LEAN SIX SIGMA CHEAT SHEETS ©

(includes concepts, formulas, references, and links)

• YELLOW BELT CHEAT SHEETS

• GREEN BELT CHEAT SHEETS

• BLACK BELT CHEAT SHEETS

Prepared by Dr. Keith Schellenberger
For Friends of Lean Six Sigma Services
Updated February 2021

Email: EMBB@LeanSixSigmaServices.net
Phone: (919) 653-8044
Skype: keith.w.schellenberger
website: http://www.LeanSixSigmaServices.net
Linked In:http://www.linkedin.com/in/keithschellenberger

Lean Six Sigma Services

TABLE OF CONTENTS

YELLOW BELT STATISTICS CHEAT SHEET

Includes formulas: what they are, when to use them, references

- **Basic Concepts:**
  - **Mean (average):** add all numbers in the list together and divide by the number of items in the list. The formula:

    $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - **Median (middle):** order the numbers and take the number in the middle. The formula:

    $$\tilde{x} = \begin{cases} \textbf{middle ordered value, if n is odd} \\ \textbf{the average of the two middle ordered values, if n is even} \end{cases}$$

    $$= \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n\ odd \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & n\ even. \end{cases}$$

  - **Mode:** The number listed most. The most frequently observed value.
  - **Variance** (average distance from the average squared):
    - Determine how far each observation is from the average. Square it. Add all squared observations together. Divide by observations -1.
    - Note this is squared as a way to get absolute value; otherwise the value would be zero.
    - Formula:

      $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

  - **Standard Deviation** (average distance from the average):
    - Square root of the variance
    - The empirical rule for normal distributions:
      - 68% are within 1 standard deviation of the mean
      - 95% are within 2 standard deviations of the mean
      - 99.7% are within 3 standard deviations of the mean
    - Formula: $s = \sqrt{s^2}$
  - **Range:** The difference from the largest to the smallest value in a set.
    - Formula: Max - Min

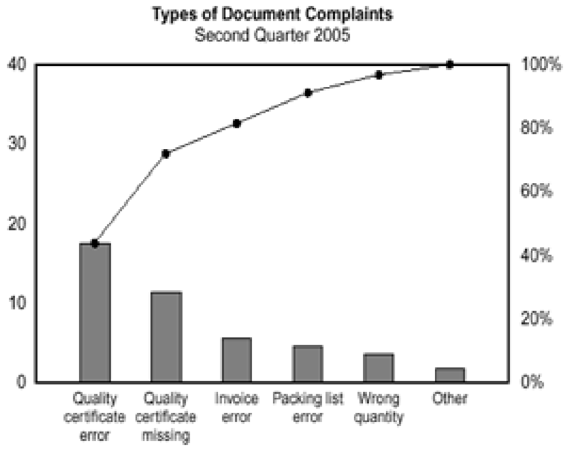
• **Confidence Intervals** (Lean Six Sigma Pocket Toolbook p. 160; www.sigmapedia.com/term.cfm?word_id=180; http://people.hofstra.edu/Stefan_waner/RealWorld/finitetopic1/confint.html): Estimated range of values which includes the true value with stated probability.

- This is driven from the standard deviation of the population. Depending on the sample size the sample standard deviation will be closer to the population standard deviation (see Central Limit Theorem under Green Belt for more details).
  - For sample sizes greater than 30 we can estimate the 95% Confidence Interval for the Population Mean = $\left(\overline{x} - 1.96\frac{s}{\sqrt{n}}, \overline{x} + 1.96\frac{s}{\sqrt{n}}\right)$

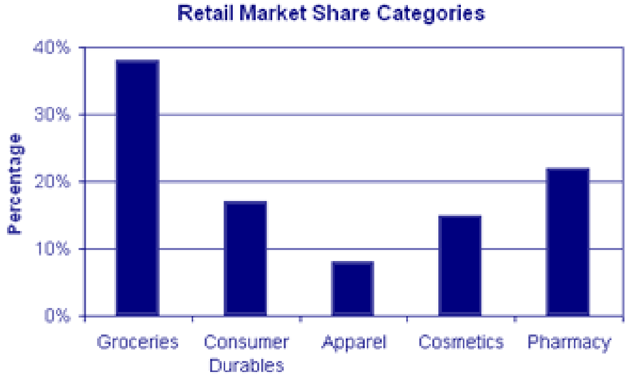
- Example:

- If we can be 95% confident the average IQ in the world is between 95 & 110 then the confidence interval for the average IQ in the world is between 95 & 110 with 95% confidence.

- **Common Charts**
  - **Pareto Chart** (http://en.wikipedia.org/wiki/Pareto_chart)
    - A Pareto chart is a bar chart (see below) ordered from category with the highest value to category with the lowest value. It also shows cumulative values in a line.
    - It is commonly used to show which areas to focus on to generate the most improvement in a Lean Six Sigma project. An example:

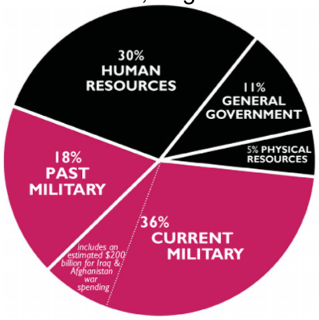


  - **Bar Chart** (http://en.wikipedia.org/wiki/Bar_chart)
    - A **bar chart** or **bar graph** is a chart with rectangular bars with lengths proportional to the values that they represent:
    - It is often used to understand relationships between certain items, to get an understanding of the relative value of the categories.



  - **Pie Chart** (http://en.wikipedia.org/wiki/Pie_chart)
    - A **pie chart** (or a **circle graph**) is a circular chart divided into sectors, illustrating relative magnitudes or frequencies. In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents. Together, the sectors create a full disk.

