charts basically fall into two groups:

- Discrete proportion/percentage where we can count the number of occurrences and also the number of non-occurrences. In such cases, the Control Chart is based on what is known as the Binomial Distribution. Under certain conditions, it is possible to approximate the Binomial Distribution with a Normal Distribution. There are two Control Chart types that use the Binomial Distribution:
    - **np-Chart:** Proportion of non-occurrences in samples of constant size.
    - **p-Chart:** Proportion non-occurrences in samples of varying size.
- Discrete count where we can count the number of occurrences of a certain “event” or characteristic, but not the number of non-occurrences. In such cases the Control Chart is based on what is known as the Poisson Distribution. There are two Control Chart types that use the Poisson Distribution:
    - **c-Chart:** Number of non-occurrences per unit/opportunity in samples/opportunities of constant size.
    - **u-Chart:** Number of non-conformities per unit/opportunity in samples/opportunities of varying size.

Selecting the correct chart depends upon the data and Figure 36 presents a simple flow chart to help ensure the right selection.

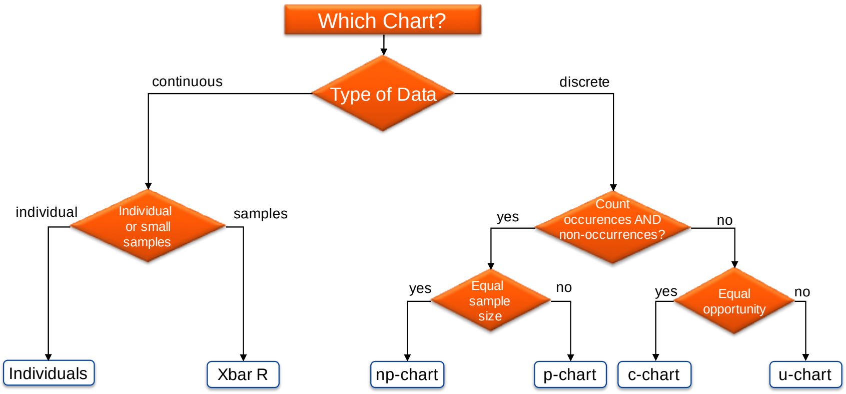


Figure 36: A flow Chart to help select the Correct Control Chart

Table 5 presents several examples to help explain the rationale behind choosing the best Control Chart.

Situation	Data Collected	Frequency	Chart Type	Rationale
Predicting customer numbers in a restaurant	% seat occupied	hourly	p	The restaurant will have a finite number of seats (say 50) hence p = number occupied /50
	Number of seat occupied	hourly	np	Same situation but np = number of seats occupied
Number of complaints resolved within five working days	Number of complaints	weekly	p	The number of complaints will vary from week to week. So the only meaningful measure is the proportion resolved (in five working days).
Site safety	Number of reported safety incidents per day worked per month	monthly	u	Cannot count the number of non-safety incidents. Also, the number of days per month worked could vary.
Predicting system yield	Percentage good output	daily	p	The system may produce vary quantities of output.

Table 5: Examples of selecting the "best" discrete Control Chart.

With Discrete Control Charts, of any type, only the first four tests for assignable causes given in Table 4 apply. Figure 37 shows an example of a c chart. This example is for an insurance company monitoring the number of claims for accidental damage on a weekly basis in a particular area.

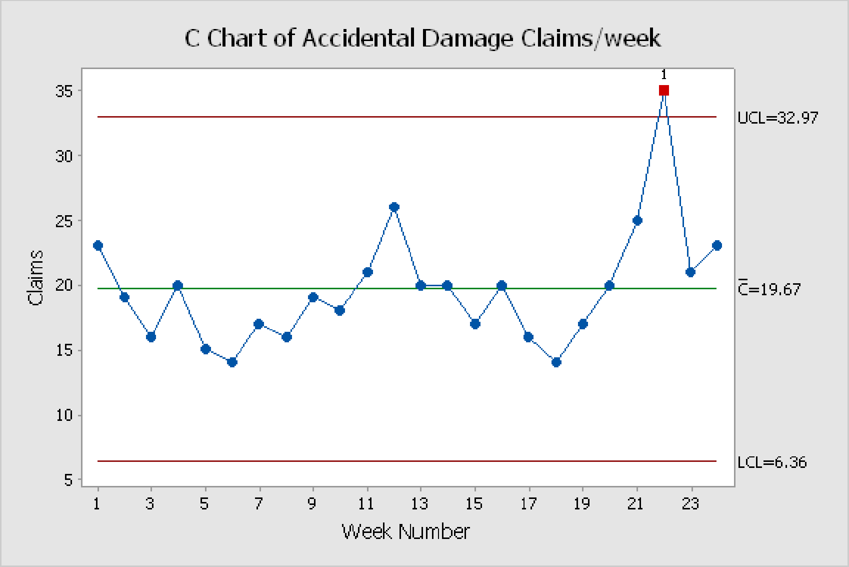


Figure 37: A c chart of accidental claims/week

What is apparent from Figure 37 is that an assignable cause is signalled in week 21 showing this number of claims for accidental damage is higher than expected. Such patterns of behaviour do exist in the Insurance industry, where people having seen friends make successful claim will also try, resulting in a reinforcing loop. What was developed for the Manufacturing industries to monitor product quality, is now being used in many surprising situation ranging from Insurance claims, through to Police forces monitoring drug related crime.

