GreyCampus

LEAN SIX SIGMA BLACK BELT
CHEAT SHEET

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

CONTENTS

ANOVA

DOE (DESIGN OF EXPERIMENTS)

- One Factor At a Time (OFAT)
- Comparison
- Randomization
- Replication
- Blocking
- Orthogonality
- Factorial experiments
- Step-by-step procedure

REGRESSION

- Linear Regression
- Non-Linear Regression
- OLS (Ordinary Least Squares)

NON-NORMAL DISTRIBUTIONS

- with respect to (wrt) Confidence Intervals
- wrt Gage R&R
- wrt T-test
- wrt ANOVA
- wrt Pearson Correlation Coefficient
- wrt Central Limit Theorem
- wrt Control Charts
- wrt Control Limits
- wrt Process Capability
- wrt Control Plans
- wrt Design Of Experiments
- wrt Regression

VARIANCE INFLATION FACTOR

LIFE TESTING & RELIABILITY

- AQL (Acceptable Quality Limit)
- AOQL (Average Outgoing Quality Limit)

QFD (QUALITY FUNCTION DEPLOYMENT)

- Critical to Quality Characteristics (CTQ)
- House of Quality (HOQ)

ANOVA

Anova

- Used for hypothesis testing when comparing multiple groups.
- Hypothesis takes the form H_o: μ1 = μ2 = μ3 = …
- 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

- 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)** 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.

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

- 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 X1, ..., Xp that may be related to y, we can use linear regression analysis to quantify the strength of the relationship between y and the Xj, to assess which Xj may have no relationship with y at all, and to identify which subsets of the Xj 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

- 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** : In most real-world applications the impact of non-normal data on this tool is negligible.
        - **T-test** : 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** : 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 Fmax Test or the Friedman Test..

- **Pearson Correlation Co-efficient**
  - ◇ 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**. 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**
  - ◇ Generalized linear models (GLMs) are used to do regression modelling for non-normal data with a minimum of extra complication compared with normal linear regression.
    1. The GLM consists of three elements.
    2. A distribution function f, from the exponential family.
    3. A linear predictor $\eta = X\beta$.
    4. A link function g such that $E(Y) = \mu = g\text{-}1(\eta)$.
  - ◇ The exact formulas depend on the underlying distribution.

- Some of the more common non-normal distributions include
  - **Weibull, Exponential, Log-normal**. Minitab's distribution identity function can test for these distributions. NOTE: Weibull can be normal if ⅄=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)

- 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:
    - $Y = \beta 0 + \beta 1\ X1 + \beta 2\ X\ 2 + \ldots + \beta k\ Xk + \varepsilon$
    - VIF can be calculated in three steps:
      - Calculate k different VIFs, one for each Xi by first running an ordinary least square regression that has Xi as a function of all the other explanatory variables in the first equation.
        - If i = 1, for example, the equation would be $X1 = \alpha 2\ X\ 2 + \ldots + \alpha k\ Xk + c0 + \varepsilon$ where c0 is a constant and e is the error term.
      - Then, calculate the VIF factor for β^ with the following formula:
        - $VIF(\hat{\beta}_i) = [1 /(1\text{-}R2i)]$
        - where R2iis the coefficient of determination of the regression equation in step one.
      - Then, Analyze the magnitude of multicollinearity by considering the size of the $VIF(\hat{\beta}_i)$.
      - A common rule of thumb is that if $VIF(\hat{\beta}_i) > 5$ 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

AQL

- 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.