- It is often used to understand relationships between certain items, to get an understanding of the relative value of the categories:



- **Line Chart** (http://en.wikipedia.org/wiki/Line_chart)
  - A **line chart** or **line graph** is a type of graph created by connecting a series of data points together with a line.
    It is the basic graph we often see showing stock values:

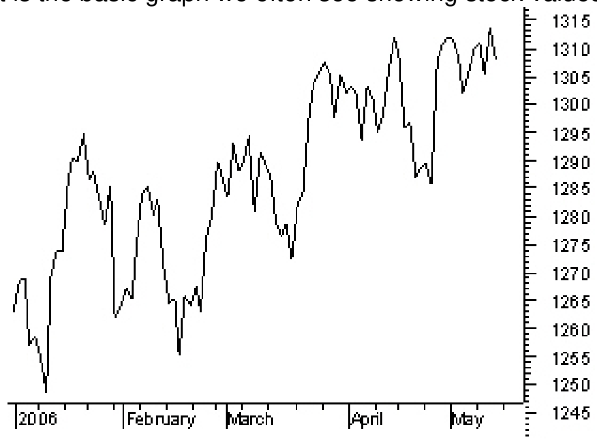
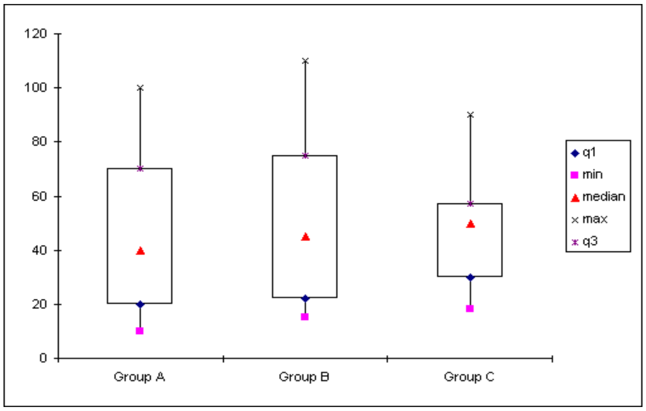


Chart by MetaStock Copyright © 2006 Investopedia.com

• **Box Plots** (The Lean Six Sigma Pocket Toolbook p.110, http://www.coventry.ac.uk/ec/~nhunt/boxplot.htm):

- Box plots (also know as Box & Whisker diagrams) graphically show distributions with:
  - The median
  - A box around the middle 50% of the range of values,
  - Tails showing the bottom 25% of values, and the top 25% of values
- Example:



• **Rolled Throughput Yield:** The number of good units produced divided by the number of total units going into the process. (http://www.isixsigma.com/dictionary/First_Time_Yield_-_FTY-167.htm)

- Algorithm:
  - Calculate the yield (number out of step/number into step) of each step.
  - Multiply these together.
- Example:
  - 100 units enter A; 90 good units are produced. The yield for process A is 90/100 = .9
    90 units go into B and 80 units leave. The yield for process B is 80/90 = .89
    80 units go into C and 75 leave. The yield for C is 75/80 = .94
    75 units got into D and 70 leave. The yield for D is 70/75 = .93
    The Rolled Throughput Yield is equal to .9*.89*.94*.93 = .70.
  - Note: First Pass Yield is the yield for any one step in the process, so the First Pass yield for step D, the last step in the process, is .93

GREEN BELT STATISTICS CHEAT SHEET

Includes formulas: what they are, when to use them, references

Lean Concepts

• **VSM Value Stream Mapping** (http://en.wikipedia.org/wiki/Value_stream_mapping)

- Value Stream Mapping is a tool used to understand a process and how much value-added and non-value added time is spent on and between each activity.
- The VSM will include a data box of key statistics, such as:

Data Box	Current State	Future	% Improvement
Value Add Time			
Non Value Add Time			
Lead Time			
People			
System			
System Touches			
Inventory			

• **TAKT Time** (http://en.wikipedia.org/wiki/Takt_time)

- Often referred to as the rate of customer demand.
- It is how often a product needs to be completed to meet customer demand.
- Formula = Effective Working Time / Average Customer Demand (for that time period)

• **Batch Size** (http://www.shmula.com/270/batch-and-queue-or-single-piece-flow)

- To keep this explanation lean I'll just write that moving to batch sizes of one generally reduces cycle time and improves throughput.

• **SMED Singe Minute Exchange of Die**

- **SMED** stands for Single Minute Exchange of Die, and covers the techniques for obtaining a changeover time of less than 10 minutes (a single digit number of minutes).
- Basically, the SMED methodology consists of **6 steps**:
    - observe the current changeover process
    - identify internal and external activities
    - convert activities from internal to external setup
    - increase efficiency of the remaining internal activities
    - optimize the Startup time
    - increase efficiency of external activities

• **Theory of Constraints** (http://en.wikipedia.org/wiki/Theory_of_constraints)

- The underlying assumption of Theory of Constraints is that organizations can be measured and controlled by variations on three measures: Throughput, Operating Expense, and Inventory. Throughput is money (or goal units) generated through sales. Operating Expense is money that goes into the system to ensure its operation on an ongoing basis. Inventory is money the system invests in order to sell its goods and services.
- Theory of Constraints is based on the premise that the rate of goal achievement is limited by at least one constraining process. Only by increasing flow through the constraint can overall throughput be increased.- Assuming the goal of the organization has been articulated (e.g., "Make money now and in the future") the steps are:
  1. IDENTIFY the constraint (the resource/policy that prevents the organization from obtaining more of the goal)
  2. Decide how to EXPLOIT the constraint (make sure the constraint's time is not wasted doing things that it should not do)

3. SUBORDINATE all other processes to above decision (align the whole system/organization to support the decision made above)
4. ELEVATE the constraint (if required/possible, permanently increase capacity of the constraint; "buy more")
5. If, as a result of these steps, the constraint has moved, return to Step 1. Don't let inertia become the constraint.

• **TPM Total Productive Maintenance** (http://en.wikipedia.org/wiki/Total_Productive_Maintenance)

- TPM is a program for planning and achieving minimal machine downtime
- Equipment and tools are literally put on "proactive" maintenance schedules to keep them running efficiently and with greatly reduced downtime
- Machine operators take far greater responsibility for their machine upkeep
- Maintenance technicians are liberated from mundane, routine maintenance, enabling them to focus on urgent repairs and proactive maintenance activities
- A solid TPM program allows you to plan your downtime and keep breakdowns to a minimum
- Without a strong TPM program, becoming truly Lean would be difficult or impossible in an environment heavily dependent on machinery
- Buy-in at the shop floor level is generally quite high as TPM is an exciting undertaking
- A robust TPM system consists of:
  - Autonomous Maintenance
  - Focused Improvement
  - Education and Training
  - Planned Maintenance
  - Quality Maintenance
  - Early Management and Initial Flow Control
  - Safety, Hygiene and Pollution Control
  - Administrative and Office TPM
- The metric used in Total Productive Maintenance environments is called **OEE** or **Overall Equipment Effectiveness**
  - OOE is measured as a percentage
  - OOE = Availability * Performance * Quality
    - Availability = % of scheduled production equipment is available for production
    - Performance = % number of parts produced out of best known production rate
    - Quality = % of good sellable parts out of total parts produced

