TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SER. A I OSA - TOM. 450

ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

APPLICATIONS OF INDUSTRIAL STATISTICS

by

Jarno Kankaanranta

TURUN YLIOPISTO
UNIVERSITY OF TURKU
Turku 2012

From
Department of Information Technology,
The Doctoral Programme of
the Faculty of Mathematics and Natural Sciences,
University of Turku
Turku, Finland

Supervised by

Research Director Aulis Tuominen
Business and Innovation Development
University of Turku
Salo, Finland

Dr. Arho Suominen
Department of Information Technology
University of Turku
Salo, Finland

Reviewed by

Professor Jussi Kantola
Department of Production
University of Vaasa
Vaasa, Finland

Dr. Jyri Salminen
Nokia Corporation
Salo, Finland

Opponent

Docent Tomas Eklund
Department of Information Technologies
Åbo Akademi University
Turku, Finland

ISBN 978-951-29-5229-8 (PRINT)
ISBN 978-951-29-5230-4 (PDF)
ISSN 0082-7002

Painosalama Oy – Turku 2012

To my children Noora, Niklas and Joonas

and to my wife Pirjo

Abstract

This dissertation examines knowledge and industrial knowledge creation processes. It looks at the way knowledge is created in industrial processes based on data, which is transformed into information and finally into knowledge. In the context of this dissertation the main tool for industrial knowledge creation are different statistical methods.

This dissertation strives to define industrial statistics. This is done using an expert opinion survey, which was sent to a number of industrial statisticians. The survey was conducted to create a definition for this field of applied statistics and to demonstrate the wide applicability of statistical methods to industrial problems. In this part of the dissertation, traditional methods of industrial statistics are introduced. As industrial statistics are the main tool for knowledge creation, the basics of statistical decision making and statistical modeling are also included.

The widely known Data Information Knowledge Wisdom (DIKW) hierarchy serves as a theoretical background for this dissertation. The way that data is transformed into information, information into knowledge and knowledge finally into wisdom is used as a theoretical frame of reference. Some scholars have, however, criticized the DIKW model. Based on these different perceptions of the knowledge creation process, a new knowledge creation process, based on statistical methods is proposed.

In the context of this dissertation, the data is a source of knowledge in industrial processes. Because of this, the mathematical categorization of data into continuous and discrete types is explained. Different methods for gathering data from processes are clarified as well. There are two methods for data gathering in this dissertation: survey methods and measurements.

The enclosed publications provide an example of the wide applicability of statistical methods in industry. In these publications data is gathered using surveys and measurements. Enclosed publications have been chosen so that in each publication, different statistical methods are employed in analyzing

of data. There are some similarities between the analysis methods used in the publications, but mainly different methods are used.

Based on this dissertation the use of statistical methods for industrial knowledge creation is strongly recommended. With statistical methods it is possible to handle large datasets and different types of statistical analysis results can easily be transformed into knowledge.

Keywords: knowledge creation, data information knowledge wisdom hierarchy, industrial statistics

Abstrakti

Tämä väitöskirja tutkii tietoa ja sitä, miten tietoa pystytään luomaan teollisista prosesseista. Tiedonluomisprosessin avulla teollisista prosesseista saatu data käsitellään erilaisten teollisuudessa käytettyjen tilastollisten menetelmien (industrial statistics) avulla ja muunnetaan informaatioksi. Informaatiosta luodaan edelleen tietoa tulkitsemalla tilastollisen analyysin tulokset.

Työssä määritellään, mitä teollisella tilastotieteellä tarkoitetaan. Kyseinen määrittely on tehty alan asiantuntijoille lähetetyn kyselyn perusteella. Teollinen tilastotiede on määritelty, jotta on pystytty luomaan käsitys tilastollisten menetelmien laajoista soveltamismahdollisuuksista erilaisissa teollisuuden ongelmissa. Samalla myös perinteiset teollisen tilastotieteen menetelmät, kuten esimerkiksi teollinen kokeensuunnittelu, tilastollinen prosessin ohjaus ja luotettavuusanalyysit on esitelty. Koska tieto on luotu datan perusteella käyttämällä tilastollista mallintamista ja päätöksentekoa, myös näiden menetelmien perusteet on esitelty.

Työn teoreettisena taustana sovelletaan laajasti käytettyä data-informaatio-tieto-viisaus hierarkiaa *(DIKW hierarchy)*. Kyseisen hierarkian perusteella on tutkittu, miten data pystytään muuntamaan informaatioksi ja miten informaation perusteella pystytään luomaan tietoa. Useat tutkijat ovat analysoineet työn teoreettisen viitekehyksenä toimivaa hierarkiaa. Data-informaatio-tieto-viisaus hierarkiaa on myös kritisoitu joidenkin tutkijoiden toimesta. Eri tutkijoiden määritelmien perusteella on tehty omat määritelmät kyseisille termeille. Työssä on kehitetty malli, joka selittää, miten teollisuudessa tilastollisten menetelmien avulla pystytään datan perusteella luomaan tietoa. Koska data on teollisuuden prosesseissa kaiken tiedon lähde, työssä on esitelty datan matemaattinen jaottelu jatkuvaan ja diskreettiin dataan, sekä erilaisia datan keruumenetelmiä.

