Quality Engineering - Class Notes (experimental)

Jonathan Rosenblatt

December 4, 2016

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

List of Examples

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. In the BI world, these are typically known as *key performance indicators* (KPI).

KPI

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. The same as 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. Inclues hypothesis testing, parameter estimation, confidence estimation, prediction, and others.

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. A.k.a. *anomaly detection*, or *novelty detection*, in the machine learning literature..

Computer Simulation Well, just what the name implies.

Control Chart A graphic visualization of the historical evolution of one (or several) CTQs. Typically augmented with some graphic decision criteria.

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 in a SCADA system[a].

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

[a]Supervisory Control and Data Acquisition refers to systems than integrate sensors and computers for fully, or semi, automated Industrial Control (ICS). SCADA and ICS recently became famous in the context of the STUXNET worm, since STUXNET target the SCADA system controlling uranium enriching centrifuges.

Extra Information. [Smoking and Cancer] If, like me, you are curious to understand how did we miss the cancerous effects of smoking, have a look at Cornfield et al. (2009).

1.2 Some History

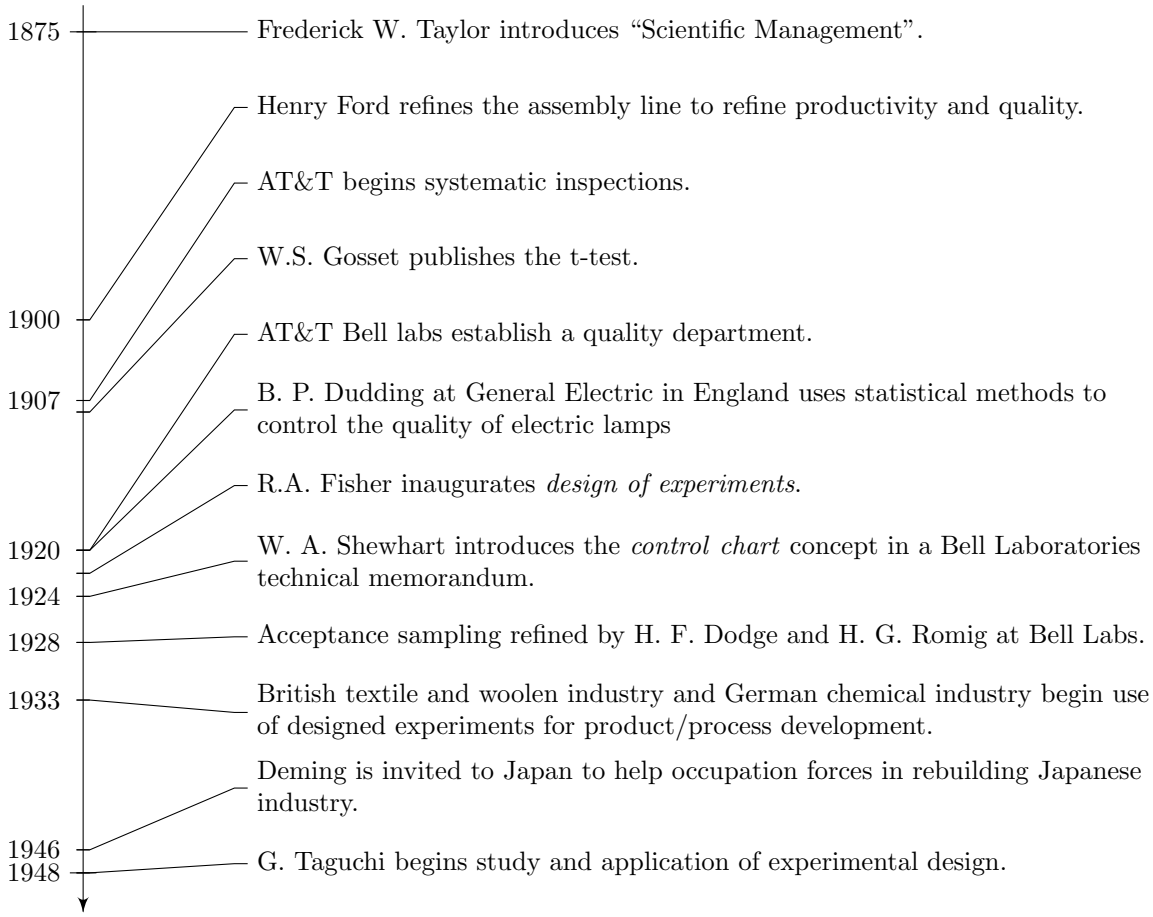


Table 1.1: Adapted from (Montgomery, 2007, Table 1.1).

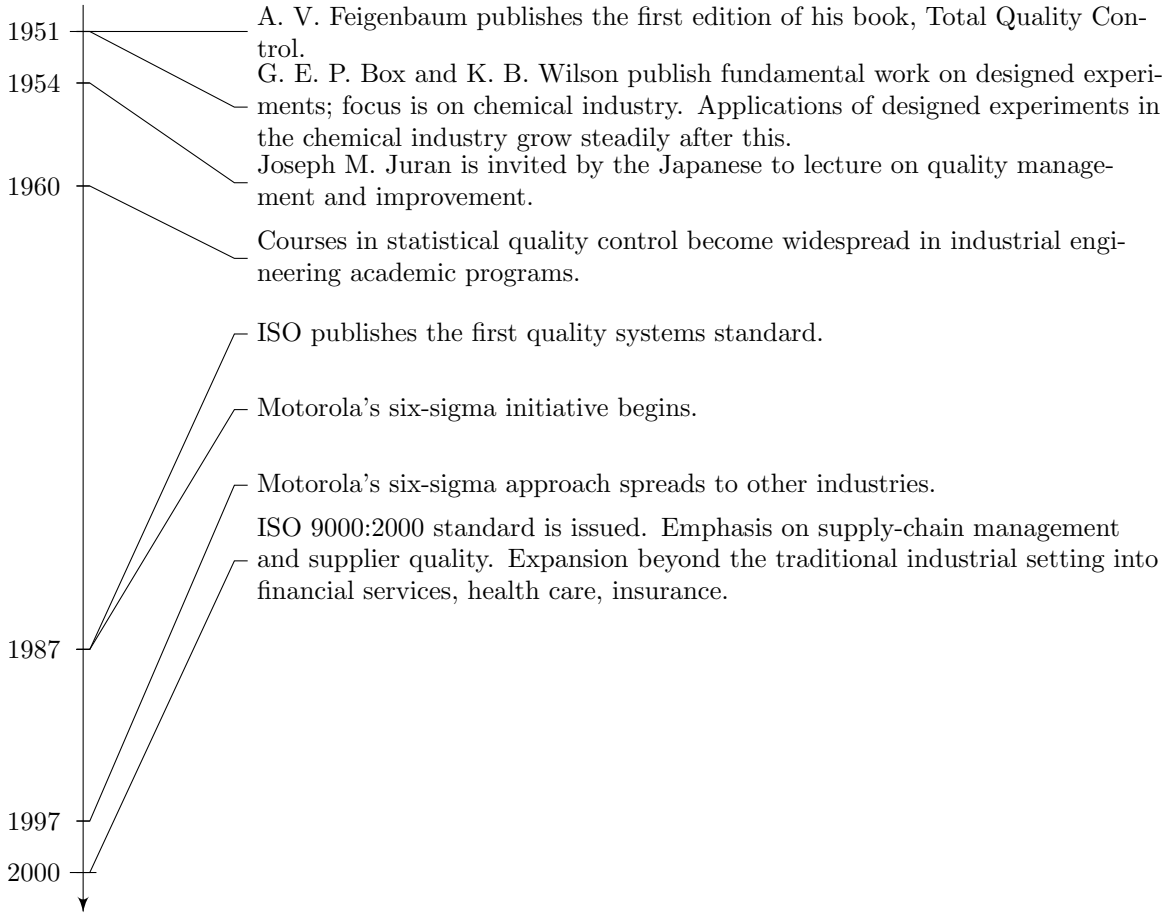


Table 1.2: Adapted from (Montgomery, 2007, Table 1.1).

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

$$CTQ \sim \mathcal{N}\left(T, \sigma^2\right),$$

we can compute:

$$\alpha = 1 - P(LSL < CTQ < USL) = \tag{1.2}$$

$$1 - P(|CTQ| < \delta) < \qquad \text{assuming } 3\sigma capability \tag{1.3}$$

$$1 - P(|CTQ| < 3\sigma) = \qquad \text{assuming normality} \tag{1.4}$$

$$0.0027. \tag{1.5}$$

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

$$P(failure) < 1 - (1 - 0.0027)^m$$

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

$$6\sigma < \delta.$$

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.

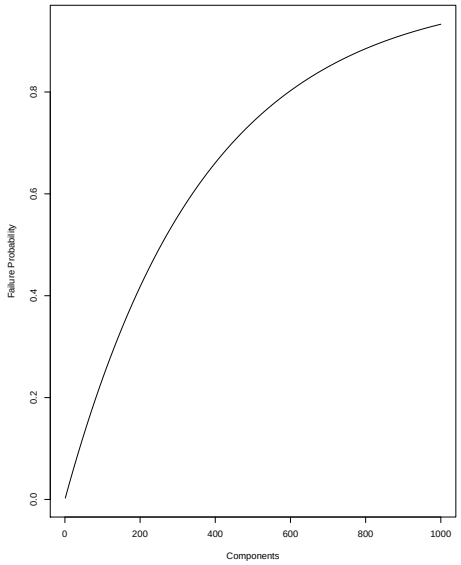


Figure 1.1: The probability of failure as a function of components under the 3-sigma standard.

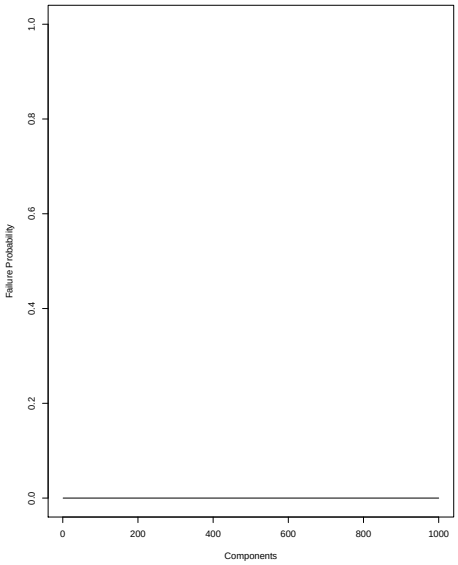


Figure 1.2: The probability of failure as a function of components under the 6-sigma standard.

Remark 1 (PPM of Six-Sigma)**.** In the six-sigma literature you will find the following table:

Sigma Level	DPMO	Process Yield
1	500000	50%
2	308537	65%
3	66807	93%
4	6210	99.4%
5	233	99.976%
6	3.4	99.9997%

Figure 1.3: Defect PPM of varying sigma processes.
Source: Souce: `http://www.whatissixsigma.net/10-things-about-six-sigma/`

As you may notice, the 3-sigma and the 6-sigma defect PPMs in the table are different than the ones in Eq.(1.2). This is explained in (Montgomery, 2007, p.29):

> When the six-sigma concept was initially developed, an assumption was made that when the process reached the six-sigma quality level, the process mean was still subject to disturbances that could cause it to shift by as much as 1.5 standard deviations off target.

This means that the convention is to compute Eq.(1.2) while replacing the $\mathbf{E}[CTQ] = T$ assumption by $|\mathbf{E}[CTQ] - T| \leq 1.5\sigma$.

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[1] 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 Information. [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 1.4 known as DMAIC: Define, Measure, Analyze, Improve, Control.

Here are some general observations on DMAIC:

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 [2]?

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.

[1] `https://portal.sii.org.il/heb/qualityauth/certificationtypes/qualitylinks/iso9001/`
[2] `http://asq.org/learn-about-quality/six-sigma/overview/dmaic.html`

Figure 1.4: The DMAIC cycle.
`http://www.sapartners.com/sigma-academy/`

1.6 Bibliographic Notes

[TODO]

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 Information.** 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 combinations 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.

Summary of Scale

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

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

where c is some constant. In the **R** function `mad()`, c is set at 1.4826 so that it estimates σ in a Gaussian population.

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

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

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

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 Information. [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 Information. [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 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.

PCA, SVD,ICA

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

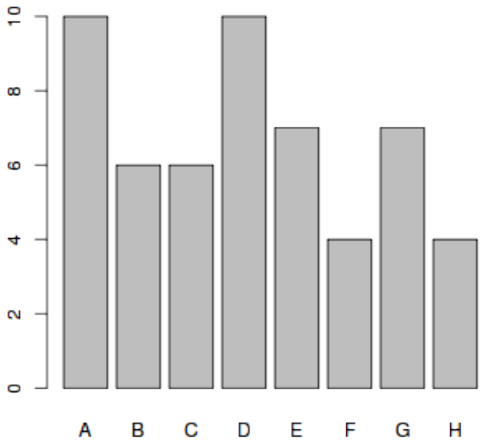


Figure 2.1: The Bar-Plot.
`http://www.r-tutor.com/elementary-statistics/qualitative-data/bar-graph`

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

Mosaic Plot

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

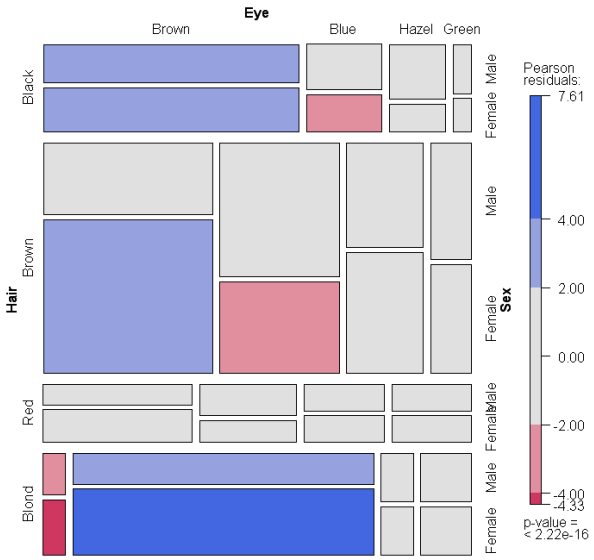


Figure 2.2: Mosaic Plot.
`http://www.statmethods.net/advgraphs/mosaic.html`

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.

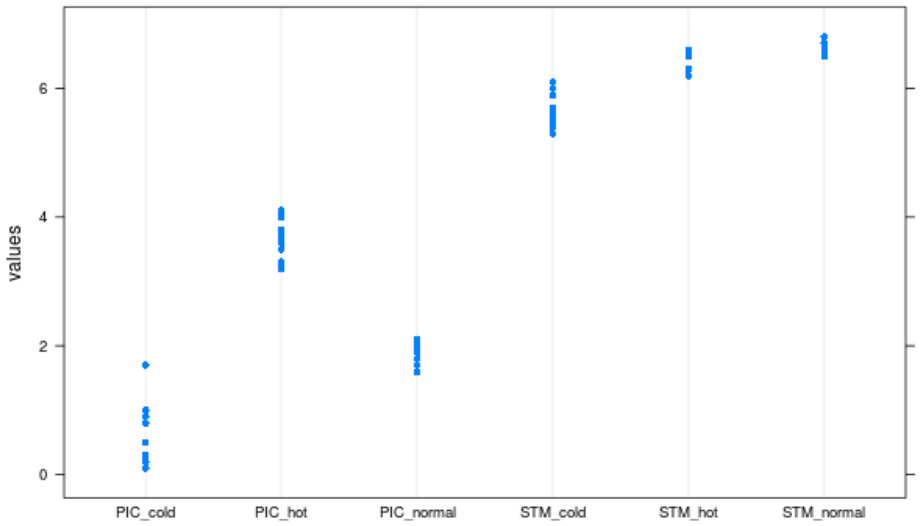


Figure 2.3: Dot Plot.
`http://stackoverflow.com/questions/15109822/r-creating-scatter-plot-from-data-frame`

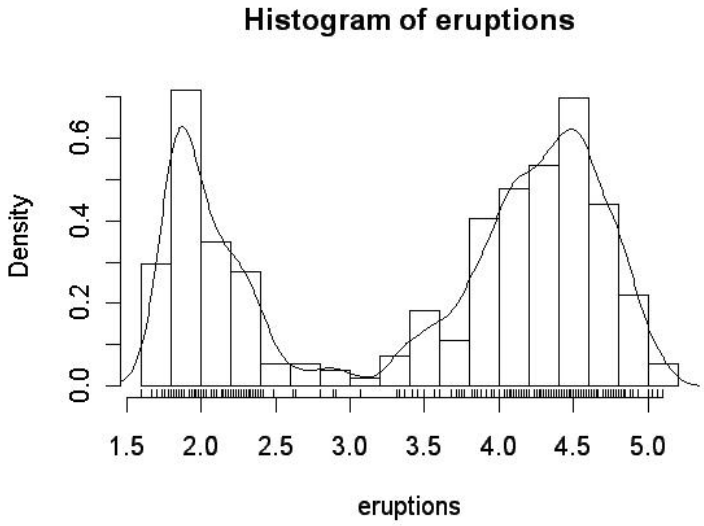


Figure 2.4: Histogram. Notice the ticks on the x axis. These are the raw data points. Make sure you always add them, with the `rug()` **R** command.
`http://compbio.pbworks.com/w/page/16252882/Basic`

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.

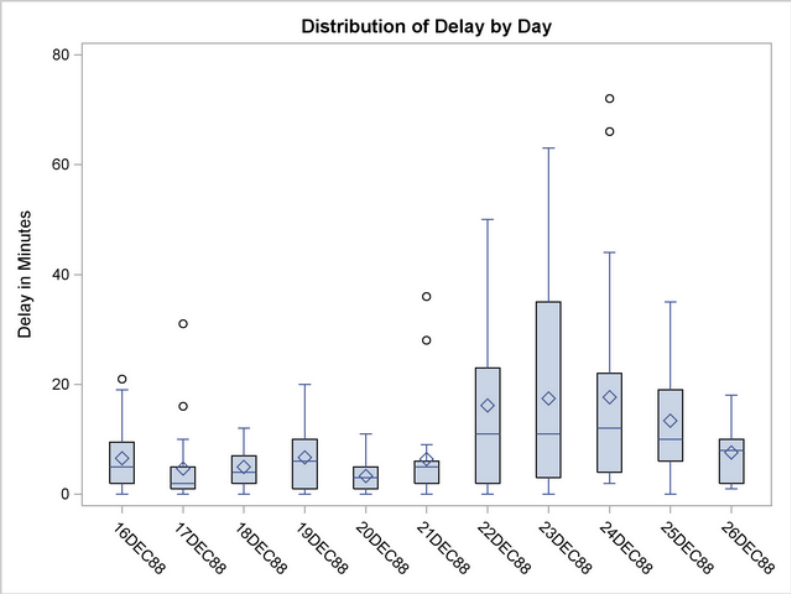


Figure 2.5: Boxplot.
`http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm`

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.