Sampling

• **Sampling** (http://www.statpac.com/surveys/sampling.htm)

- **Sample Size Calculator** (http://www.surveysystem.com/sscalc.htm , http://edis.ifas.ufl.edu/PD006)
  - To determine how large a sample you need to come to a conclusion about a population at a given confidence level.
  - Formula for creating a sample size to test a proportion:
    - $n_o = \{Z^2pq\}/\{e^2\}$
      - n_o = required sample size
      - Z = value from the Z table (found on-line or in stats books) for confidence level desired
      - p = estimated proportion of population that has the attribute we are testing for
      - q = 1-p
      - e = precision {ie: if we want our proportion to be known within 10% then set 'e' at .05 and if we set the confidence interval at 95% and the sample gives a proportion of 43%, the true value at 95% confidence is between 38% and 48%}
  - Formula for creating a sample size to test a mean:
    - $n_o = \{Z^2\sigma^2\}/\{e^2\}$

        - n_o = required sample size
        - Z = value from the Z table (found on-line or in stats books) for confidence level desired
        - σ = variance of the attribute in the population
        - e = precision in the same unit of measure as the variance
- **Single lot Sampling** (http://www.itl.nist.gov/div898/handbook/pmc/section2/pmc22.htm)
    - Single lot sampling is when your sample comes from a single lot. It is often used in manufacturing when output from a single lot is sampled for testing.
    - This may be used as a **Lot Acceptance Sampling Plan (LASP)** to determine whether or not to accept the lot:
        - **Single sampling plans:** One sample of items is selected at random from a lot and the disposition of the lot is determined from the resulting information. These plans are usually denoted as (*n,c*) plans for a sample size *n,* where the lot is rejected if there are more than *c* defectives. *These are the most common (and easiest) plans to use although not the most efficient in terms of average number of samples needed.*
- **Dual lot Sampling**
    - Dual lot sampling is when your sample comes from 2 different but similar lots. It is often used as part of an MSA or as part of hypothesis testing to determine if there are differences in the lots.
- **Continuous Sampling** (http://www.sqconline.com/csp1_enter.php4)
    - Continuous sampling is used for the inspection of products that are not in batches. The inspection is done on the production line itself, and each inspected item is tagged conforming or non-conforming. This procedure can also be employed to a sequence of batches, rather than to a sequence of items (known as **Skip Lot Sampling**).
- **Stratified Sampling** (http://www.coventry.ac.uk/ec/~nhunt/meths/strati.html , http://www.answers.com/topic/stratified-sampling-1)
    - Stratified Sampling is when the population is dived into non-overlapping subgroups or strata and a random sample is taken from each subgroup. It is often used in hypothesis testing to determine differences in subgroups.
- **Random Sampling** (http://www.stats.gla.ac.uk/steps/glossary/sampling.html)
    - Random sampling is a sampling technique where we select a group of subjects (a sample) for study from a larger group (a population). Each individual is chosen entirely by chance and each member of the population has a known, but possibly non-equal, chance of being included in the sample.

MSA

- **MSA Measurement System Analysis** (http://en.wikipedia.org/wiki/Measurement_Systems_Analysis)
    - A **Measurement System Analysis**, abbreviated **MSA**, is a specially designed experiment that seeks to identify the components of variation in the measurement.
    - Since analysis of data is key to lean six sigma ensuring your data is accurate is critical. That's what MSA does – it tests the measurements used to collect your data.
    - Common tools and techniques of Measurement Systems Analysis include: calibration studies, fixed effect ANOVA, components of variance, Attribute Gage Study, Gage R&R, **ANOVA Gage R&R**, Destructive Testing Analysis and others. The tool selected is usually determined by characteristics of the measurement system itself.
        - Gage R & R (http://en.wikipedia.org/wiki/ANOVA_Gage_R%26R , http://www.sixsigmaspc.com/dictionary/RandR-repeatability-reproducibility.html)
            - ANOVA Gauge R&R measures the amount of variability induced in measurements that comes from the measurement system itself and compares it to the total variability observed to determine the viability of the measurement system.
            - There are two important aspects on a Gauge R&R:

- **Repeatability**, Repeatability is the variation in measurements taken by a single person or instrument on the same item and under the same conditions.
- **Reproducibility**, the variability induced by the operators. It is the variation induced when different operators (or different laboratories) measure the same part.

- **Formulas (this is best done using a tool such as Minitab or JMP):**
  - $yijk = \mu$□$+ \alpha i + \beta j + \alpha\beta ij + \varepsilon ijk$
    - Y_{ijk} = observation k with part i& operator j
    - μ = population mean
    - α_i = adjustment for part i
    - β_j = adjustment for operator j
    - $\alpha\beta_{ij}$ = adjustment for part i operator j interaction
    - ε_{ijk} = observation random 'error'
  - $\sigma^2_y = \sigma^2_i + \sigma^2_j + \sigma^2_{ij} + \sigma^2_{error}$
    - σ^2_y = Variance of observation y
    - σ^2_i = Variance due to part i
    - σ^2_j = Variance due to operator j
    - σ^2_{ij} = Variance due to part I operator j interaction
    - σ^2_{error} = Variance of that observation due to random 'error'
- These formulas are used in the ANOVA Gage R &R to determine repeatability & reproducibility.

• **Kappa MSA**

- MSA analysis for discrete or attribute data.
  - Kappa (K) is defined as the proportion of agreement between raters after agreement by chance has been removed.
  - $K = \{P_{observed} – P_{chance}\}/\{1 – P_{chance}\}$
    - $P_{observed}$ = proportion of units classified in which the raters agreed
    - P_{chance} = proportion of units for which one would expect agreement by chance
  - Generally a K < .7 indicates measurement system needs improvement
  - K > .9 are considered excellent

Data Analysis

• **Statistics Error Types** (http://en.wikipedia.org/wiki/Type_I_error#Type_I_error)

- **Type 1, Alpha or α errors**
  - **Type I error**, also known as an "**error of the first kind**", an **α error**, or a "**false positive**": the error of rejecting a null hypothesis when it is actually true. Plainly speaking, it occurs when we are observing a difference when in truth there is none. An example of this would be if a test shows that a woman is pregnant when in reality she is not. Type I error can be viewed as the error of excessive credulity.
- **Type 2, Beta or β errors**
  - **Type II error**, also known as an "**error of the second kind**", a **β error**, or a "**false negative**": the error of failing to reject a null hypothesis when it is in fact not true. In other words, this is the error of failing to observe a difference when in truth there is one. An example of this would be if a test shows that a woman is not pregnant when in reality she is. Type II error can be viewed as the error of excessive skepticism.