Discrete Data Control Charts are not without their problems. Firstly, it is important to realise that discrete does mean discrete. A p chart is for attribute data where the numerator AND denominator are both discrete. It is possible to derive proportions from continuous data in both numerator and denominator. For example:

- Percentage of time a machine is broken.
- Percentage scrap when measured as a continuous variable e.g. tonnes of steel.
- Efficiency and accounting ratios % productivity, % profit etc. where ratios are based on time, £ etc.

In such cases, you should use an Individuals Control Chart.

There are also occasions where the chart just looks wrong! Typically, because there are too many assignable causes signalled. Figure 38 shows such a situation with a c chart, but similar issue can occur with u, n and np Control Charts. The basic rule is if more than 1/3 of the data points are outside the limits, the chart “doesn’t look right”.

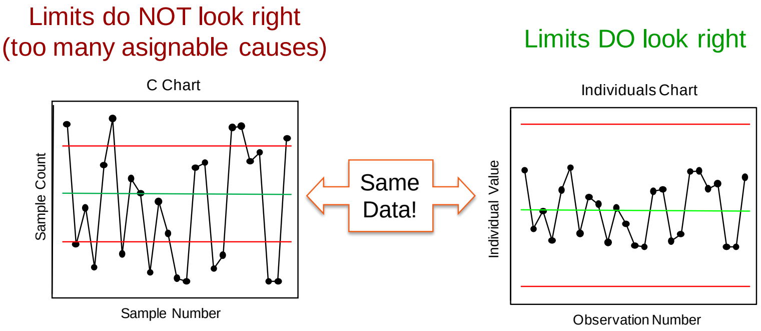


Figure 38: A c chart with too many Assignable Causes

For c and u Charts:

- Whenever the counts > 50 and the number units is relatively small the data may not be very well modelled by a Poisson distribution.

For n and np Charts:

- If the sample sizes are very large (> 1000) it is possible that the data is better modelled by a Normal distribution
- The assumption that the expected proportion is constant for each sample does not hold – so the data are not binomial

In either situation, use an individual's chart instead of a c, u, p or np chart

What Goes Wrong: The limitations of Graphical Analysis

The Achilles' heel of Graphical Analysis' is data collection. If the data collected from the system of interest is not planned, not meticulously carried out and poorly recorded, not amount of analysis will help. It really is a case of rubbish in rubbish out.

The second limitation with Graphical Analysis, is summed up by the statistician's saying that "if you torture the data for long enough it will tell you anything". If there are no patterns in the data, there are no patterns in the data if even you want them to be there. Actually the whole point of Graphical Analysis is to avoid this scenario by providing tools that will allow the true information to be extracted from the raw data.

Frequency Plots (Histograms) and Time Series (Control Charts) go hand in hand. While it is possible to extract information from just the one, usually the it is the combination of both that provides the understanding of system variation sought. It also important to keep the purpose in mind: we undertake Graphical Analysis not because it is the next on the list, but because we suspect that variation is a key phenomena of the system under investigation. A few systems are totally deterministic, such as some software intensive systems, and therefore Graphical Analysis has limited value. Most, however, are subject to variation, especially Human

Activity Systems, and in these variation, unless properly analysed can lead to an inadequate, if not incorrect understanding.

Success Criteria

Graphical Analysis is easy to ignore, easy to get wrong and easy to miss-interpret. The following are worth taking note of:

- Put effort into understanding as much as you can about the system first. Use Systems thinking tools like Multiple Cause Diagrams, Context, Sequence Diagrams, Conceptual Models etc. to find out about how the system might behave to form hypotheses which is turn can help determine what might be worth collecting data on.
- Use the Input-Output Analysis tool to identify all system inputs and outputs (and if the Sequence Diagram is good enough internal system parameters). Consider the importance of each in terms of collecting data on the likely variation.
- Construct clean Data collection plans that explain what data is to be collected from the system, how it is to be collected, when it is to be collected.
- If you are relying on other people to collect data for you make sure they understand the reason for collecting the data and why it is so important to collect as you have planned.
- Plot the data as it is collected. This can often show poor collection practice and indicate additional items to be collected.
- Wherever possible use proprietary software Tools like Minitab^{TM} to perform the data analysis.
- Wherever possible validate you finding with further data collection and Graphical Analysis.

Appendix A: Some Mathematics

This tool guide was intentionally written with a few equations and mathematical equations and proofs as possible. Mathematics, however, can add a further level of understanding about the real world, just as System Thinking can. The following provides some of the mathematics behind variation and Graphical Analysis.

Numerical Representations

While the graphical representation of large sets of data is a good communication tool, it is still not compact enough, for some situations. In these situations, numerical representations of the data are highly useful. In general, we need to know two features, the central tendency and dispersion.

Measures of Central Tendency

In general, for sets of raw data there is a value about which most of the observed values tend to group. This phenomenon is known as the central tendency. There are three methods of measuring central tendency:

1. **Mean**

The mean, sometimes called the arithmetic average, is the most common measure of central tendency. Indeed, whenever we talk about an average most people will think of the mean. Mathematical the mean of a set of *n* observations of *x* is:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

The Greek letter Σ (capital sigma) is used to indicate summation of the individual observations. Strictly speaking, we should annotate the summation sign to indicate the summation limits, that is:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

This annotation, while precise and correct, is tedious, and people often simplify it by just using Σ and thereby assuming the limits of 1 to *n*.

