Quality Engineering - Class Notes (experimental)

Jonathan Rosenblatt

November 14, 2015

Preface

This text accompanies my Quality Engineering course at the Dept. of Industrial Engineering at the Ben-Gurion University of the Negev. It has several purposes:

- Help me organize and document the course material.
- Help students during class so that they may focus on listening and not writing.
- Help students after class, so that they may self-study.

At its current state it is experimental. It can thus be expected to change from time to time, and include mistakes. I will be enormously grateful to whoever decides to share with me any mistakes found.

I also ask for the readers' forgiveness for my Wikipedia quoting style. It is highly unorthodox to cite Wikipedia as one would cite a peer reviewed publication. I do so, in this text, merely for technical convenience.

I hope the reader will find this text interesting and useful.

Contents

List of Figures

List of Definitions

Chapter 1

Introduction

Quality Engineering is the study and design of practices aimed improving the "quality" of production. Production is understood in a wide sense, and includes services as well. Quality is understood in many senses. Here are several definitions compiled verbatim from Montgomery (2007) and Wikipedia (2015e):

1. Montgomery: "The reciprocal of variability".
2. American Society for Quality: A combination of quantitative and qualitative perspectives for which each person has his or her own definition; examples of which include, "Meeting the requirements and expectations in service or product that were committed to" and "Pursuit of optimal solutions contributing to confirmed successes, fulfilling accountabilities. In technical usage, quality can have two meanings: (a) The characteristics of a product or service that bear on its ability to satisfy stated or implied needs. (b) A product or service free of deficiencies."
3. Subir Chowdhury: "Quality combines people power and process power".
4. Philip B. Crosby: "Conformance to requirements."
5. W. Edwards Deming: "The efficient production of the quality that the market expects".
6. W. Edwards Deming: "Costs go down and productivity goes up as improvement of quality is accomplished by better management of design, engineering, testing and by improvement of processes."
7. Peter Drucker: "Quality in a product or service is not what the supplier puts in. It is what the customer gets out and is willing to pay for."

8. Victor A. Elias: "Quality is the ability of performance, in each Theme of Performance, to enact a strategy."

9. ISO 9000: "Degree to which a set of inherent characteristics fulfills requirements."

10. Joseph M. Juran: "Fitness for use.".

11. Noriaki Kano and others, present a two-dimensional model of quality: "must-be quality" and "attractive quality." The former is near to "fitness for use" and the latter is what the customer would love, but has not yet thought about. Supporters characterize this model more succinctly as: "Products and services that meet or exceed customers' expectations."

12. Robert Pirsig: "The result of care."

13. Six Sigma: "Number of defects per million opportunities."

14. Genichi Taguchi: "Uniformity around a target value."

15. Genichi Taguchi: "The loss a product imposes on society after it is shipped."

16. Gerald M. Weinberg: "Value to some person".

17. Jonathan D. Rosenblatt: "The efficient fulfilment of a promise".

Collecting ideas

1. Quality is not only about production.
2. Quality is the means, not the end.
3. Quality may deal with the **design** or with **conformance** to a given design.

Almost all of the above definitions, may apply to different characteristics, we call *dimensions of quality*. Following Wikipedia (2015b) :

Dimensions of Quality

Performance Performance refers to a product's primary operating characteristics. This dimension of quality involves measurable attributes; brands can usually be ranked objectively on individual aspects of performance.

Features Features are additional characteristics that enhance the appeal of the product or service to the user.

Reliability Reliability is the likelihood that a product will not fail within a specific time period. This is a key element for users who need the product to work without fail.

Conformance Conformance is the precision with which the product or service meets the specified standards.

Durability Durability measures the length of a product's life. When the product can be repaired, estimating durability is more complicated. The item will be used until it is no longer economical to operate it. This happens when the repair rate and the associated costs increase significantly.

Serviceability Serviceability is the speed with which the product can be put into service when it breaks down, as well as the competence and the behavior of the service person.

Aesthetics Aesthetics is the subjective dimension indicating the kind of response a user has to a product. It represents the individual's personal preference.

Perceived Quality Perceived Quality is the quality attributed to a good or service based on indirect measures.

1.1 Terminology and Concepts

1.1.1 Basic Terminology

Quality Characteristics A.k.a. *Critical to Quality Characteristics* (CTQs). May be physical, sensory, or temporal properties of a process/product. Obviously related to the dimensions of quality.

Quality Engineering "The set of operational, managerial, and engineering activities that a company uses to ensure that the quality characteristics of a product are at the nominal or required levels and that the variability around these desired levels is minimum." (Montgomery, 2007)

Variables Continuous measurements of some CTQ.

Attributes Discrete measurements of some CTQ.

Target Value The desired level of a particular CTQ. A.k.a. *nominal* value.

USL & LSL Largest and smallest allowable values of a CTQ.

Specifications The set of permissible values for all CTQs. Either a set of target values, or USL-LSL intervals.

Non-conformity A non conforming product is one that fails to meet the specification.

Fallout The same as non-conformity.

Defect A non-conformity that is serious enough to affect the use of the product.

DPMO Defect per million opportunities.

PPM Parts per million. Interchangeable with DPMO.

1.1.2 Statistical Terminology

Exploratory Data Analysis (EDA) An assumption free quantitative inspection of data; "Story telling"; no inference.

Inference Data analysis with the intention of generalizing from a sample to a population.

Causal Inference Inference, with the intention of claiming causal relations between quantities under study.

Predictive Analytics Data analysis with the intention of making predictions with future data. Can be seen as inference, without aiming at causality.

Design of experiments (DOE) By far the best and most established way for causal inference. The *random assignment* of units to groups allows to interpret statistical correlations as causal.

Statistical Process Control (SPC) Data analysis with the intention of identifying anomalous behaviour with respect to history[1].

Computer Simulation Well, just what the name implies.

[1] Akin to *anomaly detection*, or *novelty detection*, in the machine learning literature.

Control Chart A graphic visualization of the historical evolution of one (or several) CTQs.

(Un)Controllable Inputs Each process has inputs that affect the behaviour of the CTQ. Some are controllable, and some are not.

Factorial Design In the language of DOE, controllable inputs are *factors*. A factorial design, is an experiment where factors are varied in order to study their effect on the CTQ.

Off/On-line process control SPC can be performed on or off line. On-line, a.k.a. *in-process control*, meaning control happens as the process evolves, and off-line meaning before it starts or after it has finished.

Engineering control A.k.a. *automatic control*, or *feedback control*. SPC that triggers an intervention that keeps the process in control.

Outgoing/Ingoing Inspection Refers to the stage at which SPC is performed. As inputs come in (ingoing), or as outputs come out (outgoing).

1.2 Some History

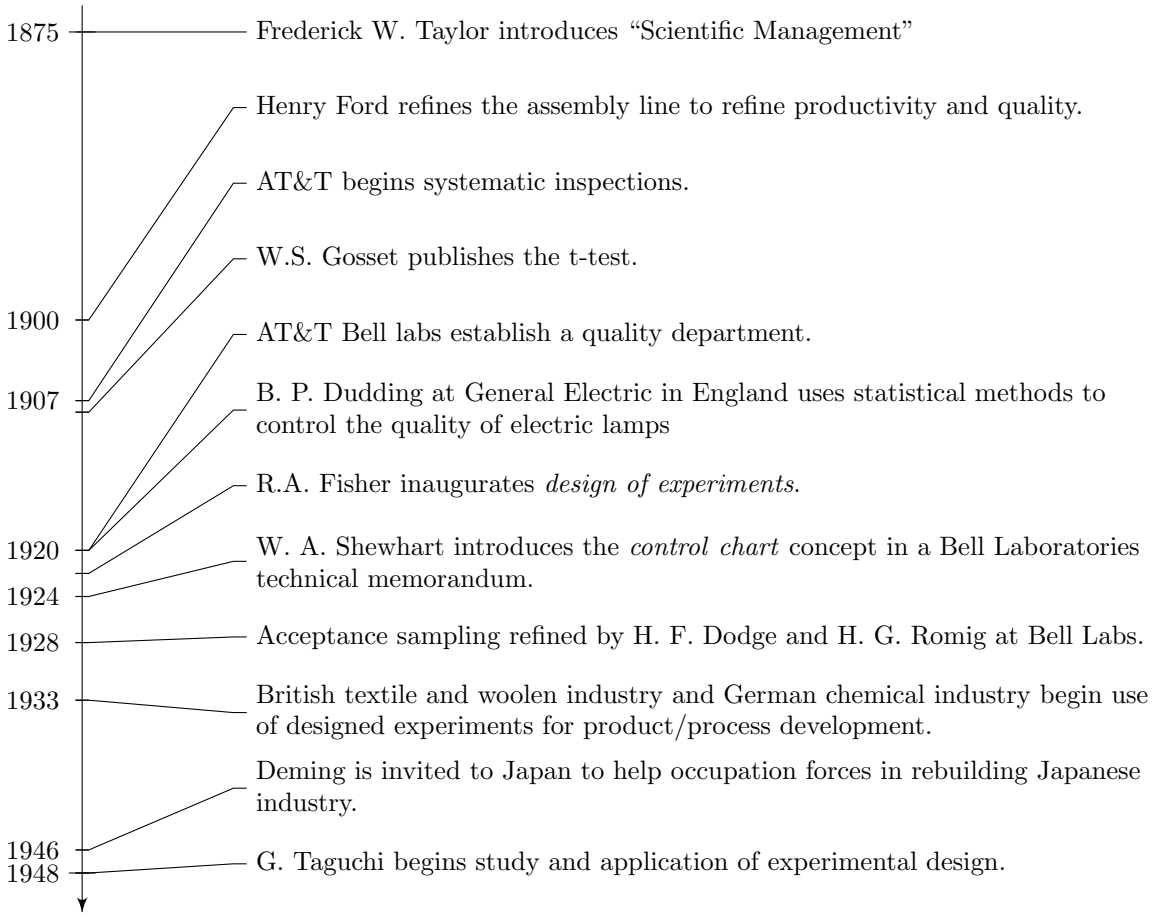


Table 1.1: Adapted from (Montgomery, 2007, Table 1.1).

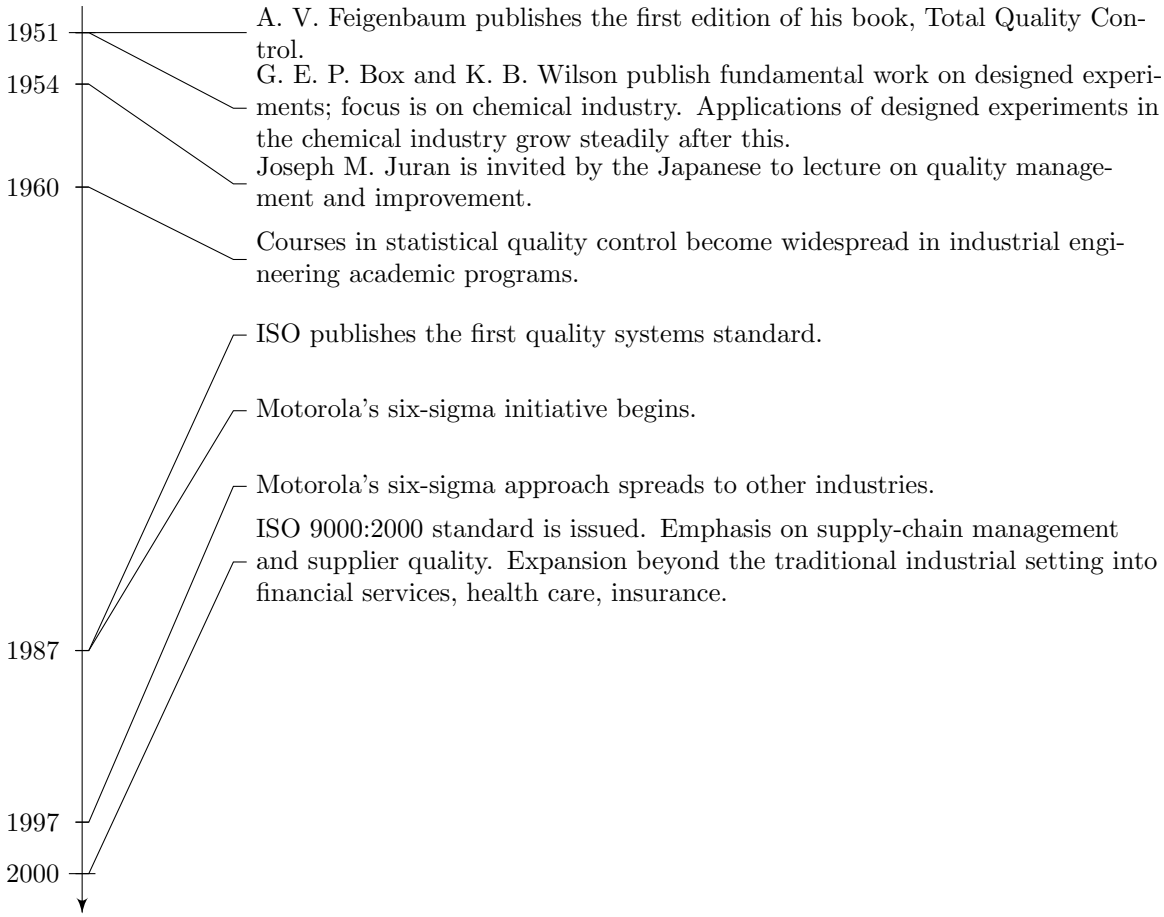


Table 1.2: Adapted from (Montgomery, 2007, Table 1.1).

1.3 Management Aspects of Improving Quality

The founding fathers of QC have many do's-and-don'ts for managers. See Montgomery (2007, Sec 1.4) for details. As usual, we collect recurring ideas:

1. The responsibility for quality rests with management.
2. QC is not a one-time project, but an on-going process. It may advance continuously, or incrementally.
3. QC is (or should be) manifested in organizational structure, training, recruitment, incentives, knowledge management, to name a few.

1.4 Programs and Initiatives

1.4.1 Zero Defects Program (ZD)

Quoting Wikipedia (2015h):

> ... a management-led program to eliminate defects in industrial production that enjoyed brief popularity in American industry from 1964 to the early 1970's. Quality expert Philip Crosby later incorporated it into his "Absolutes of Quality Management" and it enjoyed a renaissance in the American automobile industry, as a performance goal more than as a program, in the 1990s. Although applicable to any type of enterprise, it has been primarily adopted within supply chains wherever large volumes of components are being purchased (common items such as nuts and bolts are good examples).

1.4.2 Quality is Free Initiative

Quoting Montgomery (2007):

> ... in which management worked on identifying the cost of quality (or the cost of *nonquality*, as the Quality is Free devotees so cleverly put it). Indeed, identification of quality costs can be very useful, but the Quality is Free practitioners often had no idea about what to do to actually improve many types of complex industrial processes.

1.4.3 Value Engineering Program (VE)

Quoting Wikipedia (2015g):

> Value engineering (VE) is systematic method to improve the "value" of goods or products and services by using an examination of function. Value, as defined, is the ratio of function to cost. Value can therefore be increased by either improving the function or reducing the cost. It is a primary tenet of value engineering that basic functions be preserved and not be reduced as a consequence of pursuing value improvements.

1.4.4 Total Quality Management (TQM)

TQM originates in the 1980's with the ideas of Deming and Juran. It is a very wide framework that attempts at capturing the company-wide efforts required for QC. According to Montgomery (2007, p.23):

> TQM has only had **moderate success** for a variety of reasons, but frequently because there is insufficient effort devoted to widespread utilization of the technical tools of variability reduction. Many organizations saw the mission of TQM as one of training. Consequently, many TQM efforts engaged in widespread training of the workforce in the philosophy of quality improvement and a few basic methods. This training was usually placed in the hands of human resources departments, and much of it was ineffective. The **trainers often had no real idea about what methods should be taught**, and success was usually measured by the percentage of the workforce that had been "trained," not by whether any measurable impact on business results had been achieved.
>
> ... Another reason for the erratic success of TQM is that many managers and executives have regarded it as **just another "program" to improve quality**. During the 1950's and 1960's, programs such as Zero Defects and Value Engineering abounded, but they had little real impact on quality and productivity improvement.

1.4.5 Six-Sigma

Quoting Montgomery (2007):

> Products with many components typically have many opportunities for failure or defects to occur. Motorola developed the Six-Sigma program in the late 1980s as a response to the demand

for their products. The focus of six-sigma is reducing variability in key product quality characteristics to the level at which failure or defects are extremely unlikely.