• **Hypothesis Testing** (The Lean Six Sigma Pocket Toolbook p 156; The Six Sigma Memory Jogger II p 142)

When to use what test: (The Six Sigma Memory Jogger II p 144)

- If comparing a group to a specific value use a **1-sample t-test** (The Lean Six Sigma Pocket Toolbook p 162)

- Tells us if a statistical parameter (average, standard deviation, etc.) is different from a value of interest.
- Hypothesis takes the form Ho: µ = a target or known value
- This is best calculated using a template or software package. If needed the formula can be found in the reference.

- If comparing 2 independent group averages use a **2-sample t-test** (The Lean Six Sigma Pocket Toolbook p 163)
  - Used to determine if the means of 2 samples are the same.
  - Hypothesis takes the form Ho: $\mu_1 = \mu_2$

- If comparing 2 group averages with matched data use **Paired t-test** (http://www.ruf.rice.edu/~bioslabs/tools/stats/pairedttest.html)
  - The number of points in each data set must be the same, and they must be organized in pairs, in which there is a definite relationship between each pair of data points
  - If the data were taken as random samples, you must use the independent test even if the number of data points in each set is the same
  - Even if data are related in pairs, sometimes the paired t is still inappropriate
  - Here's a simple rule to determine if the paired t must not be used - if a given data point in group one could be paired with any data point in group two, you cannot use a paired t test
  - Hypothesis takes the form Ho: $\mu_1 = \mu_2$

- If comparing multiple groups use **ANOVA** (The Lean Six Sigma Pocket Toolbook p 173)
  - Hypothesis takes the form Ho: $\mu_1 = \mu_2 = \mu_3 = \dots$
  - See Black Belt section

The smaller the p-value the more likely the groups are different.

• **Pearson Correlation Co-efficient** (http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)

- In statistics, the **Pearson product-moment correlation coefficient** (sometimes referred to as the **PMCC**, and typically denoted by *r*) is a measure of the correlation (linear dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables.
  - Remember Pearson measures correlation not causation.
- A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear relationship between the variables.
- The statistic is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient can be calculated as:
  - r = {1/(n-1)}∑ [(Xi – X(avg))/(Sx)][(Yi-Y(avg))/(Sy)]
  - where
    - n = sample size
    - Xi = the value of observation I in the X plane
    - X(avg) = the average X value
    - Sx = the standard deviation of X
    - Yi = the value of observation i in the Y plane
    - Y(avg) = the average Y value
    - Sy = the standard deviation of Y

• **Central Limit Theorem** (http://en.wikipedia.org/wiki/Central_limit_theorem)

- In probability theory, the **central limit theorem** (**CLT**) states conditions under which the sum of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed.
    - Let $X_1, X_2, X_3, ..., X_n$ be a sequence of n independent and identically distributed (i.i.d) random variables each having finite values of expectation μ and variance $\sigma^2 > 0$. The central limit theorem states that as the sample size n increases, [1] [2] the distribution of the sample average of these random variables approaches the normal distribution with a mean μ and variance σ^2 / n irrespective of the shape of the original distribution.
- By using the central limit theorem we can apply tools that require a normal distribution even when the distribution of the population is non-normal. However, be careful when interpreting results if you use the CLT to analyze averages rather than samples from the direct population. Remember your analysis is based on averages which can be dangerous. In many cases it's safer to find a tool that doesn't require normally distributed data when analyzing non-normal data.
- Applying CLT to data analysis:
    - Take 10 data points from the distribution & average them. Then take 29 other samples of 10 data points averaged. Then use these 30 data averaged data points to do your analysis. This converts your data from its original distribution to a normal distribution.
    - As long as n (sample size) is large enough, the sampling distribution will be normal and the mean will be representative of the population mean μ.
    - **No matter what the parent looks like the child will look reasonably normal by 30!**
    - **Formulas:**
        - Sampling distribution for the mean has $SD = \frac{\sigma}{\sqrt{n}}$
        - For variables data, applying the 95% rule we expect μ to lie in the interval:
            - 95% Confidence Interval for the Population Mean = $\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$
            - Sample size: h (maximum error allowed) = $2\frac{s}{\sqrt{n}}$
            - $n = \left\lceil \left(\frac{2s}{h}\right)^2 \right\rceil$
        - Attribute data:
            - 95% Confidence Interval for the Population Proportion = $p - 2\sqrt{\frac{p(1-p)}{n}}, p + 2\sqrt{\frac{p(1-p)}{n}}$
            - Sample size: $n = \left\lceil \frac{4p(1-p)}{h^2} \right\rceil$

• FMEA Failure Mode and Effects Analysis
(http://en.wikipedia.org/wiki/Failure_mode_and_effects_analysis)

- This tool helps determine where to focus improvement efforts by analyzing severity of failures, probability of occurrence of an error, and likelihood of detection of an error.
- An RPN (Risk Priority Number) is computed by multiplying these factors together. Processes/Steps with the highest RPN should be addressed first.

Process Control

• **Attribute vs. Variable Data** (Black Belt memory jogger p 34, Green Belt memory jogger p78, http://www.sixsigmaspc.com/dictionary/discreteattributedata.html)

- Different Tools are best able to handle different types of date (for example control charts).
- For six sigma data analysis Discrete or Attribute data is considered the same thing. Discrete data is any data not quantified on an infinitely divisible numerical scale. Discrete data has boundaries and includes any number that must be an integer. Discrete examples include day of the week; hour of the day; age in years; counts; income; an non-numeric translated to a numeric value (i.e.: good/bad, on-time/not on-time, pass/fail, primary colors, eye color, grade, method).
- Variable or Continuous data is any data on a continuous scale. Examples include length, diameter, temperature, weight, time.