Example: The following data is collected from a system: 2.1, 3.4,5.6, 1.2, 2.5, and 6.2. The mean is given by:

$$\bar{x} = \frac{2.1+3.4+5.6+1.2+2.5+6.2}{6} = \frac{21}{6} = 3.5$$

2. Median

The median of a set of data, is the central value when the values of the data set are ranked in order of magnitude. The median is a better estimate of the true mean when small samples (four or less) are used. It is also very useful if the variation distribution is highly skewed. In both cases the “true” picture can be distorted by extreme outlying values. If there are an even number of observations, then the median is taken as the mean of the central pair.

Example: Find the median of:

i)   1, 5, 7, 2, 3, 6, 6, 5, 4

ii) 1, 5, 2, 7, 3, 6, 4, 5, 9, 2

for i) rank in magnitude order

1, 2, 3, 4, 5, 5, 6, 6, 7

thus median = 5

for ii) rank in magnitude order

1, 2, 2, 3, 4, 5, 5, 6, 7, 9

the central pair = {4,5} and hence the median 4+5/2 = 4.5

3. Mode

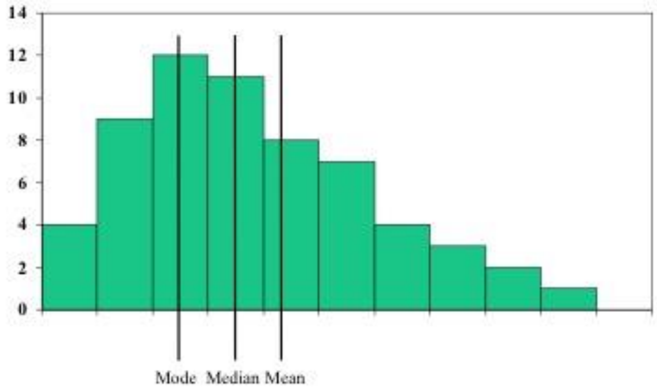
The mode is most frequently occurring value in a data set.

Example: Find the mode of following data set

2, 1, 3, 4, 3, 2, 3, 5, 7

The value 3 occurs more than any other value hence the Mode = 3

Differences between Mean, Median and Mode



If the frequency histogram is symmetric, then the Mean, Median and Mode all take the same value. If, however, the frequency histogram is non-symmetric or skewed then the Mean, Median and Mode will take different values.

Measures of Spread

Spread is the extent to which the set of data values is dispersed on either side of the central value. Spread is a measure of variability and there are three basic methods:

1. Range

The range is simply the difference between the largest and smallest values in the data set.

Range = R = largest value – smallest value

While the range is easy to determine it is susceptible to 'freak' data values in the data set. This can occur for a number of reasons, but in general it is down to human error during data collection.

Example: Find the range of 1.2, 5.4, 2.3, 3.3, 1.9

Range = 5.4 –1.2 = 4.2

Note that in this example the value 5.4 could be a “freak” since it is significantly larger than any of the other values.

2. Variance

We have seen that while the range is easy to calculate it can give a false measure of the dispersion if there is freak data. An alternate approach is to calculate, for a data set, the deviations from the mean and then calculate the average of these deviations. While this sounds a good idea it fails in practice! Consider the last example which has a mean of 2.82.

x_i	deviations from the mean
1.2	*-1.62*
5.4	*2.58*
2.3	*-0.52*
3.3	*0.48*
1.9	*-0.92*
Sum	*0.0*

The proof that the sum of the deviation from the mean will always be zero follows is.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad (1)$$

To make things easier to follow. I'm going to simplify the summation term by

$$\sum = \sum_{i=1}^{n}$$

Let us expand out the brackets in equation (1)

$$\sum(x_i - \bar{x}) = x_1 - \bar{x} + x_2 - \bar{x} + \ldots + x_n - \bar{x}$$

which can be rewritten as

$$\sum(x_i - \bar{x}) = x_1 - \sum\frac{x_i}{n} + x_2\sum\frac{x_i}{n} + \ldots + x_n - \sum\frac{x_i}{n}$$

because by definition

$$\bar{x} = \sum\frac{x_i}{n}$$

Collecting all the *x_1, x_2*, etc. together and collecting all *n* means together gives

$$\sum(x_i - \bar{x}) = (x_1 + x_2 + \ldots + x_n) - n\sum\frac{x_i}{n}$$

resulting in

$$\sum(x_i - \bar{x}) = \sum\frac{x_i}{n} - \sum\frac{x_i}{n} = 0$$

What is really needed is the absolute deviations from the mean which can be achieved by calculating the mean square deviation from the mean

$$\sigma^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

The statistic σ^2 is called the variance.

Example: Find the variance of 1.2, 5.4, 2.3, 3.3, 1.9

$$\sigma^2 = \frac{1}{5}\left((1.2-2.82)^2 + (5.4-2.82)^2 + (2.3-2.82)^2 + (3.3-2.82)^2 + (1.9-2.82)^2\right) = 2.126$$

3. Standard Deviation

The units of the variance are not the same units as the mean, but the mean^2. Thus the most common measure of dispersion is the square root of the variance or standard deviation σ.