Assume a device has m components. The failure probability of component $j \in 1, \ldots, m$ is α_j. What is the probability of the device failing, when assuming independent failures?

$$\begin{aligned} P(\text{failure}) &= P(\text{at least one failure}) \\ &= 1 - P(\text{no failure}) \\ &= 1 - \prod_{j=1}^{m}(1-\alpha_j) \end{aligned} \tag{1.1}$$

Assuming all components have the same fallout rate, we omit the index j in α_j.

The failure probability α is implied by the CTQs, and its specification limits (USL, LSL). Denoting the target value of the CTQ by T, then $USL = T + \delta$ and $LSL = T - \delta$. Three-sigma means that the production variability, σ, is small enough so that

$$3\sigma = \delta.$$

Assuming

$$CTQ \sim \mathcal{N}(T, \sigma),$$

we can compute:

$$\alpha = 1 - P(LSL < CTQ < USL) \tag{1.2}$$
$$= 1 - P(|CTQ| < \delta) \tag{1.3}$$
$$= 1 - P(|CTQ| < 3\sigma) = 0.0027. \tag{1.4}$$

The 3-sigma quality guarantee is also known as $2,700$ defective parts per million (ppm) for now obvious reasons. Plugging the 3-sigma performance in Eq.(1.1) returns PPM

$$P(\text{3-sigma failure}) < 1 - (1 - 0.0027)^m$$

Figure 1.1 illustrates the probability of failure against the number of components. For simple devices, the 3-sigma criterion may suffice. But now imagine the number of components in a car, a cellular phone, ... The 3-sigma rule is just not good enough. This is where 6-sigma requirement comes along. It implies that the production is process is so accurate that

$$6\sigma = \delta.$$

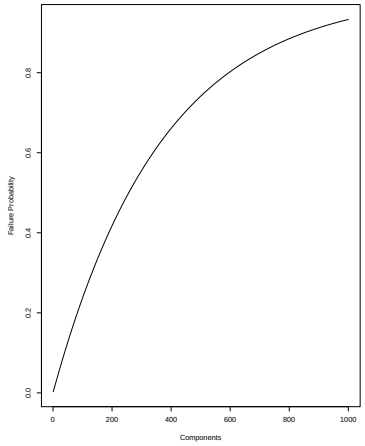


Figure 1.1: The probability of failure as a function of components under the 3-sigma standard.

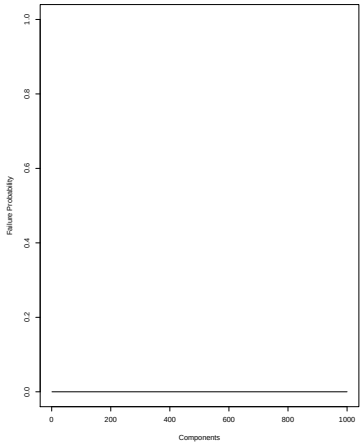


Figure 1.2: The probability of failure as a function of components under the 6-sigma standard.

Updating Eq.(1.2) we get that the defective *ppm* of 6-sigma is 0.002. This is obviously excellent news, except for the typically tremendous effort involved in achieving this level of quality.

According to Montgomery (2007), the 6-sigma methodology has gained more success than it predecessors:

> The reason for the success of six-sigma in organizations outside the traditional manufacturing sphere is that variability is everywhere, and where there is variability, there is an opportunity to improve business results.

1.4.6 Lean Systems

Quoting Wikipedia (2015c) (my own emphasis in bold):

> Essentially, lean is centered on making obvious what **adds value** by **reducing everything else**. Lean manufacturing is a management philosophy derived mostly from the Toyota Production System (TPS) (hence the term Toyotism is also prevalent) and identified as "lean" only in the 1990s.

1.4.7 Design for Six-Sigma (DFSS)

Quoting Wikipedia (2015a) (my own emphasis in bold):

It is based on the use of **statistical tools** like linear regression and enables empirical research similar to that performed in other fields, such as social science. While the tools and order used in Six Sigma require a process to be in place and functioning, DFSS has the objective of **determining the needs of customers** and the business, and driving those needs into the product solution so created. DFSS is relevant for relatively simple items / systems. It is used for product or process design in contrast with process improvement.

1.4.8 Quality Systems and Standards

The first quality standard was issued by the International Standards Organization (ISO) in 1987. Current quality standards are known as the *ISO9000 series*. These include: ISO9000

ISO9000:2000 Quality Management System-Fundamentals and Vocabulary.

ISO9001:2000 Quality Management System-Requirements.

ISO9004:2000 Quality Management System-Guidelines for Performance Improvement.

In Israel, it is the Standards Institute of Israel[2] that may give ISO9000 (like any ISO) certifications upon inspecting the candidate organization. As emphasized by Montgomery (2007, p.24), ISO9000 is a set of rules and best practices, mostly oriented at *knowledge management*. It may help to *preserve* quality, but it does not, nor does it aim to, *improve* quality. As such, it will not be the focus of our course, which will focus on *statistical tools*.

Extra Info 1. [TODO: Just-in-Time, Poka-Yoke]

1.5 DMAIC

There are many names for the process of quantitative re-evaluations of performance against a given target: *data driven decision making* (DDD), *Shewart cycle*, etc. We will focus on one such framework, illustrated in Figure 1.3 known as DMAIC: Define, Measure, Analyze, Improve, Control.

Here are some general observations on DMAIC:

[2] `https://portal.sii.org.il/heb/qualityauth/certificationtypes/qualitylinks/iso9001/`

Figure 1.3: The DMAIC cycle.
`http://www.sapartners.com/sigma-academy/`

1. It is aimed at promoting improvement and creative thinking.
2. It is not part of the six-sigma methodology, but will typically take part in its implementation.

What do the stages of DMAIC mean [3]?

Define the problem, improvement activity, opportunity for improvement, the project goals, and customer (internal and external) requirements.

Measure process performance.

Analyze the process to determine root causes of variation, poor performance (defects).

Improve process performance by addressing and eliminating the root causes.

Control the improved process and future process performance.

In the following chapter we give a set of statistical tools required for *measuring*,*analyzing* and *controlling* a process.

[3] `http://asq.org/learn-about-quality/six-sigma/overview/dmaic.html`

Chapter 2

Exploratory Data Analysis

In this chapter, we give a short review of methods for *exploratory data analysis* (EDA), a.k.a. *descriptive statistics*. Recall that our goal is an assumptions-free description of our data. EDA thus consist of computing interpretable summaries of the data, called *summary statistics*, and visualizations.

Descriptive Statistics

2.1 Summary Statistics

We now distinguish between summary statistics that apply to attributes, categorical by definition, and variables, continuous by definition.

2.1.1 Summarizing Categorical Data

Univariate

Summarizing a vector of categorical data can naturally be done by tabulating it, i.e., computing the frequency and relative frequency of each category. Clearly averages, medians, and the likes are incomputable, since categorical data has no ordering, nor does it admit simple operations such as summation.

Extra Info 2. Variability of categorical data can clearly not be measured by its variance, since it does not admit a summation operation. It is, however, possible to define different measures of variability that do apply. The *entropy* is such an example.

Entropy

Bivariate

Generalizing the univariate case to bivariate, or multivariate, one can keep tabulating. I.e., compute the frequency, and relative frequency, of combina-

tions of categories.

2.1.2 Summarizing Continuous Data

Continuous variables admit many more mathematical manipulations than categorical attributes.

Univariate

We start by presenting the most natural summaries of the data. Without going into the formal definition, we refer to them as *summary of location*. These include:

Location Summaries

Definition 1 (The Mean)**.** The *mean*, or *average*, is defined as

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

Definition 2 (The Median)**.** The median is the observation that is smaller than half of the sample and larger than half of the sample.

Definition 3 (α-Trimmed Mean)**.** The α-trimmed mean is the average of the observations left after ignoring the largest and the smallest $(100\alpha)\%$ of them.

The naïve average is the 0-trimmed mean, and the median is the 0.5-trimmed mean.

From summaries of location, we move to summaries of *scale*.

Summary of Scale

Definition 4 (The Standard Deviation)**.**

$$s(x) := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

For the following, we require the definition of the sample quantiles, themselves **not** a scale summary.

Definition 5 (α Quantile)**.** The α-quantile of the sample is the observation that is larger than $(100\alpha)\%$, and smaller then $(100(1-\alpha))\%$ of the sample.

The empirical maximum and minimum are then $x_{1.0}$ and $x_{0.0}$, respectively.

Definition 6 (The Range)**.**

$$Range(x) := \max_i\{x_i\} - \min_i\{x_i\} = x_{1.0} - x_{0.0} \tag{2.3}$$

Definition 7 (The Inter Quantile Range- IQR)**.**

$$IQR(x) := x_{0.75} - x_{0.25} \tag{2.4}$$

Definition 8 (The Median Absolute Deviation- MAD)**.**

$$MAD(x) := \{|x_i - x_{0.5}|\}_{0.5} \tag{2.5}$$

Note that the MAD may be sometimes scaled by 1.4826, so that it estimates σ. Such is the behaviour of the **R** function `mad()`.

After summaries of scale, we move to summaries of *skewness*, or *asymmetry*.

Definition 9 (Yule Skewness Measure)**.**

$$YULE(x) := \frac{\frac{1}{2}(x_{0.75} + x_{0.25}) - x_{0.5}}{IQR(x)} \tag{2.6}$$

Bivariate

From univariate data x, we move to bivariate x, y. Clearly we can apply univariate summaries component-wise. We want, however, to summarize the *joint* behaviour of the data. For this purpose, we assume that data comes in pairs, implying that x and y are of same length.

Definition 10 (Covariance)**.** The sample covariance, or *empirical* covariance is defined as

$$Cov(x, y) := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{2.7}$$

Definition 11 (Pearson's Correlation Coefficient)**.** *Pearson's Correlation Coefficient*, or *Pearson's Moment Product Correlation Coefficient*, is defined as

$$r(x, y) := \frac{(n-1)Cov(x, y)}{S(x)S(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S(x)S(y)} \tag{2.8}$$

We can dwell into the meaning and intuition underlying Pearson's correlation coefficient, but we will not. The curious reader is reffered to Rodgers and Nicewander (1988).

The next measure of association captures a more general association.

Definition 12 (Spearman's Correlation Coefficient)**.** Spearman's correlation coefficient is merely Pearson's correlation coefficient computed on the *ranks* of x and y.

We conclude by noting that *regression coefficients* are also a measure of association.

Multivariate Data

Multivariate data, both continuous (variables), and discrete (attributes), admits a vast realm of method for summary and visualization. Clearly, associations between several variables can be very complicated so that the more we try to summarize, the more information we give up. On the other hand, and unlike the univariate and bivariate case, our minds will need some type of simplification since they cannot grasp the raw data (did you ever try to imagine how $\mathbb{R}^4$ looks like?). As usual, we emphasize that our purpose is to summarize the joint association in the data. For component-wise summaries, we can always apply the univariate summaries one variable at a time.

By far the most popular measures of joint association are the covariance matrix and correlation matrix.

Definition 13 (Covariance Matrix)**.** For multivariate data consisting of $x_1, \ldots, x_j, \ldots, x_p$ vectors, each with n entries: $x_{j,1}, \ldots, x_{j,n}$, we define the (sample) covariance matrix to be a $p \times p$ matrix whose elements are the (sample) covariances between corresponding vectors:

$$\hat{\Sigma}_{k,l} := Cov(x_k, x_l). \tag{2.9}$$

Extra Info 3 (Sample Covariance Matrix)**.** The matrix $\hat{\Sigma}$ has many useful properties. The curious reader is referred to Petersen and Pedersen (2006), and references therein, for more details.

Definition 14 (Correlation Matrix)**.** For multivariate data consisting of $x_1, \ldots, x_j, \ldots, x_p$ vectors, each with n entries: $x_{j,1}, \ldots, x_{j,n}$, we define the (sample) correlation matrix to be a $p \times p$ matrix whose elements are the (Pearson) correlations between corresponding vectors:

$$\hat{R}_{k,l} := r(x_k, x_l) \tag{2.10}$$

Extra Info 4 (Multivariate Data Analysis)**.** Multivariate analysis is an important, and very actively studied field in statistics and machine learning. A non-comprehensive list of methods that belong to this realm include Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Factor

PCA, SVD,ICA

Analysis (FA), Independent Component Analysis (ICA), Dimensionality Reduction, Manifold Learning, Self Organizing Maps, etc. Ask me for reference books or courses if this topic interests you.

2.2 Visualization

2.2.1 Visualizing Categorical Data

Univariate

Much like computing summaries, there is not much to be said about visualizing univariate categorical variables. The most natural, and perhaps only visualization, is the *bar plot*, illustrated in Figure 2.1.

Remark 1 (Pie Chart)**.** About those pie charts. There is really no reason to use them. Ever[1].

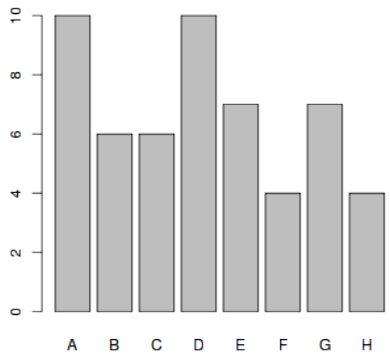


Figure 2.1: The Bar-Plot.
`http://www.r-tutor.com/elementary-statistics/qualitative-data/bar-graph`

Bivariate

Visualizing a two-way cross-table can be done using an extension of the bar-plot. Several extensions exist. By far, the most informative and rec-

[1] `http://www.businessinsider.com/pie-charts-are-the-worst-2013-6`.

ommended figure, in this author's view, is the *mosaic plot*, illustrated in Figure 2.2.

Mosaic Plot

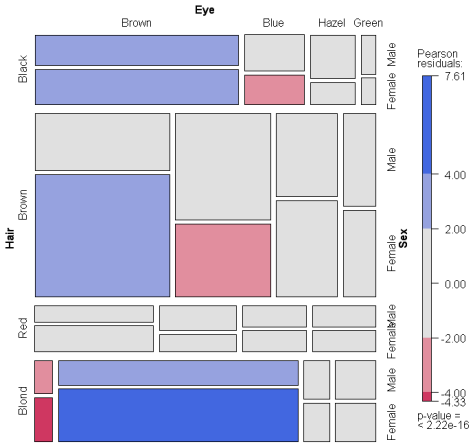


Figure 2.2: Mosaic Plot.
`http://www.statmethods.net/advgraphs/mosaic.html`

2.2.2 Visualizing Continuous Data

Univariate

Visualization of univariate continuous vectors can present the raw data, or it distribution (i.e.- discarding the indexes). The most basic visualizations are the *dotchart*, *histogram*, *boxplot*, *stem-and-leaf plot*. These are illustrated in figures 2.3, 2.4, 2.5, 2.6 respectively.

Bivariate

The simultaneous visualization of two continuous variables, can naturally be done with a *scatter plot*. More sophisticated visualization, which generalizes the histogram into two dimensions, is the *hexbin plot*. These are illustrated in figures 2.7, and 2.8, respectively.

Multivariate Data

Since we cannot possibly visualize data in more than 3-dimensions, and we clearly prefer data in 1 or 2 dimensions, the visualization of multivariate data will typically consist of summarizing the data into $1D$ or $2D$, and then applying the above mentioned visualization techniques.