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

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.

Dash-board

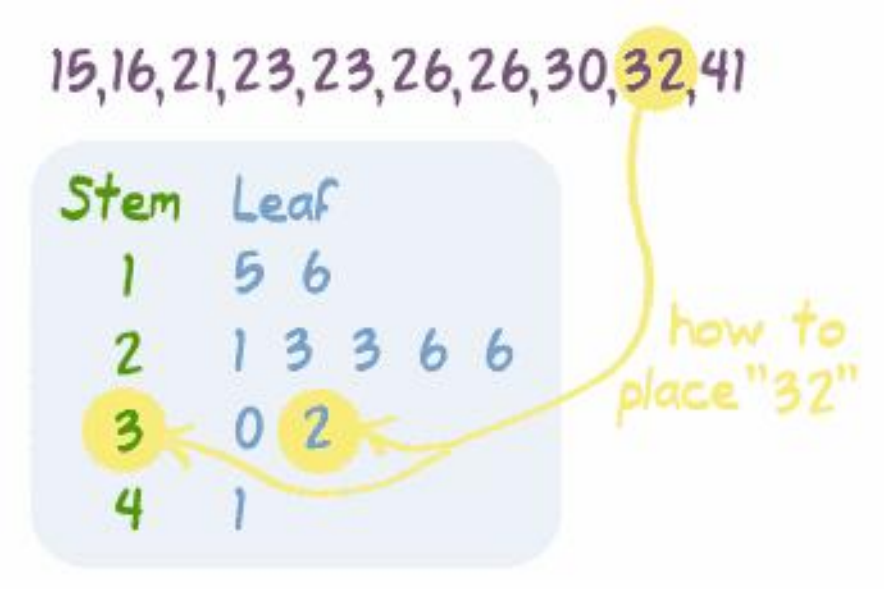


Figure 2.6: Stem-and-leaf plot.
`https://www.mathsisfun.com/data/stem-leaf-plots.html`

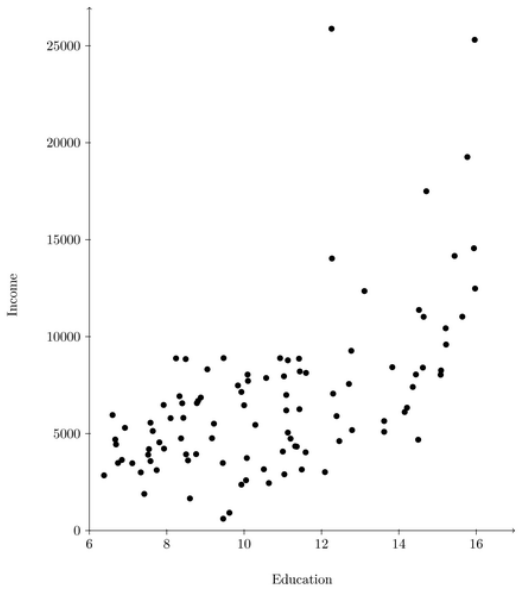


Figure 2.7: Scatter Plot.
`http://texample.net/tikz/examples/scatterplot/`

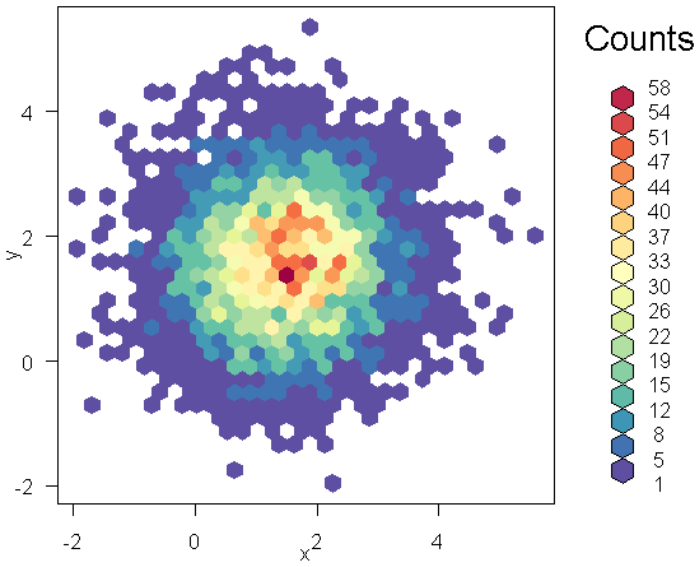


Figure 2.8: HexBin plot (a 2D histogram).
`http://www.r-bloggers.com/5-ways-to-do-2d-histograms-in-r/`

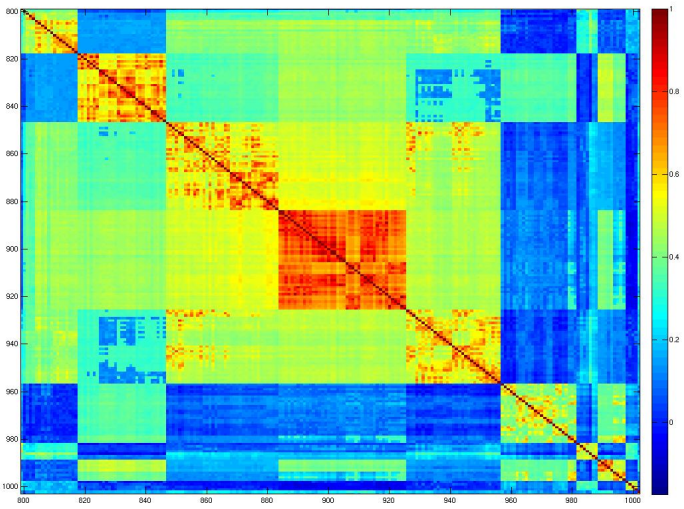


Figure 2.9: Image of covariance matrix.
`http://cs.brown.edu/courses/csci1950-g/results/final/sghosh/`

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.

Null/Alternative Hypothesis Some statement about the world we wish to test with data. The frequentist argument follows a Popperian philosophy[1]: 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*. Other terms for the alternative hypothesis are the *research hypothesis*, or simply the *signal*.

In 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. May also be though of as a *signal detector*.

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.

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.

[1]Following Karl Popper's philosophy of science, we can never know that something is true, we can only know when it is not true. Popper philosophy was motivated by the fact that no one suspected Isaac Newton's mechanics to be wrong, until relativity theory was proposed by Einstein.

	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/`

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.

Extra Information. [ROC termininology]
The engineering, statistical, and information retrieval literature, define error and precision criteria[a]:

True Positive Rate ...

Sensitivity ...

Recall ...

False Positive Rate ...

Fallout ...

Specificity

Miss Rate

Prevalence

True Negative Rate

Accuracy

Positive Predictive Value

Precision

False Omission rate

False Discovery Rate

Negative Predictive Value

Positive Likelihood Ratio

Negative Likelihood ratio

Disgnostic odds ratio

[a] `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`

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

Think about it. Can you design a test with type I error larger its power? Would you ever want such a test? Think about it using the analogy to statistical tests and criminal courts.

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.

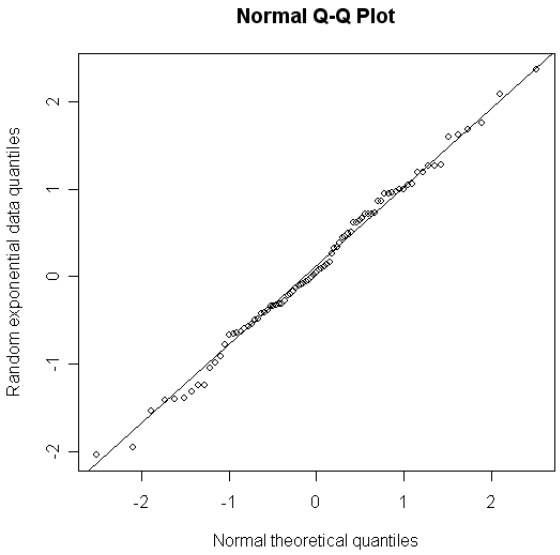


Figure 3.2: A QQplot of Gaussian distributed data.

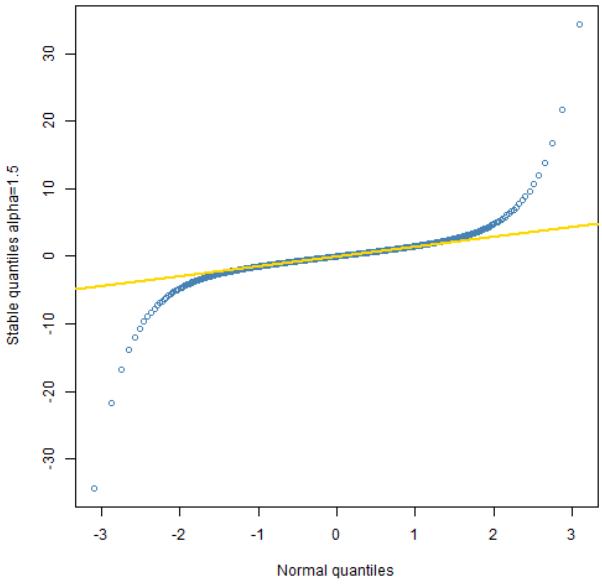


Figure 3.3: A QQnorm plot of non Gaussian distributed data.

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

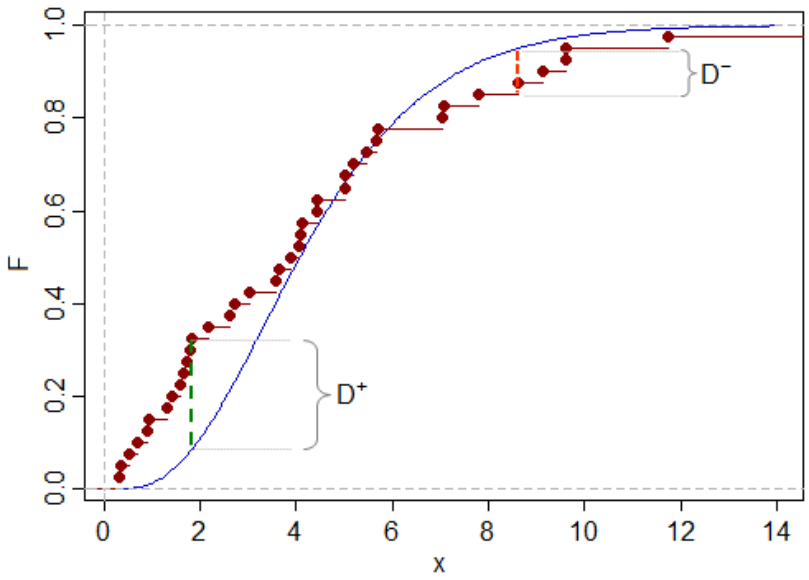


Figure 3.4: Kolmogorov-Smirnov Test. The D statistic is D_+ in the figure.

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

Kolmogorov Distribution

Extra Information. [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.

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.

Product Specification

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.

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 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.

Think about it. Can you design a process capability index with a 6-sigma baseline?

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 Section 4.1.5.

The design of C_p rests on its relation to the probability non-conformance. To explore this relation we introduce the following notation:

Collecting Notation
T, 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

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

As a sanity check, we check 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 CTQ has a normal distribution.
2. The process is centred, i.e. $\mu = T$.

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 non compliance rate, if we were completely wrong about the distribution of the CTQ. Chebyshev's inequality will be useful.

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 very very 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, since we will not need to tune the production process to some worst-case scenario, but rather to the exact properties of our production process. We can assume normality, as we typically do, but there are other alternatives that may help us to preserve the important relation between C_p and p_{NC}:

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 any other distribution you are willing to assume.

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. It is immediate to verify that $p_{NC}(C_p(q))$ is *always* $2,700ppm$. The only difficulty may be in estimating $CTQ_{0.00135}$ and $CTQ_{1-0.00135}$.

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 $\mu = T$. Again, this will break the relation between p_{NC} and C_p Eq. 4.3), even for a normally distributed CTQ.

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:

$$\Phi(-3C_{pk}) \leq p_{NC} \leq 2\Phi(-3C_{pk}). \tag{4.7}$$

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 non-compliance 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 (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 itself as a capability index: $\min\{\frac{USL-\mu}{\sigma}, \frac{\mu-LSL}{\sigma}\}$, which can be seen as capability index with respect to a 1-sigma process. Three-σ would thus simply be reported as 3, 6-σ 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: Taguchi Index

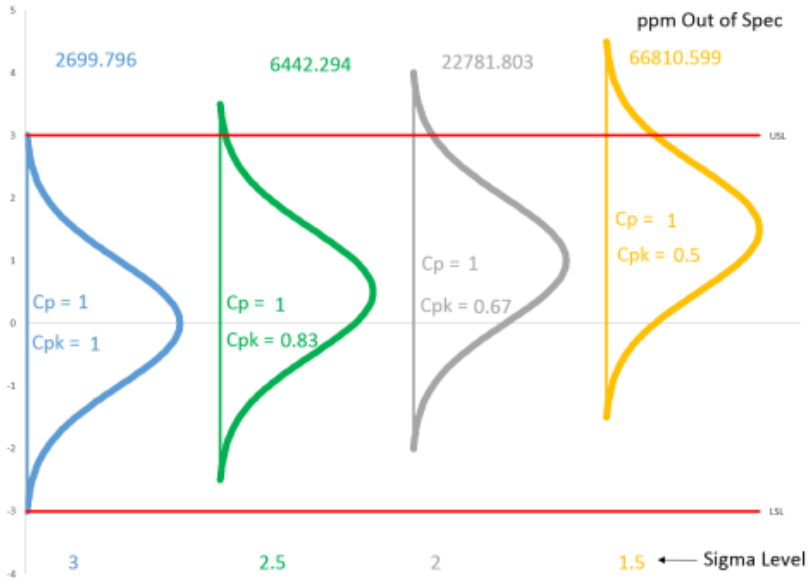


Figure 4.1: C_{pk} and C_p.
`https://www.spcforexcel.com/knowledge/process-capability/interactive-look-process-capability`.

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 $C_{pm} \leq C_p$.

Think about it. What is the relation between C_{pm} and p_{NC}?

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 are very important for small sample sizes, where point estimates unreliable. We will only present the simplest case of C_p, but results are available for all other PCRs. Being a monotone transformation of σ, we can call upon confidence intervals for the variance of a normal population, so that with probability $1 - \alpha$:

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

where

$$\hat{C}_p = \frac{USL - LSL}{6s}. \tag{4.12}$$

Eq.(4.11) is simply derived from Eq.(4.13). Intervals for the other capability indexes, are available in Montgomery (2007) and references therein.

4.1.4 Testing Hypotheses on Capability Ratios

Consider a supply contract, which requires production to have $C_p > 1.5$. It may easily be the case, that $C_p > 1.5$, 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, and when C_p is estimated as in Eq.(4.12), then

$$(n-1)\left(\frac{C_p}{\hat{C}_p}\right)^2 \sim \chi^2_{n-1}. \tag{4.13}$$

Eq.(4.13) and a little algebra gives us the rejection limits, were we to use $\hat{C}_p$ 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 $1-\alpha$ test using $\hat{C}_p$ is

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

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 the more general tests:

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

Think about it. Look at the relation between the hypotheses tested and the direction of the rejection region. Do you see why the confidence interval corresponds (only) to the two-directional hypothesis test?

Think about it. If σ is estimated from a very large sample I can trust it without bothering with hypothesis testing. Try to show this formally using what you know of the distribution of $\hat{C}_p$.

4.1.5 Other estimators of process capability

The first C_p estimator we proposed used s to estimate the σ (Eq.4.12). While a seemingly very natural estimator, it is actually not the standard in the quality engineering community. The scaled range (Def 6) of the data is more often recommended:

$$\hat{C}_p := \frac{USL - LSL}{6\,Range(x)/d_2} \tag{4.14}$$

$$\tag{4.15}$$

where d_2 is some scaling constant so that the range actually estimates σ. It may be surprising that $Range(x)$ is preferred over s to estimate σ. This is indeed a particularity of capability indexes, and it is justified by that fact that $Range(x)$ is very sensitive. Unlike s, any single defect in the production will increase $Range(x)$, thus lower $\hat{C}_p$.

The C_p estimator in Eq.(4.12), using s to estimate σ, is actually known as the $\hat{P}_p$ *performance index*. It was promoted in 1991 by the Automotive Industry Action Group (AIAG) as a capability estimator which is informative even in the presence of defects. If robustness is indeed desirable, one may consider the following estimators:

Performance Index

$$\hat{C}_p := \frac{USL - LSL}{6MAD(x)} \tag{4.16}$$

$$\hat{C}_{pk} := \min\{\frac{USL - x_{0.5}}{3MAD(x)}, \frac{x_{0.5} - LSL}{3MAD(x)}\}, \tag{4.17}$$

$$\hat{C}_{pm} := \frac{\hat{C}_p}{\sqrt{1 + (\frac{x_{0.5} - T}{MAD(x)})^2}}. \tag{4.18}$$

When using a robust capability estimator, attention should be paid to its interpretation. The production may actually be out of control, so these indexes measure not the capacity of the actual process, but rather the *potential capacity*, once the process will be brought into control.

Remark 3. At this point, I hope you are wondering why isn't p_{NC} used as a capability index. Well, it is! It is simply not called a "capability index", simply because the term is reserved to C_p, C_{pk}, C_{pm} etc.

4.2 Bibliographic Notes

This chapter is based almost entirely on Montgomery (2007).

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

5.1 The Run Chart

The simplest possible control chart is the *run chart*, in which each measured CTQ is simply plotted against the time of measurement. If strong anomalies exist, or temporal patterns, they may be already visible in the run-chart. On the down-side, the run-chart is essentially a statistical test to detect out-of-control behaviour based on a single observation at a time. Knowledge of statistical hypothesis testing suggest we can do better. Enter the $\bar{x}$-chart.