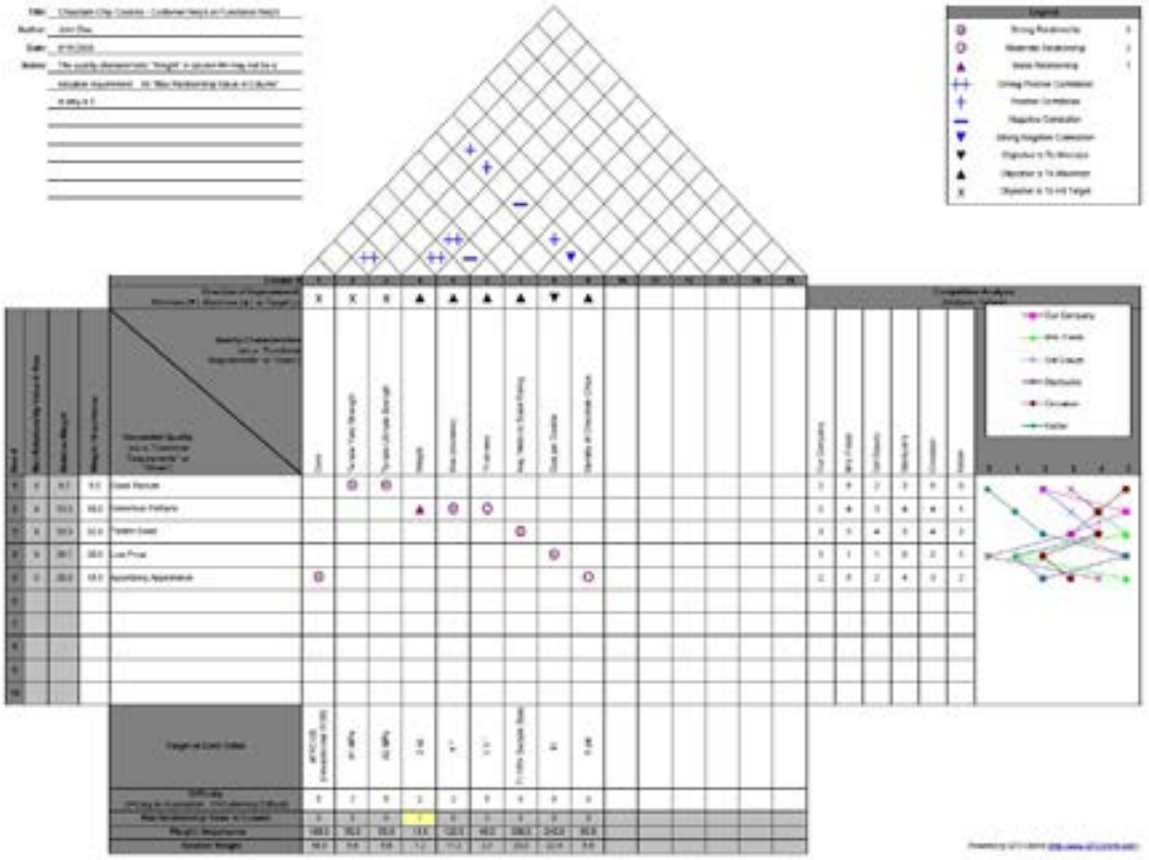
AOQL

- **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)

- **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)** 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 that
  - ◇ 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.

ABOUT GREYCAMPUS

GreyCampus is a leading provider of on-demand training that address the unique learning needs of professionals, delivered as online self-learning, live online training or in-person classroom training. Our aim is to provide quality training enabling professionals to achieve their certification and career enhancement goals. We offer training for certifications in areas of ***Big Data & Hadoop, Project Management, IT Service Management, Quality Management, Python Programming, Agile Training Coaching & Certification and Workplace Tools.***

TRAINED OVER 30,000 PROFESSIONALS

REACH ACROSS 50+ COUNTRIES

EXAM PASS RATE OF OVER 97 %

COURSES ACCREDITED BY LEADING GLOBAL BODIES

ACCREDITATIONS & ASSOCIATIONS

Project Management Institute
Registered Education Provider
Provider ID : 3871

ITIL® TRAINING ORGANIZATION ACCREDITED BY PEOPLECERT

AXELOS CERTIFIED PARTNER

PRINCE2® TRAINING ORGANIZATION ACCREDITED BY PEOPLECERT

IASSC Accredited

DISCLAIMER