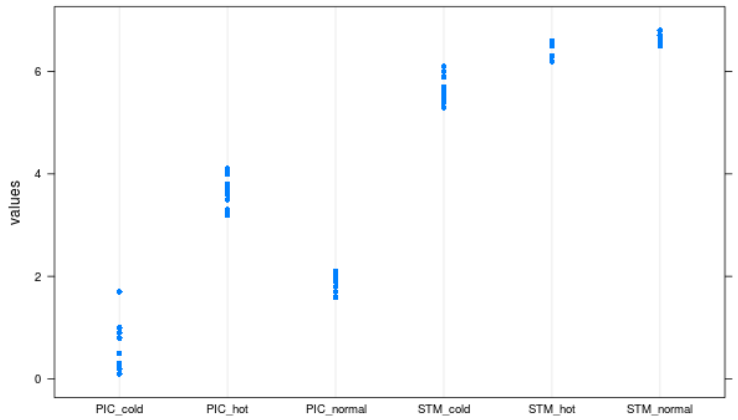


Figure 2.3: Dot Plot.
`http://stackoverflow.com/questions/15109822/r-creating-scatter-plot-from-data-frame`

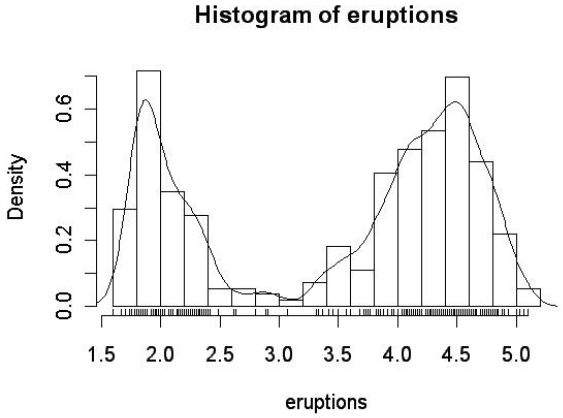


Figure 2.4: Histogram. Notice the ticks on the x axis. These are the raw data points. Make sure you always add them, with the `rug()` **R** command.
`http://compbio.pbworks.com/w/page/16252882/Basic`

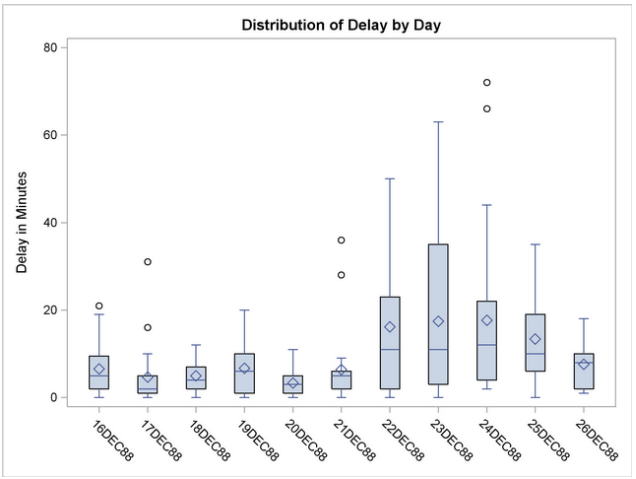


Figure 2.5: Boxplot.
`http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm`

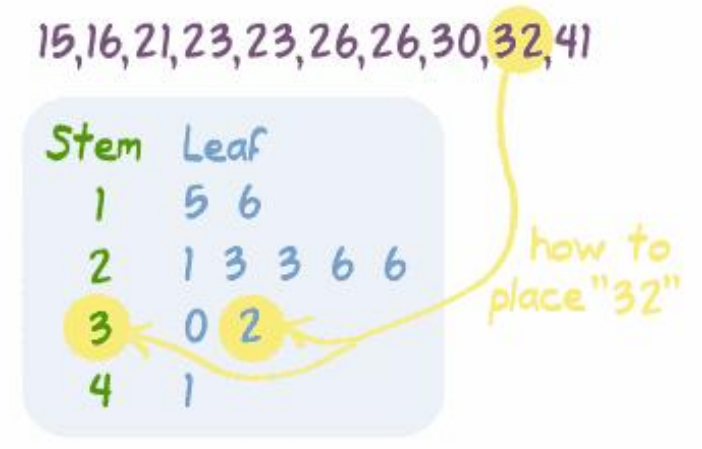


Figure 2.6: Stem-and-leaf plot.
`https://www.mathsisfun.com/data/stem-leaf-plots.html`

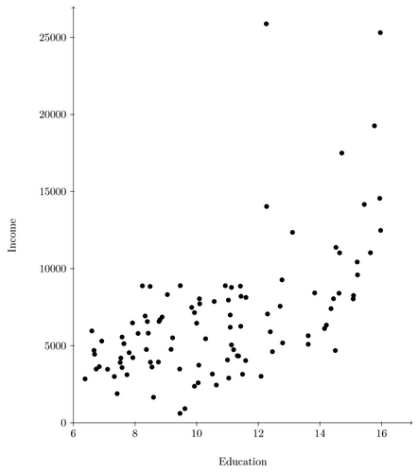


Figure 2.7: Scatter Plot.
`http://texample.net/tikz/examples/scatterplot/`

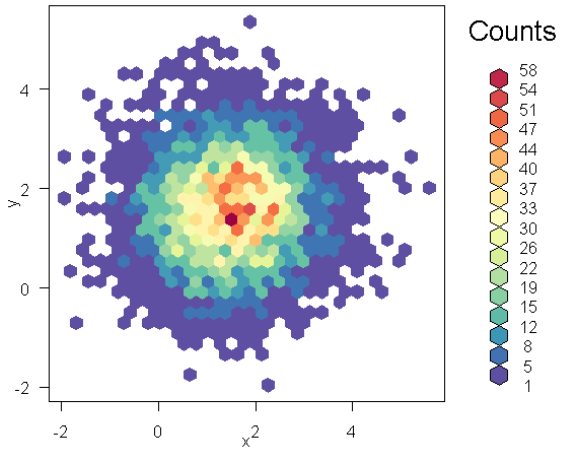


Figure 2.8: HexBin plot (a 2D histogram).
`http://www.r-bloggers.com/5-ways-to-do-2d-histograms-in-r/`

An important exception is due to the observation that a computer image, is essentially a matrix. We can thus visualize matrices, with a simple image, and in particular, covariance and correlation matrices, as illustrated in Figure 2.9.

A second exception is when the data has both continous variables and discrete attributes. Endlessly many combinations are then possible. The author strongly recommends to visit Hans Rosling's *Gap Minder* at `http://www.gapminder.org/world` for an excellent interactive visualization.

Gap Minder

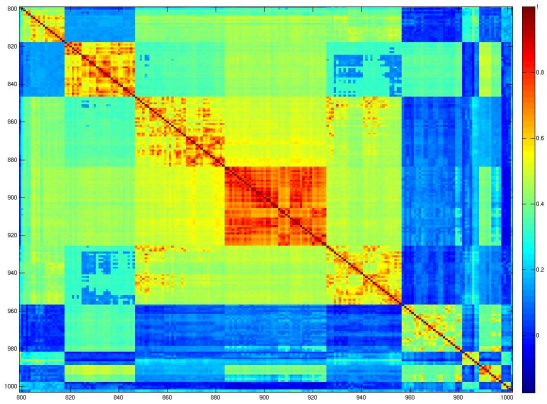


Figure 2.9: Image of covariance matrix.
`http://cs.brown.edu/courses/csci1950-g/results/final/sghosh/`

2.2.3 On-Line Visualization

For the purpose of quality control, we may often want an *on-line* visualization, and not *off-line*, as the ones previously discussed. This is the purpose of *dashboards*, illustrated in Figure 2.10.

Dashboard

Figure 2.10: Dashboard.
http://www.iconics.com/Home/Products/AnalytiX/Quality-AnalytiX.aspx

Chapter 3

Statistical Inference

The idea of extrapolating knowledge from a *sample* to a population is known as *statistical inference*. It encompasses the ideas of *parameter estimation*, *confidence intervals*, and *hypothesis testing*. We will assume the reader is familiar with these, but recall some required terminology. The QC and SPC terminology are not always consistent with statisticians' terminology. When new names are given to old ideas, we will emphasize this in the text.

In Statistical Control

Null/Alternative Hypothesis Some statement about the world we wish to test with data. The frequentist argument follows a Popperian philosophy: to show the alternative hypothesis is true, we will show that the null hypothesis is not true. In the context of quality control, the null hypothesis will be the process is *in statistical control*, while the alternative will be that it is *out of control*.

Statistical Test The procedure of inferring from data on the truthfulness of the alternative hypothesis.

Assumptions As the name suggests, these are assumptions. We stress that unlike hypothesis, assumptions are not being tested in a statistical test.

Test Statistic The function of the data to be computed for the purpose of inference. As such, it is a random variable.

Null/Alternative Distribution The distribution of the test statistic under the null/alternative hypothesis.

Type I/II error See Figure 3.1.

False/True Positive/Negative See Figure 3.1.

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error **False positive**	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error **False negative**

Figure 3.1: Type I/II error confusion table.
`https://infocus.emc.com/william_schmarzo/beware-of-false-positives/`

Rejection Region The collection of event that will lead us to reject the null hypothesis, and believe in the alternative hypothesis.

p-value A.k.a. *observed significance*. The null probability of the observed (or "more extreme") event.

Significance Level A.k.a. α. The probability of a false positive.

Power The probability of a true positive.

i.i.d. "Independent and identically distributed" (i.i.d.) is an assumption made on the sampling distribution, meaning that samples are statistically independent, and all originating from the same distribution.

The following sections of this chapter present particular statistical inference methods we will be using in the following chapters.

3.1 Goodness of Fit (GOF)

Goodness of fit (GOF) deals with the inference on the sampling distribution, a.k.a., the generative process. It can be approached via rigorous hypothesis testing, or by visualizations.

3.1.1 QQplot and QQnorm

The fundamental idea of the *quantile-quantile plot* (QQplot) is to compare the empirical quantiles in the sample, to the theoretical quantiles implied by the assumed distribution. If the theory and observations agree, we conclude our assumptions are plausible. For the particular case of testing the normality of the data, the corresponding QQplot is known as a *QQnorm plot*.

Figure 3.2 illustrates a QQnorm plot of normal distributed data, while Figure 3.3 is the same for non-normal data.

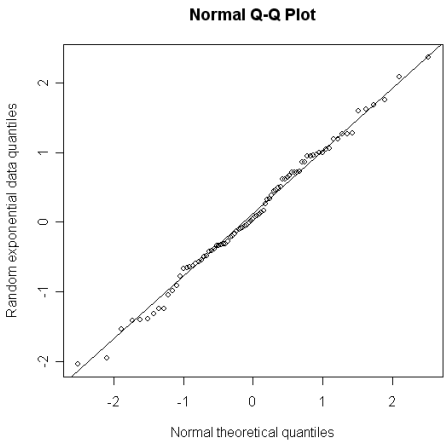


Figure 3.2: A QQplot of Gaussian distributed data.

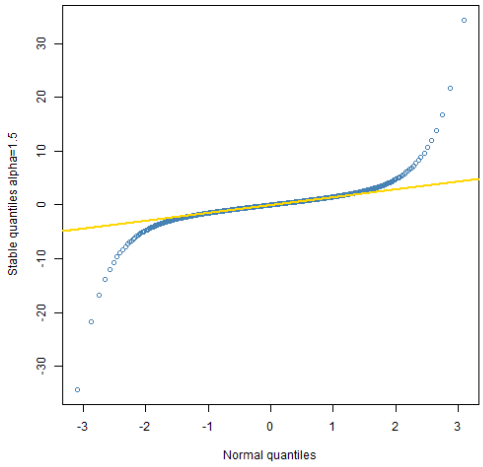


Figure 3.3: A QQnorm plot of non Gaussian distributed data.

3.1.2 Chi-Square GOF Test

The Chi-Square GOF test (not to be confused with the Chi-Square independence test), tests a hypothesis on the sampling distribution of discrete (attributes) data. Note that it is very general, since all continuous variables may be discretized, simply by binning.

Definition 15 (Chi-Square GOF Test)**.** Assume an i.i.d. sample $x_1, \ldots, x_n$. The Chi-Square GOF is a tests $H_0 : \mathbf{x}_i \sim P$ versus $H_1 : \mathbf{x}_i \not\sim P$. P is assumed to be discrete with K categories and $p_k := P(\mathbf{x}_i \in k)$. The test statistic, X^2, is defined as

$$X^2 := \sum_{k=1}^{K} \frac{(obs_k - exp_k)^2}{exp_k}, \tag{3.1}$$

where $obs_k := \#\{x_i \in k\}$, and $exp_k := p_k n$. The approximate null distribution of X^2 is χ^2_{K-1}.

3.1.3 Kolmogorov–Smirnov GOF Test

The Kolmogorov-Smirnof GOF test, tests a hypothesis on the sampling distribution of continuous (variable) data.

Definition 16 (Kolmogorov–Smirnov GOF Test)**.** Assume an i.i.d. sample $x_1, \ldots, x_n$. The Chi-Square GOF is a tests $H_0 : \mathbf{x}_i \sim P$ versus $H_1 : \mathbf{x}_i \not\sim P$. P is assumed to be continuous. The test statistic, D, is depicted in Figure 3.4.

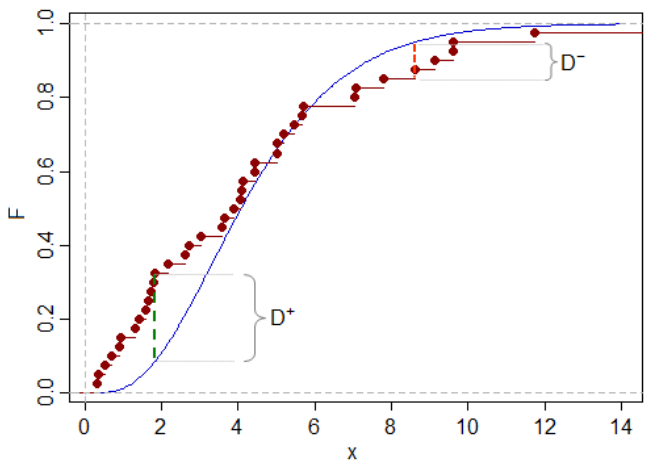


Figure 3.4: Kolmogorov-Smirnov Test. The D statistic is D_+ in the figure.

The null distribution of D is the *Kolmogorov distribution* obtained from tables.

Kolmogorov Distribution

Extra Info 5 (GOF tests)**.** There are endlessly many more GOF tests, such as Anderson-Darling, Jarque–Bera, Shapiro–Wilk, Kuiper's test, etc. Wikipedia is a good place for further reading.

Chapter 4

System Capability Analysis

In a *system capability analysis*, we essentially use statistical tools to measure the variability in a production process. This analysis can answer questions that are raised at the measuring, analyzing, and improving stages of the DMAIC cycle. The methods we will discus compare between the processes variability and its specifications. Note, however, that a simple statistical analysis of the production process, without relating to specification, may also qualify as a system capability analysis.

We naturally want production processes that adhere to specifications, and want to quantify the level of adherence. The quantification is performed by comparing the variability in a CTQ to the specification. In this chapter we assume the process's capability is fixed over time. In Chapter 9 we revisit the same problems, when allowing the process's capability to vary over time.

The particular setup discussed in this chapter is also known as *product specification*. When we use the actual time, and ordering of data samples, as in a control chart, we will no longer regard it as product specification but rather as a bona-fide capability analysis. This is the subject of Chapter 9.

Product Specification

Because capability analysis, or product specification, is essentially the study of the CTQ's distribution, it can be approached with the aforementioned statistical tools such as univariate summary statistics and visualizations presented in Sections 2.1 and 2.2. To test a particular hypothesis want to be tested on the distribution of the CTQ, we may call upon the inference tools from Chapter 3.

4.1 Process Capacity Indexes

Classical statistical devices do not incorporate the designed process's capabilities. *Process capability ratios* (PCR), or *process capability indexes*, are

merely population parameters that also depend on specifications. The first, and most basic PCR is the C_p of a particular CTQ.

Process Capability Index

Definition 17 (C_p)**.**

$$C_p := \frac{USL - LSL}{6\sigma}, \tag{4.1}$$

where σ is the standard deviation of the CTQ. Eq.(4.1) readily offers an interpretation of the C_p: it measures how accurate is our process compared to a 3-sigma process: $C_p = 1$ for 3-sigma, $C_p = 2$ for 6-sigma, etc.

Clearly C_p is a process parameter, that needs to be estimated.

Definition 18 ($\hat{C}_p$)**.**

$$\hat{C}_p := \frac{USL - LSL}{6\hat{\sigma}}, \tag{4.2}$$

where $\hat{\sigma}$ is some estimate of the standard deviation of the CTQ.

The most natural $\hat{\sigma}$ is the sample standard deviation s, but we will explore other options in the following.

There is a relation between C_p and the probability of non-conformance. To explore this relation we introduce the following notation:

Collecting Notation
$T := (USL + LSL)/2$, the target value.
$\delta := (USL - LSL)/2$, the specification tolerance. $USL = T + \delta, LSL = T - \delta$.
$\mu := \mathbf{E}[CTQ]$, the expected CTQ.
$p_{NC} := 1 - P(CTQ \in [LSL, USL])$, the probability of non compliance.