• **Control Charts** (http://www.statsoft.com/textbook/stquacon.html, http://www.itl.nist.gov/div898/handbook/pmc/pmc.htm , The Black Belt Memory Jogger p. 221, For non-normal distributions: http://www.ct-yankee.com/spc/nonnormal.html)

- A Control Chart is simply a run chart with statistically based limits.
- Within the subject of control charts Voice of the Customer (VOC) is the customer required specifications or the Upper Spec Limit (USL) and Lower Spec Limit (LSL).
- The Voice of the Process is the Upper Control Limit (UCL), and Lower Control Limit (LCL). It is 3 standard deviations from the process mean and is what the process will deliver (99.7% of the time).
- When to use what chart:
    - Variable Individual data: use **XmR** or **I** or **mR** (use moving range to measure control)
        - **XmR** or **I**
            - Called XmR since we use the Moving Range relative to X
            - Also called I since it is based on Individual data
            - This chart will work for virtually any situation
            - K = # of subgroups
            - Avg(X) = ∑X/k
            - $R_m = | (X_{i+1} – X_i) |$
            - Avg (R_m) = ∑R/(k-1)
            - E_2 is based on sample size & is in a table in the reference
            - UCL = avg(X) +E_2 * Avg (R_m)
            - LCL = avg(X) +E_2 * Avg (R_m)
        - **mR**
            - K = # of subgroups
            - Avg(X) = ∑X/k
            - $R_m = | (X_{i+1} – X_i) |$
            - Avg (R_m) = ∑R/(k-1)
            - D_3 & D_4 are based on sample size & is in a table in the reference
            - UCL = D_4 * Avg (R_m)
            - LCL = D_3 * Avg (R_m)
    - Variable data of group sizes 2-10: use **Xbar** & **R** (use range to measure control)
        - **Xbar**
            - K = # of subgroups
            - Avg(X) = Avg of X for subgroup k
            - Avg(Avg(X)) = ∑Avg(X)/k
            - Avg (R) = ∑R/k
            - A_2 is based on sample size & is in a table in the reference
            - UCL = Avg(Avg(X)) + A_2 * Avg(R)
            - LCL = Avg(Avg(X)) – A_2 * Avg(R)
        - **R**
            - K = # of subgroups
            - Avg (R) = ∑R/k
            - D_4 & D_3 are based on sample size & is in a table in the reference

- UCL = D_4 * Avg (R)
- LCL = D_3 * Avg (R)

- Variable data of group sizes 10 or more: use **Xbar** & **S** (use standard deviation to measure control)
    - **Xbar**
        - K = # of subgroups
        - Avg(X) = Avg of X for subgroup k
        - Avg(Avg(X)) = ∑Avg(X)/k
        - Avg (s) = ∑s/k
        - A_3 is based on sample size & is in a table in the reference
        - UCL = Avg(Avg(X)) + A_3 * Avg(s)
        - LCL = Avg(Avg(X)) – A_3 * Avg(s)
    - **S**
        - K = # of subgroups
        - Avg (s) = ∑s/k
        - B_3 & B_4 are based on sample size & is in a table in the reference
        - UCL = B_4 * Avg(s)
        - LCL = B_3 * Avg(s)
- Note: Defect data is a failure to meet one of the acceptance criteria; Defective data is when an entire unit fails to meet acceptance criteria. A defective unit may have multiple defects.
- Attribute Defect Data Constant sample size: use **C**
    - **C**
        - c = number of defects for each subgroup
        - k = # of subgroups
        - avg (c) = ∑c/k
        - UCL = avg (c) + 3√avg (c)
        - LCL = avg (c) - 3√avg (c)
- Attribute Defect Data Variable sample size : use **U**
    - **U**
        - u = c/n for each subgroup
        - avg (u) = ∑c/∑n
        - UCL = avg (u) + 3√(avg (u)/n)
        - LCL = avg (u) + 3√(avg (u)/n)
- Attribute Defective Data Constant sample size: use **nP** for number defective
    - **nP**
        - most reliable when sample size greater than 50
        - np = # defective units for each subgroup
        - k = # of subgroups
        - avg (np) = ∑np/k
        - UCL = avg (np) + 3√[avg(np) * {1- avg(p)}]
        - LCL = avg (np) - 3√[avg(np) * {1- avg(p)}]
- Attribute Defective Data Variable sample size: use **p** for fraction defective
    - **P**
        - most reliable when sample size greater than 50
        - p = np/n = % defective units for each subgroup
        - avg(p) = ∑np/∑n
        - UCL = avg (p) + 3√([avg(np) * {1- avg(p)}]/n)
        - LCL = avg (p) - 3√([avg(np) * {1- avg(p)}]/n)
- Process is known to be in control & want to examine small shifts: use **CUSUM** or **EWMA.**
    - **CUSUM (CUmulative SUM chart) (http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc323.htm , http://www.variation.com/cpa/help/hs108.htm)**
        - The *Cusum chart* can detect process shifts more rapidly than the XmR or XbarR. If a trend develops, it's evidence that the process has shifted.
        - To compute:

        - n = number of data points
        - Avg(x) = ∑X/n
        - $S_{(i)}$ = Cumulative Sum of the difference between the current value and the average. $S_i = S_{i-1} + (X_i - \overline{X})$
        - The cumulative sum is not the cumulative sum of the values. Instead it is the cumulative sum of differences between the values and the average. Because the average is subtracted from each value, the cumulative sum also ends at zero.
        - If $S_{(i)}$ is positive or negative for more than a few points in a row it is an indication there may be a change in the process.
    - **EWMA (Exponentially Weighted Moving Average) (http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm)**
        - The Exponentially Weighted Moving Average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time.
        - $EWMA_t = \lambda Y_t + (1 - \lambda)\, EWMA_{t-1}$ for t = 1, 2, ..., *n*.
            - $EWMA_0$ is the mean of historical data (target)
            - Y_t is the observation at time *t*
            - *n* is the number of observations to be monitored including $EWMA_0$
            - $0 < \lambda \le 1$ is a constant that determines the depth of memory of the EWMA.
            - The parameter λ determines the rate at which 'older' data enter into the calculation of the EWMA statistic. A value of $\lambda = 1$ implies that only the most recent measurement influences the EWMA (degrades to Shewhart chart). Thus, a large value of $\lambda = 1$ gives more weight to recent data and less weight to older data; a small value of λ gives more weight to older data. The value of λ is usually set between 0.2 and 0.3 (Hunter) although this choice is somewhat arbitrary. Lucas and Saccucci (1990) give tables that help the user select λ.
        - $s^2_{ewma} = (\lambda/(2-\lambda))\, s^2$
        - The center line for the control chart is the target value or $EWMA_0$. The control limits are:
            - $UCL = EWMA_0 + k s_{ewma}$
            - $LCL = EWMA_0 - k s_{ewma}$
            - where the factor *k* is either set equal 3 or chosen using the Lucas and Saccucci (1990) tables. The data are assumed to be independent and these tables also assume a normal population.
- Out of Control Conditions:
    - If 1 or more points falls outside of the UCL (Upper Control Limit) or LCL (Lower Control Limit). (i.e.: any point more than 3 standard deviation from the mean).
    - If 2 of 3 successive points fall beyond 2 standard deviations from the mean.
    - If 4 of 5 successive points fall beyond 1 standard deviations from the mean.
    - If there is a run of 6 or more points that are either successively higher or successively lower.
    - If 8 or more successive points fall on either side of the mean.
    - If 15 points in a row fall within 1 standard deviation from the mean. (Examine to determine if the process has become better controlled).