Kokoelmaväitöskirja sisältää julkaisuja, jotka pyrkivät esittelemään tilastollisten menetelmien laajoja mahdollisuuksia tiedon luomiseksi erilaisista teollisista prosesseista. Julkaisut on valittu niin, että niissä on pyritty käyttämään erilaisia tilastollisia menetelmiä datan analysoimiseen. Tosin julkaisujen välillä on myös samankaltaisuutta käytettyjen menetelmien osalta, koska niissä

on analysoitu osin samankaltaisia tilanteita. Julkaisuissa on pyritty käyttämään eri datan keruumenetelmiä, kuten kyselyitä ja mittaamista.

Tämän väitöskirjan perusteella tilastollisia analyysimenetelmiä pitäisi käyttää teollisuudessa tiedon luomisessa. Tilastollisten menetelmien avulla on mahdollista käsitellä suuria tietomääriä ja analyysitulokset pystytään helposti muuntamaan tiedoksi.

Avainsanat: tiedon luominen, data informaatio tieto viisaus hierarkia, teollinen tilastotiede

Foreword

Life is strange. When I was young and had just failed my first application to university, I could not even imagine that 18 years later I would be completing my dissertation. During these years many things have happen. The direction of my professional life has changed – twice. I have met interesting people who have greatly influenced my life, both professional and personal. I have got married and studied mechanical engineering, electronics manufacturing and even statistics, which I could never have anticipated.

First I would like to thank my supervisors, Research Director Aulis Tuominen and Dr. Arho Suominen who have helped me a lot for finalizing the dissertation. Without you this work would never have been completed. Thank you for encouraging me in creating a work that meets the expectations placed on doctoral dissertation.

I would like to address my sincere gratitude to my reviewers, Professor Jussi Kantola and Dr. Jyri Salminen for their significant effort during the review process.

Mr. Ossi Hämeenoja is the prominent figure in my professional life. He have put the spark on me, when I was working on my Master of Science thesis. After I met him, the direction of my professional life completely changed. Because of him, I became interested in industrial statistics and methods such as Six Sigma. Thank you for your encouragement, support, and friendship.

Many thanks to my co-workers in Salo, together we have created an inspiring working atmosphere. Thank you to Dr. Wukui Zheng and M.Sc. Andi Mwegerano, for co-authoring papers with me. Some of these papers are also included in this dissertation.

Thank you to my parents Erja and Jorma. When I was young you provided me with a safe and pleasant home to live. You have always encouraged me towards my objectives as well as loved and supported me and my family.

Today I have achieved something important in my professional life. Still, it is only letters after your name. What is really important is not grades or titles. For me it is my family, children and wife. Thank you for my daughter Noora and my son Joonas. When I was struggling with my research and studies you were smiling and hugging me to give me the strength I need. My wife Pirjo, thank you for your support and your help. You have made it possible for me to write this dissertation. Thank you for loving me with all my weaknesses.

Lieto 19.11.2012

Jarno Kankaanranta
University Teacher
Master of Science (tech.)

List of original publications included in the thesis

This thesis is based on the following original publications, which are referred to in the text by the Roman numerals I-V:

I. Anita Gajurel, Jarno Kankaanranta and Arho Suominen. Analyzing Mobile Phone Use: The Adoption of Technologies and Services by Young People. *In Proceedings of the IAMOT 2012 Conference*, March 2012

II. Andi Mwegerano, Jarno Kankaanranta, Ossi Hämeenoja and Arho Suominen. Perceived After Sales Service Quality: Communication within the Service Chain. *Quality Technology and Quantitative Management*, Volume 9, Issue 4, December 2012, pp. 407-419

III. Wukui Zheng, Arho Suominen, Jarno Kankaanranta and Aulis Tuominen. A New Structure of a Passive Direct Methanol Fuel Cell. *Chemical Engineering Science*, Volume 76, Issue 9, July 2012, pp. 188-191

IV. Jarno Kankaanranta, Andi Mwegerano and Ossi Hämeenoja. The Importance of Having Mobile Terminal Samples for Analyzing and Verifying Customer Issues. *African Journal of Business Management*, Volume 6, Issue 44, November 2012, pp. 11088-11094

V. Wukui Zheng, Jarno Kankaanranta and Arho Suominen. Morphological Analysis of Technologies with MDS. *Journal of Business Chemistry*, Volume 9, Issue 3, November 2012, pp. 147-160