With our new notation Eq.(4.1) is now $C_p = \frac{\delta}{3\sigma}$. Assuming $CTQ \sim \mathcal{N}(\mu, \sigma^2)$, and that the process is centred so that $\mu = T$, then C_p is related to p_{NC} via

$$p_{NC} = 2\,\Phi(-3C_p) \tag{4.3}$$

As a sanity check, we check this relation for a 3-sigma process. A 3-sigma process implies that $C_p = 1$, and Eq.(4.3) returns $p_{NC} = 0.0027$, as we have already seen in the introduction (Section 1.4.5).

Montgomery (2007) recommends the following C_p values:

	C_p Value	Implied ppm
Existing processes	1.33	66
New processes	1.50	6.8
Safety, strength, or critical parameter, existing process	1.50	6.8
Safety, strength, or critical parameter, new process	1.67	0.5
Six Sigma quality process	2.00	0.002

To derive Eq.(4.3), and thus the ppm column in the table, we had to call upon several assumptions. Namely:

1. The process is in statistical control, i.e. μ is fixed over time
2. The CTQ has a normal distribution.
3. The process is centred, i.e. $\mu = T$.
4. T is mid-way between LSL and USL.

4.1.1 Non-Conformance for a Non-Gaussian CTQ

The first assumption we will now relax is the Gaussianity of CTQ. We start by checking what is the fallout rate, if we were completely wrong about the distribution of the CTQ. Chebyshev's inequality provides a universal bound on p_{NC} for a $C_p = 1$ process.

Theorem 4.1.1 (Chebyshev's inequality)**.** *For any random variable* $\mathbf{x}$*, with* $\mu := \mathbf{E}[\mathbf{x}]$*, and* $\sigma^2 := \mathbf{E}[(\mathbf{x} - \mu)^2]$*, then*

$$P(|\mathbf{x} - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{4.4}$$

If $C_p = 1$ then $\delta = 3\sigma$. Plugging $k = 3$ in the inequality returns $p_{NC} < 0.11$. This means that a 3-sigma process, assumingly with $2,700ppm$, may actually have $111,111ppm$, if we were wrong about the Gaussianity assumption.

The moral of the story is that by assuming the correct distribution of the CTQ, we may save a lot of resources. We can assume normality, as we typically do, but there are other alternatives:

1. Transformations: it is quite possible that the CTQ is not Gaussian in its original scale, but it is Gaussian in a different scale. You should always inspect a QQnorm (Sec. 3.1.1) plot after a *log* or *sqrt* transformation.
2. Assume a different distribution: We derived $p_{NC}(C_p)$ (eq. 4.3) under a normality assumption, but it may certainly be derived for different distributions.
3. The denominator of C_p is a range that leaves 0.00135 probability of the Gaussian tail outside the range. When relaxing the normality assumption, σ is no longer related to the tail probability as it was before. We may still, however, directly plug $CTQ_{0.00135}$ and $CTQ_{1-0.00135}$ quantiles to get a CPR in the same spirit of the C_p. This is known as the $C_p(q)$ index we now define.

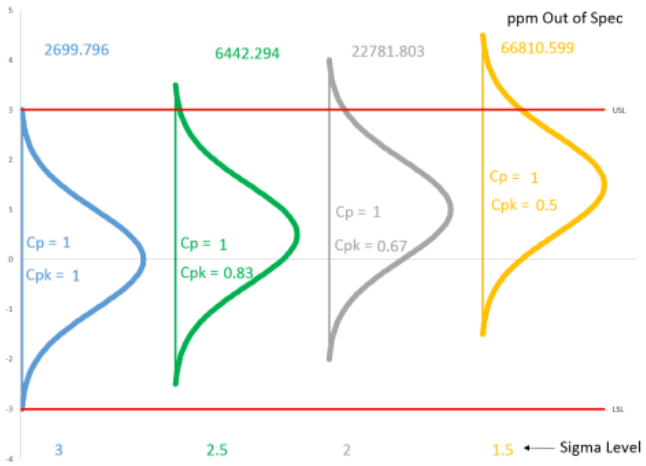


Figure 4.1: C_{pk} and C_p.
`https://www.spcforexcel.com/knowledge/process-capability/interactive-look-process-capability`.

Definition 19 ($C_p(q)$)**.**

$$C_p(q) := \frac{USL - LSL}{CTQ_{1-0.00135} - CTQ_{0.00135}} \tag{4.5}$$

4.1.2 Process Capability of a Non-Centred Process

We will now relax the assumption of $\mu = T$, while still in statistical control.

Definition 20 (C_{pk})**.**

$$C_{pk} := \min\{C_{pu}, C_{pl}\} \tag{4.6}$$

where $C_{pu} := \frac{USL-\mu}{3\sigma}$ and $C_{pl} := \frac{\mu-LSL}{3\sigma}$.

For a non-centred process, this definition is more informative on the probability of non-compliance. Indeed, Eq.(4.3) will not hold, but we can derive an updated version:

$$p_{NC} \approx \Phi(-3C_{pk}). \tag{4.7}$$

Generally, $C_{pk} \leq C_p$, with equality holding for centred processes ($\mu = T$). An illustration is given in Figure 4.1.

The C_{pk} index is motivated by conserving the relation between the index and the fallout rate p_{NC}, which is not captured by C_p when the process is not

centred. Indeed if C_p can be interpreted as "accuracy compared to a (centred) 3-sigma process", then C_{pk} can be interpreted as "accuracy compared to a (equally non-centred) 3-sigma process". The 3-sigma process is used as a benchmark for historical reasons. In practice, it is actually very common to report the sigma-level as a capability index. This implies the 1-sigma process as a benchmark: $\min\{\frac{USL-\mu}{\sigma}, \frac{\mu-LSL}{\sigma}\}$. Three-sigma would thus be reported as 3, 6-sigma as 6, etc.

Another index, known as C_{pm}, or the *Taguchi capability index*, also deals with the non centring but slightly differently. It is motivated by the observation that for $\mu = T$, then $\sigma = \sqrt{\mathbf{E}\left[(CTQ - T)^2\right]}$. This relation does not hold if $\mu \neq T$, which leads us to the following definition:

Definition 21 (C_{pm})**.**

$$C_{pm} := \frac{USL - LSL}{6\sqrt{\mathbf{E}\left[(CTQ - T)^2\right]}} \tag{4.8}$$

$$= \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - T)^2}} \tag{4.9}$$

$$= \frac{C_p}{\sqrt{1 + (\frac{\mu - T}{\sigma})^2}}. \tag{4.10}$$

Eq.(4.10) readily shows that just like the C_{pk}, then $C_{pm} \leq C_p$.

4.1.3 Interval Estimation for Capability Indexes

Since the various process capability indexes are merely population parameters, we can also construct confidence intervals (CIs) for them, which may be very informative if sample sizes are small. The simplest case is that of C_p. Being a monotone transformation of σ, we can call upon confidence intervals for the variance of a normal population, so that with probability $1 - \alpha$:

$$C_p \in \left[\hat{C}_p\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}}, \hat{C}_p\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\right]. \tag{4.11}$$

This equation is simply derived from Eq.(4.12). Intervals for the other capability indexes, are available in Montgomery (2007) and references therein.

4.1.4 Testing Hypotheses on Capability

Consider a supply contract, which requires production to have $C_p > 1.5$. It may easily be the case, that the $C_p > 1.5$, even if $\hat{C}_p < 1.5$, especially if the

sample size is small. It thus makes a lot of sense, to design hypothesis tests on process capabilities. We observe that for an i.i.d. sample from a Gaussian population, where $\hat{C}_p$ is estimated with s, then

$$(n-1)\left(\frac{C_p}{\hat{C}_p}\right)^2 \sim \chi^2_{n-1} \tag{4.12}$$

so that $(n-1)\left(\frac{C_p}{\hat{C}_p}\right)^2$ may serve as test statistic.

Example 1 (C_p test for 6-sigma compliance)**.**

$$H_0 : C_p \leq 2$$
$$H_1 : C_p > 2$$
$$(n-1)\left(\frac{2}{\hat{C}_p}\right)^2 \overset{H_0}{\sim} \chi^2_{n-1}$$

so that the $1-\alpha$ rejection region in $\hat{C}_p$ scale is

$$\hat{C}_p > \sqrt{\frac{4(n-1)}{\chi^2_{n-1,\alpha}}}.$$

Note that we should not be testing this hypothesis with the confidence interval in Eq.(4.11) because this particular hypothesis is directional. Now for a more general case:

$$H_0 : C_p \leq a; H_1 : C_p > a \Rightarrow \text{reject if } \hat{C}_p > \sqrt{\frac{a^2(n-1)}{\chi^2_{n-1,\alpha}}},$$
$$H_0 : C_p \geq a; H_1 : C_p < a \Rightarrow \text{reject if } \hat{C}_p < \sqrt{\frac{a^2(n-1)}{\chi^2_{n-1,1-\alpha}}},$$
$$H_0 : C_p = a; H_1 : C_p \neq a \Rightarrow \text{reject if } \hat{C}_p < \sqrt{\frac{a^2(n-1)}{\chi^2_{n-1,1-\alpha/2}}} \text{ or } \hat{C}_p > \sqrt{\frac{a^2(n-1)}{\chi^2_{n-1,\alpha/2}}}$$

4.1.5 Process Performance Indices

Process *performance* indices measure compliance to specification of a process out of statistical control. These include the P_p and P_{pk} indices. Besides mentioning their existence, we will not give them further attention, since we adopt Montgomery (2007)'s view that their use is strongly discouraged.

Remark 2. At this point, I hope you are wondering why isn't the fallout rate, p_{NC}, not used as a capability index. Well, it is! 1It is simply not called a "capability index", simply because the term is reserved to C_p, C_{pk}, C_{pm} etc.

Chapter 5

Statistical Process Control

Statistical process control (SPC), a.k.a. *change detection*, or *novelty detection*, deals with the quantitative analysis of a "process", which may be a production line, a service, or any other repeated operation. As such, SPC may be found in the Analyze, Improve, and Control stages of the DMAIC cycle. The purpose of the SPC, in the terms coined by Shewhart, is to seperate the variability in the process into *assignable* causes of variation and *chance* causes of variation. Assignable are also known as *special* causes, or simply *signal*. Chance causes are also known as *common* causes of variation, or *haphazard* variability, or simply *noise*.

Change Detection

Causes of variation

A process is said to be in *statistical control* if all its variation is attributable to chance causes. If this is not the case, we call it *out of control* and we will seek the assignable causes, so that we may reduce variability by removing them. All the statistical tools of chapters 2 and 3 may be called upon for this endeavour but in this chapter we focus on one particular such tool- the *control chart*. We start with the *Shewhart control chart*, in which each value is charted using different data, from different periods.

Shewhart Chart

Extra Info 6 (A mathematically rigorous treatment of SPC)**.** The contents of this chapter is mostly derived from Montgomery (2007). For a more mathematically rigorous treatment of the topic see Basseville et al. (1993). For an **R** oriented exposition of the topic, see Qiu (2013).

5.1 A soft start. The $\bar{x}$-chart

We demonstrate the concepts and utility of control charts with the simplest, yet most popular of them all, the $\bar{x}$-chart. The chart borrows its name from the fact that it is essentially a visualization of the time evolution of the average ($\bar{x}$) of the CTQ. The chart is also augmented with visual aids that

help in determining if the process is *in control*, i.e., if it is consistent with its own history.

Remark 3 (Control Charts and Capability Analysis)**.** Control charts have no information on the specifications of the process, merely on its own history. Process capability analysis may, however, benefit from the ideas of control charts, as we explain in Chapter 9.

An illustration of a $\bar{x}$-chartis given in Figure 5.1. The ingredients of this chart are the centerline, lower and upper control limits (LCL, UCL), and $\bar{x}$ evolving in time. If at each period $t = 1, \ldots, \tau$ we compute the average of n samples, we denote

$$\bar{x}_t := 1/n \sum_{i=1}^{n} x_{it}.$$

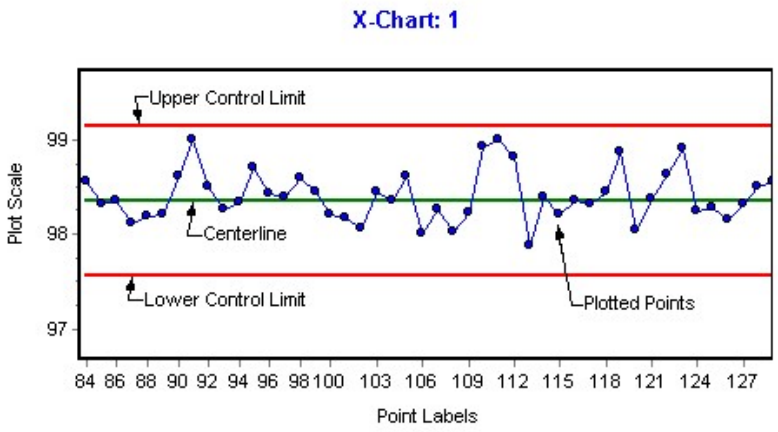


Figure 5.1: $\bar{x}$-chart.
`https://mvpprograms.com/help/P-mvpstats/spc/WhatAreControlCharts`

Figure 5.1 makes it evident $\bar{x}$-chartrequires us to make several design decisions. A standard design decision is setting the center line as the grand average of the process:

$$\hat{\mu}_0 = 1/\tau \sum_{t=1}^{\tau} \bar{x}_t, \tag{5.1}$$

where μ_0 denotes the in-control mean of the process. Notation originates from treating the in-control process as a null hypothesis, as it should be thought of.

If it is unclear to you, how may we compute the grand average of a process that is still evolving and has not finished, you are right! We thus introduce the idea of *Phase I* and *Phase II*. Initially we assume the process it out of control, we identify and remove assignable causes of variation, until we are left with a "well-behaved" subset of data points, we believe to be in-control. We call this Phase I, and we use it to initialize required quantities such as the centre line. Eq.(5.1) thus implies that in Phase I we were left with τ samples assumingly in statistical control. After the chart has been calibrated, and major assignable sources of variability removed, we can finally start monitoring the process, known as Phase II.

Phase I/II

Other design decisions to be made are:

1. UCL and LCL (Do not confuse with USL and LSL!).
2. Sample size in each sample (n).
3. The within period sampling scheme, known as *rational groupings*.
4. The between-period sampling scheme, notably the *frequency of samples*.
5. Other stopping rules.

These design decisions ultimately govern the error rates of the chart, which in turn, incur financial costs. For now we will restrict attention to type I/II error rates, until Section 5.4 where we consider these choices as economical optimization problems.

For ease of exposition, control chart design is demonstrated for the $\bar{x}$-chart, but equally applies to other control charts, presented in Section 5.5. We start by a type I error rate analysis. Denote α_t the false alarm probability at period t. How do our design choices affect α_t?

$$\begin{aligned}
\alpha_t &:= 1 - P_{H_0}(\bar{x}_t \in [LCL, UCL]) && (5.2)\\
&= 2P_{H_0}(\bar{x}_t < LCL) && (5.3)\\
&= 2P_{H_0}(Z < \frac{LCL - \mu}{\sigma_{\bar{x}}}) && (5.4)\\
&= 2P_{H_0}(Z < -L) && (5.5)\\
&= 2\Phi(-L) && (5.6)
\end{aligned}$$

The above follows from assuming that $UCL := \mu_0 + L\sigma_{\bar{x}}, LCL := \mu_0 - L\sigma_{\bar{x}}$, $\mathbf{x}_{it} \sim \mathcal{N}(\mu, \sigma^2)$, and denoting $\sigma_{\bar{x}} := \frac{\sigma}{\sqrt{n}}$. A typical design choice is $L = 3$, known as *3-sigma control limits*, implying a false alarm rate of $\alpha_t = 0.0027$. Since we assumed the process is fixed over time, then so is α_t and we can simply write $\alpha_t = \alpha$.

3-Sigma Control Limits

Remark 4 (3-Sigma Control Limits vs. 3-Sigma Capability)**.** Do not confuse these two similar ideas. 3-Sigma Control Limits is a statement on the false alarm rate of a process with respect to its own **history**. 3-Sigma Capability is a statement on the non-compliance rate of a process with respect to its **specification**.