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

Example: Find the standard deviation of 1.2, 5.4, 2.3, 3.3, 1.9

$$\sigma = \sqrt{\frac{1}{5}\left((1.2-2.82)^2 + (5.4-2.82)^2 + (2.3-2.82)^2 + (3.3-2.82)^2 + (1.9-2.82)^2\right)} = 1.458$$

Sampling

As stated earlier, the data set that contains all the possible observations is called the population. Populations are often too large, or unknown because they contain future events and observations. Faced with such situations we take a sample of the population, with the intent that characteristics of the sample estimate that characteristics of the population.

The numerical statistics (e.g. mean, standard deviation) of the population will in general be different from the statistics of the sample. It is normal convention to define

μ as the mean of the population

and

σ as the standard deviation of the population

while

$\bar{x}$ is the mean of the sample

s is the standard deviation of the sample.

In practice $\overline{x} \neq \mu$ and $s \neq \sigma$ but estimates of μ and σ can be found from a sample of *n* items from the population.

$$\mu \approx \overline{x} = \frac{\Sigma x_i}{n}$$

and

$$\sigma \approx \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n-1}}$$

Probability

Probability is the study of events that happen by chance. This may at first appear rather strange that chance may be of interest in understanding variation, but it is the way in which randomness can be predicted. Formally, the probability of an event is defined as:

If a trial can result in any one of *n* exhaustively mutually exclusive, and equally likely events, and *m* of these are regarded as successes, then the probability P that any one trial will result in a success is

$$P = \frac{m}{n}$$

Exhaustive means *n* and only *n* possible results or outcomes. Mutually exclusive means that any event, once it has occurred excludes all other possible events. The usual way of denoting the probability of an event *E* is *P(E).*

Example: There are 4 aces in a pack of 52 playing cards, what is the probability of drawing an ace?

$$P(ace) = \frac{Number\,of\;\; successes\,(i.e.\,number\,of\;\; aces)}{Number\,of\;\; outcomes(i.e.\,number\,of\;\; cards)} = \frac{4}{52} = \frac{1}{13}$$

If an event will certainly occur it has a probability of 1. Likewise, if an event will certainly not occur it has a probability of 0. Thus if we know the probability that an event will occur is *P(E),* and the probability it will not occur is *Q(E)*, then

$$P(E) + Q(E) = 1$$

Example: The chance of drawing a "diamond" from a pack of 52 cards is 13/52 or ¼. Thus the probability of not drawing a "diamond" is 39/52 or ¾ which can be obtained from

$$P(diamond) + P(not\ diamond) = 1$$

$$1/4 + P(not\ diamond) = 1$$

In the above examples it is possible by common sense to determine the various probabilities. In practice this is harder to achieve. In such cases, the probabilities can often be determined from the collection of data.

Example: In a survey it was found, that out of 3000 babies born, 1575 where girls. Thus it can be estimated that the probability that any particular baby will be a girl is

$$\frac{1575}{3000} = 0.525$$

or 52.5%.

In practice, both approaches are used depending upon the problem at hand. The study of probability and therefore chance is based upon two basic laws.

Addition law of Probability (OR)

In some situations we are interested in the probability that either one of two possible outcomes *A* or *B* will occur. Here we wish to determine the probabilities that either *A* OR *B* will occur. This is achieved by adding together the respective probabilities

$$P(A\,or\;B) = P(A) + P(B)$$

Example: The probability of throwing a 5 or 6 with a fair dice is

$$P(5\,or\;6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Multiplication Law of Probability (AND)

The multiplication law is used to determine the probability of successive events occurring. This is achieved by multiplying together the respective probabilities. There are, however two situations. One where the probability of an event is affected by what has happened previously. This is called conditional probability. Here the probability that an event E_2 when another event E_1 has already occurred is

$$P(E_2 | E_1)$$

The probability that E_1 *AND* E_2 occur is

$$P(E_1\;and\;E_2) = P(E_1) \times P(E_2 | E_1)$$

Example: What is the probability of drawing 2 successive aces from a pack of 52 playing cards?

The probability of drawing an ace first is

$$P(ace1)=\frac{4}{42}=\frac{1}{13}$$

The probability of drawing the second ace is affected by the fact that a card has already been taken from the pack

$$P(ace2)=\frac{3}{51}$$

Hence the probability of drawing two successive aces is

$$P(ace1\ and\ ace2)=\frac{4}{52}\times\frac{3}{51}=\frac{1}{221}$$

The other situation is where the two events are independent, that is E_2 is not affected by any previous event E_1

Example: If, in the previous example, the first ace had been replaced the probability that they both occur is

$$P(E_1andE_2)=P(E_1)\times P(E_2)$$

$$P(ace1\ and\ ace2)=\frac{4}{52}\times\frac{4}{52}=\frac{16}{2704}=\frac{1}{169}$$

Sampling with and without replacement

These last two situations have an important role in sampling. Usually sampling takes place without replacement, but if the population is large, replacement makes little difference.