Contents

Abbreviations

ANOVA	Analysis of Variance
AOI	Automatic Optical Inspection
CUSUM	Cumulative Sum
DFSS	Design for Six Sigma
DIKW	Data Information Knowledge Wisdom
DMAIC	Define, Measure, Analyze, Improve and Control (Six Sigma road-map)
DoE	Design of Experiment
EWMA	Exponentially Weighted Moving Average
FA	Factor Analysis
LCL	Lower Control Limit
MANOVA	Multivariate Analysis of Variance
MDS	Multidimensional Scaling
MEWMA	Multivariate Exponentially Weighted Moving Average
OFAT	One Factor at the Time
OVAT	One Variable at the Time
PCA	Principal Component Analysis
RSM	Response Surface Methodology
SECI	Socialization Externalization Combination Internalization
SIPOC	Suppliers, Inputs, Process, Outputs, Customer
SPC	Statistical Process Control
UCL	Upper Control Limit

Figures

Tables

1. Introduction

The amount of data in diverse areas of industry is rapidly increasing. Different quantities (temperature, pressure, etc.) are easily measured from each step of a multistage process. This expansion is similar to the development of enlarged capability for data storage, such as databanks and the increased size of hard drives. Correspondingly, the data processing capability of computers has been increasing. At the same time a variety of comprehensive statistical and mathematical softwares are available for data analysis. These softwares provide a wide range of, nearly automated analysis methods. With them users are able to complete the analysis of extremely difficult problems in a small amount of time. But still, despite all the increased capabilities and storage space, there is a problem – the data in itself does not provide any competitive advance. Data in itself does not tell anything about the stage of the process or the opinion of the customer. Data has always been transformed into information and further on knowledge. For this reason, the industrial knowledge creation or knowledge process and the use of statistical methods in it needs to be examined. This introductory section describes the motivation to write this dissertation. The research questions and approach are also outlined as separate sub-sections. The section concludes in a description of the structure of the dissertation.

1.1. Motivation

The purpose of this thesis is to examine how statistical methods could be used in industry – generating information based on data and transforming this information into knowledge. This field is extensive since each statistical method could be applied into real life problems in industry. Thus this dissertation is limited to handling only situations with certain types of data, resulting in the title: Applications of Industrial Statistics. In the publication part of the dissertation, each publication concentrates on different types of data, intended to elaborate on a current situation and relevant problem within industrial statistics.

In this dissertation, different possibilities of applying statistics in industry are studied. Nowadays, there are different kinds of industrial processes where statistical methods are applicable. These processes produce vast amounts of data, which should be used some way. In the context of this dissertation industrial statistics as a term should be understood broadly. Industrial statistics are not only concerned with analysing the data from a process. Data could also come from a survey or have been previously collected to a dataset which was not intended for statistical analysis.

There is no reason to collect data just for the sake of the data itself; it must be collected because it has value and not because all our competitors collect data as well. When dealing with data, we always need to understand the fact that, information based on data has to be generated and that this information can be transformed into knowledge. Statistics, as well as methods that apply statistics, like Six Sigma, Design of Experiment (DoE) and Statistical Process Control (SPC) are very useful for the knowledge process. These methods help clarify the function of the examined process – for instance whether the process is under control, or what are the significant factors that explain the function of the process. More generally statistical methods are useful for knowledge creation, because it is possible to produce some statistics like mean, standard deviation and variance that characterize the entire batch of data. This is in accordance with Cleveland (1982) who agrees that mathematical formulas and theorems consist of a lot of information about data.

Ever since working on his Master of Science thesis, the author has been interested in statistics, and especially applied statistics. This Master of Science thesis was the

beginning of a whole new professional career. During the entire academic career of the author, the teaching and research of industrial statistics have been the key elements to motivate this dissertation. The possibility to apply statistical methods to industrial processes has played an important role during the research for this dissertation. There is nothing new about applying statistical methods in industry, but this dissertation has provided a possibility to examine how statistical methods create information based on data in industry.

1.2. Research Questions

This dissertation focuses on knowledge creation in industrial processes. In this context, the term process should be understood broadly: processes where data is gathered using a survey are processes just as much as manufacturing processes are. Knowledge is based on data which is transformed into information. This dissertation studies the knowledge creation process that transforms data into information and finally into knowledge. The tools for knowledge creation are different statistical analysis methods. The applicability of statistical methods in industry is extensive. It could be argued that everything in statistics could be applied in the field of industry. For this reason, only a number of statistical methods are introduced. These methods are used for knowledge creation in the enclosed publications. In addition, the scope of this thesis is not to explain every statistical method, as this kind of information about basic statistical analysis methods is easily studied from several books like Walpole et al. (2002) as well as Wonnacot and Wonnacot (1990).