- **VOC Voice Of the Customer** (http://en.wikipedia.org/wiki/Voice_of_the_customer)

- Every lean Six Sigma project needs to understand the Voice of the Customer. You can not truly exit define phase without an understanding of what's important to the customer.
- In Lean only process that add value to the customer are value-add processes. You need to understand the customer (i.e.: what would the end-customer pay for) to understand if a step is technically a value-added step in a Lean analysis.
- In control charts the VOC is translated to Upper Spec Limit (USL) & Lower Spec Limit (LSL) (see control charts above).

• **Control Limits**

- See control charts above. Control Limits are the 'voice of the process' and based on the standard deviation and variability of the process.

• **Process Capability** (The Black Belt Memory Jogger p 95)

- Measures the capability of a process to meet customer specifications.
- The higher the Cp, Cpk, Pp, or Ppk, the better, the less variance there is in the process.
- Short-Term Capability indices:
    - Used if long-term data is not available or if process has changed recently.
    - Uses short term process variation to determine process capability
    - Cp tells you if the process is capable: If Cp > 1 then the process is capable (the process spread fits within tolerance). While Cpk also tells you if the process is centered; if Cpk < 1 while Cp > 1 then the process is not centered and should first be centered.
    - Cp
        - The ratio of customer-specified tolerance to 6 standard deviations of the short-term process variation.
        - Gives an indication of what the process could perform to if the mean of the data was centred between the specification limits.
        - Sometimes referred to as process potential

      • Formula: Cp = $\frac{USL-LSL}{6s_{MR}}$
    - Cpk
        - The ratio of the distance between the process average and the closest specification limit to 3 standard deviations of short term process variation.
        - More realistic measure of process than C_p because it uses data mean.

      • Formula: Cpk = min $\left\{\frac{USL-\bar{x}}{3s_{MR}}, \frac{\bar{x}-LSL}{3s_{MR}}\right\}$

- Long Term Capability indices:
    - Used if long-term data available & representative of the current process.
    - Uses long-term process standard variation
    - Pp
        - The ratio of customer-specified tolerance to 6 standard deviations of the long-term process variation.
        - Gives an indication of what the process could perform to if the mean of the data was centred between the specification limits.
        - Sometimes referred to as process potential

      • Formula: Pp = $\frac{USL-LSL}{6s}$
    - Ppk
        - The ratio of the distance between the process average and the closest specification limit to 3 standard deviations of long term process variation.
        - More realistic measure of process than P_p because it uses data mean.

- Formula: Ppk = min $\left\{ \frac{USL - \bar{x}}{3s}, \frac{\bar{x} - LSL}{3s} \right\}$

- **Control plan** (http://it.toolbox.com/wiki/index.php/Control_Plan)
  - The control plan is the plan used to control the process. Within six sigma it is used in the control phase & after project closure to ensure project improvements are sustained.
  - There a various control plan templates, but what needs to be captured is what is being measured; how it is used, what sampling method is used, who owns the control chart, where it is located, and what conditions constitute loss of control or improvement that require corrective action.
  - Usually Cp, Cpk, Pp, and Ppk (see process capability above) are measured as part of the control plan, and often process control charts (see control charts above) are the key control chart used.

BLACK BELT STATISTICS CHEAT SHEET

Includes formulas: what they are, when to use them, references

- **ANOVA** (http://en.wikipedia.org/wiki/ANOVA , The Lean Six Sigma Pocket Toolbook p 173)
  - Used for hypothesis testing when comparing multiple groups.
    - Hypothesis takes the form H_0: $\mu_1 = \mu_2 = \mu_3 = \dots$
  - In its simplest form ANOVA gives a statistical test of whether the means of several groups are all equal, and therefore generalizes Student's two-sample *t*-test to more than two groups.
  - It is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables.

- **DOE Design Of Experiments** (http://en.wikipedia.org/wiki/Design_of_experiments)
  - Design of Experiments is using statistical tools (such as ANOVA above and regression below) to be able to determine the importance of different factors with a minimal amount of data. It is used when you have many different factors that may impact results (i.e.: many x's that impact Y in the classic Y=f(x) formula).
  - By collecting the data, organizing it, and analyzing it using DoE methodology you do not have to study **One Factor At a Time (OFAT)** {http://en.wikipedia.org/wiki/One-factor-at-a-time_method } to isolate which factors have the biggest impact.
  - DoE's are best carried out using software specifically designed for DoE (such as Minitab or JMP).
  - Most important ideas of experimental design:
    - **Comparison**
      - In many fields of study it is hard to reproduce measured results exactly. Comparisons between treatments are much more reproducible and are usually preferable. Often one compares against a standard or traditional treatment that acts as baseline.
    - **Randomization**
      - There is an extensive body of mathematical theory that explores the consequences of making the allocation of units to treatments by means of some random mechanism such as tables of random numbers, or the use of randomization devices such as playing cards or dice. Provided the sample size is adequate, the risks associated with random allocation (such as failing to obtain a representative sample in a survey, or having a serious imbalance in a key characteristic between a treatment group and a control group) are calculable and hence can be managed down to an acceptable level. Random does *not* mean haphazard, and great care must be taken that appropriate random methods are used.
    - **Replication**
      - Measurements are usually subject to variation, both between repeated measurements and between replicated items or processes. Multiple measurements of replicated items are necessary so the variation can be estimated.
    - **Blocking**
      - Blocking is the arrangement of experimental units into groups (blocks) that are similar to one another. Blocking reduces known but irrelevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study.
    - **Orthogonality**
      - Orthogonality concerns the forms of comparison (contrasts) that can be legitimately and efficiently carried out. Contrasts can be represented by vectors and sets of orthogonal contrasts are uncorrelated and independently distributed if the data are normal. Because of this independence, each orthogonal treatment provides different information to the others. If there are *T* treatments and *T* – 1 orthogonal contrasts, all the information that can be captured from the experiment is obtainable from the set of contrasts.