5.2 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 4 (Control Charts and Capability Analysis)**.** While seemingly very similar ideas, there is a fundamental difference between capability analysis and process control: process control compares to the *statistical regularity in the past*, while capability analysis compares to *specification*. Process capability and control charting ideas may be compounded, 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* (μ_0), lower and upper control limits (LCL, UCL), and the statistic $\bar{x}_t$ evolving in time. If at each

period t we compute the average of the n samples of the period. We denote

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

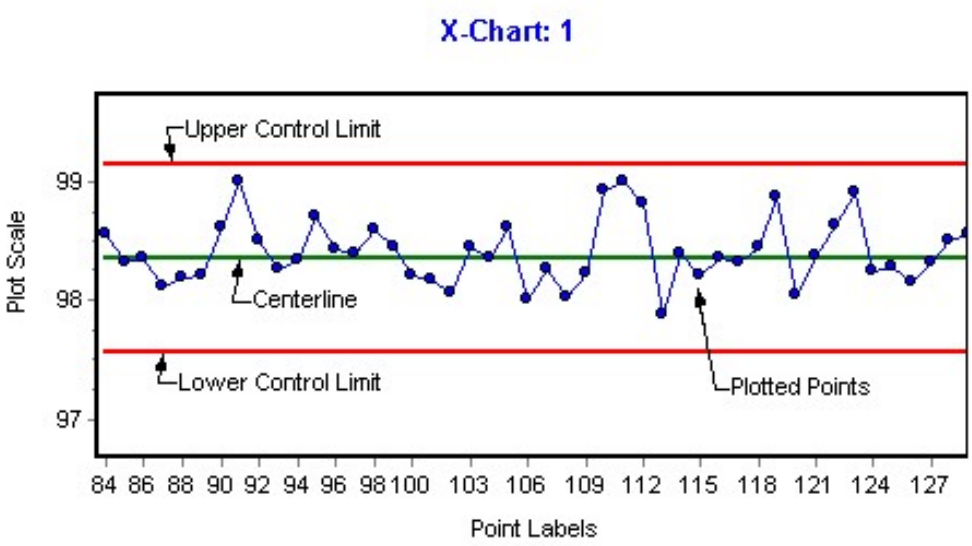


Figure 5.1: $\bar{x}$-chart.
`https://mvpprograms.com/help/P-mvpstats/spc/WhatAreControlCharts`

Phase I/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 ($\hat{\mu}_0$) and stadard errors $\sigma_{\bar{x}}$. After the chart has been calibrated, and major assignable sources of variability removed, we can finally start monitoring the process, known as *Phase II*.

Figure 5.1 makes it evident $\bar{x}$-chartrequires us to make several design decisions. A standard design decision is setting the centerline as average of a subset of τ periods that we believe to be in control. Phase I may provide us with such a sample.

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

where μ_0 denotes the in-control expectation of the process, and summation is over the τ in Phase I we believe to be in control. Notation originates from treating the in-control process as a null hypothesis, as it should be thought of.

Back to the design decisions we make when designing a control chart.

Design decisions

1. Centerline (μ_0).
2. Upper and lower confidence limits: UCL and LCL (do not confuse with USL and LSL!).
3. Sample size in each sample, denoted n.
4. The within period sampling scheme, known as *rational groupings*.

5. The between-period sampling scheme, notably the *frequency of samples*, denoted h.

6. 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 5.7 where we consider these choices as an economical optimization problem.

For ease of exposition, control chart design is demonstrated for the $\bar{x}$-chart, but equally applies to other control charts, presented in Section 5.3.

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 choosing $UCL := \mu_0 + L\sigma_{\bar{x}}, LCL := \mu_0 - L\sigma_{\bar{x}}$, $\mathbf{x}_{it} \sim \mathcal{N}(\mu, \sigma^2)$, and $\sigma_{\bar{x}}$ being the standard deviation of the statistic being monitored, which we can estimate, for example, from Phase I. For instance: $\widehat{\sigma^2_{\bar{x}}} = \frac{1}{\tau - 1}\sum_t^\tau (\bar{x}_t - \hat{\mu_0})^2$ In this case, $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. 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

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

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

and the rest follow from the distribution of $\bar{x}_t$ 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.: $\mathbf{x} \sim_{H_1} \mathcal{N}(\mu_1, \sigma^2); \mu_1 = \mu_0 + k\sigma_{\bar{x}}$, we plot in Figure 5.2, the detection power of a 3-sigma $\bar{x}$-chartas a function of k. This is known in the statistical literature as a *power function*, and in the engineering literature as the *operator characteristic* (OC).

Operator Characteristic

Remark 5 (3-Sigma rule)**.** At the risk of stating the obvious, the "sigma" in the 3-Sigma rule will always be that of the statistic being monitored. If we are monitoring $\bar{x}_t$, we will use $\pm 3\sigma_{\bar{x}}$. In particular, if $n = 1$, then obviously $\sigma_{\bar{x}} = \sigma_x$. More generally, if we are monitoring some $T(x)_t$, we will use $\pm 3\sqrt{Var\left[T(x)_t\right]}$.

Remark 6 (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. It tells you nothing on the probability of non-compliance. 3-Sigma Capability is a statement on the probability of a unit to be defect. Monitoring production via its capability is very rare, as we would like the alarms to sounds long before defect units leave the production line.

Extra Information. [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 Reciever Operator Charachteristic (ROC). The curious reader is referred to Wikipedia (2015f) for more information.

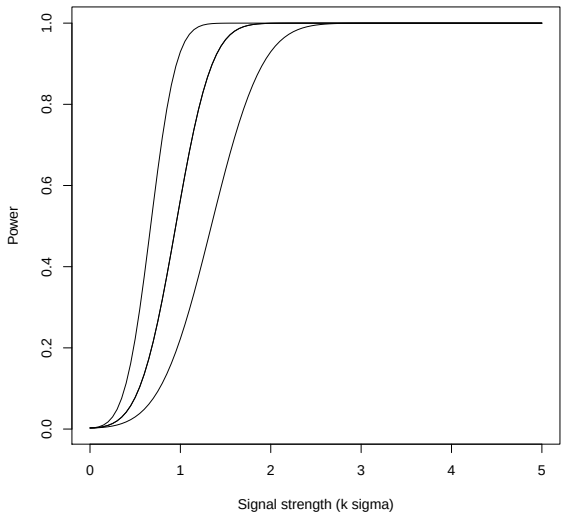


Figure 5.2: Power function of the 3-sigma $\bar{x}$-chartwith $n = 5, 10, 20$ and $\mu_1 = \mu_0 + k\sigma$.

A very important 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 statistically independent and α_t, π_t fixed in time, 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

$$ARL_0 = 1/\alpha, \tag{5.8}$$
$$ARL_1 = 1/\pi. \tag{5.9}$$

ARL is measured in periods. We can convert to time units by multiplying the ARL by the duration of sampling interval (h). 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 7 (ARL more important than type-I error)**.** 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 will occur with certainty if the process runs long enough, then it is actually the ARL that is more informative than type-I errors when designing a control chart.

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 h. Brilliant again! Nothing to lose, except time and money.

4. Change the sampling scheme within period. We elaborate on this in Section 5.2.2.

[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 5.2.3.

6. Pool together more periods. See Section 5.4.

7. Measure many CTQs simultaneously. See Section 5.5.

5.2.1 Control Limits and the Alarm Rate

Ceteris paribus, 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, roughly one year if 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 ARLs.

Extra Information. [Non Normal CTQ] If normality of $\bar{x}_t$ can be assumed, then one may estimate μ_0 and $\sigma_{\bar{x}}$ from phase I, and set LCL and UCL by finding the L that solves $2\Phi(-L) = \alpha$, for some desired α. If normality cannot be assumed, there are many ways to go about. Here are some options:

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

2. 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).

3. Use empirical quantiles: If phase I has 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 $P(\mathbf{x} \notin [\hat{\mathbf{x}}_{\alpha/2}, \hat{\mathbf{x}}_{1-\alpha/2}]) \approx \alpha$. This is a **super practical** way to go about. The only downside of this avenue, is that you do not always have enough in-control data from phase I.

5.2.2 Rational Groupings

Recall that at each period we compute the average of n samples. To fix ideas, think of a period being a day of production. How should we draw samples in this period? 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.

5.2.3 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 regions. It is quite common to define *warning limits* 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*, and it is a very efficient way to detect anomalies. Another approach is to define multiple sets of stopping rules.

Adaptive Sampling

WECO Rules

1. One or more points[a] outside of the 3-sigma control limits.
2. Two of three consecutive points outside the 2-sigma 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 centerline.

[a]By "point" we mean the computed statistic at each period: $\bar{x}_t$.

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. The good news is the analysis in Eq.(5.10) does not apply to WECO, because the rules not independent but rather highly dependent. A can compute the ARO_0 of WECO in a quick simulation.

Extra Information. [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 review.

5.3 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.

5.3.1 R Chart

Where $\bar{x}$ is replaced by the range, $\max_i\{x_{it}\} - \min_i\{x_{it}\}$. This chart is sensitive to many changes in the distribution of the CTQ; the variance in particular.

5.3.2 s Chart

Where $\bar{x}$ is replaced by s. Sensitive to variability changes. This is an important and useful chart which we will revisit in Section 5.6.2.

5.3.3 s^2 Chart

Like the s chart, only in variance-scale.

5.3.4 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. This is very useful if you allow the distribution of the CTQ to vary with some covariate, and you want to detect a change in this relation. To fix ideas, think that your CTQ depends on the temperature at the time of production. The distribution of the CTQ will thus vary with the temperature, but the break in their relation is cause for alarm.

5.3.5 Derivative Chart

If the break of control has occurred between periods ("the night shift guys broke it!"), the change in an $\bar{x}$-chartmay carry more information (i.e., more power) than the value of $\bar{x}_t$ itself. We can thus chart, not $\bar{x}_t$ itself, but rather, the *change* in $\bar{x}_t$ between periods. It is quite possible $\bar{x}_t$ seems perfectly ok, but that $\bar{x}_t - \bar{x}_{t-1}$ seems highly irregular. Generalizing this idea, we can monitoring the *derivative* of some statistic to detect this kinds of changes.

5.3.6 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., categorical CTQs.

5.3.7 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.

5.3.8 u Chart

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

5.4 Pooling Information Over Periods

Assume an out-of-control process is a very mild shift of the mean controlled-process (μ). A power analysis may suggest that this shift is hard to detect, 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 8 (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 9 (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 $n = 1$.

5.4.1 Moving Average Chart (MA)

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

Definition 22 (MA)**.** The *moving average* (MA) in a window of w periods ending at period t, is defined as

$$M_t := \frac{\bar{x}_t + \cdots + \bar{x}_{t-w+1}}{w}. \tag{5.11}$$

Assuming $\bar{x}_t \sim \mathcal{N}(\mu, \sigma^2_{\bar{x}})$ then clearly

$$M_t \sim \mathcal{N}\left(\mu, \sigma^2_{M_t} = \frac{\sigma^2_{\bar{x}}}{w}\right). \tag{5.12}$$

The control limits on M_t are typically

$$UCL := \mu_0 + L\sigma_{M_t} = \mu_0 + L\,\frac{\sigma_{\bar{x}}}{\sqrt{w}}, \tag{5.13}$$

$$LCL := \mu_0 - L\sigma_{M_t} = \mu_0 - L\,\frac{\sigma_{\bar{x}}}{\sqrt{w}}. \tag{5.14}$$

For $L = 3$ 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, possibly using a power analysis as a function of w.

5.4.2 Exponentially Weighted Moving Average Chart (EWMA)

The moving average gives all observations in the window 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

$$z_t := \lambda\bar{x}_t + (1 - \lambda)z_{t-1}. \tag{5.15}$$

By recursive substitution, we have

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

Assuming $\bar{x}_t \sim \mathcal{N}(\mu, \sigma^2_{\bar{x}})$ then

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

$$\sigma^2_{z_t} = \sigma^2_{\bar{x}} \left(\frac{\lambda}{2-\lambda}\right)(1-(1-\lambda)^{2t}). \tag{5.18}$$

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

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

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.

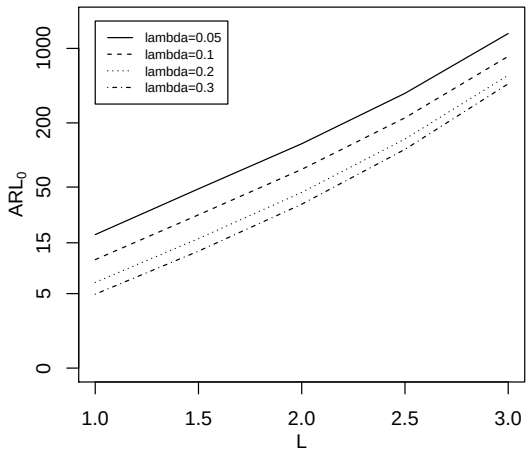


Figure 5.3: ARL_0 for EWMA.
Code from `http://users.phhp.ufl.edu/pqiu/research/book/spc/r-codes/fig53.r`

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.

5.4.3 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

$$C_t := \sum_{j=0}^{t} (\bar{x}_j - \mu_0) = (\bar{x}_t - \mu_0) + C_{t-1}. \tag{5.20}$$

Assuming $x_t \sim \mathcal{N}(\mu, \sigma^2_{\bar{x}})$ then if under control then $C_t \sim \mathcal{N}(\mu_0, t\sigma^2_{\bar{x}})$, we could thus set

$$\begin{aligned} UCL &:= \mu_0 + L\sigma_{C_t} = \mu_0 + L\sqrt{t\sigma^2_{\bar{x}}}, \\ LCL &:= \mu_0 - L\sigma_{C_t} = \mu_0 - L\sqrt{t\sigma^2_{\bar{x}}}, \end{aligned} \tag{5.21}$$

and $L = 3$ as usual. You may encounter these limits in your favourite software (`qcc` package in **R**), but it less often discussed in the literature. Montgomery (2007) for example discuses very different limits. The discrepancy is explained in the following Extra Info.

Extra Information. CUSUMs were introduced by Page (1954). 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. 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*. The limits in Page (1954) are far from intuitive. This author has found Ritov (1990) to be the best explanation of Page (1954), but do not expect an easy reading...

V-Mask

Adaptive Contol Limits

5.5 Multivariate Control Charts for Location

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.

5.5.1 Mass Univariate Control

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

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

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

5.5.2 Hotteling's T^2

Hotteling's T^2 statistic is a generalization of the t-test due to Hotelling (1931). To emphasize the relation to the t-test we write the classical t-statistic in the following weird form:

$$t^2(x_t) = n(\bar{x}_t - \mu_0)(\hat{\sigma_x}^2)^{-1}(\bar{x}_t - \mu_0). \tag{5.23}$$

This notation readily extends to the multivariate case. For p CTQs, then $\bar{x}_t$ and μ_0 are p-length vectors, and $\hat{\sigma}^2$ is replaced with the $p \times p$ covariance matrix $\hat{\Sigma}$. Both μ_0 and Σ can be estimated from Phase I.

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

$$T^2(x_t) := n(\bar{x}_t - \hat{\mu}_0)'\hat{\Sigma}^{-1}(\bar{x}_t - \hat{\mu}_0). \tag{5.24}$$

To derive the control limits, we will be assuming that $\mathbf{x}_{it}$ is p-variate Gaussian distributed, $\mathbf{x}_{it} \sim \mathcal{N}(\mu_0, \Sigma)$. Put differently, we assume that the i'th sample at period t is a p-vector, such that its j'th coordinate $\mathbf{x}_{ijt}$ is univariate Gaussian, with mean $\mu_{0,j}$ variance $\Sigma_{j,j}$, and covariance with some other CTQ, $\mathbf{x}_{ij't}$, is given by $\mathbf{Cov}[\mathbf{x}_{ijt}, \mathbf{x}_{ij't}] = \Sigma_{j,j'}$.

The sampling distribution of T^2 may depend on how μ_0 and Σ are estimated. For our purposes, where μ_0 and Σ are estimated in Phase I, we can safely use the following approximation

$$T^2 \overset{H_0}{\rightsquigarrow} \chi^2_p. \tag{5.25}$$

We can thus construct the control limit for this scenario:

$$UCL := \chi^2_{1-\alpha,p}. \tag{5.26}$$

Think about it. Why is there no LCL? Think of a two-sided t-test ...

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

Remark 10 (Multivariate t or Multivariate Z?)**.** For simplicity, I do not distinguish between a t statistic and a Z statistic. Since it is implied that variances are estimated in Phase I, then all t-statistics are actually Z statistics. If variance were to be estimated in each period, then obviously my t-statistics would be proper t-statistics, and the reader should really consult (Qiu, 2013, Ch.7) for details.

5.5.3 Drill Down

Your control chart just went outside the control limits and an alarm sounded. The next natural question- what made the process go out of control? In the univariate case, you should call your specialists team. In the multivariate case, there is another analysis stage you can/should do: *drill down*. By this we mean that we look an each of the p processes separately to see which processes triggered the alarm.

The most natural way to drill-down, is to look at the raw univariate processes and see if any is out of the control limits. Is it possible that a Hotelling alarm was sounded, but no single process seems out of control. Absolutely! This was the whole point for which we used a multivariate control: because the signal is so mild that any single process had no power.

Extra Information. [Many name to the drill-down] In the electrical engineering literature, the "drill down" is also known as *signal identification* that follows the initial *signal detection*. In the analysis of variance literature (ANOVA) this is a *post-hoc* test that follows an *omnibus test*. In the multiple testing literature, this is simply a multiple test.

Extra Information. [Consonant tests] As previously stated, it is possible that a multivariate alarm went off but no single variable seems out of control. This should not surprise you, since the whole motivation for doing the multivariate analysis was to draw power from several processes at the same time. The statistical literature does provide sets of multivariate and univariate tests that have to agree, meaning that it is impossible for the multivariate test to be rejected, without at least one univariate test to be rejected. This is known as *consonant* tests. For more on that, the reader is referred to Goeman and Solari (2011).

5.6 Multivariate Control- Extensions

Hotelling's T^2 is perhaps the first and most used multivariate test statistic. Just like the univariate t-test, however, it is hardly the only multivariate test ...

5.6.1 Temporal Pooling of Multivariate Charts

The basic problem: can we gain power by pooling multivariate data over several periods? Sure!

Definition 25 (Moving Sum Hotelling)**.** The *Moving Sum Hotelling* chart in a window of w periods ending at t, is defined as

$$M_t := T^2(X_t) + \cdots + T^2(X_{t-w+1}) \tag{5.27}$$

where T^2 is Hotelling's test statistic X_t is the matrix of n samples on p measurements at period t.

As before, we will typically, but not necessarily, use $\hat{\Sigma}$ from period I in T^2 .