In this dissertation, knowledge is based on data which is transformed into information. This transformation happens through statistical analysis. In statistics, the choice of the correct analysis method is usually related to the measurement scales based on Stevens (1946). For instance there are several possibilities for measures of central tendency. It is possible to use the mode with all scales of data (nominal, ordinal, interval and ratio). The median is a valid measure of central tendency in all measurement scales - excepting the nominal scale – and the mean can be used when dealing with the interval scale and the ratio scale. (Helenius, 1995; Devore, 2011) The aforementioned is criticized as well, according to several scholars such

as Velleman and Wilkinson (1993). Khurshid and Sahai (1993) studied several academics who have criticized Stevens (1946) and stated that the proper statistical analysis method depends on the population distribution rather than Steven's scales.

Based on the backgrounds and origins of the study, the following research questions were formulated:

> *What is the conceptual structure used to define the data to knowledge process in industry like?*
>
> *What are the relationships between data, information and knowledge in the industrial knowledge creation process?*
>
> *How does different data challenge the industrial knowledge creation process?*

This work concentrates on knowledge creation as a theoretical background, while statistical methods serve as tools for that purpose. This kind of a situation could be discovered in many companies. Companies are nowadays dealing with larger and larger amounts of data. They are also gathering different kinds of data to characterize the function of their processes. Statisticians that serve in industry (industrial statisticians) create information when they analyse the data. Finally, information is transformed into knowledge when results are interpreted. More generally, this dissertation looks at statistically operated industrial knowledge creation process.

Since the studied area is wide, the following limitations are placed. This dissertation concentrates on data, not on methods and it is not a theoretical examination on statistics. Thus, this dissertation does not create any new statistical theory. This scope was selected due to several reasons. The key element in statistical data analysis is typically the measurement scale or the distribution of the data. For example, the discrete or continuous type of data defines whether it is possible to use parametric methods or non-parametric methods. The enclosed publications have been selected so that they are all based on different types of data, but there are still some similarities between the datasets used in them. The use of different kinds of data was done because it demonstrates the wide range of applicable areas of statistics in industry. The goal is to give a solid view of a wide range of applicability of statistical methods in industry. It is similarly crucial to transform this data into infor-

mation and further on into knowledge. The background of the researcher as an engineer rather than a statistician also plays a significant role.

1.3. Research Method

The two main approaches to scientific research are the qualitative and quantitative methods. In the quantitative approach, data that either is or could be easily turned into numbers is analysed. The analysing process for this kind of data includes, for example, complex statistical analysis methods. Qualitative data consist of words that describe or maybe categorize something. (Robson, 2002) There are other and more specific distinction between these two approaches and a lot of debate in this field about the differences between these approaches. The intention is not to participate on this discussion. One classification that explains the differences between qualitative and quantitative research could be found in Halfpenny (1979). Mahoney and Goertz (2006) agree that qualitative and quantitative researches are alternative cultures and both have their own values, beliefs and norms. Bryman (1988) argues that quantitative research deals with hard and confidential data, uses structured research strategy and the relationship between researcher and researched phenomenon is distant. On the other hand qualitative research deals with rich and deep data, uses unstructured methods and the relationship between researcher and researched phenomenon is close. Both quantitative and qualitative analysis methods are used in this dissertation. Quantitative methods are used while analysing the data in enclosed publications. The examination of knowledge creation process is qualitative, because it is mainly based on one's own understanding.

Marshall and Rossman (2006) explain that the conceptual framework links the research questions into larger theoretical concepts and links the study to the examined field. Botha (1989) argues that the conceptual framework integrates words into larger systems. A well-known Data Information Knowledge Wisdom hierarchy (DIKW hierarchy) serves as a conceptual framework for knowledge creation - how data is transformed into information and further to knowledge using statistical analysis methods.

According to Marshall and Rossman (2006) there are generally four possible purposes for a study: to explore, to explain, to describe or to emancipate a phenome-

non. The purpose of an exploratory study is to investigate little-understood phenomena and identify or discover essential categories of meaning as well as generate hypotheses for future studies. Explanatory studies explain patterns related to the researched phenomenon. They can also identify probable associations, improving our understanding of the examined phenomenon. A descriptive type of study, on the other hand, documents and describes the examined phenomenon. Emancipatory studies create chances and will engage to social action. (Marshall & Rossman, 2006) This dissertation mainly fits the description of an explanatory and descriptive type of study. At the beginning of this dissertation the terms data, information, knowledge and wisdom are defined in order to elaborate on how these terms are related to each other. This is important as it generates the basis for industrial knowledge creation using statistical methods. This is followed by a conceptual understanding of the notion of industrial statistics.