A power analysis for our design choices follows the same lines. Denote by H_1 the out-of-control distribution, β_t the type-II error rate, and $\pi_t = 1 - \beta_t$ the power, at period t. We then have

$$\pi_t := 1 - P_{H_1}(\bar{x}_t \in [LCL, UCL]) \tag{5.7}$$

and the rest follow from the distribution of $\bar{x}$ when the process is out of control. Since the out-of-control shift is (asumingly) stable, we can again omit the time index and write $\pi = \pi_t$. Assuming the out-of-control process is a shift of magnitude $k\sigma$, i.e.: $\mathbf{x} \sim_{H_1} \mathcal{N}(\mu_1, \sigma^2); \mu_1 = \mu_0 + k\sigma$, we plot in Figure 5.2, the detection power of a 3-sigma $\bar{x}$-chart, as a function of k. This is known in the statistical literature as a *power function*, and in the engineering literature as the *true positive rate* operator characteristic (TPR-OR).

Operator Characteristic

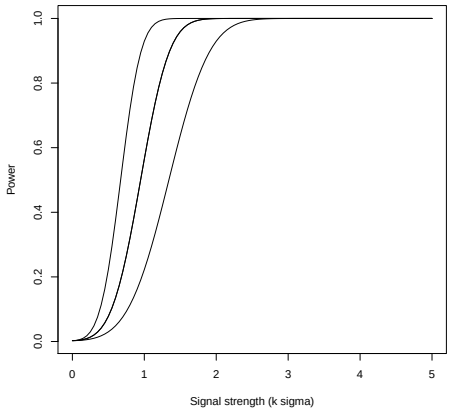


Figure 5.2: Power function of the 3-sigma $\bar{x}$-chartwith $n = 5, 10, 20$ and $\mu_1 = \mu_0 + k\sigma$.

Extra Info 7 (Operator Characteristics)**.** Many operator characteristics have been proposed to study the performance of control charts, statistical tests, or binary classifiers in general. You may be already familiar with some, such as the ROC-curve. The curious reader is referred to Wikipedia (2015f) for more information.

An important related quantity is the *average run length* (ARL), which is the expected number of periods between two crossings of control limits, i.e., the expected periods between alarms. We denote by ARL_0 the ARL when the process is under statistical control, and ARL_1 otherwise[1]. For Shewhart charts, where $\bar{x}_t$ are typically statistically independent, then clearly the number of periods until a crossing is geometrically distributed. Using the expectation of a geometric random variable we can conclude that

ARL

$$ARL_0 = 1/\alpha, \tag{5.8}$$

$$ARL_1 = 1/\pi. \tag{5.9}$$

Clearly we can convert to time units by multiplying the ARL by the duration of sampling interval. This is known as the *average time to signal* (ATS). It is quite common to design a control chart so that it achieves a particular ATS_0.

ATS

Remark 5 (ARLs more important than type-I errors)**.** In the case of Shewhart charts, there is a simple mapping between ARL and type I error rates. This need not be the case for general control charts. Since type I errors are certain, if the process runs long enough, then it is actually the ARL the design parameter of importance, and not the type-I error.

Now assume that we are unhappy with our control chart. It simply makes too many false alarms, or takes too long to detect loss of statistical control. What can we do about it? Well, this is exactly the same question as when increasing the power or lowering the type I error of a statistical hypothesis test. This is obviously no coincidence, since control charts are nothing but a statistical test! Here are some action courses:

1. Increase L. This is the same as shrinking the rejection region: it will decrease the false alarm rate, at the cost of power.

2. Increase n. Brilliant! Statistically, there is nothing to lose. It may, however, cost time and money.

3. Increase the sampling frequency. Brilliant again! Nothing to lose, except time and money...

4. Change the sampling scheme within period. We elaborate on this in Section 5.1.2.

[1]Note that it is implied that the process has a *stable* distribution, even though it is out if control.

5. Add other stopping rules: this acts just like growing the rejection region. It will increase power, at the cost of type I error. We elaborate in Section 5.1.3.

6. Pool together more time points or more CTQs. We elaborate on this in sections 5.2 and 5.3, respectively.

5.1.1 Control Limits and the Alarm Rate

As previously discussed, L governs the tradeoff between type I and type II errors, or sensitivity versus specificity. It is very common to set $L = 3$. For a normally distributed CTQ, this implies $2,700$ false alarms per million periods. This also implies an ARL_0 of $1/\alpha \approx 370$ periods, which is conveniently, about a year when sampling once a day. We may obviously, discard this $L = 3$ convention, and directly set UCL and LCL so they guarantee some desirable false alarm rate, or ARL.

If normality of $\bar{x}_t$ can be assumed, then one may estimate σ from phase I, and set LCL and UCL by finding the L that solves $2\Phi(-L) = \alpha$. If normality cannot be assumed, there are many ways to go about. Here are some options:

1. Increase n: even if $\mathbf{x}_{it}$ is non normal, for large enough n, then $\bar{x}_t$ will be via the central limit theorem (CLT).

2. Use empirical quantiles: If phase I was returned enough data, then we may estimate $\mathbf{x}_{\alpha/2}$ and $\mathbf{x}_{1-\alpha/2}$ using the empirical quantiles of phase I. The false alarm rate will be α since $P(\mathbf{x} \notin [\mathbf{x}_{\alpha/2}, \mathbf{x}_{1-\alpha/2}]) = \alpha$.

3. If some other distribution can be assumed then we may compute the false alarm rate of particular limits either analytically, or computationally (by simulation).

5.1.2 Rational Groupings

Recall that at each period we compute the average of n samples. How should we draw these samples? At the same time from the same machine? At different times from the same machine? Many configurations are possible, and the correct approach depends on the type of out-of-control behaviour one seeks. *Rational groupings* merely reminds us to sample "rationally" in each period. Quoting Montgomery (2007)'s words of caution:

> ... we can often make any process appear to be in statistical control just by stretching out the interval between observations in the sample.

5.1.3 Other Stopping Rules

The assumption that we may only create alarms if $\bar{x}$ exceeds some control limits is needlessly restrictive. A first relaxation is by allowing multiple control regions. It is quite common to define *warning limits* which only call for inspection, and *action limits*. Each may have its own alarm rate. We may even change the sapling scheme if limits are breached. Increasing the sampling rate once the warning limits have been breached is known as *adaptive sampling*, or *variable sampling*.

Adaptive Sampling

Another approach is to define multiple sets of stopping rules. Here is an example:

1. One or more points outside of the 3-sigma control limits.
2. Two of three consecutive points outside the 2-sigma warning limits but still inside the 3-sigma control limits.
3. Four of five consecutive points beyond the 1-sigma limits.
4. A run of 8 consecutive points on one side of the center line.

The above set of rules is known as the Western Electric Rules, a.k.a. , the *WECO* rules. Augmenting the set of rules is the same as increasing a rejection region. It adds more sensitivity, at the cost of false alarms. If the rules are properly selected, the gain in sensitivity is worth the increase in false alarms.

WECO

As a quick exercise, we may compute α for m independent rules, each with α^* type I error:

$$\alpha = 1 - (1 - \alpha^*)^m. \tag{5.10}$$

Having 4 rules, like WECO, each at $\alpha^* = 0.0027$ implies that we actually have $\alpha = 0.01$ and $ARL_0 \approx 93$. For daily sampling of an in-control process, this means an alarm every quarter, and not every year.

Extra Info 8 (Stopping Rules)**.** There are many sets of stopping rules. These include WECO, Nelson, AIAG, Juran, Hughes, Duncan, Gitlow, Westgard, and more. See `http://www.quinn-curtis.com/spcnamedrulesets.htm` for a quick review.

5.2 Pooling Information Over Periods

Assume an out-of-control process is simply a mild shift of the controlled-process. This shift may be hard to detect in Shewhart chart, especially if n

is not too large (as seen in Figure 5.2). If the shift persists over periods, we may gain power, i.e., sensitivity, by pooling several periods together. We now present several ways to pool information from history. These are typically applied in Phase II, where out-of-control processes are expected to have only mild shifts, and not major ones as in Phase I.

Remark 6 (No longer Shewhart)**.** The name *Shewhart control chart* is reserved to charts plotting one period at a time. When several periods are pooled together, we will no longer call this "Shewhart".

Remark 7 (One observation at a time)**.** The following charts have a continuous flavour. As such, it is both favourable, and common, to compute them using one observation at a time, meaning that $n = 1$.

5.2.1 Run Test Chart

[TODO]

5.2.2 Moving Average Chart (MA)

One way to pool information from different periods is by a *moving average*.

Definition 22 (Moving Average- MA)**.** The moving average in a window w, at period t, is defined as

$$M_t := \frac{x_t + \cdots + x_{t-w+1}}{w} \tag{5.11}$$

Assuming $x_t \sim \mathcal{N}(\mu, \sigma_x^2)$ then clearly

$$M_t \sim \mathcal{N}\left(\mu, \frac{\sigma_x^2}{w}\right). \tag{5.12}$$

The control limits on M_t are typically

$$UCL := \mu_0 + 3\sigma_{M_t} = \mu_0 + 3\frac{\sigma_x}{\sqrt{w}} \tag{5.13}$$

$$UCL := \mu_0 - 3\sigma_{M_t} = \mu_0 - 3\frac{\sigma_x}{\sqrt{w}}. \tag{5.14}$$

The false alarm rate of this criterion is trivially $\alpha = 0.0027$. The ARL_0 is no longer simple to compute. This is because the pooling of periods has compromised independence between periods, and Eqs.(5.8,5.9) are no longer valid. Do not despair as the ARL may still be computed. You can always use simulation to compute it, or try using the `spc` **R** package.

We are free to choose the magnitude of w. If w is too small, there is no real pooling from history. At the limit, where $w = 1$, we are back to the classical Shewhart chart. If w is too large, then each new observation has very small importance, and it may take a long time to detect a shift. Which is the right intermediate value of w, is left for you to decide.

5.2.3 Exponentially Weighted Moving Average Chart (EWMA)

The moving average gives all observations the same importance. We want to change this, giving more importance to new observations so that we may capture drifts quickly when they occur. The *Exponentially Weighted Moving Average* (EWMA), a.k.a. the *geometric moving average* (GMA), does just that.

GMA

Definition 23 (EWMA)**.** For a fixed $\lambda \in [0, 1]$, the *exponentially weighted moving average* (EWMA) is defined as

$$z_t := \lambda x_t + (1 - \lambda) z_{t-1} \tag{5.15}$$

By recursive substitution, we have

$$z_t = \lambda \sum_{j=0}^{t-1} (1 - \lambda)^j x_{t-j} + (1 - \lambda)^t z_0, \tag{5.16}$$

and

$$z_t \sim \mathcal{N}\left(\mu_0, \sigma_{z_t}^2\right), \tag{5.17}$$

$$\sigma_{z_t}^2 = \sigma_x^2 \left(\frac{\lambda}{2 - \lambda}\right) (1 - (1 - \lambda)^{2t}). \tag{5.18}$$

Eq.(5.17) may be used to construct control limits for EWMA: It is however, more economic to observe that for large λ and t: $(1 - (1 - \lambda)^{2t}) \approx 1$ so that we may use

$$\begin{aligned} UCL &:= \mu_0 + 3\sigma_{z_t} \approx \mu_0 + L \sqrt{\sigma_x^2 \left(\frac{\lambda}{2 - \lambda}\right)}, \\ UCL &:= \mu_0 - 3\sigma_{z_t} \approx \mu_0 - L \sqrt{\sigma_x^2 \left(\frac{\lambda}{2 - \lambda}\right)}, \end{aligned} \tag{5.19}$$

with $L = 3$ being the typical choice. By now, you should immediately know what is the false alarm rate of these limits. By now, you should also know

that because of the dependence between z_t's, computing the ARL is not as simple as for Shewhart charts. The `xewma.arl()` **R** function, in package `spc`, permits doing so easily. Its output for various λ and L is illustrated in Figure 5.3.

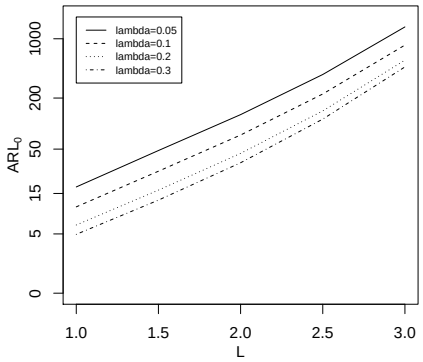


Figure 5.3: ARL_0 for EWMA.
Code from `http://users.phhp.ufl.edu/pqiu/research/book/spc/r-codes/fig53.r`

In the MA chart, we used the choice of w to balance between quick response (small w) and sensitivity (large w). EWMA has no window-width parameter, since it looks into all of history. On the other hand, we can control it by choosing λ. Large λ gives more importance to the present. At the limit, $\lambda = 1$, EWMA collapses to a standard Shewhart chart.

5.2.4 Filtered Derivative Chart

[TODO]

5.2.5 CUSUM Chart

The *cumulative sum* chart is similar to the EWMA in that it pools information from the history. The CUSUM simply sums all past deviations from the centre line. If the process is in control, deviation will cancel each other, and their sum will vary around 0. If the process is out of control, a drift will

appear. The statistic to be plotted is

$$C_t := \sum_{j=0}^{t}(x_j - \mu_0) = C_{t-1} + (x_t - \mu_0) \tag{5.20}$$

Observing that when under control then $C_t \sim \mathcal{N}(\mu_0, t\sigma_x^2)$, we could set

$$\begin{aligned} UCL &:= \mu_0 + 3\sigma_{C_t} = \mu_0 + L\sqrt{t\sigma_x^2}, \\ UCL &:= \mu_0 - 3\sigma_{C_t} = \mu_0 - L\sqrt{t\sigma_x^2}. \end{aligned} \tag{5.21}$$

You may encounter these limits in your favourite software (`qcc` package in **R** ?), but it less often discussed in the literature. This is because CUSUMs were introduced by Page (1954), which offered different control limits. Montgomery (2007) adopts Page's view and presents limits in two forms: the *decision interval* (DI) form, a.k.a. the *tabular* form, and the graphical form known as a *V-mask*. These two control limits are equivalent. Before we present them, we try to offer some intuition for the difference between the limits in Eq.(5.21) and those of Page (1954).

The fundamental difference between the control limits of Page (1954), and the ones presented until now, is that Page designed limits for the particular history of each process, while the limits until now, including Eq.(5.21) do not adapt to the particular history of the process. As such, Page's control limits are said to be *adaptive*.

Adaptive Contol Limits

[TODO: explain V-mask and tabular]

5.2.6 Shiryaev-Roberts Procedure

[TODO]

5.2.7 Combined Shewhart and Running Window Charts

Well sure- if you want to enjoy the quick detection of Shewhart charts, and the sensitivity of running windows charts, you may certainly design charts that marry these ideas. False alarm rates and ARLs should be computed mutatis mutandis.

5.3 Multivariate Control Charts

Example 2 (Intensive Care Unit)**.** Consider an intensive care unit. The CTQs are the patient's blood pressure, temperature, etc. We want to sound

an alarm if the patient's condition deteriorates. Clearly, we can apply the univariate methodology above on each CTQ. It is possible, that the deterioration is mild, so that it is not picked up by any CTQ individually (low power), but could have been noticed were we to aggregate signal over various CTQs. This is the concern of the current section.

5.3.1 Mass Univariate Control

A first natural approach is to raise an alarm when **any** of the processes exceeds its respective control limits. For m independent processes, with false alarm rate α each, then the joint false alarm rate is

$$\alpha^* = 1 - (1 - \alpha)^m. \tag{5.22}$$

Clearly we could set $\alpha = 1 - \sqrt[m]{1 - \alpha^*}$, so that the joint false alarm rate is under control, but we would not be enjoying the added sensitivity of pooling many CTQs together.

5.3.2 Hotteling's T^2

Hotteling's T^2 statistic is a generalization of the t-test. To see this we write the t-statistic in the following weird form:

$$t_t(x) = (\bar{x}_t - \mu_0)(s_t^2(x))^{-1}(\bar{x}_t - \mu_0). \tag{5.23}$$

This notation readily extends to the multivariate case. For p CTQs, then $\bar{x}_t$ and μ_0 are p-length vectors, and $s_t^2(x)$ is replaced with the $p \times p$ matrix $\hat{\Sigma}_t(x)$.

Definition 24 (Hotteling's T^2)**.**

$$T_t^2 := n(\bar{x}_t - \hat{\mu}_0)'\hat{\Sigma}_t^{-1}(\bar{x}_t - \hat{\mu}_0). \tag{5.24}$$

To derive the control limits, we will be assuming that $\mathbf{x}_{it}$ is p-variate Gaussian distributed, $\mathcal{N}(\mu_0, \Sigma_{p\times p})$. We need to be very clear, however, on how μ_0 and Σ_t^{-1} are estimated. We thus consider the following scenario:

1. T^2 is computed at Phase II with a single observation at a time ($n = 1$)
2. μ_0, Σ_t^{-1} were estimated from m observations in Phase I.