Since $T^2(X_t)$ is approximately χ^2_p distributed, periods are independent, and sums of Chi-square distributions are still Chi-square with the summed degrees of freedom[2], we have that for a process in control

$$M_t \sim \chi^2_{wp}. \tag{5.28}$$

This readily suggests that the UCL for this chart is $\chi^2_{1-\alpha,wp}$.

Think about it. Is the moving sum Hotelling a Shewhart chart? Can you design more multivariate charts with period pooling? Can you derive their control limits?

Example 3 (Vine Health)**.** Consider vine crops. The crops are monitored for their health. The health measurements (the CTQs) may be considerably affected by irrigation, weather, etc. An unhealthy vine is thus one that behaves differently than the rest. The signal for out-of-control we seek is thus not in the mean CTQ of the vine, but rather in its correlation with others. We would like a multivariate version of the s chart to monitor the health of our vines.

5.6.2 Multivariate s^2 Chart

As discussed in the vine health example (Example 3), there are cases where the out-of-control behaviour is manifested in the covariance between CTQs, and not in their mean. Recalling the Exploratory Data Analysis Section (2), the most popular measure of multivariate relation is the *covariance matrix*, $\hat{\Sigma}$ (Definition 13). To monitor changes in the correlations, we would like to compare the current covariance to its historic values, Σ_0. We would thus like to monitor $\hat{\Sigma}_t - \Sigma_0$, which is a $p \times p$ matrix. While we already know how to plot a matrix, plotting the evolution of a matrix in time, as is required for a control chart, is no simple task[3]. Without going into the details of multivariate statistics, the fundamental idea is to summarize the matrix by a single number, so that it may be easily plotted, and control limits computed. We now provide several functions which can be thought of as *norm* functions, i.e., functions that measure the "distance" from $\hat{\Sigma}_t$ to Σ_0.

[2]Trust me on this one.

[3]Think about it. If a matrix is an image, then the evolution of a matrix in time is, well, simply a movie.

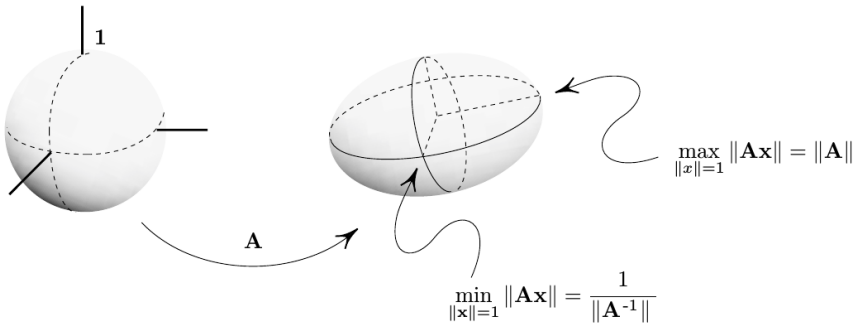


Figure 5.4: On the left: the unit circle in $\mathbb{R}^3$. On the right: the ellipse is the image of the unit circle after transformed by matrix A. The operator norm is the distance between the origin and the furthest point on the ellipse.
Source: Meyer (2001).

Definition 26 (Frobenius norm)**.** The *Frobenius norm* of matrix A, a.k.a. the *Hilbert-Schmidt norm*, or *Schur norm* is defined as:

Hilbert-Schmidt norm, Schur norm

$$\|A\|_{Frob} := \sqrt{\sum_{ij} A_{ij}^2} \tag{5.29}$$

As such, it merely treats the matrix as a stacked vector, and computes the Euclidean distance of the vector from zero.

Using the Frobenius (matrix) norm for the process control, we would be plotting the time evolution of $\|\hat{\Sigma}_t - \Sigma_0\|_{Frob}$.

Definition 27 (Spectral norm)**.** Denoting the Euclidean distance of a vector x from zero by $\|x\|_{Euc}$, we can denote a matrix's *spectral norm*, a.k.a. *operator norm*, or *induced norm*, by:

Operator norm, Induced norm

$$\|A\|_{Spec} := \max_x \{\|Ax\|_{Euc}; \|x\|_{Euc} = 1\}. \tag{5.30}$$

The geometry of this definition is depicted in Figure 5.4.

Using the Spectral (matrix) norm for the process control, we would be plotting the time evolution of $\|\hat{\Sigma}_t - \Sigma_0\|_{Spec}$.

We will not present control limits for the generalized s charts, but we do remind the reader that given a long enough Phase I, we can always look at the past values of the statistic for choosing the control limits.

Think about it. The drill-down after a multivariate s^2-chart went out of control is more complicated than the drill-down after a multivariate $\bar{x}$-chartwent out of control. Why is that?

Think about it. Like the Hotelling chart, there is no LCL, for these charts, only UCL. I.e., we only reject for large values of the matrix norms. Why is that?

Extra Information. [Matrices and graphs] Since a matrix defines a *weighted graph* (and vise-versa), a.k.a. a *network*, any network similarity measure can be used to measure the distances between matrices, and can thus be used to summarize the covariance matrix and plot in a control chart.

5.7 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 11 (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:

$$\max_{n,L,h}\{\mathbf{E}\left[C/T\right]\} \tag{5.31}$$

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 hours between sample periods.

The parameters governing the solution are:

Income$_0$ Brute income per production cycle when in control, without process control expenses. Denoted V_0.

Income$_1$ Brute income per production cycle when out of control, after accounting for recall, legal damages, etc., but without process control expenses. Denoted V_1.

Fixed sample cost Fixed cost of sampling. Denoted a_1.

Variable sample cost Variable cost sampling. Denoted a_2.

Cost of True Positive The cost to investigate and find an assignable cause; a_3.

Cost of False Positive The cost to investigate a false alarm; a_3'.

Think about it. How will the optimal n, L, h change if the parameters of the problems are changed? If V_1 is increased? If a_3 is decreased?

For the purpose of this course, we will content ourselves with the above informal sensitivity analysis. For completeness, the precise optimization problem is given in the next Extra Info.

Extra Information. We now need to establish how $\mathbf{E}[C/T]$ 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), $\mu_1 = \mu_0 \pm k\sigma_x$.
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
$$\tau := \frac{1-(1+\lambda h)e^{-\lambda h}}{\lambda(1-e^{-\lambda h})}. \tag{5.32}$$
6. The power to detect an OC is
$$\pi := \Phi(-L - k/\sqrt{n}) + (1 - \Phi(L - k/\sqrt{n})). \tag{5.33}$$
7. The false alarm rate
$$\alpha := 2\Phi(-L). \tag{5.34}$$
8. Because of the Poisson process assumption, $\mathbf{E}[C/T] = \mathbf{E}[C]/\mathbf{E}[T]$.
9. The expected cycle length:
$$\mathbf{E}[T] = \frac{1}{\lambda} + \frac{h}{\pi} - \tau + D. \tag{5.35}$$
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
$$\mathbf{E}[C] = V_0\frac{1}{\lambda} + V_1\left(\frac{h}{\pi} - \tau + D\right) - a_3 - \frac{a_3' e^{-\lambda h}}{1-e^{-\lambda h}} - (a_1 + a_2 n)\frac{\mathbf{E}[T]}{h}.$$

Given all the above, we may now plug Eq.(5.31) into our favourite numerical solver to find the optimal h, L, n.
[TODO: add timeline]

5.8 Bibliographic Notes

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). A quick digest review may be found in Natrella (2010). For multivariate process control, see for example Ge and Song (2012). For a technical discussion of multivariate statistics see Anderson (2003). For linear algebra, in particular matrix norms, see Meyer (2001). For some recent advances on high-dimension multivariate tests see Srivastava (2013) and refer-

ences therein. An excellent(!) reference for many useful, and regrettably overlooked, statistical techniques, see Wilcox (2005).

Chapter 6

Design of Experiments

This chapter is devoted to the matter of designing studies, and follows the lines of Cox and Reid (2000) and Cox and Donnelly (2011). As we will see, empirical studies come in many forms and shapes. The most fundamental distinction is between *observational studies* and *designed experiments*. This distinction is performed along our ability to control the studied effects, and will have crucial implications on the interpretation of results, and in particular, on causality claims.

Observational Studies

6.1 Challenges in Empirical Studies

The challenges in all empirical studies, be it designed experiments, or observational are:

Systematic errors Can be thought of as bias, even thought much more general than the narrow definition in statistical theory. Refers to the distortion of conclusions due to confusion sources that do not cancel out in the long run.

Non-systematics errors Can be thought of as the "noise". If the non-systematic errors are large, then noise may mask the signal.

Uncertainty in conclusions Only in the formalism of mathematics and logic things are true or false with certainty. In natural sciences we have assumptions, and confidence levels. Each conclusions should thus be accompanied by the level of certainty in which it is stated.

Efficiency The above goals may certainly be achieved given unlimited resources. Achieving them quickly and cheaply, is a worthy goal, and a veritable art.

Range of validity Do our conclusions hold in all countries? Weathers? Altitudes? Races? How general are our conclusions?

6.1.1 Dealing with Systematic Errors

Systematic errors arise from two sources: either the study design is such that it does not measure what we think it measures, or leakage of personal judgment.

The first source of systematic errors can be dealt at the design stage, or at the analysis stage. Much emphasis is given in the literature to unbiased estimation, but reality is that a good design can save a lot of trouble at the analysis stage. Later in this chapter, we will discuss in detail the means to avoid systematic errors.

6.1.2 Dealing with Non-Systematic Errors

Informally speaking, non-systematic errors, i.e. noise, is reduced by:

Compare like with like Comparing similar units. This is done by the sampling scheme.

Replications Repeat a measurement enough times, and the non-systematic error will average out.

6.2 DOE Preliminaries

We will now discuss *designed experiments*. Observational studies are briefly discussed in Section 6.12.

When designing a product (remember DFSS 1.4.7), or once a control chart has signaled an alert, we will want to know what *controllable inputs* influenced our process, and how to reduce its variability. In our SPC terminology, we will want to know what are the *causal effects* of our controllable inputs, on our *CTQ*. The theory of discovering these effects is the theory of *design of experiments* (DOE), which uses a a slightly different terminology which we will define in the next section.

6.2.1 Terminology

Many, if not most of the following terms, originate in R.A. Fisher's seminal book "The Design of Experiments" (Fisher, 1960). As such, the DOE literature is rich in agricultural terms, due to its historical origins. When old ideas get new names, we try to emphasize this in the text.

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

Experimental Unit Entity on which a measurement or an observation is made.

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*, or *cell*.

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 $\{-1, 1\}$.
2. Dummy coding: where levels are encoded with $\{0, 1\}$.

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

Response The CTQ in the SPC literature. The y variable in the regression literature. The *response* in the DOE literature.

Main Effect Change in the expected response between two factor–levels. We emphasize that effects, unlike simple population parameters, imply a causal relationship. Akin to the *assignable causes* in the SPC literature, and β's in the regression literature.

Interaction Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors. Part of the *assignable causes* in the SPC literature.

Noise A.k.a. *error*. The variability in the response that cannot be attributed to any factor. This is the *common* variability in the SPC literature, and the ε in the regression literature.

Replication A single repetition of an entire experiment, i.e., a single measurement of the response in all experimental conditions.

Repeats Repeated measurements on the response under the same conditions, i.e., under the same treatment. A.k.a. *repeat tests*.

Covariate An variable that influences the response but is unaffected by any other experimental factors. We cannot change at will, but we can *control/account for them*.

Non Specific Factor A variable that we suspect to affect the response, but we can only vaguely define, thus impossible to measure. As a consequence, we will not care to study its effect, but rather just remove it (e.g. by blocking). Think of "life style" as an example of a non-specific factor.

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*. The blocking in the design needs to be accounted at the time of the data analysis.

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

Several matters should be emphasized before we dive in.

Randomization is the random assignment of units to treatments. It is fundamental to our purpose because the idea of an *effect* implies *causality*. Since we seek to intervene in the production to reduce variability, we only care about causal effects. It is the mechanism of *randomization*, that allows us to conclude that the correlations we find are indeed causal, and not merely statistical. For a treatment of causal inference in non-randomized experiments, see the Bibliographic Notes section.

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 meetings, planning, and expertise 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 in real life problems rarely obey text-books. 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 for several reasons. First, because Cox and Reid (2000) do so. Second, because these are two different thing. The ANOVA framework may be replaced by the framework of *linear models*, *mixed models*, *variance components*, and possibly others analysis frameworks. There is a vast literature focusing on the analysis. If asked, this author may recommend Hocking (1985), which presents both the ANOVA terminology, and the linear models terminology.

ANOVA

6.3 Systematic Errors in DOE

Dealing with systematic errors is possibly the greatest concern in causal inference. Be it a medical intervention, an industrial production process, an economic policy, or a social welfare plan- all of these critically rely on a fair assessment of the magnitude of the effect of the intervention. Bias is disastrous.

Bias originates from two sources. The first is some unwelcome property of the data collecting process. The second is by the leakage of some personal judgment into the measuring or analysis.

Example 4 (Judgment leaking into measurements)**.** Consider cancer patients assigning themselves to the treatment or placebo groups. This *self selection* will clearly bias the reported effect of the new drug (who would willingly choose the placebo?!?).

Self Selection

Example 5 (Judgment leaking into analysis)**.** Return to the new-drug experiment. This time, imagine the data analysis is done by an employee of the pharmaceutical company, knowing that "Drug A" is the new drug his company has been working on, which will grant him a considerable bonus if shown efficient. Can he be expected to perform a fair analysis?

Two ways to deal with such bias:

Balanced Design In the simplest interpretation of "balance" we mean a design where an equal number of units is assigned to each treatment. More generally, we will keep "balance" by *randomization*, and implying some symmetry in the combinatorial design of the experiment.

Bliding By blinding we mean that personnel is unaware of the factors being studied. In regular blinding, subject are unaware of the treatment they are being administered. In *double blinding*, the experimenters and analysis are unaware of the treatments they are studying.

6.4 Non Systematic Errors in DOE

The idea that random samples come with (common causes of) variability, i.e. noise, should not be new to the reader. In this section, we will try to decompose variability into it sources, and learn several techniques to reduce them. Starting with a motivating example, which you should use to fix ideas as you progress along this section.

Example 6 (Web Design)**.** Consider the problem of optimizing a web site, where individuals performance is measured by conversion rate (the probability of a new user to signup, purchase, etc). In the DOE language, the conversion is the *response*. A user (or ip address) is an *experimental unit*. The site's attributes are the *factors*. A particular site's design, i.e. a combination of factors, is the *treatment*. The users' attributes are *covariates*. The user's life-style, a *non-specific factor*. If the site's attributes affect differently users with different attributes, we say there is an *interaction*.

Example 7 (Two Competing Diets)**.** Consider the problem of comparing two nutrition diets. Clearly, the effect of a diet on life expectancy strongly interacts with both genetics and life style. It would be thus very nice if we could block experimental units, so that each block has a homogenous genetics and life-style. That is why **twins** are so popular when designing experiments!

6.4.1 Gage R&R Studies

R&R stands for *repeatability* and *reproducibility*. These experiments are aimed at assessing the irreducible non-systematic errors in measurement and *range of validity* of conclusions. As such, in a typical gauge R&R experiment, measurement will be made under the same conditions in different labs, by different technicians, etc.

6.4.2 Completely Randomized Designs

In the simplest of designs, all experimental units are randomly assigned to treatments. This is typically easy to implement. In Example 6 this would imply randomly assigning users to interfaces. In Example 7 this would mean ignoring the fact that the experimental units come in naturally blocked groups of two, and randomly assigning all people to treatments. If you suspect, as you should, that in both examples we may reduce variability by grouping units together, keep reading.

6.4.3 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.

In the twins example, nature has provided us with natural blocks of two. If comparing two treatments, we would naturally randomize treatments within twin-pair. This is known as a *randomized complete block design*. If we had three treatments, or more, to compare, we cannot completely randomize. This is knows as an *incomplete block design*. There are several approaches to incomplete block designs, but we refer the reader to (Cox and Reid, 2000, Sec.4.2) for details.

Complete Block Design

> **Think about it.** [Paired samples] While we try to avoid matter of data-analysis, it should be obvious to the reader that a two-group randomized block design should be analyzed as a **paired sample**. If adding a Gaussianity assumption, we have a paired-sample **t-test**.

In our web-design example nature has not provided natural homogenous blocks, but we can create them ourselves. We could block individuals along, say, the age covariate. We would then randomly assign users to layouts, only **within** age groups, so that all layouts are presented to each age-group. If each age group has as many users as there are different site designs, this is a randomized complete block design.

If there are more layouts than users per age group, this is an incomplete block design. Another idea is to use each user as his own block, by introducing a temporal dimension to the experiment. We could show all layouts to the same person. This particular type of blocking is known as a *crossover* design, discussed in Section 6.4.4.

> **Think about it.** With twins, you can have a complete block design comparing only two treatments. How would you go about to compare 3 treatments? And $k > 2$ treatments?

In summary, blocking is done in order to create homogeneous groups. Put differently: compare like-with-like. One can block in along factors, covariates, or non-specific factors. One can then

compare treatments only within blocks (e.g. the paired sample t-test), thus gaining accuracy by avoiding the variability between blocks. A good blocking strategy is such where blocks are very homogeneous within themselves, but very different between themselves, at least with respect to the response being studied. Is there any downside to blocking? Not statistically, maybe economically.

6.4.4 Crossover Design

Return to the diet comparison of Example 7. Now assume that 5 diets are being compared, and requiring quintuplets[1] will severely limit the available sample size. If the response of different people to the diets is strongly affected by genetics and\or lifestyle, it may be the case that by (randomly) assigning diets to individuals the diets effect will be completely masked by the variability between subjects. To solve this, we could give each subject all 5 diets. Each subject would act as his own block. Diet effect would be analyzed within subjects, and hopefully, having removed the variability between subjects, the diet effect would reveal itself.

Enter a new complication. What if the response to a particular diet is affected by the previous diet? This is known as a *carryover effect*, or *residual effect*, which may bias the main layout effects of interest. If all subject receive diet B after A, then the estimated effect of B will be biased by the carryover effect of A. A way to "average out" this carryover effect is to balance the treatment sequences. If each subject is randomly assigned to one of the 5! diet arrangements this balance would be achieved.

Carry-over Effect

The idea of administering all treatments to each unit, in a random sequence that balances carryover effects is known as a *fully randomized crossover design*. Table 6.1 demonstrates the $3! = 6$ sequences for a 3 treatment problem.

Fully Randomized Crossover

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

6.5 Efficiency in Design– Factorial Designs

In the previous section we discussed methods of reducing the non-systematic errors in the response to each treatment, so that we can better reveal differences between treatments. In the current section we try to break treatments to their building blocks; *factors*. By doing so we will gain both economical and statistical efficiency. Each sample will give us information on the effect of not one, but several *factors*. We will also be able to learn if the effect of factors are additive, or non additive, and we will be able to extrapolate from observed treatments, to unobserved treatments.