1.4. Structure of the Dissertation

In chapter two the differences between four common, but still different terms, data, information, knowledge and wisdom, are examined. This examination is done by firstly defining those terms. Following this, the interrelationships between these terms are studied. This is done by using the commonly known DIKW-hierarchy. In this hierarchy data is on the lowest level, followed by information, then knowledge and finally wisdom on the highest level. This chapter does not take part in the philosophical discussion concerning the terms; it only provides some working definitions and tries to explain how data is transformed into information, knowledge and finally wisdom using statistical methods and experts.

The third chapter introduces some typical types of data in industry. A classification of industrial data based on the way it has been collected rather than qualitative or quantitative as well as continuous or discrete is also suggested. Analysis methods described in this chapter are mainly the same ones as those which have been used in enclosed publications.

In the fourth chapter the concept and traditional methods of industrial statistics, as well as future challenges in industrial statistics are explained and examined. This is done based on a survey which was sent to the experts of industrial statistics and a

literature review. Respondents for this survey were chosen based on publications in the field of industrial statistics. In this chapter some typical analysis methods for each type of data are introduced as well.

In the last chapter, chapter five, the final conclusions from this work are drawn. This chapter also explains the relevance of the included publications and clarifies the findings based on the research. This chapter correspondingly summarizes the wide applicability of statistical methods in industry. Finally suggestions on how statistics should be used in industrial knowledge creation are presented. The study is concluded with several suggestions for future research.

2. Data Information Knowledge and Wisdom

The definitions of data, information, knowledge and wisdom have been contemplated by different philosophers. For example Plato, Polanyi (1969; 1966), Dretske (1981) and Lehrer (1990) have examined knowing. The classical formulation of the definition of knowledge is from Plato[1]. In the case of wisdom, there have likewise been many attempts by a number of philosophers to understand it. Among others, there are for example philosophical approaches, implicit-theoretical approaches and explicit-theoretical approaches. (Sternberg, 2003) These philosophical contemplations, however, have been excluded from this dissertation, resulting in a more direct or engineer-like, approach to defining the concepts.

Previously mentioned limitations on this dissertations are correspondingly placed because there is no consensus on the definitions (Levitin & Redman, 1998) and it is

[1] Classical definition of knowledge is from one of the Plato's dialogues called Theaetetus

possible that one person's knowledge might be another person's data (Zeleny, 1987) and one person's knowledge might be another person's information (Lee & Yang, 2000). Authors like, Rowley (2007) and Zins (2007) collected different definitions of data, information and knowledge from different scholars in an attempt to form a coherent theory. It is also noticeable that the purpose of this dissertation is not to take part in the philosophical debate about epistemology[2] or wisdom. The purpose of this dissertation is to present a number of reasons why statistical methods should be used in industry and to offer some illustration on how information is formed and how knowledge is created in industry using statistical methods.

2.1. From Data to Knowledge and Wisdom

In the following chapters terms the data, information and knowledge are described to provide working definitions for this dissertation. This section is mainly based on the work of scholars like Ackoff (1989), Cleveland (1982) and Zeleny (1987) who have examined the interrelationships between the terms. In some sections newer definitions are from academics like Lee and Yang (2000) and Zins (2007) even though they have not researched the interconnections between the terms. Definition by statistician, Fisher (1925) has also been included.

2.1.1. Data

> *"In God we trust; all others must bring data"*
>
> *Dr. W. Edwards Deming*[3]

This quote from one of the prominent scholars in statistical quality control, Dr. Deming, illustrates the importance of the data. It means that it is not possible to trust anything without data. In the context of this dissertation, this could be understood in the way that data is the sole source of information and finally knowledge.

[2] Epistemology is the theory of knowledge (Brown & Duguid, 2002)
[3] This citation is from http://www.quotesdaddy.com/author/W.+Edwards+Deming/3 accessed on 28.03.2012

One of the earliest academics, who examined the relationship and interconnections between data, information and knowledge was Zeleny (1987). He describes data with the simple metaphor "know-nothing" and clarifies that data when it is a component of his hierarchy, could be generated without human interpretation. Ackoff (1989) examined the same structure almost simultaneously and his result was a more specific definition for data. He argues that data represents the properties of objects, events and their environments. It could be perceived that data is a product of observation, which is similar with sensing. On the other hand, Davenport and Prusak (2000) describe data as discrete and objective details about events. Data is useful in an organizational framework, because it typically describes structured records like transaction.

Summarising the previously mentioned definitions of data it is possible to conclude that data characterises the properties of objects and events. Data has similarities with sensing and describes events. In organizational framework data typically consists of records about events.

2.1.2. Information

"Information is not knowledge"

Nobel laureate Albert Einstein[4]

At the beginning of this chapter there is a quotation from Nobel laureate Albert Einstein. In this quotation, he makes a clear separation between information and knowledge. Just as there are many definitions for data, there are many definitions for information as well. The following definitions are mainly from scholars who have researched relationships and differences between those terms.