Eq.(5.24) thus yields

$$T_t^2 := (x_t - \hat{\mu}_0)'\hat{\Sigma}_t^{-1}(x_t - \hat{\mu}_0). \tag{5.25}$$

and

$$T_t^2 \sim_{H_0} F_{p,m-p}\frac{p(m-1)(m+1)}{m(m-p)} \overset{m\to\infty}{\rightsquigarrow} \chi_p^2. \tag{5.26}$$

We can thus construct the control limit for this scenario:

$$UCL := F_{1-\alpha,p,m-p}\frac{p(m-1)(m+1)}{m(m-p)} \approx \chi_{1-\alpha,p}^2. \tag{5.27}$$

Note that the T^2 statistic is *non directional*: it will increase in the presence of both positive and negative drift.

We may consider the distribution of T^2 under many configurations: where $\mu_0, \hat{\Sigma}_t^{-1}$ are estimated at Phase II, where we take n observation in each period, etc. Conveniently enough, if the estimation of μ_0 and Σ is based on many observations (compared to p) then under all scenarios

$$T_t^2 \rightsquigarrow \chi_p^2, \tag{5.28}$$

and thus

$$UCL \approx \chi_{1-\alpha,p}^2. \tag{5.29}$$

Extra Info 9. For exact results and references on the various scenarios described, we refer the reader to (Qiu, 2013, Ch.7).

Since the above limits have an (approximate) type-I error rate of α, and the periods are typically independent, then we can readily apply Eq.(5.8) to compute ARL_0.

5.3.3 Other Pooling Statistics

Hotteling's T^2 implicitly targets a weak signal on many coordinates. In the signal detection literature, this is known as a *dense* signal. In our intensive care unit example (Example 2), it may be the case the a patient's deterioration is mildly manifested in many CTQs. It may however, be the case, that deterioration is manifested in a small subset of CTQs. This is further emphasized in Example 3.

Example 3 (Cyber Monitoring System)**.** Consider a server farm. All servers dump their status into logs. These include CPU loads, temperature, network I/O, etc. The administrator is worried about an imminent attack, and it thus parsing the logs for CTQs, and inspecting the on-line status on his dashboard. He knows that the cyber-attacker is no amateur, so that if any fingerprint is left in the logs, it will be very subtle.

The intensive care example, and the cyber security example, motivate our search for a statistic, that unlike Hotteling's T^2 , is sensitive to *sparse* signals. I.e., a signal that is manifested in very few of the p CTQs. For this purpose, we merely offer several candidate multivariate statistics, with references where appropriate.

Sparse Signal

Max pooling Where we control the process using the max over CTQs. Useful when we expect signal in a handful of CTQs.

Higher Criticism Where pooling is performed using the Higher Criticism statistic. Appropriate for an intermediate *rare-weak* sparsity pattern. See details in (Jin et al., 2005).

Skewness Statistic Where pooling is performed using the empirical third moment. Appropriate for an intermediate sparsity pattern. See details in (Jin et al., 2005).

5.4 Economical Design of Control Charts

Up until now, our design of control charts was driven by type-I error rates, and ARLs. Economical consideration were merely implied. In this section, economical consideration take the driver's seat. We present a toy model, to demonstrate the economical optimization of design parameters in a $\bar{x}$-chart. Before beginning, a few remarks are in order.

Remark 8 (Economical Design of Control Charts)**.**

1. According to Montgomery (2007)

   > Saniga and Shirland (1977) and Chiu and Wetherill (1975) report that **very few practitioners** have implemented economic models for the design of control charts.

   Hmmmm.. Have things changed since 1977?

2. A comprehensive theoretical analysis of the optimization of a quality control system may be found in Girshick and Rubin (1952). Again, Montgomery (2007) is skeptic:

   > The optimal control rules are difficult to derive, as they depend on the solution to complex integral equations. Consequently, **the model's use in practice has been very limited**.

In light of the above skepticism, and following the lines of Duncan (1956), we aim at the modest goal of an economical optimal $\bar{x}$-chart. Our target function is optimizing the expected income per hour, with respect to the design parameters:

$$\max_{n,L,h}\{\mathbf{E}\left[C/T\right]\} \tag{5.30}$$

where C is the income between two productions halts, i.e., a *cycle*; T is the cycle duration; n is the number of samples per period; L governs the control limits via $UCL := \mu_0 + L\sigma_{\bar{x}} = \mu_0 + L\sigma_x/\sqrt{n}$; h is the time interval between sample periods.

We now need to establish how $\mathbf{E}\left[C/T\right]$ is related to n, L, h. Here is our set of assumptions and notation:

1. When in control (IC), production is centred on μ_0, assumed known.
2. When out of control (OC), $\mu_1 = \mu_0 \pm \delta\sigma_x$.
3. When OC, production may proceed (!).
4. Search and repair costs are not part of C.
5. OCs occur as a Poisson process, with rate λ events per hour. The expected time from a sampling to an OC events is thus
$$\tau := \frac{1-(1+\lambda h)e^{-\lambda h}}{\lambda(1-e^{-\lambda h})} \tag{5.31}$$
6. The power to detect an OC is
$$\pi := \Phi(-L-\delta/\sqrt{n}) + (1-\Phi(L-\delta/\sqrt{n})) \tag{5.32}$$
7. The false alarm rate
$$\alpha := 2\Phi(-L) \tag{5.33}$$
8. Because of the Poisson process assumption, $\mathbf{E}\left[C/T\right] = \mathbf{E}\left[C\right]/\mathbf{E}\left[T\right]$.
9. The expected cycle length:
$$\mathbf{E}\left[T\right] = \frac{1}{\lambda} + \frac{h}{\pi} - \tau + D \tag{5.34}$$
where $\frac{1}{\lambda}$ is time IC; $\frac{h}{\pi} - \tau$ is the time the process is OC until detection; D is a fixed time to identify the assignable cause.

10. The expected income per cycle

$$\mathbf{E}[C] = V_0\frac{1}{\lambda} + V_1\left(\frac{h}{\pi} - \tau + D\right) - a_3 - \frac{a_3' e^{-\lambda h}}{1-e^{-\lambda h}} - (a_1 + a_2 n)\frac{\mathbf{E}[T]}{h}$$

where V_0 is the net income per cycle when IC; V_1 is the net income when OC; $(a_1 + a_2 n)$ is the fixed and variable cost of taking a sample; a_3 is the cost of finding an assignable cause; a_3' is the cost of investigating a false alarm.

Given all the above, we may now plug Eq.(5.30) into our favourite numerical solver to find the optimal h, L, n.

5.5 Shewhart Charts With Other Test Statistics

We have been focusing on the $\bar{x}$-chartfor ease of exposition. There are, however, many cases where the mean is not an appropriate test statistic. Examples include:

1. A discrete CTQ, where only the number of non-compliances can be counted.

2. Where the departure from statistical control is not only a shift in μ.

The following charts are designed for those cases. Practically all of the ideas presented for the $\bar{x}$-chartmay be adapted to these other test statistics after appropriate adaptations. The reader is referred to Montgomery (2007) for the details.

5.5.1 R Chart

Where $\bar{x}$ is replaced by the range. Sensitive to variability changes. Popularized by its ease of implementation, in the pre-PC age.

5.5.2 s Chart

Where $\bar{x}$ is replaced by s. Sensitive to variability changes. Usually more sensitive than the range.

5.5.3 s^2 Chart

Like the s chart, only in variance-scale.

5.5.4 p and np Chart

Where $\bar{x}$ is replaced by the proportion (p), or number (np), of non-conforming units. Appropriate for attributes, i.e., discrete QTCs.

5.5.5 c Chart

Like a np chart, but where the number of nonconforming units is replaced with the total number of nonconformances, allowing multiple defects per unit.

5.5.6 u Chart

Like the c chart, but allowing a variable number of units per sample (varying n).

5.5.7 Regression Control Chart

In a *regression control chart*, the test statistic can be a regression coefficient. When compounded with multivariate charts, a regression control chart may accommodate several regression coefficient, or the residuals.

5.6 Extensions of Control Charts

We have been discussing the simplest of $\bar{x}$-chartchart, for ease of exposition. Several immediate extensions are possible. The reader is referred to Montgomery (2007) for details.

One sided charts Just like hypothesis testing, where we may consider non-directional or directional hypotheses, we may consider directional control chart. Obviously the control limits and the test statistic may need appropriate updates.

Autocorrelation and time series models The assumption of independence between sampling periods may be relaxed, by adopting a quantitative or qualitative model for temporal dependence.

Running windows The ideas of pooling information over periods with a MA, EWMA, and CUSUM may be extended to all type of Shewhart charts. Also, there are endlessly many period pooling schemes. You

are free to pick a set of weights, or convolve your data with a *causal filter*[2] can provide a pooling scheme.

Multiple Charts When inspecting a process, rarely does one inspect a single chart at a time. A typical dashboard would include several charts running in parallel, as depicted in Figure 2.10.

5.6.1 Cuescore Charts

[TODO]

[2] `https://en.wikipedia.org/wiki/Causal_filter`

Chapter 6

Design of Experiments

This chapter is devoted to the matter of designing experiments, and follows the lines of Cox and Reid (2000). A control chart may be seen as an on-line experiment alerting us when the milk goes sour, but it will not tell us why. When designing a product (remember DFSS 1.4.7), or once a control chart has signalled an alert, we will want to know what has influences our production, how to remove variability, thus optimizing production. In our SPC terminology, we will want to know what are the *causal effects* of our *controllable inputs* (or *factors*), on our *CTQ* (or *response*). The theory of discovering these effects is the theory of *design of experiments* (DOE). Its goal is to *screen* factors with no effect, to estimate effect sizes, and find optimal factor-level combinations; all these as efficiently as possibly.

Roughly speaking, the challenges in designing good experiment are:

1. Efficiency: extract the most information per sample.
2. Signal to noise: remove variability that might mask factor effects.
3. Bias: avoid uncontrolled factor-effects from being "absorbed" into controlled factor effects.

Before we dig in, several matters should be emphasized:

Randomization Randomization is fundamental to our purpose. This because the idea of an *effect* implies causality. Any inference we make, is causal, which is the inference we need for controlling a process. It is the mechanism of randomization, that allows us to conclude that inferred correlations are causal, and not merely statistical. For a treatment of causal inference in *observational data*, i.e., without randomization, see Rosenbaum (2002).

Pre-experiment In this text we take it for granted that the purpose of the experiment is well known, and the candidate factors defined. We are fully aware, as should be the reader, that in application this is a non-trivial luxury. Indeed, a lot of planning, and domain-knowledge go into the selection of factors, their candidate levels, etc.

Power Analysis Part of the pre-experiment may include a power analysis. The pre-experiment power analysis will typically be very approximate, and rely on many assumptions. It is still important, as it gives an idea of the feasibility of an experiment, and avoids wasting resources.

No Textbook Solution We will present many design ideas and principles, yet it should be emphasized that real life problems rarely obey textbooks. You should thus feel free, and even obliged, to think about your particular problem and adapt the experiment as you best see fit.

Data Analysis In this text, we only discuss the **design** of the experiment, and not the **analysis** of the data. This is a non-standard choice as DOE is typically presented alongside the *analysis of variance* (ANOVA) framework. We decouple the two since: Cox and Reid (2000) do so, these are two different thing, and finally, because the ANOVA framework may be easily replaced by the framework of *linear models*, *mixed models*, *variance components*, and possibly others. There is a vast literature focusing on the analysis method. If asked, this author may recommend Hocking (1985), which presents both the ANOVA terminology, and the linear models terminology. That book, however, may be hard to come by, so feel free to ask me for other references if required.

ANOVA

Linear Models, Mixed Models, Variance Components

6.1 Terminology

The following list is compiled from Mason et al. (2003). Many, if not most of the following terms, originate in R.A. Fisher's seminal book "The Design of Experiments" (Fisher, 1960). As usual, when old ideas get new names, we try to emphasize this in the text.

Experimental Unit Entity on which a measurement or an observation is made; sometimes refers to the actual measurement or observation.

Homogenous Experimental Unit Units that are as uniform as possible on all characteristics that could affect the response.

Factors A controllable experimental variable that is thought to influence the response. In the language of SPC: *a controllable input.*

Level Specific value of a factor.

Treatment The particular factor-level combination applied to an experimental unit. A.k.a. *manipulation*.

Factor Encodings The numerical encoding of factor levels. Of minor importance for designing. Of major importance for analysis. Two level factor encodings include:

1. Effect coding: where levels are encoded with $-1, 1$ returning orthogonal design matrices for balanced designs.
2. Dummy coding: where levels are encoded with $0, 1$ returning easily interpretable coefficients.

Experimental Region All possible factor–level combinations for which experimentation is possible. A.k.a. *factor space*, and *design region*.

Design Matrix A matrix description of an experiment that is useful for constructing and analyzing experiments.

Response The CTQ.

Main Effect Change in the expected response between two factor–levels. We emphasize that effects, unlike simple population parameters, imply a causal relationship.

Interaction Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors.

Replication Repetition of an entire experiment or a portion of an experiment under two or more sets of conditions.

Covariate An uncontrollable variable that influences the response but is unaffected by any other experimental factors.

Design Complete specification of experimental test runs, including blocking, randomization, repeat tests, replication, and the assignment of factor–level combinations to experimental units.

Blocking Blocking, or *grouping*, is an experimental design technique that removes excess variation by grouping experimental units or test runs so that those units or test runs within a block are more homogeneous than those in different blocks. Blocking attributes are also known as *non specific factors*.

Non Specific Factors

Confounding When the design is such that several effects cannot be told apart. A.k.a. *aliasing*.

Repeat Tests Two or more observations that have the same levels for all the factors.

Balance Some symmetry in the combinatorial design of the experiment. In its simplest interpretation, a design where an equal number of units is assigned to each treatment.

Orthogonality Special simplifications of analysis and achievement of efficiency consequent on such *balance*.

6.2 Dealing with Variability

The idea that random samples come with variability, or noise, should not be new to the reader (but see Extra 10 below.) In this section, we will try to decompose variability into it sources, and learn several techniques to reduce the noise sources. Starting with a motivating example.

Example 4 (Movie Ratings)**.** Consider the problem of ranking movies along their rating. For this purpose, individuals are asked to rate each movie they have seen. A movie's rating is thus influenced by several factors: the movie's quality, the viewer's general tastes, the viewer's particular tastes to that type of movie, the viewer's mood at the time of watching, other factors. How can we accurately rate a movie, in the presence of these variability sources? The movie's quality is a controllable input, thus a factor. The viewer's strictness is not controllable, but observable, thus may be introduced as a covariate. The viewer's affinity to that type of movie is an interaction, since both movie type and viewer, are observable. The viewer's mood is unobservable, but we certainly acknowledge its existence. Any other variance source, will be captured by the error term of the model.

Example 4 teaches us that by an informed experimental design and analysis, we may reduce noise. Either by moving it to the signal (with covariates), or by trying to reduce it. These are the ideas discussed in this section.

Extra Info 10 (PAC learning)**.** It may be shocking for a statistician, but not all data is assumed noisy. Try reading about *PAC learning* in the machine learning literature for a counter example.

6.2.1 Gage R&R Studies

R&R stands for *repeatability* and *reproducibility*. In the context of quality control[1] repeatability is the variability under repeated measurement, and reproducibility is the variability when the same measurement is performed elsewhere (different lab, technician, etc.). Gage R&R experiments consist of performing several *repeat tests*, and different replications, in order to assess R&R, which can be thought of the assessment of the precision of the experiment.

6.2.2 Randomized Block Designs

The idea of blocking is to replace the complete *randomization scheme* by a restricted randomization scheme so that variability can be reduced without introducing bias. The restricted randomization is created by *grouping*, or *blocking* groups of experimental units, and randomizing allocation within the group.

Randomized Complete Block Designs (RCB)

The simplest of approaches to reducing uncontrollable variability sources is to *block*, or *group* similar observations together. If blocks come in pairs of observations, such as eyes, twins, etc. we call the RCB design a *matched pairs design*, which should be familiar to you from the analysis with paired t-tests.

Matched Pairs

In a general *complete block design*, when given k treatments, we form n blocks of k homogenous individuals and randomly assign them to treatments.