Example 8 (Web design revisited)**.** Back to the web-design from Example 6, assume now that treatments, i.e. layouts, differ in the location of a button, the colour of the heading and the size of the logo. If each has two possible levels, we deal with 3 factors with 2 levels each, totalling in 8 different layouts. Instead of estimating the effect of the 8 layouts, we estimate the effect of each of the 3 factors. Given a budget of, say, $n = 8$ observations, we would have one observation per

[1] Five identical siblings.

treatment. The magic of a factorial design is that we have 4 observations for each factor level. If we treat a web-design as the sum of its factor effects, we have gained accuracy

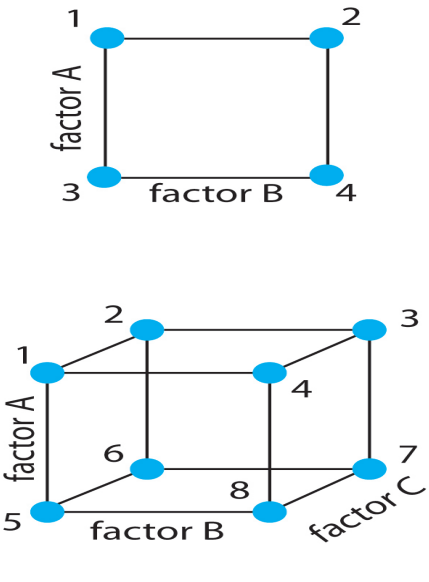
Think about it. Is the magic of the factorial design a free lunch? No. The lunch has a price. The price is that we have assumed that a site is the sum of its components, without interactions. Without this assumption, there is no statistical accuracy gained from a factorial design. There is however, the benefit that we can actually verify the validity of the additivity assumption.

6.5.1 Full Factorial Designs

A *full factorial*, or *complete factorial* design, is one where all factor-level combinations are replicated the same number of times. The factors $A, B, C, ...$ each with numbers of levels $l_A, l_B, l_C, \ldots$, define $l_A \times l_B \times l_C \times \ldots$ treatments. By far, the most common of the full-factorial designs is where all factors have two levels: $l_A = 2, l_B = 2, \ldots$. These is known as 2^k-design, where all combinations of the k factors is measured in each replication.

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 replication 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.



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`

With this 2^2 design, we may recover several effects:

Main effect of A The effect of varying A from $(-)$ to $(+)$.

Main effect of B The effect of varying B from $(-)$ to $(+)$.

Main effects with interaction The effect of varying both A and B from $AB = (--)$ to $AB = (++)$.

We typically denote the (mean) response for each treatment by the $\mu_{\odot}$ notation, where $\odot$ encodes the applied treatment by stating the conditions that were at their $+$ state:

Treatment	A	B	Mean
1	+	-	μ_a
2	+	+	μ_{ab}
3	-	-	$\mu_{(1)}$
4	-	+	μ_b

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. Interaction

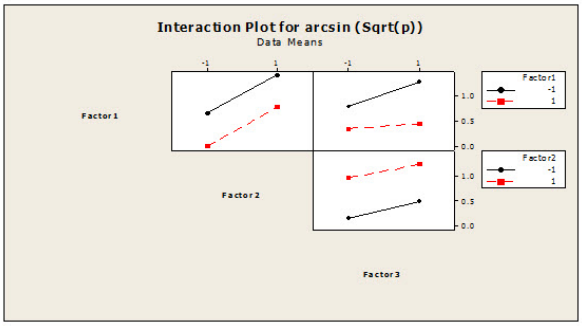


Figure 6.2: Interactions plot.
`http://blog.minitab.com/blog/statistics-in-the-field/optimizing-attribute-responses-using-desig`

Remark 12 (Popularity of 2^k)**.** 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.

Extra Information. [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 the effect would have been, had we 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 Continuous Factors Section 6.6.

3^k Designs

The name 3^k design is rather self explanatory. If a continuous variable is discretized to 3 levels, one may actually estimate and check for non-linear main-effects, which is impossible with 2^k designs. Then again, more than 2 levels are rarely treated as full factorial experiments. This is because 3 level factors typically appear when aiming at optimizing the factor level combination, for which the *response surface* methodology of Section **??** is more economical.

Full Factorial vs. One-Factor-at-a-Time

Importantly, a factorial design is better than k experiments with one factor at a time. This is because:

1. Factorial experiments make better (statistical) use of each sample unit for estimating main effects.

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

A classical illustration of the second point, is given in Figure 6.3. The figure depicts a one-factor-at-a-time optimization sequence (from A to H). Since the factors are not varied simultaneously, the experiments is unable to identify that the response is not a linear surface. Put differently, he is unable to estimate an interaction.

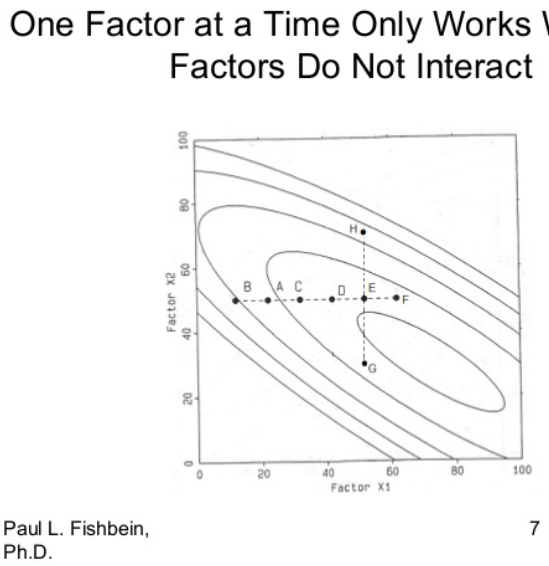


Figure 6.3: Optimizing factors, one at a time.
`http://www.slideshare.net/PaulFishbein/optimization-without-statistical-doe-2015-03-23-46817073`

6.5.2 Fractional Factorial

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, the information we give up is the least interesting.

Example 9 (From 2^2 to 2^{2-1})**.** As a first toy 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, which are enumerated in Table 6.2 along with the extractable information in each elimination.

The lesson from Example 9 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 design the experiment by going over all the $\binom{2^k}{2^k - 2^{k-p}}$ possible eliminations like in Example 9, but rather choose from pre-selected designs. Table 6.3, generated

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

with the `FrF2()` in the `FrF2` **R** package, is an optimal 2^{5-2} 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 6.3: 2^{5-2} design.

Extra Information. [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).

6.6 Continuous Factors

If aiming at screening factors, the fact that a factor is continuous is immaterial. One may sample a continuous factor at two points, and treat it as a 2 level factor. This is not the case, however, if aiming at characterizing the full response surface, for which the factorial designs are no longer appropriate.

When dealing with *continuous factors*, or *quantitative factors*, we have many more analysis strategies than when dealing with qualitative factors. All the following designs are more efficient than non-linear factorial designs such as 3^k. They gain efficiency by being *sequential* and not "one-shot" like the factorial designs.

Central Composite Design Aimed at finding optimal factor combinations. At each stage we decide which factors have a non-linear effect and add levels to these factors. Chooses the level do be added using a quadratic surface assumption.

Bayesian Optimization Like the Central Composite design, but replaces the quadratic surface assumption with a more general functional form[2].

Response Surface Methdology As the name suggests, aimed at recovering the response surface, and not only optimal combinations.

6.7 Taguchi Methods

Taguchi methods is a collective name for the philosophy, design, and analysis methods in industrial production applications promoted by Genichi Taguchi in 1970's Japan. Focusing on the design principles, we summarize Taguchi's emphasis:

1. Achieving low variability is no-less important, and more challenging then achieving a target value. Log variability ($\log s$) is thus often used as the response.

2. Factors which can be controlled only in a lab, but not in production, are deliberately varied in the lab.

3. An ample use of Plackett–Burman designs, and its extensions, to screen large numbers of factors. A *Plackett-Burman Design* is a fractional factorial design optimized for **screening**. It is a very economical design when studying the **main effects** of a **large number** of factors.

Plackett Burman

6.8 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 10 (Design for 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 **linear** model, the figure depict and intuition suggests, that it is preferable to spread the sampling points as far as possible.

Example 11 (Design for non linear regression)**.** Figure 6.5 demonstrates the effect of the different location of the sampling points (x) on the quality of the estimated regression line, in a **nonlinear** model. It may seems that unlike the linear case (Example 10) optimality is achieved in some intermediate spread of the xs.

Now for some facts, 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. For continuous factors with **linear** effects, the optimal design implies spreading the sampled factors levels as much as possible. This is no longer true for **non-linear** effects where the optimal design may depend on the unknown true effects.

[2] A Gaussian process prior, i.e., a function which's gradients are assumed to havea Gaussian distribution.

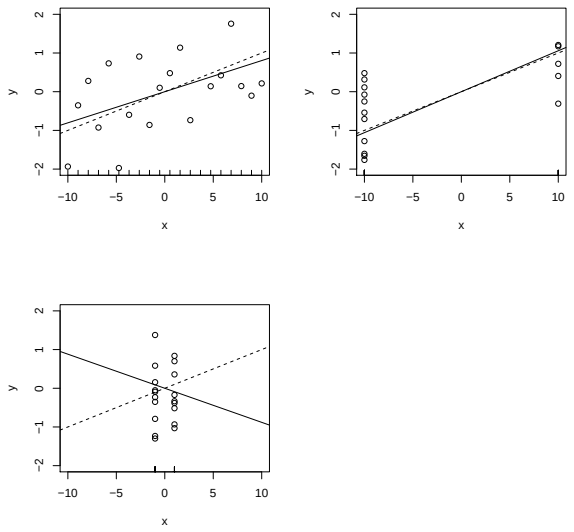


Figure 6.4: Design for linear regression. Different panels show different designs. True function as a dashed line. Estimated function as a full line.

6.8.1 Space Filling Design

The most natural of designs, which is particularly suitable when we have no a-priori assumption on the functional relation ($f(x)$) between the (continuous) factors 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 $f(x)$ is made available, then a space filling design is typically sub optimal (see Example 10).

Extra Information. [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).

6.8.2 Covariance Optimality

When estimating the effect of a single continuous factor, we would like a design that gives us the most information per observation on some effect β. This is the same as minimizing the variance of the estimator, $\min\{Var[\hat{\beta}]\}$, with respect to the design. In the case of linear regression with a single coefficient we know that

$$Var[\hat{\beta}] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{6.1}$$

Minimizing $Var[\hat{\beta}]$ is thus the same as spreading the x's as far as possible, in accordance with the intuition from Example 10. Now recall that in a multivariate linear regression problem with an $n \times p$ design matrix, then

$$Var[\hat{\beta}] = \sigma_\varepsilon^2 (X'X)^{-1}. \tag{6.2}$$

In this multivariate case, there may be several notions of "maximal spread". Denoting

$$M := (X'X)^{-1},$$

we can define:

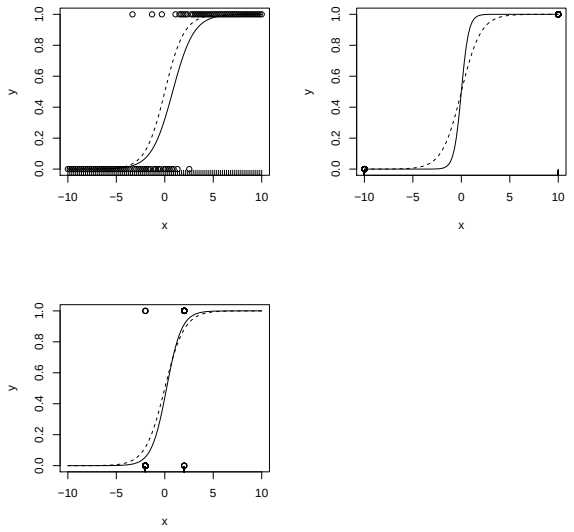


Figure 6.5: Design for non linear regression. Different panels show different designs. True function as a dashed line. Estimated function as a full line.

Definition 28 (A-Optimality)**.** A design is said to be *A-optimal* if it minimizes the average univariate variance. Formally: $\min\{\text{Tr}(M)\}$.

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 29 (D-Optimality)**.** A design is said to be *D-optimal* if it minimizes the volume of the confidence region for β. Formally: $\min\{\det(M)\}$.

Both A-optimality and D-optimality implicitly target linear models, such as in Example 10, because they aim at spreading the x's. From Example 11 we know this to be sub-optimal for non-linear models. For non-linear designs, one may opt for space filing designs (6.8.1), or consult Pukelsheim (1993).

Extra Information. [Other Optimality Criteria] There are as many optimality criteria as there are matrix norms. For instance, recall why not optimize the Spectral-Norm or the Frobenius-Norm of M, which we defined for the multivariate s^2 chart (Section 5.6.2). For a more detailed review, see Wikipedia (2015d). For a mathematical rigorous, and through treatment, see Pukelsheim (1993).

6.9 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, assume the null hypothesis is true and tested with the (ridiculous) test, where

each observation is tested as it comes. We will be making n tests, each at level α. If observations independent, the probability of a type I error is the familiar $1-(1-\alpha)^n > \alpha$.

We remind the reader that we already met sequential designs: These are the Central Composite Design, Bayesian Optimization and Response Surface Methodology, from the Continuous Factor section (6.6).

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 Information. [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.

6.9.1 Pooled Designs

Pooled designs, a.k.a. as *group testing*, or *hierarchical design*, appears when it is cheaper to measure the response for a group of experimental units simultaneously, rather then for each single one. It originated in the context of identifying soldiers with syphilis in WWII. It was the Harvard Economist Robert Dorfman that suggested that one may mix blood samples from several soldiers and then repeat the same in the subgroups that show the presence of syphilis (Dorfman, 1943). The method became immensely popular in the DNA analysis age, where technology allows biologists to test for the presence of particular protein in a sample from a group of cells (e.g. a blood sample).

Pooled designs broadly classify into two classes:

Adaptive group testing Where the groupings, and the groups to be tested depend on the previous results in the same experiment (see also *adaptive designs*).

Non-adaptive group testing Where the experiment consists of sequential group tests, but the design is fixed independently of the experimental results.

Think about it. Try to use the idea of group testing, and airborne-chemical sensors, to design a luggage inspection process at the airport.

Extra Information. [Bloom filters] If you are familiar with data structures and particularly *Bloom Filters*, then you will probably note that a Bloom filter is a data structure that groups objects in a way that facilitates lookup using group testing.

6.10 $A\backslash B$ Testing

Our web-site optimization example is interesting not only because it accommodates so many DOE practices, but it is also a practical problem of great interest. For historical reasons, the experimentation with several web-site layouts is not called a factorial experiment (which it is), but rather an *$A\backslash B$-test*.

Some particular characteristics of $A\backslash B$ testing include:

1. Users come by the millions. Sample size is rarely an issue, so that avoiding systematic errors is more important than avoiding non-systematic errors which can be averaged out.

2. Browser cookies, or login details define blocks. Technology permits to optimize the size for each block separately. If blocking is to very specific subgroups (female users, running Linux, that purchased on Amazon, etc.), then sample size, and thus variability, may become an issue again.

3. Sites have many layout parameters (locations, colours, sizes,...). Studying all these combinations may result on a formidable task. Bayesian Optimization and Central Composite designs will be preferred over full factorials.

6.11 Computer Experiments

Example 12 (Designing Wings)**.** Consider the problem of designing an air-craft'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 airplane, 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.

The wind design example (12) 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 factor levels 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 levels, 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. The analysis of computer experiment is very different than real-life experiment since we have no non-systematic errors. Space filling designs (Sec. 6.8.1), and Bayesian Optimization (6.6) being a particularly prevalent choices. See Sacks et al. (1989) or Santner et al. (2013) for further details.

6.12 Observational Studies

In this section we abandon designed experiments, in which we were able to randomize and choose the factors levels to be measured. In *observational studies*, we have no such controls. As such, in observational studies there are no *factors*, but rather, only *covariates*.

Example 13 (Post-sale testing)**.** Consider the notorious Galaxy Note7[3], or any car, plane, smartphone, or software. Bugs and malfunctions will always be found after launch[4]. One way to discover what triggers the malfunction is to try and replicate it in a controlled environment, i.e., in a designed experiment. A second, possibly more efficient way, is to collect post-sale failure data. This is why your software asks you to send crash reports. This is why plane and engine manufacturers equip their engines with sensors that constantly usage reports to the engine manufacturer headquarters.

[3]That would spontaneously burst into flames, until Samsung recalled all devices.

[4]In the pharmaceutical industry, this is known as the *Phase IV* of a clinical trial.

Observational studies classify into:

Cross Sectional Where each sample unit, i.e. individual, is observed in a single point in time.

Prospective Where each sample unit is observed on various occasions in the future, until some outcome occurs.

Retrospective Where measurements are recovered on sample units from the past, only if some outcome has occurred.

Cross sectional studies are the simplest type of observational study, a.k.a. a *simple random sample*. There really isn't much to say about them, except a vivid warning against causal inference from such studies (see 6.12.5).

[Simple Random Sample]

In *prospective studies*, a.k.a. *longitudinal data*, or *cohort study*, sampling units are selected and then measured over time alongside covariates.

[Longitudinal Data]

For the study of rare outcomes, prospective studies may be very wasteful. *Retrospective studies*, a.k.a. *case-control studies* remedy this by collecting units with the desired outcome, and only then matching them with controls and recovering their event history. In the words of Cox and Donnelly (2011): *"... a prospective study looks for the effects of causes whereas a retrospective study examines the causes of effects"*. This is clearly not a random sample in the population, so it is unclear that inference in retrospective studies is valid.

Case-Control Study

To demonstrate the difference in inference between the sampling schemes consider a binary outcome $Y \in \{0,1\}$ and a binary covariate $X \in \{0,1\}$. In terms of the post-sale problem in Example 13, X can stand for a device's version indicator, and Y a failure indicator. The effect of X on Y, can be quantified by the *odds ratio*. To define the odds ratio, we first define the *odds*.

Definition 30 (Odds)**.** The odds of two events, $Y = 1$, and $Y = 0$, is defined as

$$Odds := P(Y = 1)/P(Y = 0). \tag{6.3}$$

The *odds*, just like *probabilities* is an uncertainty measure. It quantifies the ratio of "successes" per "fail". In some communities, the odds are actually more popular than probabilities (think horse racing). In our post-sale example, the odds measures the ratio of working devices per faulty device (not per *produced* device like the probability measure).