Cleveland (1982) was one of the earliest scholars who examined the interrelationship and structure between data, information and knowledge. He uses a simple conceptualization for information. He argues that information is facts and ideas that are known by somebody at a given time. He divides information in six dimensions

[4] Citation is from http://www.brainyquote.com/quotes/quotes/a/alberteins163057.html accessed on 28.03.2012

(**Table 2-1**). According to Cleveland (1982) information is expandable: more and more information is provided and people are struggling to reduce the information overload and the uncertainty about what they should do. Information is compressible so many complex data structures can be squeezed into simpler ones. Mathematical formulas and theorems consist of a lot of information about data. Substitutability means that information is able to replace capital, labor or physical materials. As an example automation and robotics have transformed workers in a factory from non-information workers to information workers. Transportability of information makes it possible to transport resources quickly. Information leaks, and the more it leaks the more we get, so information is diffusive. Information is shared when a person passes an idea or a fact to another person, so information is shareable.

dimension	definition
expandable	More and more information is provided and people people are trying to reduce the information overload.
compressible	Complex data structures are possible to compress into sigle ones. Theorems and formulas consists of lot of information about the data.
substitutable	Information could replace capital, labor or physical materials.
tranportable	Recources could be transported quickly
diffusive	Information is leaking. When information leaks we get more
shareable	Information is shared when someone tells an idea or fact to someone else

Table 2-1: The six dimensions of information according to Cleveland (1982)

However, Zeleny (1987) has also examined the same field and ended up with the use of a simple metaphor to describe information, "know-how". According to Zeleny (1987) information as a part of his hierarchy could be generated without human interpretation. Almost simultaneously Ackoff (1989) defined information as a data that is transformed into information through analysis. Ackoff (1989) argued that this processing is in many cases statistical or arithmetical. Usually information contains answers to questions like "who", "what", "where", "when" and "how many".

A Japanese scholar, Nonaka (1994) agreed that information is a flow of messages. This definition has similarities with the definition of Davenport and Prusak (2000). According to Davenport and Prusak (2000) information can be defined as a message, which is usually audible or visible in the form of a document. It is possible to describe information as "data that makes a difference". Because information is a message it has an impact on the receiver's judgments and behaviour. Information has a meaning and data is transformed into information when meaning is added. Davenport and Prusak (2000) described five methods of transforming data into information. Those methods are as follows:

- *Calculated: the data may have been analysed mathematically or statistically*
- *Categorized: we know the units of analysis or key components of the data*
- *Condensed: the data may have been summarized in a more concise form*
- *Contextualized: we know for what purpose the data was gathered*
- *Corrected: errors have been removed from the data*

The first point of Daventports and Pruzaks (2000) list, presented above, is interesting. This "calculated" point clarifies the assumption that mathematics and statistics could be used for information creation. This supports the assumption made at the very beginning of this dissertation – statistics create information based on data.

Two rather different definitions of information are from Shannon (1948a; 1948b) and Fisher (1925). These two scholars have not examined the knowledge hierarchy. They have studied it in the context of their own discipline. Fisher was a statistician and Shannon researched the theory of communication. For Shannon (1948a; 1948b) information is related to messages sent to communication channel by an information source. Fisher (1925) defined information as a kind of mathematical measure. According to his publication information is "intrinsic accuracy". He explains

that the intrinsic accuracy of the error curve is equally perceived as the amount of information in a single observation belonging to such a distribution.

To conclude this section discussing the definitions of information as presented by a number of academics, information is something that is formed through mathematical or statistical analysis. Information is audible or visible messages that are sent into an information channel. Information has an effect on the receiver's acts and judgments and in the statistical discipline it can be defined as "intrinsic accuracy".

2.1.3. Knowledge and Wisdom

"Knowledge is power"

Sir Francis Bacon[5]

The citation at the beginning of this chapter needs no explanation: knowledge is power. Data and information, examined in the previous chapters (Chapters 2.1.1 and 2.1.2), are not enough by themselves – they only form the lower levels of DIKW-hierarchy (Chapter 2.3). Because knowledge is so important, it has to be created. Knowledge is created in the knowledge process. In this dissertation the transformation from data into information is based on statistical analysis, and further, the creation of knowledge is realized when analysis results are interpreted. This is the main reason why the interconnection between data, information and knowledge are studied in the following chapters.