    - **Factorial experiments**
        - Use of factorial experiments instead of the one-factor-at-a-time method. These are efficient at evaluating the effects and possible interactions of several factors (independent variables).
- **Step-by-step procedure** in effective design of an experiment. (Note this is taken from the link above; not every step may be needed for your experiment. Software packages often have tutorials showing how to do a DoE specifically with their application.)
    - Select the problem
    - Determine dependent variable(s)
    - Determine independent variables
    - Determine number of levels of independent variables
    - Determine possible combinations
    - Determine number of observations
    - Redesign
    - Randomize
    - Meet ethical & legal requirements
    - Develop Mathematical Model
    - Collect Data
    - Reduce Data
    - Verify Data

- **Regression**
    - **Linear Regression** (http://en.wikipedia.org/wiki/Linear_regression)
        - Linear regression attempts to use a straight line to determine a formula for a variable (y) from one or more factors (Xs).
        - This is best done using a software package such as excel, Minitab, or JMP.
        - Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:
            - If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.
            - If we have a variable *y* and a number of variables X_1, ..., X_p that may be related to *y*, we can use linear regression analysis to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*, so that once one of them is known, the others are no longer informative.
        - Linear regression models are often fit using the least squares approach, but may also be fit in other ways, such as by minimizing the "lack of fit" in some other norm, or by minimizing a penalized version of the least squares loss function as in ridge regression. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, while the terms "least squares" and *linear model* are closely linked, they are not synonymous.
    - **Non-linear Regression** (http://en.wikipedia.org/wiki/Non-linear_regression)
        - Non-linear regression attempts to determine a formula for a variable (y) from one or more factors (Xs), but it differs from linear regression because it allows the relationship to be something other than a straight line.
        - This is best done using a software package such as an excel add-on, Minitab, or JMP.
        - examples of nonlinear functions include exponential functions, logarithmic functions, trigonometric functions, power functions, Gaussian function, and Lorentzian curves. Some functions, such as the exponential or logarithmic functions, can be transformed so that they are linear. When so transformed, standard linear regression can be performed but must be applied with caution:

    - Some nonlinear regression problems can be moved to a linear domain by a suitable transformation of the model formulation.
        - For example, consider the nonlinear regression problem (ignoring the error): $y = ae^{bx}$.
        - If we take a logarithm of both sides, it becomes: $\ln (y) = \ln (a) + bx$.
        - Therefore, estimation of the unknown parameters by a linear regression of ln(*y*) on *x*, a computation that does not require iterative optimization. However, use of a linear transformation requires caution. The influences of the data values will change, as will the error structure of the model and the interpretation of any inferential results. These may not be desired effects. On the other hand, depending on what the largest source of error is, a linear transformation may distribute your errors in a normal fashion, so the choice to perform a linear transformation must be informed by modeling considerations.
- In general, there is no closed-form expression for the best-fitting parameters, as there is in linear regression. Usually numerical optimization algorithms are applied to determine the best-fitting parameters.
- The best-fit curve is often assumed to be that which minimizes the sum of squared residuals. This is the **(ordinary) least squares (OLS)** approach. However, in cases where the dependent variable does not have constant variance a sum of weighted squared residuals may be minimized; see weighted least squares. Each weight should ideally be equal to the reciprocal of the variance of the observation, but weights may be recomputed on each iteration, in an iteratively weighted least squares algorithm.

• **Non-normal distributions**

- There are as many different distributions as there are populations. Deming said 'there a no perfect models, but there are useful models'. That's the question you need to ask relative to your problem: What tools & techniques will work well with the distribution for the population I have?
- Cycle time data can not go below zero, and therefore is never truly normal. It invariably has a longer tail to the right than to the left.
- Often times when a population has non-normal data it can be stratified into segments that have approximately normal distributions. And you can assume normality after you have determined these subpopulations and pulled data from the subpopulations to test separately.
- When doing data analysis I would first determine if the data is normal. If it is not:
    - Consider how important the distribution is to the tools you plan to use. Of the tools in this cheat sheet these ones are affected if the distribution is non-normal:
        - **Confidence Intervals**: concept is the same, but you can not use the 1.96 as the multiplier for a 95% CI.
        - **Gage R&R** (http://elsmar.com/Forums/showthread.php?t=8597): In most real-world applications the impact of non-normal data on this tool is negligible.
        - **T-test** (http://www.qimacros.com/qiwizard/tukey-quick-test-excel.html): t-test assumes normality. Consider:
            - If data is close to normal you may be safe using a t-test.
            - Depending on other factors you may be able to use the Central Limit Theorem and normalize the data.
            - You may prefer a hypothesis test that doesn't assume normality such as the Tukey Test or Moods Median Test.
        - **ANOVA** (http://lssacademy.com/2007/06/15/dealing-with-non-normal-data/ http://www-rohan.sdsu.edu/%7Ecdlin/677/ANOVA_Assumptions.ppt): ANOVA assumes normality. Consider:
            - If data is close to normal you may be safe using ANOVA.
            - Depending on other factors you may be able to use the Central Limit Theorem and normalize the data.

  - You may prefer a hypothesis test that doesn't assume normality such as Levene's Test or Brown-Forsythe Test or the F_{max} Test or the Friedman Test..
- **Pearson Correlation Co-efficient** (http://en.wikipedia.org/wiki/Correlation)
  - Pearson is usually considered accurate for even non-normal data, but there are other tools that are specifically designed to handle outliers. These correlation coefficients generally perform worse than Pearson's if no outliers are present. But in the event of many extreme outliers consider Chi-square, Point biserial correlation, Spearman's p, Kendall's T, or Goodman and Kruskal's lambda.
- **Central Limit Theorem** can be used to transform data from a non-normal to a normal distribution.
- **Control Charts, Control Limits, Process Capability,** and **Control Plans** (http://www.ct-yankee.com/spc/nonnormal.html). Standard Control Charts, Control Limits, and Process Capability metrics assume normality. Consider:
  - If data is close to normal you may be safe using a standard control chart.
  - Depending on other factors you may be able to use the Central Limit Theorem and normalize the data.
  - Your statistical package may allow for a test based on the distribution you have. For example Minitab supports Weibull distribution and will compute the capability six pack including the control charts for a Weibull distribution.
  - You may use the transformation formulas to use a control chart designed for non-normal data.
    - UCL = .99865 quantile
    - Center line = median
    - LCL = .00135 quantile
    - This revision does affect out of control conditions as well as process capability measurements.
- **Design Of Experiments**
  - DOE assumes normal data.
  - If data is close to normal you may be safe using DOE.
  - Depending on other factors you may be able to use the Central Limit Theorem and normalize the data.
- **Regression** (http://www.statsci.org/glm/intro.html)
  - **Generalized linear models (GLMs)** (http://en.wikipedia.org/wiki/Generalized_linear_model)are used to do regression modelling for non-normal data with a minimum of extra complication compared with normal linear regression.
    - The GLM consists of three elements.
1. A distribution function *f*, from the exponential family.
2. A linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$.
3. A link function *g* such that $E(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\eta)$.
    - The exact formulas depend on the underlying distribution.