The odds ratio is the ratio of odds under two conditions:

Definition 31 (Odds Ratio)**.**

$$OR := \frac{P(Y = 1|X = 1)/P(Y = 0|X = 1)}{P(Y = 1|X = 0)/P(Y = 0|X = 0)}. \tag{6.4}$$

Just in order to emphasize what the OR is **not**, we also define the *relative risk*, which is perhaps a more natural effect measure:

Definition 32 (Relative Risk)**.**

$$RR := \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)}. \tag{6.5}$$

You may verify that if X has not effect on Y, i.e. $P(Y|X) = P(Y)$, for all X, then both $OR = 1$ and $RR = 1$. We now show how the OR can be estimated from the various designs.

6.12.1 Prospective Study

A prospective study means that we decide how many samples of each version to take and wait until some fail, and then compare the failure probabilities. Formally this means we directly estimate $P(Y = 1|X = 1)$ and $P(Y = 1|X = 0)$. Estimating OR is thus trivial.

6.12.2 Retrospective Study

A retrospective study means that we sample a fixed number of working and broken devices. We can thus estimate the frequency of each version given the outcome: $P(X|Y)$. To relate it to the OR, we will need to relate the estimable $P(X|Y)$, to the desired $P(Y|X)$ via Bayes' Theorem:

$$P(Y|X) = P(X|Y)P(Y)/P(X) \tag{6.6}$$

Several applications of Eq.(6.6) yields that

$$OR = \frac{P(X=1|Y=1)/P(X=1|Y=0)}{P(X=0|Y=1)/P(X=0|Y=0)}. \tag{6.7}$$

Eq.(6.7) is very good news. It means that we even if we sample retrospectively, we can estimate the OR as if we sampled prospectively.

Think about it. Can we estimate other effect measures such as the RR from a retrospective study?

Extra Information. [Effect parameters as distances] The OR can be estimated from both prospective and retrospective studies because it is a *symmetric* effect measure. As such, it can be seen as a *distance* between distributions. The RR, while more easily interpretable, is not symmetric, and thus cannot be thought of as a distance.

6.12.3 Cross-Section

A cross-section study is **not** a probable sampling scheme for our post-purchase example. It implies that we sample randomly in a "pile" of devices of all versions and states. We thus get an estimate of the proportion of each X, Y combination, and need to relate it to the difference in failure probabilities. Formally, we estimate all $P(Y, X)$'s. Via Bayes' Theorem we then have

$$OR = \frac{P(Y=1, X=1)P(Y=0, X=0)}{P(Y=1, X=0)P(Y=0, X=1)},$$

which shows that the OR can indeed be estimated from a cross-section study.

Think about it. Can you estimate the RR from a cross section study?

Think about it. In regression analysis we typically analyze bias and variance while conditioning on the design matrix, X. Is such a practice consistent with the designed experiment or the observational view of a study? If observational, is it cross sectional? Prospective? Retrospective?

6.12.4 Special Sampling Schemes

Cross Sectional studies can satisfy the popular i.i.d. assumption from introductory statistics courses. Prospective and retrospective studies no longer satisfy this assumption, because samples within subject are dependent. These are still fairly simple sampling schemes. Here is a more complicated example.

Example 14 (Respondent Driven Sampling)**.** Consider the following chain-referral sampling scheme. A Facebook questionnaire is passed by rewarding respondents some credit if they name other respondents. The probability of being sampled is thus not fixed for everyone, but rather a function of the number of your facebook friends. Now imagine the questionnaire is meant to estimate the awareness to product ZZZ. It seems reasonable to assume that the more friends you have, the more likely you are to be sampled, and also, the more likely to be familiar with ZZZ. This sampling scheme will thus lead to an upward biased estimate of ZZZ's popularity.

If the non-simple sampling scheme in Example 14 is not accounted for at the analysis stage, we will clearly have biased popularity estimates. To deal with over-representation we will want to weight each sample by its probability of being sampled.

Definition 33 (Horowitz Thompson Estimator)**.** If each unit i has probability π_i of being sampled and measured response y_i, then an unbiased estimate of the population mean of y is given by

$$\frac{\sum_{i \in S} y_i/\pi_i}{\sum_{i \in S} 1/\pi_i}, \tag{6.8}$$

where $i \in S$ means that i has actually been sampled.

You may verify that the Horowitz-Thompson estimator in Definition 33 return the usual sample mean, if the sampling is simple, thus all π_i are equal.

6.12.5 Causal Inference

Causal inference is typically the Holy-Grail of empirical research. Causality may be inferred from a designed experiment, or from an observational study. As already stated on several occasions, inferring causality from observational studies is much riskier than from designed experiments. This is because in observational studies we have two sources of uncertainty, that are canceled by the randomization mechanism of designed experiments:

1. **Non-controlled variables** that affect both the covariates and the outcome (think of stress in the smoking example in Extra Info 1.1.2).

2. **Reverse Causality**: if a correlation is found between A and B, does A cause B, or B cause A?

It is not impossible to infer causality from observational studies, but the price to pay is added assumptions. The first, and probably harmless assumption is that causality works forward in time. For this reason a prospective study is a good way to go. Retrospective studies also allow causal inference, but with some more assumptions. Even cross-section studies allow causal inference, but many assumptions will be required. A set of assumptions, is the underlying scientific theory. In physics this is a helpful guideline, but in the social sciences, this is merely a call for endless debates. See the Bibliographic notes section for some references on causal inference from observational studies.

6.13 Bibliographic Notes

Our general discussion of observational studies, designed experiments, causal inference, etc, is derived from Cox and Donnelly (2011). For an overview of DOE see Cox and Reid (2000), Mason et al. (2003), Everitt and Skrondal (2010). Another nice and freely available resource is a Penn State course on the topic[5]. Some seminal references in the field include Fisher (1960) and Box et al. (1978). For respondent driven sampling see Berchenko et al. (2013). For causal inference in non-designed experiments see Rosenbaum (2002). For optimal designs see Pukelsheim (1993). For analysis of data: linear models, ANOVA, etc. there are endlessly many books. This author's recommendations include Hocking (1985), Greene (2003). For the relation between factorial designs and coding theory in computer science, see Hill (1986). For design and analysis of computer experiments, see Sacks et al. (1989) or Santner et al. (2013).

[5] `https://onlinecourses.science.psu.edu/stat503/node/1`

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 5.7. 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 13 (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 14 (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 15 (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 6.9.

7.4 Bibliographic Notes

[TODO]

Chapter 8

Reliability Analysis

The attempt to define the difference between *reliability* and *quality* will certainly fail given the intentional ambiguity in our definition of *quality* (Chapter 1). For our purposes, however, this terminological matter will not matter, since we will simply define reliability analysis to be the analysis of the *time* to *failure*. We will also assume that "time" and "failure" are well defined and agreed upon.

We intuitively understand "more reliable" to mean "lasts longer". We should also consider, however, the case of a product that is designed to fail after some time, thus forcing the consumer to buy a new one. Some may say that a major hi-tech company named after a fruit employs this practice. Be it true or not, I hope we can agree that good knowledge of your product's life expectancy is a desirable.

Reliability analysis involves the study of a probabilistic property of our product- its *survival*. Any probabilistic model will require calibration to reality via data. This chapter thus introduces both the probability calculus typically used for reliability analysis, and some statistical considerations involved when calibrating these models.

8.1 Probabilistic Analysis

8.1.1 A Static View

Let $\mathbf{x}_j \in \{0,1\}, j = 1,\ldots,p$ denote the state of the j'th component of a system, and $x = (x_1,\ldots,x_p)$.

Definition 34 (Structure Function)**.** The *structure function*, $\Phi = \Phi(x) : x \mapsto \{0,1\}$, is an indicator function of the state of the system. A failure indicated by 0.

Remark 16 (Φ)**.** We apologize to the reader for using Φ to denote both the $\mathcal{N}(0,1)$ CDF, and the structure function. We do so to stay in accordance with reliability literature, and since no collisions are created in this chapter by doing so.

Definition 35 (Series System)**.** A *series system*, or *serial system*, is one where all components need to function for the system to function:

$$\Phi(x) = \prod_{j=1}^{p} x_j.$$

A reliability diagram of a series system is given in Figure 8.1.

Definition 36 (Parallel System)**.** A *parallel system* is one where all components need to fail for the system to fail:

$$\Phi(x) = 1 - \prod_{j=1}^{p}(1 - x_j) = \coprod_{j=1}^{p} x_j.$$

Figure 8.1: Series system.

A reliability diagram of a parallel system is given in Figure 8.2.

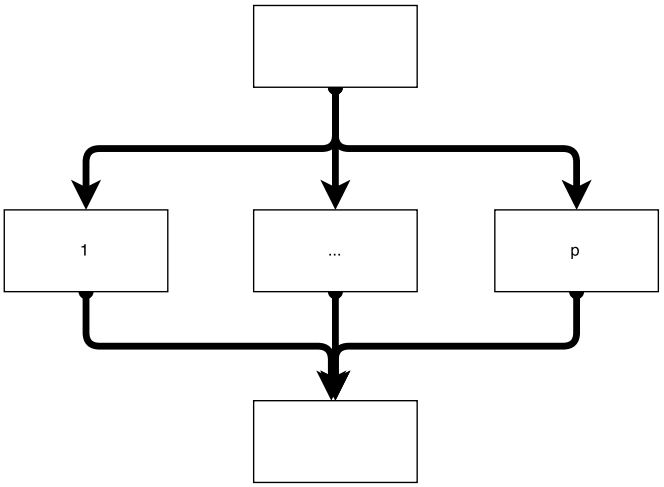


Figure 8.2: Parallel system.

Definition 37 (k-out-of-p System)**.** A *k-out-of-p* system is one where at least k components need to function for the system to function:

$$\Phi(x) = I_{\{\sum_{j=1}^{p} x_j \geq k\}}.$$

A reliability diagram of a k-out-of-p system is not provided, since it is not very friendly. Try thinking of a 2-out-of-3 system to see why.

My mistake! The previous definition is different than the one I gave in class. I initially defined it via failing components, while now it is defined via functioning components. Please note this when comparing your notes to mine.

Definition 38 (Monotone System)**.** A system is said to be *monotone* if $\Phi(x_1, \ldots, x_p)$ is non decreasing in all components.

The definition of monotonicity captures the idea that you cannot improve a system's state by breaking components. This seems rather natural (I am still looking for a counter example).

Definition 39 (Reliability)**.** We define the *reliabity of component j* to be

$$p_j := P(\mathbf{x}_j = 1),$$

and the *reliability of the system*

$$S_\Phi = S_{\Phi(x)} := P(\Phi(x) = 1).$$

Example 15 (Reliability of a series system)**.** For $\Phi(x)$ a series system, assuming independent components, we have

$$S_\Phi = \prod_{j=1}^{p} p_j.$$

Example 16 (Reliability of a parallel system)**.** For $\Phi(x)$ a parallel system, assuming independent components, we have

$$S_\Phi = 1 - \prod_{j=1}^{p}(1-p_j) = \coprod_{j=1}^{p} p_j.$$

Example 17 (Reliability of a k-out-of-p system)**.** For $\Phi(x)$ a k-out-of-p system, assuming independent components with equal reliability ($p_i = p_0$), we have

$$S_\Phi = \sum_{i=k}^{p} \binom{p}{i} p_0^i (1-p_0)^{p-i}.$$

State enumeration method

To compute the reliability of more complex structures, the brute-force approach is the *state enumeration method*. This method simply relies on summation of the probabilities of the states for which the system functions.

$$S_\Phi = \sum_x \Phi(x) P(\mathbf{x} = x).$$

Factoring method

The *factoring method*, a.k.a. *pivot-decomposition method*, relies on two ingredients: (a) conditioning on the state of some components greatly simplifies the structure, and (b) the total probability argument. Combining the two we have:

$$S_\Phi = p_j S_{\Phi|x_j=1} + (1-p_j) S_{\Phi|x_j=0},$$

where $S_{\Phi|x_j=1}$ denotes the reliability of the structure Φ conditional on $x_j = 1$. The following example demonstrates the power of the factoring method.

Example 18 (Bridge Structure)**.** Consider structure in Figure 8.3. To compute the reliability, we will call upon the factoring method while conditioning on the state of component 3:

$$S_\Phi = p_3 S_{\Phi|x_3=1} + (1-p_3) S_{\Phi|x_3=0}.$$

Now note that when $x_3 = 1$ then we have a series structure of parallel structures, while when $x_3 = 0$ we have a parallel structure of series structures.:

$$S_{\Phi|x_3=1} = (p_1 \coprod p_2)(p_4 \coprod p_5),$$
$$S_{\Phi|x_3=0} = p_1 p_4 \coprod p_2 p_5,$$

so that

$$S_\Phi = p_3(p_1 \coprod p_2)(p_4 \coprod p_5) + (1-p_3)(p_1 p_4 \coprod p_2 p_5).$$

Example 18 demonstrates a single application of the factoring method. Clearly, it can be applied recursively for more complicated systems.

The example also demonstrates a more general principle. Namely, that redundancy is preferable at the component level, and not at the system's level. Put differently- when designing a backup, and the resources allow a fill copy of the original system, we are better of by designing a component-wise backup, than a single backup system. Put formally:

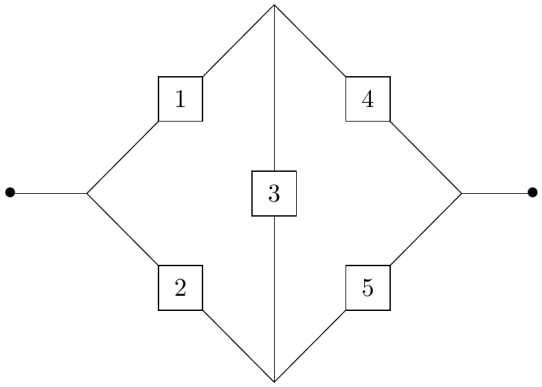


FIGURE 2.5. Bridge Structure

Figure 8.3: Structure of a bridge system. Source: (Aven and Jensen, 1999, Fig.2.5)

Theorem 8.1.1 (Component-wise redundancy)**.** *For a monotone structure* Φ*,*

$$S_{\Phi(x \coprod y)} \geq S_{\Phi(x)} \coprod S_{\Phi(y)} \tag{8.1}$$

where $x \coprod y$ *denotes a component-wise backup:* $(x_1 \coprod y_1, \ldots, x_p \coprod y_p)$*.*

Extra Information. [Reliability analysis of complex systems] Except for simple systems, of the type we presented, the computation of the reliability of a complex system may be a formidable task. For complicated real-life systems, *min-cut–max-flow* algorithms, or *inclusion-exclusion* type algorithms are employed. For more details, see Aven and Jensen (1999).

Reliability Importance Measures

The Strength Of The Chain Is In The Weakest Link.

This is obviously a profound observation in reliability analysis. In order to identify the weakest link we require some measure of reliability importance.

Definition 40 (Improvement potential)**.** The *improvement potential* is defined as the change in a system's reliability, if we could force a component to function indefinitely. Formally, we denote $\Phi^{(j)}$ to be a system where component j cannot fail: $\Phi^{(j)} := \Phi_{(p_j=1)}$. We then define the improvement potential with respect to component j to be

$$I_j := S_{\Phi^{(j)}} - S_\Phi. \tag{8.2}$$

Definition 41 (Birenbaum's measure)**.** *Birenbaum's measure* is defined as the change in a system's reliability, if we infinitesimally improve the reliability of component j. Formally

$$I_j := \frac{\partial}{\partial p_j} S_\Phi. \tag{8.3}$$

Clearly any such importance measure, once computed, may serve to decide which component should be treated to improve reliability.

8.1.2 A Time Dynamic View

The reliability of each component (p_j), typically changes in time, and so does the reliability of the whole system. In the following, $\mathbf{t}$ will typically stand for the time to malfunction. It is thus assumed to be **continuous** and **non-negative**.

Definition 42 (CDF)**.** The cumulative distribution function (CDF) of a random variable $\mathbf{t}$ at a point t is given by

$$F_{\mathbf{t}}(t) := P(\mathbf{t} < t). \tag{8.4}$$

Definition 43 (PDF)**.** The probability density function (PDF) of a continuous random variable $\mathbf{t}$ at a point t is given by

$$p_{\mathbf{t}}(t) := \frac{\partial}{\partial t} F_{\mathbf{t}}(t). \tag{8.5}$$

Definition 44 (Survival Function)**.** The survival function of a random variable $\mathbf{t}$ at a point t is given by

$$S_{\mathbf{t}}(t) := P(\mathbf{t} > t) = 1 - F_{\mathbf{t}}(t). \tag{8.6}$$

By definition, it follows that if $\mathbf{t}_j$ is the time to failure of component j, then

$$p_j(t) = S_{\mathbf{t}_j}(t).$$

If $\mathbf{t}_\Phi$ is the time to failure of a system Φ, then we may write $S_\Phi(t) = S_{\mathbf{t}_\Phi}(t)$.

Example 19 (Survival of a series system)**.** For a series system Φ, the reliability of the system at time t is given by

$$S_\Phi(t) = \prod_{j=1}^{p} p_j(t).$$

Example 20 (Survival of a parallel system)**.** For a parallel system Φ, the reliability of the system at time t is given by

$$S_\Phi(t) = 1 - \prod_{j=1}^{p} (1 - p_j(t)) = \coprod_{j=1}^{p} p_j(t).$$

Another way to present a distribution, no less informative than the previous ones, is by the *hazard function*, which is the "probability of surviving just another instant".

Definition 45 (Failure Rate)**.** The *hazard function*, or *failure rate*, of a random variable $\mathbf{t}$ at a point t is given by Hazard Function

$$h_{\mathbf{t}}(t) := \lim_{dt \to 0} \frac{P(\mathbf{t} \in [t, t+dt) | \mathbf{t} \geq t)}{dt} \tag{8.7}$$

$$= \frac{p_{\mathbf{t}}(t)}{S_{\mathbf{t}}(t)} \tag{8.8}$$

$$= -\frac{\partial}{\partial t} \log S_{\mathbf{t}}(t). \tag{8.9}$$

Definition 46 (Cumulative Risk)**.** The *cumulative hazard*, a.k.a. the *cumulative risk*, of a random variable $\mathbf{t}$ at a point t is given by Cumulative Hazrd

$$H_{\mathbf{t}}(t) := \int_0^t h_{\mathbf{t}}(t) \tag{8.10}$$

$$\Rightarrow S_{\mathbf{t}}(t) = \exp(-H_{\mathbf{t}}(t)). \tag{8.11}$$

Eq.(8.11) readily shows that a distribution is well defined by its hazards.