Like data and information, knowledge has multiple definitions. Knowledge is created when useful information is selected and organized into knowledge (Cleveland, 1982). According to Ackoff (1989) knowledge could be achieved in two different ways. It could be transferred to another person by the one who possesses it, through instruction, or can be extracted from experience. In both of these cases the acquisition of knowledge is considered learning. Learning happens when one's competence grows over time or with the number of attempts. As an example computers are "taught" when they are instructed and programmed to do something. This has similarities with education. Statistical knowledge is taught to statisticians, who then

[5] Citation from Box, Hunter and Hunter (2005)

use it to create new knowledge. Ackoff (1989) continues his definitions of knowledge with the simple metaphor "know how" – for instance, the knowledge about how a system works makes it possible to control a system. Zeleny (1987) does not agree with Ackoff's (1989) metaphor. He defines knowledge using the metaphor "know-what". Zeleny (1987) agrees that when data is a part of his hierarchy it cannot be generated without human interpretation.

Dretske (1981) as well as Nonaka and Takeuchi (1995) argued that according to western philosophers knowledge is "justified true beliefs". On the other hand, philosopher Edmund Gettier (1963) wrote a famous article where he demonstrated why this classical definition for knowledge is flawed and should not be used.

Newer definitions for knowledge could be found from academics like Davenport and Prusak (2000), Lee and Yang (2000) and Zins (2007). Davenport and Prusak (2000) claim that knowledge is not neat and simple. They argue that most people have a natural sense that it is broader, deeper and richer than data or information. They provide a definition that knowledge originates and is applied in the minds of knowers. Knowledge is expert insight that makes possible to evaluate and incorporate new experiences and information. It is a fluid mix of framed experience, values and contextual information. Knowers apply and incorporate knowledge in their minds. In an organizational context knowledge is usually embedded in routines, processes, practises and norms or some repositories and documents. Lee and Yang (2000) outlined that knowledge is more than just information. Information is transformed into knowledge when someone reads, understands, interprets and applies information to a specific work function. It is possible to make knowledge visible when an experienced person brings lessons learned into practise. Zins (2007) claims that it is also common to understand knowledge as a product of a synthesis that is in the mind of a knowing person. Based on the aforementioned he argues that, knowledge only exists in the mind of a knowing person.

Scholars like Nonaka (1994) as well as Nonaka and Takeuchi (1995) have argued that there are two dimensions in knowledge: tacit knowledge and explicit knowledge. Another definition of the dimensions of knowledge comes from Johnson et al. (2002). They suggest that knowledge is four-dimensional. Those dimensions are knowledge about facts or "know-what", knowledge about principles and the laws of nature or "know-why", the third is the ability to do something or

"know-how" and last one is who knows what and who knows what to do or simply "know-who".

Despite the previously mentioned discrimination made by Johnson et al. (2002) this dissertation follows the classical tacit knowledge, explicit knowledge categorization. Several scholars like Nonaka (1994), Smith (2001), Koskinen et al. (2003) and Nonaka and von Krogh (2009) have stated that one of the earliest epistemological definitions of tacit knowledge comes from philosopher Michael Polayi (1966).

> *"We can know more than we can tell"*
>
> *Michael Polanyi (1966)*

Tacit knowledge is created by the individuals and it is based on the experience of individuals. It could be expressed in human actions, evaluations, attitudes, points of views etc. Tacit knowledge is usually difficult to express in words. This means that experts are not able to express how they have reached their conclusions and how they make their decisions. From a technical point of view, tacit knowledge could be established when a person practises a specific skill that is developed bit by bit. (Nonaka & Takeuchi, 1995; Koskinen et al., 2003; Smith, 2001)

One synonym for explicit knowledge could be codified knowledge. Explicit knowledge is transformed formally using systematic language. This kind of knowledge is usually acquired from schools or universities. Explicit knowledge could be easily expressed in words, numbers, books, manuals and mathematical expressions. It can be easily shared and communicated. This dimension of knowledge is acquired through education or structured study. (Koskinen et al., 2003; Smith, 2001)

The purpose of this dissertation is to examine how knowledge is created based on data using statistical methods and for that reason wisdom is outside the scope of this dissertation. The definition of wisdom is also the most philosophical question. These are the reasons why the concept of wisdom has not been as thoroughly examined as those of data, information and knowledge.

In the classical DIKW-hierarchy (Chapter 2.3) the highest level on top of knowledge, is wisdom (Rowley, 2007; Frické, 2009). In the DIKW-hierarchy there are relationships between information, knowledge and wisdom. Wisdom is

achieved when knowledge is integrated into the whole. Wisdom is more than its parts. It is context dependent and could not be created without human interpretation. (Cleveland, 1982) However, one of the earliest scholars, Zeleny (1987) simply describes wisdom with the metaphor "know-why". He also argues that wisdom is based on knowledge and when it is a part of his hierarchy it could not be created without human interpretation. Zeleny (1987) explains that wisdom is beyond knowledge because it allows comparisons with regard to "know-what" and "know-why". The mental task that transforms knowledge to wisdom is judgment. (Zeleny, 1987) Another scholar Ackoff (1989) argues that wisdom always adds value.