Latin Square Design

When homogenous groups are defined by two qualitative criteria, we would like to create blocks that are balanced, so that the treatment's effect is not biased. The following example demonstrates how *latin squares* create blocks that are balanced with respect to these two qualitative criteria (which we do not call "factors" since we are not interested in their effects).

Example 5 (Agricultural Yield Study)**.** Consider an farmer growing corn. He wishes to study the effect of 7 candidate fertilizers (single factor with 7 levels). He is aware that the location of the plot may affect yield, due to slightly different sunlight, irrigation, altitude, etc. He thus assumes he has

[1] Beware that these words may be used with different meanings by different communities.

to deal with two extra variance sources: row and column of the plot. He could treat the row and column as two factors, with 7 levels each, be he does not care to estimate the row/column effect, but merely to avoid bias. He will thus try to look for an allocation of fertilizer to rows and columns so that any row/column effect will be averaged out. The Latin Square design of Table 6.2.2 does just that.

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

Table 6.1: Latin Square: A 7-factor latin square, generated with the `MOLS()` function of the `crossdes` **R** package. The number in each cell denotes the treatment (the fertilizer in Example 5).

Extensions of the Latin Square

1. **Latin Hypercube**: When balancing more than variability sources (such as row/column), we will call upon *latin hypercube designs*, a.k.a. *orthogonal latin squares*.

Orthogonal Latin Squares

2. **Greco-Latin Square**: A latin-hypercube with three variability sources.

Incomplete Block Designs

Incomplete block designs arise where there are less observations per-block than treatments. A classical example is when blocking with twins. There are several approaches to this matter, but we refer the reader to (Cox and Reid, 2000, Sec.4.2) for details.

Crossover Design

Lets return to the movie rating example (4). We can easily agree that tastes may vary considerably between individuals. We would like to balance individuals' tastes. As previously mentioned, we could opt for a blocking strategy so that we cancel within-subject variability. Alternatively, we may balance

the viewing periods over subjects. This will not remove within-subject variability, but it will avoid a "mood bias".

The more general phenomenon is dependence in the noise between trials. Consider the same subject undergoing different treatments, or adjacent fields. Clearly, one treatment may affect the response to another treatment. This is known as *carry over*, or *residual effect*. *Crossover* designs are such that all possible treatment adjacencies are considered, so that the carry over effect averages out.

Carry Over Effect

In a *fully randomized crossover design*, each unit is randomly allocated to a sequence of treatments. If treatments are combinations of several factors, it is not uncommon to use latin-hypercubes to generate the sequences. Table 6.2.2 demonstrates possible sequences for a single, 3-level factor.

	1	2	3
1	1	2	3
2	2	3	1
3	3	1	2
4	1	3	2
5	2	1	3
6	3	2	1

Table 6.2: Crossover design: a balanced sequence of administration of 3 treatments, generated with the `des.MOLS()` function of the `crossdes` **R** package.

6.3 Factorial Designs

Until now we discussed some set of treatments. It is quite common, is not certain, that the many treatments are merely combination of a small number of *factors* with a small number of *levels* each.

6.3.1 Full Factorial Designs

A *full factorial*, or *complete factorial* design, is one where all factor-level combinations are replicated the same number of times. By far the most common design, in which k factors have 2 levels, is named a 2^k-design. A 2^k design with n repeats will necessitate $2^k \times n$ experimental units. The 2^1 design has become known as A/B testing. At this point it should be emphasized that a full factorial design is much better than k experiments with one factor at a time. This is because:

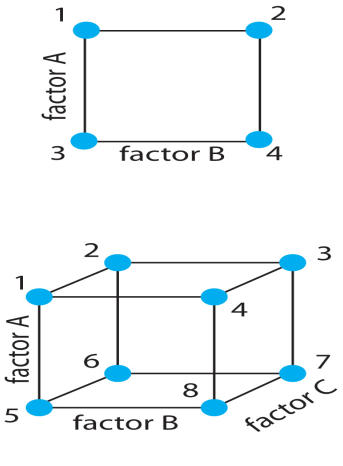
1. Factorial experiments are much more efficient at estimating main effects.

2. Factorial experiments allow the estimation of interactions between factors.

We will now present several designs for factorial experiments. Recall that factors define treatments, so that all the following designs may enjoy the tools from Section 6.2 in order to reduce variability and bias.

2^k design

Consider two factors denoted A and B. Adopt the effect coding so that we encode their levels by $-1, 1$. The design matrix of a single run is depicted in Figure 6.1 (top right) along with a visualization of the design (top left). Allowing n observations per condition, the experiment will include $4n$ observations, which will be randomized between conditions. With this 2^2 design,



factor levels

trial	A	B
1	+	−
2	+	+
3	−	−
4	−	+

factor levels

trial	A	B	C
1	+	−	−
2	+	−	+
3	+	+	+
4	+	+	−
5	−	−	−
6	−	−	+
7	−	+	+
8	−	+	−

Figure 6.1: Full factorial designs: 2^2 and 2^3.
`http://chemwiki.ucdavis.edu/Analytical_Chemistry/Analytical_Chemistry_2.0/14_Developing_a_Standard_Method`

we may recover several effects: The effect of varying A from $-$ to $+$: the *main effect of* A. The effect of varying B from $-$ to $+$: the *main effect of* B. The effect of varying both A and B from $-$ to $+$.

We typically denote $\mu_{(1)}$ the expected response for treatment $A = -, B = -$; μ_a for $A = +, B = -$; μ_b for $A = -, B = +$; μ_{ab} for $A = +, B = +$. We can thus identify the following parameters: the global mean μ, the main

effect of A τ^A, the main effect of B τ^B, and an interaction between A and B τ^{AB}:

$$\tau^A := \frac{1}{2}\left((\mu_{ab} + \mu_a)/2 - (\mu_{(1)} + \mu_b)/2\right), \tag{6.1}$$

$$\tau^B := \frac{1}{2}\left(-(\mu_{ab} + \mu_a)/2 + (\mu_{(1)} + \mu_b)/2\right), \tag{6.2}$$

$$\tau^{AB} := \frac{1}{4}\left(\mu_{ab} - \mu_a + \mu_{(1)} - \mu_b\right). \tag{6.3}$$

Interaction

Slightly intruding into the realm of data analysis, a visualization of interactions is known as the *interaction plot*, depicted in Figure 6.2. The upper left panel demonstrates a lack of interaction (think why), while the upper right panel depicts an interaction.

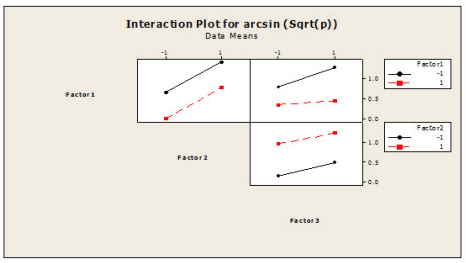


Figure 6.2: Interactions plot.
`http://blog.minitab.com/blog/statistics-in-the-field/optimizing-attribute-responses-using-design-of-experiments-doe-part-2`

Remark 9 (Screening Experiments)**.** The 2^k designs are probably the most popular full factorial designs. This may be attributed to the fact that many factors studied really have two levels, but more plausibly, since these are merely *screening* experiments. Once non related factors have been screened, the experimenter may proceed from the 2^k design to more elaborate ones.

Remark 10 (Intermediate Factor Levels)**.** In a 2^k design, a factor may actually be a continuous controllable input which was restricted to two values for convenience. After estimating the effect of the factor, we may want to know what would the effect have been, were we to set it on some intermediate level. It is customary to assume that a main effect acts linearly in-between experimental conditions, yet you should remember that there is nothing in the data to support this. For a more rigorous approach, see the Response Surface Methodology section (6.4.1).

Remark 11 (Crossed Treatments)**.** Both crossover designs and full factorial designs are *crossed* in that all factor combinations are sampled. The difference is that in a full-factorial, a factor combination is applied simultaneously to an experiment unit, and in a crossover design, the application is sequential (Everitt and Skrondal, 2010).

3^k Designs

I think that the name 3^k design is rather self explanatory. Then again, more than 2 levels are rarely treated as factorial experiments. This is because 3 level factors typically appear when aiming at optimizing the factor combination, for which the *response surface* methodology of Section 6.4.1 is more economical.

6.3.2 Fractional Factorial

The full factorial designs are the simplest designs to setup and interpret. A major drawback, are the resources required when k is large. This is where the *fractional factorial*, or *partial factorial* designs kick in. The fundamental idea is to design a full factorial, but skip a couple experimental conditions. If conditions to skip are wisely selected, only information on higher order interactions will be compromised.

Example 6 (From 2^2 to $2^{(2-1)}$)**.** As a first example, we will try to save some time and money by eliminating particular conditions of the 2^2 design in Figure 6.1. As the name may suggest, a 2^{2-1} design, has 2 experimental conditions in each run. There are thus $\binom{4}{2} = 6$ possible eliminations.

Elimination	Problem
1,2	No information on a.
1,3	No information on b.
1,4	a aliased with b aliased with ab.
2,3	a aliased with b aliased with ab.
2,4	No information on b.
3,4	No information on a.

Table 6.3: Aliasing in a 2^{2-1} design: All possible eliminations from the 2^2 design that lead to a 2^{2-1} design.

The lesson from Example 6 is that in a fractional factorial our savings in time and money, come at the cost of the information that can be drawn from

the experiment. The idea behind partial factorial experiments, is that by an informed choice of the conditions skipped, we can choose what information to give up. The information lost, is known as the *alias structure*.

Alias Structure

In practice, we will rarely do the actual elimination of conditions, but rather revert to pre-selected designs. Table 6.3.2, generated with the `FrF2()` in the `FrF2` **R** package, is an optimal 2^{3-1} design. Using the `design.info()` function of that same package, we know that the aliasing structure of this design is $a = bd = ce, b = ad, c = ae, d = ab, e = ac$. We will not go into the details of how the aliasing structure is computed, but rather refer the reader to Cox and Reid (2000).

	A	B	C	D	E
1	-1	1	-1	-1	1
2	1	-1	1	-1	1
3	-1	-1	-1	1	1
4	-1	1	1	-1	-1
5	-1	-1	1	1	-1
6	1	1	1	1	1
7	1	-1	-1	-1	-1
8	1	1	-1	1	-1

Table 6.4: 2^{5-2} design.

Definition 25 (Resolution of a Design)**.** As we have already seen, there are $\binom{2^k}{2^{k-p}}$ possible eliminations that convert a 2^k design to a 2^{k-p} design. We call the *resolution of the design*, the lowest order effect that is aliased. In Table 6.3.2, the resolution is 1, making it a rather unattractive design. Resolutions below 3 typically considered not informative, and above 5, considered wasteful.

Extra Info 11 (Coding Theory)**.** There is a close relationship between design of experiments and coding theory in computer science. A possible reference on the matter is Hill (1986), or Hedayat et al. (1999).

6.3.3 Split Plot Design

A *split plot*, or *split unit* design, is essentially a factorial design with blocking. It takes its name from farming plots: where these are used for blocking and split among the various treatments. The canonical example is due to (Montgomery, 2012, Sec.12.4), and consists of studying two production factors with blocking by day. For more on split plots see (Cox and Reid, 2000, Sec.6.4).

6.4 Continuous Factors

When dealing with continuous factors, or *quantitative factors*, we have many more analysis strategies than when dealing with qualitative factors. No matter the analysis strategy, it is still important of choosing the right factor level combinations to study.

Clearly, we cannot identify nonlinearities when sampling a factor only a two levels. We may thus opt for a full 3^k factorial design to fit a non-linear surface to the data. The factor encoding would typically be $\{-1, 0, 1\}$, A full 3^k factorial design may however be needlessly expensive. A more common approach in industrial application is the *central composite design*, where a 2^k design is augmented with well chosen sampling points. For more details see (Cox and Reid, 2000, Sec.6.6)

Central Composite Design

6.4.1 Response Surface Methdology

As the name suggests, *response surface methdology* deals with the estimation of the response surface to the levels of continuous factors. Response surfaces are typically assume to be (approximately) quadratic, and experimentation typically conducted in stages. First screening factors, then fitting a surface, and finding optimal factor levels.

6.4.2 Taguchi Methods

Taguchi methods is a collective name for the philosophy, design, and analysis methods in industrial applications promoted by Genichi Taguchi in 1970's Japan. Focusing on the design principles, we may note the following particularities of Taguchi's method:

1. Achieving low variability is more challenging then achieving a target value.
2. Factors which can be controlled in a lab, but not in production, are deliberately varied. Typically in split plot designs (6.3.3).
3. Systematic use of latin hypercubes to study main effects and two-way interactions. Particularly *Plackett–Burman designs*.

   Plackett Burman Design
4. Log variability is often used as the response.

6.4.3 Optimal Designs

Our discussion until now has been informal with respect to the desirable qualities of a design. We used the idea of "balance" and "orthogonality" to avoid bias and unwelcome variability. In this section, we try to formalize the notion of a "good design". We start with some motivating examples.

Example 7 (Design for non linear regression)**.** Figure 6.3 demonstrates the effect of the different location of the sampling points (x) on the quality of the estimated regression line, in a **linear** model. As the figure depicts, it is preferable to spread the sampling points as far as possible, as intuition may suggest.

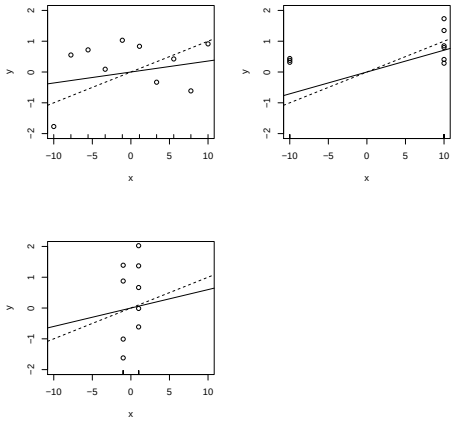


Figure 6.3: Design for linear regression. Different panels show different designs. True function as a dashed line. Estimated function as a full line.

Example 8 (Design for non linear regression)**.** Figure 6.4 demonstrates the effect of the different location of the sampling points (x) on the quality of the estimated regression line, in a **nonlinear** model. The figure is non conclusive as to the best design, but it may seem that unlike the linear case (Example 7) optimality is achieved in some non trivial sampling scheme.

Now are some facts that are supported by the previous examples:

1. The idea of "balancing" as a design criterion is very useful with discrete factors, but limited with continuous factors.

2. The optimal design may depend on the unknown generative model. Luckily, for linear models, this is not the case, and an optimal design will be so for all values of the generative parameter.

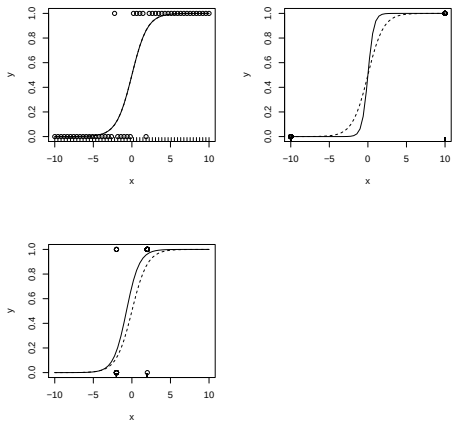


Figure 6.4: Design for non linear regression. Different panels show different designs. True function as a dashed line. Estimated function as a full line.

6.4.4 Space Filling Design

The most natural of designs, which is particularly suitable when we have no a-priori assumption on the functional relation between the (continuous) factors, $f : x \mapsto y$, and the response is known as a *space filling design*. As the name suggests, in a space filling design we aim at filling the factor space. Lacking any a-priori information, the filling will typically be as uniform as possible. We note however, that once information on f is made available, then a space filling design is typically sub optimal (see Example 7).

Extra Info 12 (Space Filling and Hashing)**.** If you are familiar with the idea of *hashing functions*, then you may see the similarity between space filling and the *uniformity* property of hash functions. For a more rigorous discussion, see Hill (1986).

6.4.5 Covariance Optimality

We have already noted the the optimality of the design depends on the data generating process. In this section it is made obvious that optimality will also depend on the analysis method we choose.

When estimating the effect of a single continuous factor, we would like a design that gives us the most information per observation. This is the same as minimizing the variance of the estimator: $\min\{Var(\hat{\beta})\}$.

When generalizing to several continuous factors, with several effect, matters are more subtle. This is because, having several parameters, we may

consider several target criteria such as minimal average variance, or minimal worst-case variance. Matters further complicate when effect estimates are correlated, as will be promptly explained.