Theorem 8.1.2 (Failure rate of a series system)**.** *The failure rate of a series system of independent components* Φ *is given by the sum of the failure rates of its components*

$$h_\Phi(t) = \sum_{j=1}^{p} h_{\mathbf{t}_j}(t) \tag{8.12}$$

The proof is immediate using the cumulative risk. The failure rate of a parallel system, does not admit such a nice closed form as we will soon see in Example 23.

Example 21 (Exponential Hazard)**.** The simplest distribution when discussing hazards is the exponential. Recalling the for non-negative t:

$$p_{\mathbf{t}}(t) = \lambda e^{-\lambda t}, \tag{8.13}$$
$$F_{\mathbf{t}}(t) = 1 - e^{-\lambda t}, \tag{8.14}$$

so that

$$S_{\mathbf{t}}(t) = e^{-\lambda t}, \tag{8.15}$$
$$h_{\mathbf{t}}(t) = \lambda. \tag{8.16}$$

The exponential is the only distribution with constant hazard which makes it very easy to analyze. The constant hazard is due to the *memoryless* property. Look at Eq.(8.7) and think why.

Example 22 (Failure rate of a series of exponential components)**.** The failure rate of a series system Φ, of p independent components each with exponentially distributed failure times, is simply

$$h_\Phi(t) = \sum_{j=1}^{p} \lambda_j, \forall t \geq 0 \tag{8.17}$$

where $\lambda_j = \lambda_j(t)$ is the rate of each component.

This is obviously the simplest system possible for reliability analysis, which stems from the fact that a minimum of exponentials is exponential with the sum of rates.

The following example, seemingly very simple, provides tremendous insight into the complexities of reliability analysis.

Example 23 (Failure rate of a two exponential-component parallel-system)**.** Consider a system of two independent, parallel, exponential components, with failure times $\mathbf{t}_j \sim \exp(\lambda_j); j = 1, 2$. The failure rate is given by

$$h_\Phi(t) = \frac{\lambda_1 e^{-\lambda_1 t} + \lambda_2 e^{-\lambda_2 t} - (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2)t}}{e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1+\lambda_2)t}} \tag{8.18}$$

Why is Example 23 so important? Because it demonstrates that even in a simple system, with the simplest components, the reliability is not so simple to compute (as a function of the components' reliability). Indeed, even though the component-wise hazards are fixed in time, the system's hazard is not fixed, and not even monotone in time (Figure 8.4).

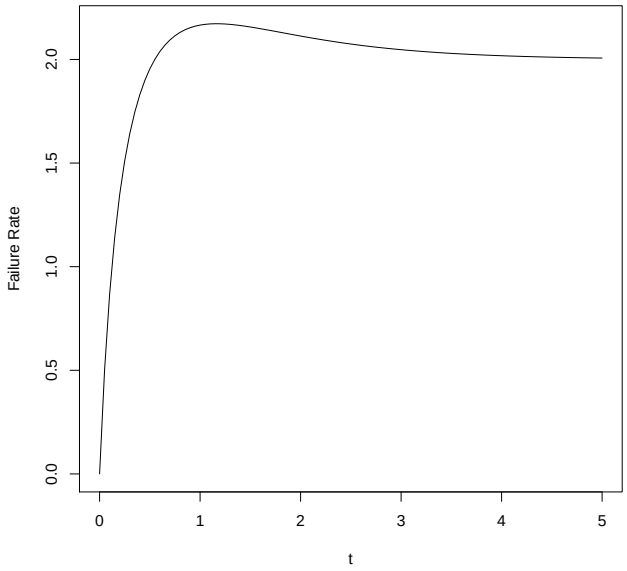


Figure 8.4: Failure rate of the parallel exponential component system.

Example 24 (Weibull Hazard)**.** The Weibull distribution is very common in reliability analysis since it is tractable generalization of the exponential distribution with many nice properties. It can be constructed by $\mathbf{t} := \lambda \mathbf{u}^{1/k}$, where $\mathbf{u} \sim \exp(1)$. This implies that for non negative t:

$$p_{\mathbf{t}}(t) = \frac{k}{\lambda}\left(\frac{\mathbf{t}}{\lambda}\right)^{k-1} e^{-(\mathbf{t}/\lambda)^k}, \tag{8.19}$$

$$F_{\mathbf{t}}(t) = 1 - e^{-(\mathbf{t}/\lambda)^k}, \tag{8.20}$$

so that

$$S_{\mathbf{t}}(t) = e^{-(\mathbf{t}/\lambda)^k}, \tag{8.21}$$

$$h_{\mathbf{t}}(t) = \frac{k}{\lambda}\left(\frac{\mathbf{t}}{\lambda}\right)^{k-1}. \tag{8.22}$$

Elementary analysis shows that the hazard function of the Weibull may be increasing or decreasing in time ($\mathbf{t}$), depending on k, but it is always monotone.

Example 25 (Empirical risk rates)**.** When examining empirical risk rates of true devices, we almost always notice a *bathtub* structure, such as in Figure 8.5. This shape captures the idea that products tend to fail more when they are brand new, or as they are very old, while their failure rates are fairly stable in the "mid-life". In this text, we will not be providing a particular distribution which has this property. We refer the reader to Nadarajah (2008) for examples of distributions which have the bathtub property. Bathtub

Aging

The idea of *aging* is that failure rate may vary over time. It is an important concept in reliability, as demonstrated by the empirical bathtub failure rate (Figure 8.5). Instead of checking if a particular textbook distribution has some ageing property, we instead analyze classes of distributions with the desired notion of ageing. Our goal will ultimately be to understand the ageing of a whole system, as a function of the ageing of its components.

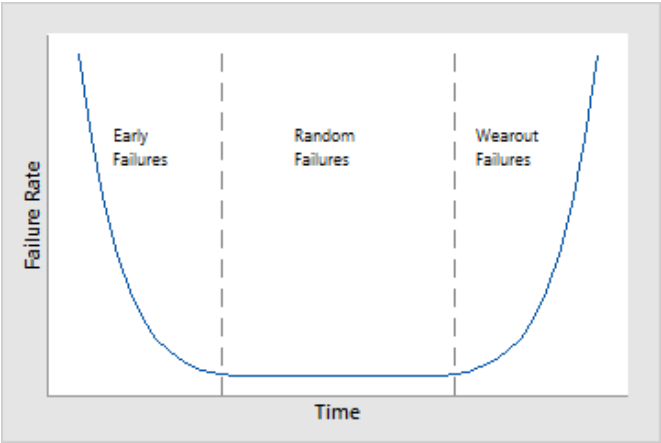


Figure 8.5: Bathtub curve of empirical failure rates.
`http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/reliability/distributions-in-reliability-analysis/hazard-functions/`

Definition 47 (IFR)**.** We call a failure time distribution to be in the *increasing failure rate* (IFR) ageing class, if it has a non decreasing failure rate.

Definition 48 (IFRA)**.** We call a failure time distribution to be in the *increasing failure rate average* (IFRA) ageing class, if $H_{\mathbf{t}}(t)/t$ is non decreasing in t.

The intuition underlying IFRA relies on understanding $H_{\mathbf{t}}(t)/t$, which can be seen as the average risk from ignition to time t. IFRA thus means that even if the risk decreases at some point in time, the average risk still increases.

Definition 49 (NBU)**.** We call a failure time distribution to be in the *new better then used* (NBU) ageing class, if $S_{\mathbf{t}}(t_1 + t_2) \leq S_{\mathbf{t}}(t_2)S_{\mathbf{t}}(t_1)$.

Definition 50 (NBUE)**.** Define the *expected residual life*, $\mu(t)$, to be

$$\mu(t) := \mathbf{E}\left[\mathbf{t} - t | \mathbf{t} > t\right].$$

Expected Residual Life

We call a failure time distribution to be in the *new better then used in expectation* (NBUE) ageing class, if $\mu(t) \leq \mu(0)$.

NBUE can be easily understood, since it means that a component's expected life is maximal when it is brand new.

Theorem 8.1.3. *$IFR \Rightarrow IFRA \Rightarrow NBU \Rightarrow NBUE$.*

The following theorem states a relation between the ageing properties of particular components, and that of the whole system. In particular it states that for the (very wide) class of monotone systems, then the IFRA property is conserved. This should be contrasted with the IFR property, which is not conserved, as demonstrated by the parallel system in Example 23.

Theorem 8.1.4 (IFRA closure theorem)**.** *If the independent components of a monotone system are IFRA, then so is the whole system.*

Series systems are a extremely small and particular subset of monotone systems. It does provide, however, an example of systems where not only IFRA is preserved, but also the stronger IFR. The following corollary follows immediately from Theorem 8.1.2 and the fact that a sum of monotone functions is monotone.

Corollary 8.1.1 (IFR closure for series systems)**.** *A series system of independent IFR components is IFR.*

Now consider a two-component system, where one component kicks-in when the first fails. We will call this an *offline backup*. The survival times of the components in an offline backup system are clearly dependent. It turns out that for such a system of IFR components, does conserve the IFR property, as seen in the following theorem.

Offline Backup

Theorem 8.1.5 (Convolution of IFR)**.** *For two independent random variables,* $\mathbf{x}$ *and* $\mathbf{y}$*, both in the IFR ageing class, then so is* $\mathbf{x}+\mathbf{y}$*.*

The theorem is called the convolution theorem, because the distribution of a sum of independent random variables, is the convolution of their distributions.

Extra Information. [IFR and log-concave] The IFR requirement, is essentially the same as log-concavity of the density function. This immediately implies many properties of the class, including the convolution theorem above. See Bagnoli and Bergstrom (2005).

Example 26 (IFR of Gamma)**.** The Gamma (and thus the Erlang) distribution is in the IFR aging class, since it is the sum of exponentials, each IFR.

Example 27 (Series system of offline backups)**.** What can we say about the ageing class of a series systems of offline backup systems? It turns out that if the components are IFR, then so will the whole system. This is immediate from Corollary 8.1.1 and Theorem 8.1.5.

8.2 Statistical Analysis

The probabilistic analysis of the previous section is great fun and all, but like any probabilistic problem, is has to be calibrated to real life. This is where data, and statistics come in. Indeed, given any particular probabilistic model, we may write the likelihood problem, and call upon maximum likelihood principles for estimation.

Failure data and models introduces particular statistical challenges:

Identifiability It is typically hard, if not impossible, to estimate the reliability of particular components, from the reliability of the whole system.

Censoring A major concern with reliability data, is that in any finite length experiment, some events will just not have happened yet; their failure time will thus be *censored*. Ironically- the more reliable a component, the less data we will have to estimate its reliability.

Lab versus real-life conditions Reliable components take very long time to fail. We will thus extrapolate from harsh lab conditions to real-life operating conditions. This requires the introduction of covariates.

Failure Distribution like any statistical model, we will need to commit to some failure time sampling distribution.

8.2.1 Identifiability

Example 28 (Likelihood estimation of a series system)**.** Assume a series system Φ with p independent, exponential components with rates $(\lambda_1,\ldots,\lambda_p)$. We have n observations on the failure times of the system $t_1,\ldots,t_n$. How can we estimate the failure rates? To use a likelihood approach, we need the data's sampling distribution. Denoting the failure time of the j'th component of the ith device with $\mathbf{t}_{i,j}$, we have that $\mathbf{t}_{i,j}\sim\exp(\lambda_j)$ by assumption. Since the system is serial, then $\mathbf{t}_i=\min_j(\mathbf{t}_{i,1},\ldots,\mathbf{t}_{i,p})$. By the properties of the exponential distribution $\mathbf{t}_i\sim\exp(\lambda)$, where

$\lambda := \sum_{j=1}^{p} \lambda_j$, as we have already seen with the failure rate. It follows that $p_{\mathbf{t}_i}(t) = \lambda \exp(-\lambda t)$. We may then write the likelihood function, maximize it with respect to λ and discover, as we already know, that

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}.$$

We are now left with the problem of recovering $(\lambda_1, \ldots, \lambda_p)$ from λ. Can we do it? On the face of it- no. Which should not surprise us, since the mere knowledge of a device failure, is not very informative on the particular component that failed, which we would need to estimate $(\lambda_1, \ldots, \lambda_p)$.

Example 28 teaches us that unless further assumptions are introduced, the estimation of the component-wise failure rates requires information on the component-wise failure times.

8.2.2 Censored Events

Consider several components being analyze for their reliability. Ironically, we actually want them to fail. If they do not, they do not convey information on their reliability. In the event that a component has not failed, we clearly cannot register its failure time. Omitting this component from the sample will upward bias the estimated reliability. These events are called *censored* observations. There are several types of censoring which depend on the design of the study and the type of event recorded. They are all dealt with careful though on the sampling distribution of the data, and the probability of a censoring event.

Design considerations:

Type I occurs when the design is such that the sampling time is fixed a-priori. This implies that the number of failures (thus censoring) events is random.

Type II occurs when the design is such that the number of failures (thus censoring) is fixed a-priori. This implies that the sampling duration is random.

If all we know on a censored event is that the actual lifetime is larger then the observation period, we call this a *non-informative* censoring. Both designs will then lead to the same modelling of the censoring event, which is now described.

The likelihood function of non-censored events is given by

$$\mathcal{L}_i = p_{\mathbf{t}}(t_i) = S_{\mathbf{t}}(t_i) h_{\mathbf{t}}(t_i), \tag{8.23}$$

and the likelihood of a censored event, under the *non-informative* assumption, is given by

$$\mathcal{L}_i = S_{\mathbf{t}}(t_i). \tag{8.24}$$

Unifying the two cases assuming independent observations, using an indicator for non-censoring, c_i, and taking logs we have

$$\begin{aligned} L &= \log \mathcal{L} && (8.25)\\ &= \log \prod_{i=1}^{n} \mathcal{L}_i && (8.26)\\ &= \log \prod_{i=1}^{n} S_{\mathbf{t}}(t_i) h_{\mathbf{t}}(t_i)^{c_i} && (8.27)\\ &= \sum_{i=1}^{n} [c_i \log h_{\mathbf{t}}(t_i) - H_{\mathbf{t}}(t_i)]. && (8.28) \end{aligned}$$

Example 29 (Censored exponential lifetimes)**.** Recalling that the failure rates of exponential lifetimes are fixed, we have that the likelihood of censored exponential lifetimes is given by

$$\sum_{i=1}^{n}[c_i \log \lambda - \lambda t_i].$$

The maximum likelihood estimator of λ is thus

$$\widehat{\lambda} = \frac{\sum c_i}{\sum t_i}. \tag{8.29}$$

Eq.(8.29) lends itself to a nice interpretation. The nominator is the total number of failures. The denominator is the total *exposure time*. The estimated failure rate is thus the number of failures per unit of exposure time.

Extra Information. The previous result is obvious is you consider failures as events which come as a Poisson process, which is implied from the exponential times assumption. The process is run for $\sum t_i$ time, and the total event count is $\sum c_i$. The trivial estimator for the rate of the process, is $\sum c_i / \sum t_i$.

8.2.3 Accelerated Life Models

An *accelerate life* model assumes that covariates rescale time. For instance, the lab may produce conditions where time advances ten times faster than in real-life operating conditions. To introduce the model, we start with a simple two group example.

Example 30 (Two group accelerated life)**.** Consider two groups indexed by a single dummy variable $x \in \{0, 1\}$. Assuming an accelerated life effect we have

$$S_{\mathbf{t}|x=1}(t) = S_{\mathbf{t}|x=0}(t/exp(\beta)) = S_{\mathbf{t}|x=0}(t/\gamma),$$

where γ is simply shorthand notation for $exp(\beta)$. If $\gamma = 1/2$, this means that time for group $x = 1$ advances twice as fast as for group $x_i = 0$.

We now generalize the idea for multiple covariates. If $\mathbf{t}_1$ is the (random) time to failure under conditions $x_1 = (x_{1,1}, \ldots, x_{1,p})$, and $\mathbf{t}_0$ under conditions $x_0 = (x_{0,1}, \ldots, x_{0,p})$, we have

$$S_{\mathbf{t}_1}(t) = S_{\mathbf{t}_0}(t/\gamma), \tag{8.30}$$

where $\gamma = e^{(x_1 - x_0)'\beta}$, meaning that the conditions x_1 are such that time is accelerated by $1/\gamma$ compared to the base conditions x_0. Example 30 is recovered by setting $x_1 = 1$ and $x_0 = 0$. If $\gamma > 1$, time under conditions x_1 advances slower than under x_0, and the product is expected to live longer. The converse holds if $\gamma < 1$.

An equivalent formulation of an accelerated life model, which also explains the appearance of the exponent in the time rescaling, is the following

$$\log \mathbf{t}_i = x_i'\beta + \varepsilon_i, \tag{8.31}$$

for some error term ε_i. For interpretation, we again denote $e^{x_i'\beta} = \gamma_i$, and infer that x_i accelerates time by $1/\gamma_i$ compared to some base rate where $e^{x_i'\beta} = 1$. This formulation is more tractable for mathematical manipulation, but conceals the nice interpretation which motivates the model's name.

Eq.(8.31) readily reveals how we can easily estimate the effects (β) of an accelerate life model. We simply take the log of the survival times ($\log t_i$), and assuming the particular distribution of ε_i, we may estimate β using maximum likelihood.

For future use, we also note that the relation between hazed function under the accelerated life assumption is given by

$$h_{\mathbf{t}|x=x_1}(t) = h_{\mathbf{t}|x=x_0}(t/\gamma)/\gamma \tag{8.32}$$

where as usual $\gamma = e^{(x_1-x_0)'\beta}$.

Example 31 (Accelerated life with Gaussian noise)**.** The maximum likelihood estimation of β when assuming that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, collapses to a simple linear regression when the dependent variable is simply $\log t_i$.

Extra Information. [Tobit regression] Assuming $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, as in the previous example, implies that failure times are *log normal* distributed. This approach is known as *Tobit* regression.

Extra Information. [Accelerated life with extreme value noise] Assuming that the density of ε is given by

$$p(\varepsilon) = e^{(\varepsilon - e^{\varepsilon})},$$

which is known as an *extreme value distribution*, then failure times have an exponential distribution, and estimation of β collapses to an exponential regression problem.

Extreme Value Distribution

Extra Information. Popular accelerated time models that capture effects of temperature include the *Arrhenius model*, and the *Eyring model*. For models for stress, voltage, humidity, and other life accelerating covariates, see (Natrella, 2010, Sec.8.1.5)

Arrhenius Eyring

8.2.4 Proportional Hazard Models

The *proportional hazard*, or *proportional risk* class of models, assumes that covariates multiply not time, but rather failure rates. Put differently, accelerated life acts linearly on time, thus non-linearly on hazards. Proportional hazards acts linear on hazards, thus non linearly on time. Qualitatively, both either accelerate (decelerate) time. Quantitatively, the exact amount of acceleration (deceleration) may differ. The choice between these models typically depends on the underlying physical theory, and on the ease of computation and interpretation.