- Some of the more common non-normal distributions include (http://www.quantitativeskills.com/sisa/rojo/distribs.htm):
  - **Weibull, Exponential, Log-normal**. Minitab's distribution identity function can test for these distributions. NOTE: Weibull can be normal if 1=λ & k=5.
  - **Gamma, Poisson, Chi-squared, Beta, Bi-modal, Binomial, Student-t**. (NOTE: Student-t is very close to normal but has a longer tail).
- Some other distributions include Laplace, Logistic, Multinomial, Negative Binomial, Erlang, Maxwell-Boltzmann, Inverse-gamma, Dirichlet, Wishart, Cauchy, Snedecor F, Uniform, Bernoulli, Geometric, Hypergeometric, Triangular, Rectangular.

• **Variance inflation factor (VIF)** (http://en.wikipedia.org/wiki/Variance_inflation_factor)

- The VIF measures how much the interaction between independent variables impact the dependent variable. (I.e. going back to the Y = f(x) equation how much do the different x's interact with each other to determine Y.)
  - Consider the following regression equation with k independent variables:
    - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$
    - VIF can be calculated in three steps:
      - Calculate k different VIFs, one for each X_i by first running an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation.
        - If $i = 1$, for example, the equation would be $X_1 = \alpha_2 X_2 + \ldots + \alpha_k X_k + c_0 + \varepsilon$ where c_0 is a constant and e is the error term.
      - Then, calculate the VIF factor for β^ with the following formula:
        - $\mathrm{VIF}(\hat{\beta}_i) = [1/(1-R^2_i)]$
        - where R^2_i is the coefficient of determination of the regression equation in step one.
      - Then, Analyze the magnitude of multicollinearity by considering the size of the $\mathrm{VIF}(\hat{\beta}_i)$.
      - A common rule of thumb is that if $\mathrm{VIF}(\hat{\beta}_i) > 5$ then multicollinearity is high. Some software calculates the tolerance which is just the reciprocal of the VIF. The choice of which formula to use is mostly a personal preference of the researcher.
- The square root of the variance inflation factor tells you how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other independent variables in the equation.

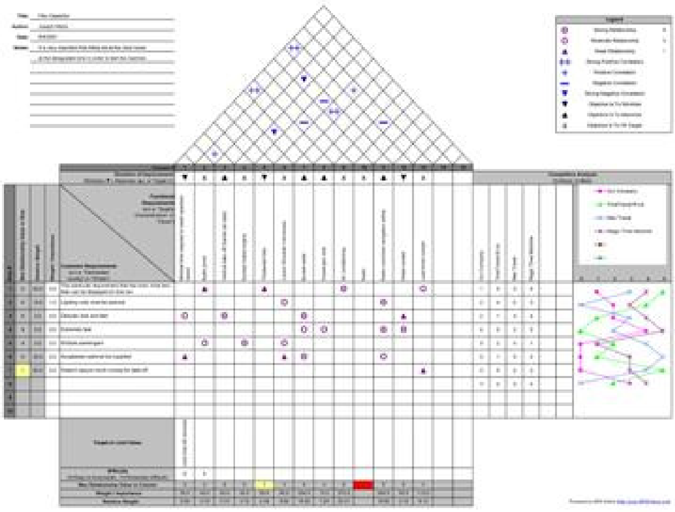
• **Life testing & reliability** (http://en.wikipedia.org/wiki/Reliability_%28engineering%29 ,p 24-21 Quality Control Handbook 3rd edition)

- **AQL** (http://en.wikipedia.org/wiki/Acceptable_quality_limit)
  - The **acceptable quality limit (AQL)** is the worst-case quality level, in percentage or ratio, that is still considered acceptable.
  - In a quality control procedure, a process is said to be at an acceptable quality level if the appropriate statistic used to construct a control chart does not fall outside the bounds of the acceptable quality limits. Otherwise, the process is said to be at a rejectable control level.
- **AOQL** (http://www.variation.com/spa/help/hs103.htm)
  - **The Average Outgoing Quality Limit (AOQL)** of a sampling plan is maximum value on the AOQ curve. It is applicable for defective units, defects per unit, and defects per quantity. It is expressed as either a defective rate (fraction defective, percent defective, dpm) or as a defect rate (defects per unit, defects per 100 units, dpm). The AOQ curve gives the average outgoing quality (left axis) as a function of the incoming quality (bottom axis). The AOQL is the maximum or worst possible defective or defect rate for the average outgoing quality. Regardless of the incoming quality, the defective or defect rate going to the customer should be no greater than the AOQL over an extended period of time. Individual lots might be worst than the AOQL but over the long run, the quality should not be worse than the AOQL.
  - The AOQ curve and AOQL assume rejected lots are 100% inspected, and is only applicable to this situation. They also assume the inspection is reasonably effective at removing defectives or defects (90% effective or more).

• **QFD (Quality Function Deployment)** (http://en.wikipedia.org/wiki/Quality_function_deployment)

- **Quality function deployment (QFD)** is a "method to transform user demands into design quality, to deploy the functions forming quality, and to deploy methods for achieving the design quality into subsystems and component parts, and ultimately to specific elements of the manufacturing process."
- QFD is designed to help planners focus on characteristics of a new or existing product or service from the viewpoints of market segments, company, or technology-development needs. The technique yields graphs and matrices.
- QFD helps transform customer needs (the voice of the customer [VOC]) into engineering characteristics (and appropriate test methods) for a product or service, prioritizing each product or service characteristic while simultaneously setting development targets for product or service.
- QFD Steps:

  - 1: Understand Customer and Technical Requirements
  - 2: Translate Technical Requirements to **Critical to Quality Characteristics (CTQs)**
  - Build to those CTQs.
- **House of Quality (HOQ)** (http://www.qfdonline.com/templates/)is the most common tool when using QFD:



  - Steps to complete HOQ
    - Put customer wants & needs on the far left
    - Put importance to the customer on the far right
    - Put technical requirements on the top under the hat
    - Rate the importance of the technical requirements to each customer need & put that value in the cell in the body of the tool
    - Rank the relationships of each technical requirement to the importance to the customer & put that in the bottom in the appropriate column
    - Technically evaluate your companies' abilities in each technical area to your competition and put that in the appropriate column in the very bottom.
    - Fill in the hat correlating the different technical requirements according to strength of correlation in the appropriate cell.
    - Analyze your house for conclusions to incorporate into the design.