Example: If in the drawing two aces example, there had been 52000 playing cards in the pack and 4000 aces, then the probability of drawing two aces without replacement is

$$\frac{4000}{52000}\times\frac{3999}{51999}=0.005916$$

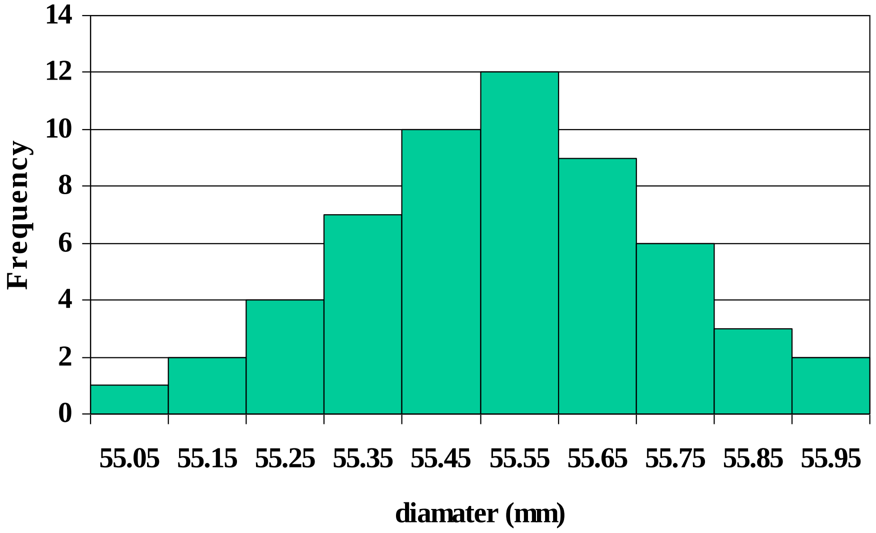
and with replacement

$$\frac{4000}{52000}\times\frac{4000}{52000}=0.005917$$

That is negligible difference.

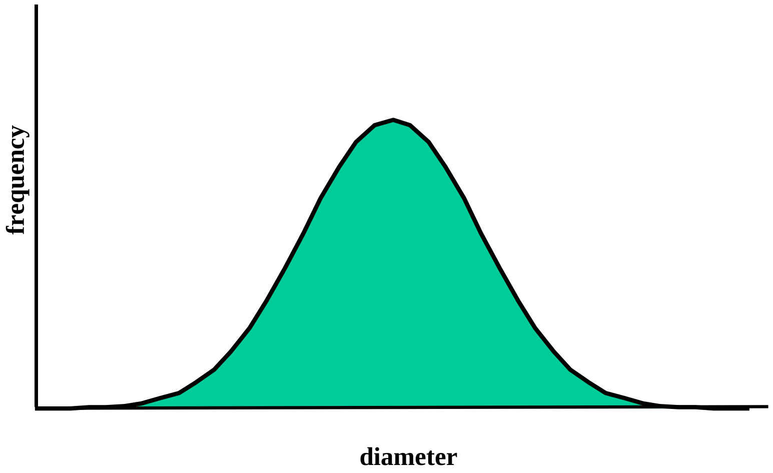
Probability Distributions

A fundamental concept that is used as the basis for six sigma is the idea of a probability distribution. A probability distribution is a mathematical model that relates the values of a variable with the probability of observing the values. There are many different types of distribution that have been developed to model real world situations. Consider the earlier piston example.



The frequency histogram here looks discrete only because of the size of class used to allocate the various measurements. If we reduced the class size the histogram would become “smoother”. If this process were continued, we would arrive at smooth frequency histogram.

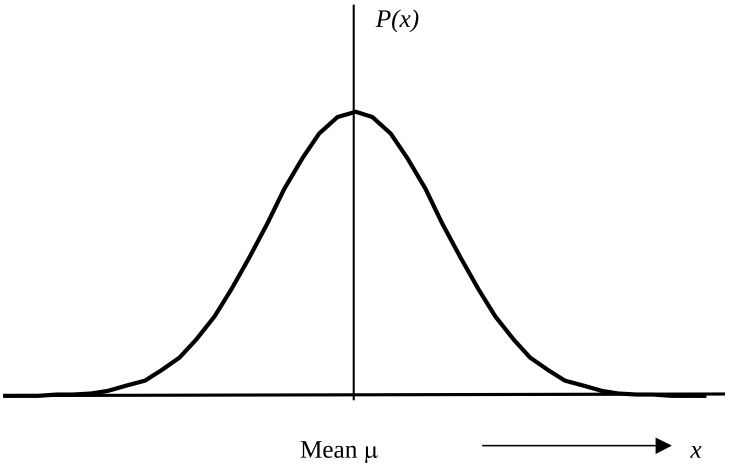
In such case the probability function *P(x)* is continuous and to distinguish it from a discrete distribution it is frequently called the probability density function.



There are a number of “standard” probability density functions that can be used to model real world situations, and the most common is the Normal or Gaussian distribution.

The Normal Distribution

The normal distribution occurs very frequently in practice. It has a mean of μ and a standard deviation of σ and has the characteristic “bell” shape as shown below.