Starting with a two group example

Example 32 (Proportional hazards in a two group model)**.** Consider two groups indexed by a single dummy variable $x \in \{0, 1\}$. Assuming proportional hazard effect we have

$$h_{\mathbf{t}|x=1}(t) = h_{\mathbf{t}|x=0}(t)e^{\beta} = h_{\mathbf{t}|x=0}(t)\gamma,$$

where γ is again shorthand notation for e^{β}. If $\gamma = 1/2$, this means that at any point in time, group $x = 1$ suffers half the risk of group $x = 0$.

Now for the general case, which compares the risk function under conditions $x = x_1$ to the risk under some base operating conditions $x = x_0$.

$$h_{\mathbf{t}|x_1}(t) := h_{\mathbf{t}|x_0}(t)e^{(x_1-x_0)'\beta} = h_{\mathbf{t}|x_0}(t)\gamma, \tag{8.33}$$

where $h_{\mathbf{t}|x_0}(t)$ is some assumed baseline hazard rate, and γ is short notation for $e^{(x_1-x_0)'\beta}$. Example 32 is recovered by setting $x_1 = 1$ and $x_0 = 0$.

The linear rescaling of the risk in the proportional hazard model, implies the following relation between survival functions

$$S_{\mathbf{t}|x_1}(t) = S_{\mathbf{t}|x_0}(t)^{\exp((x_1-x_0)'\beta)}. \tag{8.34}$$

To see this recall Eq.(8.9):

$$\begin{aligned} h_{\mathbf{t}|x_1}(t) &= -\frac{\partial}{\partial t}\log S_{\mathbf{t}|x_1}(t) \\ &= -\frac{\partial}{\partial t}\log S_{\mathbf{t}|x_0}(t)^{\exp((x_1-x_0)'\beta)} \\ &= e^{(x_1-x_0)'\beta} - \frac{\partial}{\partial t}\log S_{\mathbf{t}|x_0}(t) \\ &= h_{\mathbf{t}|x_0}(t)e^{(x_1-x_0)'\beta} \end{aligned}$$

as required.

Example 33 (Comparing survival rates in the two group model)**.** We now compare the proportional hazard assumption, versus the accelerated time assumption in the simple two group model. Under the accelerated time model, we have (by assumption): $S_{\mathbf{t}|x=1}(t) = S_{\mathbf{t}|x=0}(t/\gamma)$. Under the proportional hazard model, we have (by Eq.(8.34)) $S_{\mathbf{t}|x=1}(t) = (S_{\mathbf{t}|x=0}(t))^{\gamma}$. To visualize the difference between the groups we choose some arbitrary distribution (Weibull) and rescale it with $\gamma = 2$ under the different models. The result is depicted in Figure 8.6, from which we see that the same assumed effect $\gamma = e^{\beta} = 2$, acts in opposite directions under the different models. This is also evident when examining the simple exponential case, using Eq.(8.33) and Eq.(8.32): under the proportional hazard the risk increases, whereas under accelerated time the risk decreases.

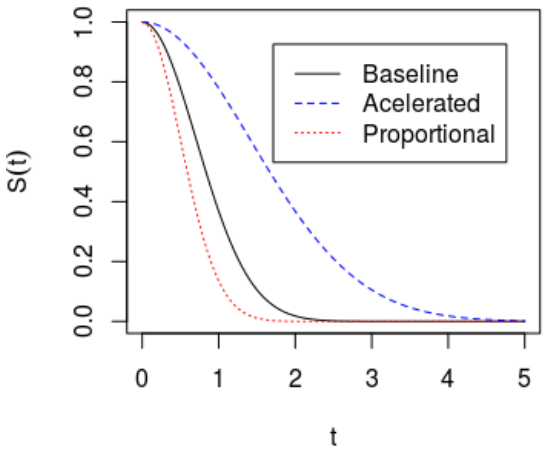


Figure 8.6

Example 33 demonstrates that the different models rescale time in different ways (and even in opposite directions!). The example may help you in interpreting the coefficients (β) of the model you chose. In particular it means that a factor with a positive effect ($\hat{\beta} > 0$) means accelerated ageing under proportional hazard, and decelerated ageing under accelerated time.

Example 34 (Accelerated life and proportional hazard for exponential failure times)**.** Recall that the exponential distribution has a fixed failure rate (over time). Using Eq.(8.32) for the accelerated life, and Eq.(8.33) for the proportional hazards, we see that the hazard after any such time rescaling, is still constant in time. The exponential distribution, rescaled under any of these models, will still return an exponential distribution. To se why this is not trivial, you can you your favourite (non negative, continuous) distribution, and check if the hazard function remains in the same class of distributions after a proportional hazard or accelerated life rescaling of time.

Example 35 (Two groups with exponential baseline)**.** To demonstrate the previous example, we return to our two group model. We assume: (a) a proportional hazard effect. (b) an exponential failure time distribution as baseline. From (b) we have that $\mathbf{t}_{x=0} \sim exp(\lambda) \Rightarrow h_{\mathbf{t}|x=0}(t) = \lambda$. From (a) we have that $h_{\mathbf{t}|x=1}(t) = \gamma h_{\mathbf{t}|x=0}(t) = \gamma\lambda \Rightarrow \mathbf{t}_{x=1} \sim exp(\gamma\lambda)$. We may now collect survival data under the two operating condition, and estimate (γ, λ), say, with a maximum likelihood estimator.

Extra Information. [General Hazard Rate Model] The effects of covariates on the failure time distribution, may be modelled in many ways. The two models presented are probably the most popular, but may certainly be extended. For a more detailed discussion, see Cox and Oakes (1984).

8.2.5 Choosing the Base Failure Rate

In all the above models, we are free to choose the base failure rate: $S_{\mathbf{t}_0}(t)$ in Eq.(8.30), or ε in Eq.(8.31), or $h_{\mathbf{t}|x_0}(t)$ in Eq.(8.33). Three possible approaches include:

1. Assume a **parametric** model, such as exponential times, Weibull times, etc.
2. Assume a **semi-parametric** model, which can be simply seen as a flexible class of distributions, that has no particular parametric representation. In reliability analysis, the *piece-wise constant* hazard is a popular choice.
3. Do not assume anything on the distribution, known as a **non-parametric** approach.

If we assume a particular parametric model, then we may gather failure time data, write the likelihood function, and return failure rate estimates, and covariate effects. We now focus on the more flexible framework of semi-parametric modelling.

8.2.6 The Parametric Case

A parametric model fitting to failure data, is simply a maximum likelihood problem. Examples 28, 31, and 8.2.3 demonstrate this.

8.2.7 The Semi Parametric Case

We now relax the explicit failure time distribution assumption, and adopt a more flexible semi-parametric distribution class, known as the *piecewise exponential class*. Consider the proportional hazard model:

Piecewise Exponenitial

$$h_{\mathbf{t}|x_1}(t) := h_{\mathbf{t}|x_0}(t) \times \exp((x_1 - x_0)'\beta). \tag{8.35}$$

The model clearly requires some baseline failure rate $h_{\mathbf{t}|x_0}(t)$. A flexible, yet not too flexible assumptions, is that the failure rate is constant in some time intervals:

$$h_0(t) = h_j \quad \text{if} \quad t \in [\tau_{j-1}, \tau_j) \tag{8.36}$$

This class of distributions has $J(J-1)$ parameters: $(\tau_1, \ldots, \tau_{J-1}, h_1, \ldots, h_J)$. We are free to choose J. Large J are very flexible classes, but will require a lot of failure data to estimate. Small J are less flexible, but require less data to estimate. At the limits, when $J = 1$, we are back to exponential failure times. At the other limit, where $J \to \infty$, we have an absurdly flexible distribution class, which requires impossibly large amounts of data to estimate.

Since the failure rate is piece-wise constant, the distribution class is known as *piece-wise exponential*. It is a rather flexible class of distributions. Figure 8.7 depicts the approximation of the Weibull survival function, using a piece-wise constant hazard function, with $J = 3$ and appropriate selected parameters.

Piecewise Exponential

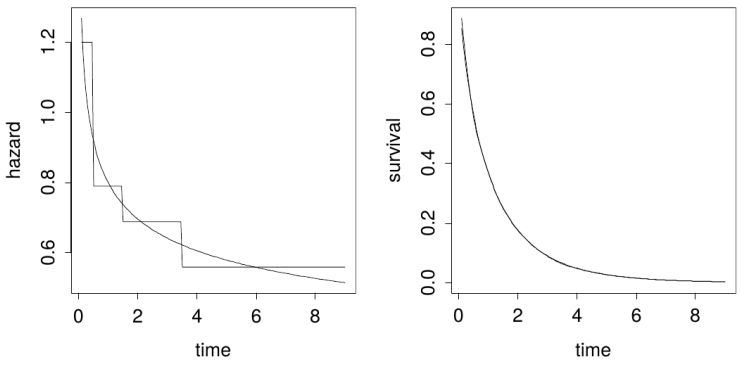


Figure 8.7: Piecewise Exponential aproximation of the Weibull distribution.

The piecewise-constant hazard model is very convenient to analyze under the proportional hazard assumption:

$$h_x(t) = h_{x_0}(t)e^{(x-x_0)'\beta} = h_j e^{(x-x_0)'\beta}. \tag{8.37}$$

There are $p + 2J$ parameters to estimate. This can be done directly using maximum likelihood, or by casting the problem as several separate Poisson regression problem. This has the benefit that the problem may be immediately solved with any statistical software suite, with existing numerical solvers. We will currently not pursue this avenue, and refer the reader to the bibliographic notes.

8.3 Collecting the pieces

In this chapter we have seen the probabilistic reliability analysis, where we assumed components' reliabilities are known. We then proceeded to the statistical problem of estimating reliabilities from failure data. We now glue collect these pieces to sketch a realistic analysis workflow, which would look roughly as follows:

1. Estimate reliability parameters by collecting component-wise failure data. Data is collected in a lab so that you may accelerate time and rescale to realistic operating conditions.

2. Perform the probabilistic analysis of the whole system using the estimated parameters.

Remark 17 (Interplay between probability and statistics)**.** There is obviously an interplay between the above stages: In the statistical analysis, we want the least possible set of assumptions; For the probabilistic analysis, the more we can assume, the more we can say about the system as a whole.

Remark 18 (What is a component?)**.** The notion of a "system" and a "component" is not well defined. Indeed, for some purposes you may consider a system, as a component in a larger system. For the statistical problem, a component is probably the smallest unit you can collect data on. At times, the smallest unit, may be the system as a whole.

8.4 Repairable systems

In this section we return the a probabilistic analysis. Unlike the previous sections, we now study systems which may be repaired after they fail. We will thus want to study the state of the system at time t. This state may depend on the number of failures and repairs performed on the system up to t.

We start with the introduction of some quantities of interest.

Definition 51 (Point availability)**.** The point availability at time t, denoted $A(t)$ is defined as

$$A(t) := \mathbf{E}\left[\Phi_t\right] = P(\Phi_t = 1). \tag{8.38}$$

Definition 52 (Interval reliability)**.** When considering the availability in some time interval J, and denoting by N_J the number of system failures in the interval, we may study

$$P(N_j \leq k) \tag{8.39}$$
$$M(J) := \mathbf{E}\left[N_j\right] \tag{8.40}$$
$$A(J) := P(\Phi_t = 1), \forall t \in J. \tag{8.41}$$

Definition 53 (Interval downtime)**.** Denoting by $Y_J = \int_J (1 - \Phi_t) dt$ the downtime during interval J, we may study

$$P(Y_j \leq y) \tag{8.42}$$
$$A^D(J) := \frac{\mathbf{E}\left[Y_J\right]}{|J|} \tag{8.43}$$

Limiting measures The particular time t, or interval J are usually not of real importance in the sense that all times and intervals are equally important. We will thus typically be interested in the above performance measures for in some *steady state* of the systems, so that the measure is representative of all t (or J), and thus no longer depends on t (or J). The typical approach for this is to study the limit of the performance measure, which implies the system has reached it's steady state. Formally, this means studying $\lim_{t\to\infty}$ of the above measures.

We now start with the analysis of a *single component* system, which we later complicate into *multiple component systems*. The required theory is that of stochastic processes, in particular *counting processes*. The reader is referred to the bibliographic notes for rigorous proofs and details.

8.4.1 Single component systems

For a single component, $\Phi_t = \mathbf{x}(t)$. If the component fails, it is replaced or repaired. We denote by T_k and R_k and the (random) time of the k'th run, and repair, respectively. We assume $T_k \sim F$ and $R_k \sim G$, independent.

Definition 54 (MTTF)**.** We denote by $\mu_F = \mathbf{E}[T_k]$, the *mean time to failure* (MTTF).

Definition 55 (MTTR)**.** We denote by $\mu_G = \mathbf{E}[R_k]$, the *mean time to repair* (MTTR).

Obviously, MTTR and MTTF are important characteristics of the single-component system.

Theorem 8.4.1 (Stable point availability)**.** *As* $t \to \infty$

$$A(t) \to \frac{\mu_F}{\mu_F + \mu_G}. \tag{8.44}$$

Theorem 8.4.2 (Stable failures per unit of time)**.** *As* $t \to \infty$*, then with probability one*

$$\frac{N_t}{t} \to \frac{1}{\mu_F + \mu_g} \tag{8.45}$$

Theorem 8.4.3 (Stable unavailability)**.** *As* $t \to \infty$*, then with probability one*

$$A^D([0,t]) = \frac{Y_t}{t} \to \frac{\mu_G}{\mu_F + \mu_G} \tag{8.46}$$

8.4.2 Multiple component systems

We will now want to study the availability of a system of multiple repairable components. The performance of the single-component system still apply, but the analysis now has to account for the fact that the state of the systems depends on the state of n repairable components, assumingly independent. By indexing the components with i, we denote $T_{i,k}, R_{i,k}$ for the uptime and repair time of the k'th failure of the i'th component. Their distributions are F_i and G_i respectively. The system failures up to time t is still $N(t)$, but not we also allow for component-wise processes $N_i(t)$, with expectations $M(t)$, and $M_i(t)$.

Denoting $A_i(t)$ the availability of component i at time t, $A(t)$ the n-vector of reliabilities, and $A_\Phi(t)$ the whole system's reliability.

[TODO:Complete from (Aven and Jensen, 1999, Sec.4.3)]

8.5 Bibliographic Notes

An light introductory discussion, may be found in Nahmias and Olsen (2015). The probabilistic analysis in this text is adapted from Aven and Jensen (1999). The seminal reference probably being Barlow and Proschan (1965). The statistical analysis is adapted from German Rodriguez's Generalized-Linear-Models class notes[1] and (Natrella, 2010, Ch.8). For more on the statistical analysis, see Cox and Oakes (1984), Kalbfleisch and Prentice (2002), or Klein and Moeschberger (2005).

[1] `http://data.princeton.edu/wws509/notes/c7.pdf`.

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

$\coprod_{i=1}^n$ The coproduct operator: $\coprod_{i=1}^n x_i := 1 - (1 - x_1) \times \cdots \times (1 - x_n)$.

$\#\{A\}$ 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 $1 \times p$ row vector x, then x' is a $p \times 1$ column vector.

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

$*$ The convolution operator: $f * g = (f * g)(t) = \int f(s)g(t - s)ds$.

f^{*n} The convolution power: n convolutions of f with itself.

$\sigma_{(j)}(A)$ The j'th largest singular value of matrix A.

$\sigma_{max}(A)$ The largest singular value of matrix A.

$\lambda_{max}(A)$ The largest eigenvalue of a symmetric positive definite matrix A.

Appendix B

R

Exploratory See, for instance, Venables and Ripley (2002), or any of the endless free web resources.

Inference The same as exploratory.

SPC For SPC with R see the `qcc`, `spc` packages, the appendix in Qiu (2013), and here `http://blog.yhathq.com/posts/quality-control-in-r.html`.

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

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

Bibliography

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, Hoboken, NJ, 2003.

T. Aven and U. Jensen. *Stochastic Models in Reliability*. Springer, 1999.

M. Bagnoli and T. Bergstrom. Log-Concave Probability and Its Applications. *Economic Theory*, 26(2):445–469, 2005.

R. E. Barlow and F. Proschan. *Mathematical Theory of Reliability*. Wiley, 1965.

M. Basseville, I. V. Nikiforov, and others. *Detection of Abrupt Changes: Theory and Application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

Y. Berchenko, J. Rosenblatt, and S. D. W. Frost. Modeling and Analysing Respondent Driven Sampling as a Counting Process. *arXiv:1304.3505*, 2013.

G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, 1978.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *International Journal of Epidemiology*, 38(5):1175–1191, 2009.

D. R. Cox and C. A. Donnelly. *Principles of Applied Statistics*. Cambridge University Press, 2011. Google-Books-ID: Uel0MgEACAAJ.

D. R. Cox and D. Oakes. *Analysis of Survival Data*. CRC Press, 1984.

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

R. Dorfman. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

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.

Z. Ge and Z. Song. *Multivariate Statistical Process Control: Process Monitoring Methods and Applications*. Springer Science & Business Media, 2012.

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

J. J. Goeman and A. Solari. Multiple Testing for Exploratory Research. *Statistical Science*, 26 (4):584–597, 2011.

W. H. Greene. *Econometric Analysis*. Pearson Education India, 2003.

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.

H. Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2 (3):360–378, 1931.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 2002.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media, 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.

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual*. SIAM: Society for Industrial and Applied Mathematics, Philadelphia, 2001.

D. C. Montgomery. *Introduction to Statistical Quality Control*. John Wiley & Sons, 2007.

S. Nadarajah. Bathtub-shaped failure rate functions. *Quality & Quantity*, 43(5):855–863, 2008.

S. Nahmias and T. L. Olsen. *Production and Operations Analysis: Seventh Edition*. Waveland Press, 2015.

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.

F. Pukelsheim. *Optimal Design of Experiments*. SIAM, 1993.

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

Y. Ritov. Decision Theoretic Optimality of the Cusum Procedure. *The Annals of Statistics*, 18 (3):1464–1469, 1990.

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.

M. S. Srivastava. On testing the equality of mean vectors in high dimension. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 17(1):31–56, 2013.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Statistics and Computing. Springer New York, New York, NY, 2002.

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].

R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 2005.