Definition 26 (Error Covariance Matrix)**.** For a p-vector of effects β, we define the $p \times p$ *error covariance* matrix $M(\hat{\beta}, \beta)$ to be

$$M(\hat{\beta}, \beta)_{i,j} := \mathbf{E}\left[(\hat{\beta}_i - \beta_i)((\hat{\beta}_j - \beta_j))\right] \tag{6.4}$$

Denoting $M = M(\hat{\beta}, \beta)$, we readily note that for unbiased $\hat{\beta}$, then the diagonal of M is simply the variances of $\hat{\beta}$, and the off-diagonal are the covariances. We also remark that if $p = 1$, then M is merely the scalar variance.

We now try to generalize the idea of "minimal variance" to the multivariate case.

Definition 27 (A-Optimality)**.** A design is said to be *A optimal* if it minimizes the trace of M.

A-optimality has an intuitive interpretation: since the average variance is proportional to the trace, then A-optimality is actually minimizing the average variance over effects.

Alas, A-optimality does not account for covariances. In an extreme scenario, if we have several copies of the same variable, the more copies we have, the more importance that variable will be given by A-optimality. The most popular optimality criterion is known as *D-optimality*, and does not suffer from this phenomenon.

Definition 28 (D-Optimality)**.** A design is said to be *D optimal* if it minimizes the determinant of M.

If you are familiar with the geometrical interpretation of the determinant, this definition may not surprise you. If not, then one way to think of D-optimality is via confidence regions. D-optimality has the property that in the multivariate Gaussian case, a D-optimal design will return confidence regions for β with smallest volume.

Extra Info 13 (Other Optimality Criteria)**.** There are as many optimality criteria as there are matrix norms. For a more detailed review, see Wikipedia (2015d).

The non-linear model example (8) suggests that the optimal design may depend on the underlying (unknown) effect β. This is obviously bad news

since if we had knowledge of β, we would not need an experiment to estimate it. There are, however, some good news. First, for linear models and least-squares estimate, then M will not depend[2] on β, and neither will the optimal design. Second, when M does depend on β, we will simply do some initial small experiment, and then optimize the design based on initial results.

6.5 Sequential Designs

Consider a clinical trial with a treatment and control group. Now assume the medicine being tested is a miracle cure with immediate improvement. Do we really need to keep administering placebos to the control group, just because that was the initial experimental design? This is where sequential designs come in. Interestingly, the initial application of a sequential design was not in drug testing, but rather in a military context (Wald, 1945).

The problem with sequential testing, is the *type-I error inflation*, which is simply a *multiplicity problem*. To see this, think about an endless sequential experiment. Also assume the null hypothesis is true. Can we agree that a regular (non sequential) experiment will not reject H_0? Can we also agree that a sequential experiment will necessarily reject H_0 at some stage?

Multiplicity

In it simplest version, a sequential design allows early stopping for rejection of the null, or for futility (non-rejection). In more elaborate schemes then not only is early stopping allowed, but also the redesign of the experiment. This is known as *adaptive design*. The crux, as usual, is not inflating the type-I error, or introducing bias, by redesigning.

Adaptive Design

Extra Info 14 (Active Learning)**.** In the machine learning literature, the idea of adaptive design of experiments is known as *active learning*, where the emphasis is less on adaptive-testing, but rather on adaptive-estimation.

6.6 Computer Experiments

Example 9 (Designing Wings)**.** Consider the problem of designing an aircraft's wing. We would like to know how the wing's attributes, i.e., factors, govern its lift. We could obviously conduct real-life experiments by varying the wing's attributes, building the wing, flying the air-plane, and recording results. Needless to say how expensive this process is. It is much more reasonable to program the differential equations that govern the lift to a computer, fix several factors values, and solve the equations. This is what *computer experiments* are all about.

[2]This is known as the *equivariant in law* property.

The wind design example (9) demonstrates the following points:

1. Computer experiments are essentially numerical solutions to complicated systems of equations.
2. Because solutions take a lot of time, only a small finite set of values may be evaluated.
3. The "response" to each treatment, is deterministic.
4. The problem of interest is in reconstructing the response at non measured factor values, so that optimal values may be identified.

It is thus not uncommon to call upon DOE theory for choosing the factor combinations to be experimented with. Space filling designs (Sec. 6.4.4) being a particularly prevalent choice. The analysis of computer experiment is very different than real-life experiment since we have no noise component. See Sacks et al. (1989) or Santner et al. (2013) for further details.

Chapter 7

Acceptance Sampling

We can improve quality (read- conformance to specification) by introducing an inspection stage in our process. Clearly, a full inspection is time consuming. It may also be destructive (you don't want to re-package ice-cream after checking its texture ...). No-inspection may be appropriate if you don't particularly care about your brand, or if production has very high capability indices. A reasonable, intermediate approach, is a partial random inspection, known as *acceptance sampling*. As the name suggests, in acceptance sampling, one samples, then checks, then accepts (or not).

Acceptance sampling can be seen as a control chart monitoring that triggers active intervention in the production. As such ,it is a crude type of *engineering control* (Sec. 1.1.2). The intervention is obvious. The monitoring is based on some continuous (variable) or discrete (attribute) of a sample of units from a *batch*, a.k.a. , a *lot*. Seen as a feedback control, it is not surprising that when designing an acceptance sampling scheme, we have similar decisions as when designing a control chart:

1. What is a batch? Just like choosing the sampling frequency in a Shewart chart. We would like homogenous batches, i.e., with low inner variability. A box, a shipment, a day's production, are typical batches.

2. Within batch sampling scheme: just like rational grouping in Shewart chart. Typical approaches include *single sampling plans*, *double*, *multiple*, and *sequential sampling plans*. This can be seen as the design of an experiment to be performed on each batch.

3. How many units? Just like choosing the sample size in a Shewart chart.

4. Decision cutoff: Just like setting control limits in a Shewart chart.

We can readily see that the design of an acceptance sampling scheme is very similar to the design of a control chart. We may construct an full blown economical optimization problem to design the sampling, as we did in Section 5.4. Just like control charts, however, it is more common to design sampling schemes using "first-order" power considerations. For this reason, the *power function* will play a crucial role.

7.1 Acceptance Sampling Terminology

Adapted from Natrella (2010).

LASP A *lot acceptance sampling plan*, ultimately, a statistical test at the end of which we either accept a batch. In this text we typically use the *batch acceptance sampling scheme* for the same purpose.

AQL The *acceptable quality level*, or *acceptable quality limit*, is the highest proportion of defects acceptable to the producer.

LTPD The *lot tolerance percent defective* is the highest proportion of defects acceptable to the consumer. Clearly, $AQL < LTPD$. LTPD is also known as *rejectable quality level* (RQL), and *limiting quality level* (LQL).

OC Curve The *operating characteristic curve* is the power function of an LSAP.

Type-A and Type-B OC Curves A *Type-A OC curve* is one computed assuming sampling from batches is done without replacement. Conversely, a *Type-B OC curve* is computed assuming sampling with replacement.

Producer's Risk The *producer's risk* is throwing away good batches. Formally, this is the probability of rejecting a batch with less than AQL defects. We consider there type-I errors.

Consumer's Risk The *consumer's risk* is accepting bad batches. Formally, this is the probability of accepting a batch with more than LTPD defects. We consider there type-II errors.

Rectifying Inspection An LASP where lots are not rejected but rather rectified.

7.2 Single Sampling Scheme

In the simplest LASP we base our decisions on a single random sample from each batch. This obviously facilitates the statistical analysis of the properties of this LASP.

Type-B Power Function

When sampling n units from a batch with a proportion of p defects, then the number of defects $\mathbf{x} \sim Binom(n, p)$. If we reject a batch when more than c defects are found, then the power function of a type-B LASP is given by

$$\pi_{n,c}(p) = P(\mathbf{x} \geq c) = \sum_{k=c}^{n} \binom{n}{k} p^k (1-p)^{1-k}. \tag{7.1}$$

Eq.(7.1) may be evaluated manually, or with the `pbinom()` **R** function.

Just like any other hypothesis test, it is common practice to set n, c so that control both the consumer's risk ($\beta_{n,c} = 1 - \pi_{n,c}$) and the producer's risk ($\alpha_{n,c}$). By adopting a the hypothesis testing philosophy, we solve n, c so that

$$\min\{n : \pi_{n,c} \geq \pi_0 \quad and \quad \alpha_{n,c} \leq \alpha_0\}. \tag{7.2}$$

For relating the LASP terminology to this problem, we need to observe that

$$\alpha_{n,c} = \pi_{n,c}(p = AQL)$$

and

$$\pi_{n,c} = \pi_{n,c}(p = LTPD).$$

For a producer who does not want to reject batches where $AQL = 10\%$ defects, with more than $\alpha_0 = 10\%$; and a consumer who does not want to accept batches where $LTPD = 30\%$, with less than $\pi_0 = 80\%$, we have that their LASP would take $n = 33$ samples, and reject a batch whenever the $\mathbf{x} > 4$, when $n = 21$.

Remark 12 (Approximate Power Calculations)**.** The problem to solve in Eq.(7.2) requires some non trivial iterations because of the discrete nature. It is quite more convenient to replace the exact form of Eq.(7.1) with a normal approximation, so that Eq.(7.2) has a closed form solution.

Type-A Power Function

It is quite wired that we would sample with replacement from a batch. It is quite more probable that we used the replacement assumption, only as an approximation because n is small compared to the batch size N. If this is not the case, the binomial distribution in Eq.(7.1) should be replaced with the Hypergeometric distribution. For all practical purposes, this means using the `phyper()` **R** function, instead of `pbinom()`.

7.2.1 Double Sampling Scheme

In a double sampling scheme, we first example n_1 units. We may then decide to accept, reject, or sample another n_2 units. After those n_2 samples, we can accept or reject. The idea of a power function remains the same, even if calculations are slightly more cumbersome. Here is our our policy: For x_1 computed on the first n_1 samples: If $x_1 < a_1$ then accept the batch; If $x_1 \geq c_1$ then reject the batch; Otherwise, compute x_2 with $n_1 + n_2$ samples. If $x_2 < a_2$ accept the batch; If $x_2 \geq c_2$ then reject the batch.

For brevity, we denote all the design parameters of the scheme by $\gamma := (n_1, n_2, c_1, c_2, a_1, a_2)$. The power function of such a scheme would thus be:

$$\pi_\gamma := P(\{\mathbf{x}_1 \geq c_1\} \cup \{\mathbf{x}_1 \in [a_1, c_1], \mathbf{x}_2 \geq c_2\}) \tag{7.3}$$

$$= P(\mathbf{x}_1 \geq c_1) + \sum_{k=a_1}^{c_1} P(\mathbf{x}_1 = k, \mathbf{x}_2 - \mathbf{x}_1 \geq c_2 - k) \tag{7.4}$$

$$= P(\mathbf{x}_1 \geq c_1) + \sum_{k=a_1}^{c_1} P(\mathbf{x}_1 = k) P(\mathbf{x}_2 - \mathbf{x}_1 \geq c_2 - k). \tag{7.5}$$

We may now use the fact that $\mathbf{x}_1 \sim Binom(n_1, p)$ and that $\mathbf{x}_2 - \mathbf{x}_1 \sim Binom(n_2, p)$, and quickly compute the power in **R** .

Remark 13 (Redundancy)**.** Unlike the single stage LASP, where we have two equations with two variables, in the two-stage case there are many γ configurations that will achieve given consumer and producer risks (α_0, π_0). The choice of the particular configuration should depend on the type of signal we expect. For quick detection of strong signal (large p), choose small n_1. For sensitive detection of subtle signal, choose large n_1.

Remark 14 (No Free Lunch)**.** While it may seem that a two stage LASP is always better than a single stage LASP, this is not the case. To see why, consider a weak signal (p close to AQL). We may need all $n_1 + n_2$ samples to get decent power. The first stage then add nothing except logistic complications.

7.3 Sequential Scheme

At this point you should be thinking: why only two stages? Clearly we may reject or accept a sample as each unit comes in. This is exactly what Sequential LASPs are all about. We will not give the details, except the observation that this is merely a type of sequential experiment as described in Section 6.5.

Chapter 8

Reliability Analysis

Chapter 9

Revisiting System Capability Analysis

[TODO]

9.1 System Capability with Control Charts

9.2 System Capability with Designed Experiments

Appendix A

Notation

In this text we use the following notation conventions:

x A column vector, or scalar, as implied by the text.

$:=$ An assignment, or definition. $A := a$ means that A is defined to be a.

$\prod_{i=1}^{n}$ The product operator: $\prod_{i=1}^{n} x_i := x_1 \times \cdots \times x_n$

$\#\{A\}$ The count operator. Returns the number of elements of the set A. Also known as the *cardinality*.

$\Phi(t)$ The standard Gaussian CDF at t: $\Phi(t) := P(Z < t)$.

$\phi(t)$ The standard Gaussian density at t: $\phi(t) := \frac{\partial}{\partial t}\Phi(t)$.

x' We use $'$ for the transpose operation. For a $1 \times p$ vector x, then x' is $p \times 1$.

$\mathbf{x}_n \rightsquigarrow P$ Convergence in distribution: for large enough n, then $\mathbf{x}_n$ is distributed like P.

Appendix B

R

For DOE with **R** , see `https://cran.r-project.org/web/views/ExperimentalDesign.html`.

Bibliography

M. Basseville, I. V. Nikiforov, and others. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

D. R. Cox and N. Reid. *The Theory of the Design of Experiments*. CRC Press, 2000.

A. J. Duncan. The Economic Design of X Charts Used to Maintain Current Control of a Process. *Journal of the American Statistical Association*, 51 (274):228–242, 1956.

B. S. Everitt and A. Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, UK ; New York, 2010.

S. R. A. Fisher. *The design of experiments*, volume 12. Oliver and Boyd Edinburgh, 1960.

M. A. Girshick and H. Rubin. A Bayes Approach to a Quality Control Model. *The Annals of Mathematical Statistics*, 23(1):114–125, 1952.

A. S. Hedayat, N. J. A. Sloane, and J. Stufken. *Orthogonal Arrays: Theory and Applications*. Springer Science & Business Media, 1999.

R. Hill. *A First Course in Coding Theory*. Clarendon Press, 1986.

R. R. Hocking. *The analysis of linear models*. Brooks/Cole Pub Co, 1985.

J. Jin, J.-L. Starck, D. L. Donoho, N. Aghanim, and O. Forni. Cosmological non-Gaussian Signature Detection: Comparing Performance of Different Statistical Tests. *EURASIP J. Appl. Signal Process.*, 2005:2470–2485, 2005.

R. L. Mason, R. F. Gunst, and J. L. Hess. *Statistical design and analysis of experiments: with applications to engineering and science*, volume 474. John Wiley & Sons, 2003.

D. Montgomery. *Design and Analysis of Experiments*. Wiley, Hoboken, NJ, 2012.

D. C. Montgomery. *Introduction to statistical quality control*. John Wiley & Sons, 2007.

M. Natrella. *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH, 2010.

E. S. Page. Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115, 1954.

K. B. Petersen and M. S. Pedersen. *The matrix cookbook*. Citeseer, 2006.

P. Qiu. *Introduction to Statistical Process Control*. Chapman and Hall/CRC, Boca Raton, 2013.

J. L. Rodgers and W. A. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988.

P. R. Rosenbaum. *Observational studies*. Springer, 2002.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer Science & Business Media, 2013.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Wikipedia. *Design for Six Sigma — Wikipedia, The Free Encyclopedia*. 2015a. [Online; accessed 29-October-2015].

Wikipedia. *Eight dimensions of quality — Wikipedia, The Free Encyclopedia*. 2015b. [Online; accessed 28-October-2015].

Wikipedia. *Lean manufacturing — Wikipedia, The Free Encyclopedia*. 2015c. [Online; accessed 29-October-2015].

Wikipedia. *Optimal design — Wikipedia, The Free Encyclopedia*. 2015d. [Online; accessed 13-November-2015].

Wikipedia. *Quality (business) — Wikipedia, The Free Encyclopedia*. 2015e. [Online; accessed 28-October-2015].

Wikipedia. *Receiver operating characteristic — Wikipedia, The Free Encyclopedia*. 2015f. [Online; accessed 6-November-2015].

Wikipedia. *Value engineering — Wikipedia, The Free Encyclopedia*. 2015g. [Online; accessed 29-October-2015].

Wikipedia. *Zero Defects — Wikipedia, The Free Encyclopedia*. 2015h. [Online; accessed 29-October-2015].