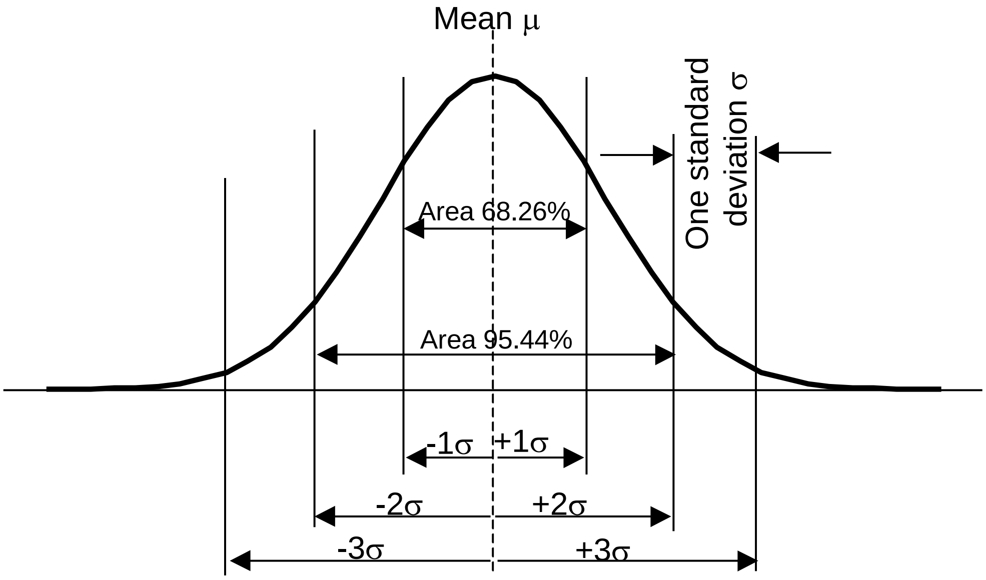


Like all probability distributions the area under the curve is 1.0. A very important feature of the Normal distribution is that the area under the curve at

$$\mu \pm \sigma = 0.6826 \, or \; 68.26\%$$

and within

$$\mu \pm 2\sigma = 0.9544 \;\; or \; 95.44\%$$



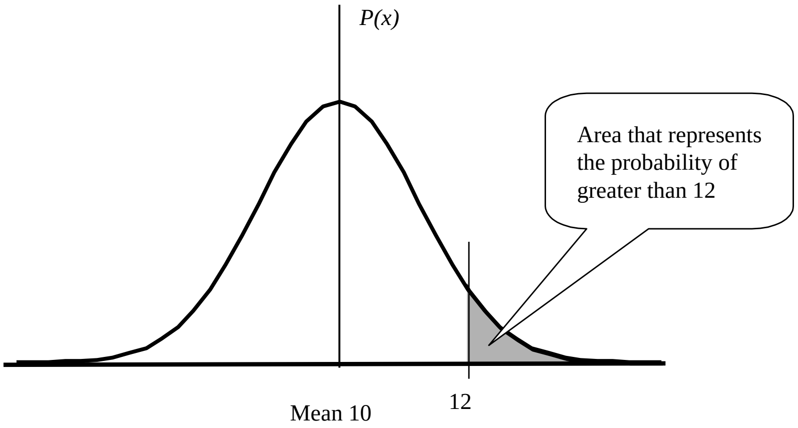
This property is extremely useful since if some characteristic is known to be distributed normally, predictions can be made about the quantities that lie within certain ranges. The one problem that does arrive is that for different situations, although a characteristic is known to be normal, some tedious mathematics has to be performed to make the required predictions. To overcome this issue a transformation can used to convert from a situation with a mean of μ and a standard deviation of σ to standard form where the mean is 0 and the standard deviation is 1. The transformation is given by

$$Z = \frac{x-\mu}{\sigma}$$

This version of the normal distribution is called the Standard Normal Distribution and tables exist for areas under the curve. A copy of the table is given in Table A1.

Example: A characteristic is found to be normally distributed, with a mean of 10 and a standard deviation of 4. What is the probability that characteristic is greater than 12?

It is always advisable to draw a picture of the situation:



Converting to standard form.

$$Z = \frac{12 - 10}{4} = 0.5$$

and from Table A1, *a Z value of 0.5 provides a probability (area) of 0.3085. Hence, the probability that the characteristic will be greater than 12 is 0.3085.*

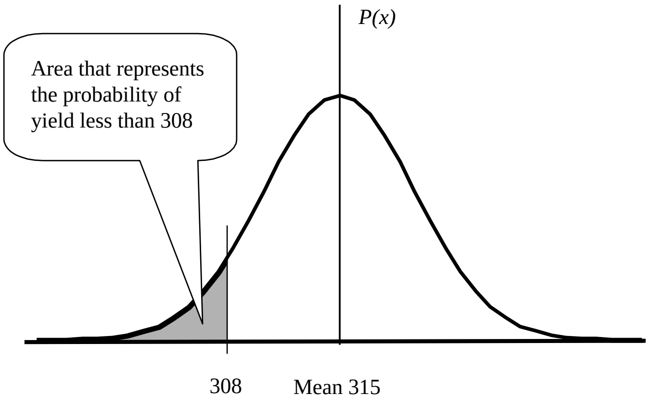
Example: A chemical is manufactured by a batch process, the average yield per batch is 315kg with a standard deviation of 4.1kg. Assuming a normal distribution determine the probability of the yield

i) being less than 308kg

ii) being in the range of 315 + 8 kg.

Part i)

As stated earlier, it is always best to draw a picture of the area you are after.



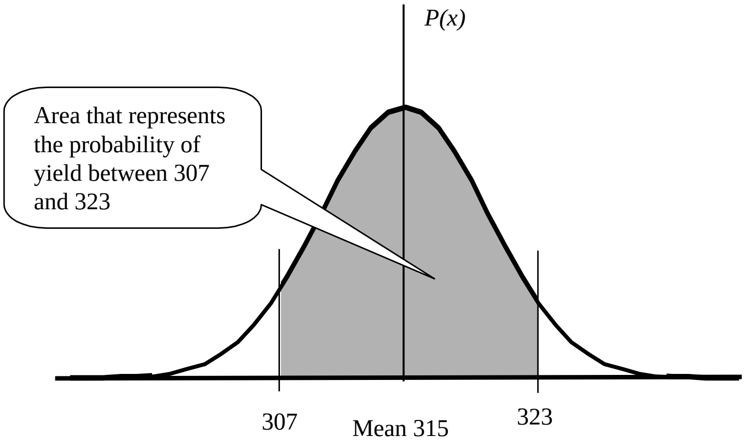
Applying the transformation to the standard form

$$Z = \frac{308 - 315}{4.1} = -1.707$$

In this case the Z value is negative, but the table in Appendix A is for positive values only. In such cases we ignore the negative sign and read the probability value for Z = 1.707 which is between 0.0446 (1.700) and 0.0436 (1.71). Assuming a linear relationship gives 0.0439. So the probability of a yield less than 308 is 0.0439.

Part ii)

Again, the starting point is to draw a picture of what is required.



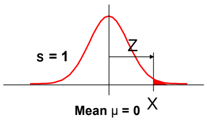
The table gives the areas in each tail:

For the right hand tail $Z = \dfrac{307 - 315}{4.1} = 1.951$

For the left hand tail $Z = \dfrac{323 - 315}{4.1} = 1.951$

From the tables 1.951 gives 0.0255 and so the area required is

$$1 - 2 \times 0.0255 = 0.949$$



The Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
3.1	.000968	.000936	.000904	.000874	.000845	.000816	.000789	.000762	.000736	.000711
3.2	.000687	.000664	.000641	.000619	.000598	.000577	.000557	.000538	.000519	.000501
3.3	.000483	.000467	.000450	.000434	.000419	.000404	.000390	.000376	.000362	.000350
3.4	.000337	.000325	.000313	.000302	.000291	.000280	.000270	.000260	.000251	.000242
3.5	.000233	.000224	.000216	.000208	.000200	.000193	.000185	.000179	.000172	.000165
3.6	.000159	.000153	.000147	.000142	.000136	.000131	.000126	.000121	.000117	.000112
3.7	1.08E-4	1.04E-4	9.96E-5	9.58E-5	9.20E-5	8.84E-5	8.50E-5	8.16E-5	7.84E-5	7.53E-5
3.8	7.24E-5	6.95E-5	6.67E-5	6.41E-5	6.15E-5	5.91E-5	5.67E-5	5.44E-5	5.22E-5	5.01E-5
3.9	4.81E-5	4.62E-5	4.43E-5	4.25E-5	4.08E-5	3.91E-5	3.75E-5	3.60E-5	3.45E-5	3.31E-5

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
4.0	3.17E-5	3.04E-5	2.91E-5	2.79E-5	2.67E-5	2.56E-5	2.45E-5	2.35E-5	2.25E-5	2.16E-5
4.1	2.07E-5	1.98E-5	1.90E-5	1.81E-5	1.74E-5	1.66E-5	1.59E-5	1.52E-5	1.46E-5	1.40E-5
4.2	1.34E-5	1.28E-5	1.22E-5	1.17E-5	1.12E-5	1.07E-5	1.02E-5	9.78E-6	9.35E-6	8.94E-6
4.3	8.55E-6	8.17E-6	7.81E-6	7.46E-6	7.13E-6	6.81E-6	6.51E-6	6.22E-6	5.94E-6	5.67E-6
4.4	5.42E-6	5.17E-6	4.94E-6	4.72E-6	4.50E-6	4.30E-6	4.10E-6	3.91E-6	3.74E-6	3.56E-6
4.5	3.40E-6	3.24E-6	3.09E-6	2.95E-6	2.82E-6	2.68E-6	2.56E-6	2.44E-6	2.33E-6	2.22E-6
4.6	2.11E-6	2.02E-6	1.92E-6	1.83E-6	1.74E-6	1.66E-6	1.58E-6	1.51E-6	1.44E-6	1.37E-6
4.7	1.30E-6	1.24E-6	1.18E-6	1.12E-6	1.07E-6	1.02E-6	9.69E-7	9.22E-7	8.78E-7	8.35E-7
4.8	7.94E-7	7.56E-7	7.19E-7	6.84E-7	6.50E-7	6.18E-7	5.88E-7	5.59E-7	5.31E-7	5.05E-7
4.9	4.80E-7	4.56E-7	4.33E-7	4.12E-7	3.91E-7	3.72E-7	3.53E-7	3.35E-7	3.18E-7	3.02E-7
5.0	2.87E-7	2.73E-7	2.59E-7	2.46E-7	2.33E-7	2.21E-7	2.10E-7	1.99E-7	1.89E-7	1.79E-7
5.1	1.70E-7	1.61E-7	1.53E-7	1.45E-7	1.38E-7	1.30E-7	1.24E-7	1.17E-7	1.11E-7	1.05E-7
5.2	9.98E-8	9.46E-8	8.96E-8	8.49E-8	8.04E-8	7.62E-8	7.22E-8	6.84E-8	6.47E-8	6.13E-8
5.3	5.80E-8	5.49E-8	5.20E-8	4.92E-8	4.66E-8	4.41E-8	4.17E-8	3.95E-8	3.73E-8	3.53E-8
5.4	3.34E-8	3.16E-8	2.99E-8	2.82E-8	2.67E-8	2.52E-8	2.39E-8	2.26E-8	2.13E-8	2.01E-8
5.5	1.90E-8	1.80E-8	1.70E-8	1.61E-8	1.52E-8	1.43E-8	1.35E-8	1.28E-8	1.21E-8	1.14E-8
5.6	1.07E-8	1.01E-8	9.57E-9	9.04E-9	8.53E-9	8.04E-9	7.59E-9	7.16E-9	6.75E-9	6.37E-9
5.7	6.01E-9	5.67E-9	5.34E-9	5.04E-9	4.75E-9	4.48E-9	4.22E-9	3.98E-9	3.75E-9	3.53E-9
5.8	3.33E-9	3.13E-9	2.95E-9	2.78E-9	2.62E-9	2.47E-9	2.32E-9	2.19E-9	2.06E-9	1.94E-9
5.9	1.82E-9	1.72E-9	1.62E-9	1.52E-9	1.43E-9	1.35E-9	1.27E-9	1.19E-9	1.12E-9	1.05E-9
6.0	9.90E-10	9.31E-10	8.75E-10	8.23E-10	7.73E-10	7.27E-10	6.83E-10	6.42E-10	6.03E-10	5.67E-10
6.1	5.32E-10	5.00E-10	4.70E-10	4.41E-10	4.14E-10	3.89E-10	3.65E-10	3.43E-10	3.22E-10	3.02E-10
6.2	2.83E-10	2.66E-10	2.50E-10	2.34E-10	2.20E-10	2.06E-10	1.93E-10	1.81E-10	1.70E-10	1.59E-10
6.3	1.49E-10	1.40E-10	1.31E-10	1.23E-10	1.15E-10	1.08E-10	1.01E-10	9.49E-11	8.89E-11	8.33E-11
6.4	7.80E-11	7.31E-11	6.85E-11	6.41E-11	6.00E-11	5.62E-11	5.26E-11	4.92E-11	4.61E-11	4.31E-11
6.5	4.04E-11	3.78E-11	3.53E-11	3.30E-11	3.09E-11	2.89E-11	2.70E-11	2.53E-11	2.36E-11	2.21E-11
6.6	2.07E-11	1.93E-11	1.81E-11	1.69E-11	1.58E-11	1.47E-11	1.38E-11	1.29E-11	1.20E-11	1.12E-11
6.7	1.05E-11	9.79E-12	9.14E-12	8.53E-12	7.96E-12	7.43E-12	6.94E-12	6.48E-12	6.04E-12	5.64E-12
6.8	5.26E-12	4.91E-12	4.58E-12	4.27E-12	3.98E-12	3.71E-12	3.46E-12	3.23E-12	3.01E-12	2.81E-12
6.9	2.62E-12	2.44E-12	2.27E-12	2.12E-12	1.97E-12	1.84E-12	1.71E-12	1.59E-12	1.49E-12	1.38E-12
7.0	1.29E-12	1.20E-12	1.12E-12	1.04E-12	9.68E-13	9.01E-13	8.38E-13	7.80E-13	7.26E-13	6.75E-13
7.1	6.28E-13	5.84E-13	5.43E-13	5.05E-13	4.70E-13	4.37E-13	4.06E-13	3.78E-13	3.51E-13	3.26E-13
7.2	3.03E-13	2.82E-13	2.62E-13	2.43E-13	2.26E-13	2.10E-13	1.95E-13	1.81E-13	1.68E-13	1.56E-13
7.3	1.45E-13	1.35E-13	1.25E-13	1.16E-13	1.08E-13	9.99E-14	9.27E-14	8.60E-14	7.98E-14	7.40E-14
7.4	6.86E-14	6.37E-14	5.90E-14	5.47E-14	5.07E-14	4.70E-14	4.36E-14	4.04E-14	3.75E-14	3.47E-14
7.5	3.22E-14	2.98E-14	2.76E-14	2.56E-14	2.37E-14	2.19E-14	2.03E-14	1.88E-14	1.74E-14	1.61E-14
7.6	1.49E-14	1.38E-14	1.28E-14	1.18E-14	1.10E-14	1.01E-14	9.38E-15	8.68E-15	8.03E-15	7.42E-15
7.7	6.86E-15	6.35E-15	5.87E-15	5.43E-15	5.02E-15	4.64E-15	4.29E-15	3.96E-15	3.66E-15	3.38E-15
7.8	3.12E-15	2.89E-15	2.67E-15	2.46E-15	2.27E-15	2.10E-15	1.94E-15	1.79E-15	1.65E-15	1.53E-15
7.9	1.41E-15	1.30E-15	1.20E-15	1.11E-15	1.02E-15	9.42E-16	8.69E-16	8.01E-16	7.39E-16	6.82E-16

Table A1: The Standard Normal Distribution