Rowley (2006) writes about the different definitions of wisdom among philosophy and other sciences. Based on this literature review she argues that wisdom is the capacity to put into action the most appropriate behaviour when the knowledge and what does the most good are taken into account.

To summarize the definitions for knowledge and wisdom, knowledge is broader, deeper and richer than data or information and it is two-dimensional. Knowledge is created when information is read, understood or interpreted. Wisdom is based on knowledge and the transformation from knowledge into wisdom requires judgment. Wisdom is the ability to put most appropriate behaviour into action when knowledge and what does the most good are taken into account.

2.2. Definition of Data Information Knowledge and Wisdom

In the previous Chapters (2.1.1, 2.1.2 and 2.1.3) the terms data, information, knowledge and wisdom are outlined. This description is formed based, among others, on the work of several scholars and philosophers. In this chapter the characterization of those terms is finally complete. The views of different scholars (in alphabetical order) have been collected into the following **Table 2-2**.

	DATA	INFORMATION	KNOWLEDGE	WISDOM
Ackoff (1989)	properties of objects, events and their environments	transformed data (mathematical or statistical analysis)	information transformed into instructions	adds value requires judgement
Cleveland (1982)		facts and ideas (mathematical formulas and theorems consist of)	selected and organized usefull information	knowledge integrated into a whole more usefull than the sum of its parts
Davenport and Prusak (2000)	discrete and objective details about happenings	data that makes a difference	expert insight makes posible to evaluate, incorporate new experiences and information	
Dretske (1981)			justified true beliefs (traditional definition according to western philosphers)	
Fisher (1925)		intrinsic accuracy		
Lee and Yang (2000)		more than information created when someone reads, undestands and applies information		
Nonaka and Takeuchi (1995)		flow of messages	justified true beliefs (traditional definition according to western philosphers)	
Rowley (2006)				capasity to put in action the most appropriate behaviour when knowledge and what does the most good is taken into a action
Shannon (1948a;1948b)		messages sent to communication channel information source		
Zeleny (1987)	know-nothing	know-how	know-what	know-why
Zins (2007)			product of a synthesis that is a mind of a knowing person	

Table 2-2: Different definitions for Data, Information, Knowledge and Wisdom[6]

Based on the literature review above the working definitions for data, information, knowledge and wisdom in this dissertation are proposed. The intention is not to take part in the debate concerning the definitions. There might be more suitable definitions, from other philosophers, information theoreticians and even among the engineers. These definitions are mainly based on the work of other scholars and the author's understanding of what those terms mean in an industrial context.

In this dissertation data represents the source it has been gathered from. Data is gathered systematically for analysis. It is possible that data is gathered for some

[6] Traditional definition of knowledge according western philosophers (Dretske, 1981; Nonaka & Takeuchi, 1995)

other purposes and analysed – taken into use afterwards. This definition has similarities to Davenport and Prusak (2000) (Chapter 2.1.1). They define data as discrete and as observations about events.

Information is defined as data that is transformed into information. This definition has similarities with Davenport and Prusak (2000) (Chapter 2.1.2). "Data that makes a difference" could be understood as data that is transformed into information. The transformation is often done using the statistical methods. These statistical methods could be for instance DoE or regression analyses. These are examples of methods that create information about the significant factors that characterize how the response, the process output is fluctuating as a function of significant factors.

The definition of knowledge loosely follows the definition of Lee and Yang (2000) (Chapter 2.1.3). They argued that knowledge is more than information and knowledge is created when someone reads, understands, interprets and applies information to some specific work function. In the context of this dissertation knowledge is created when someone understands, reads or interprets information such as statistical significances or p-values and statistical models into a common language.

The formulation of the definition of wisdom in this dissertation follows loosely Zeleny's (1987) metaphor, "know-why" and the work of Rowley (2006); the ability to put into action the most appropriate behaviour, when knowledge is taken into account (Chapter 2.1.3). Wisdom is achieved when our behaviour is based on knowledge and we know why something is happening in the examined process.

2.3. Knowledge Hierarchy

Zeleny (1987) and Ackoff (1989) were some of the first to conceptualize the DIKW-hierarchy. Since then many scholars have written about this structure. It is said that the origin of this hierarchy is the T. S. Eliot's poem The Rock in 1934. (Rowley, 2007; Frické, 2009)

> *"…Where is the life we have lost in living?*
> *Where is the wisdom we have lost in knowledge?*
> *Where is the knowledge we have lost in information?…"*
>
> *Nobel laureate T. S. Eliot*[7]

In the poem The Rock Eliot (2002) presents a hierarchy where information is in the lowest level and above that there is knowledge. The highest level in this hierarchy, on top of knowledge is wisdom. Correspondingly Zeleny (1987) describes the hierarchical organization between data, information, knowledge and wisdom and explais the differences between these terms. Data and information are piecemeal, partial and atomized by their nature. The two highest levels of the hierarchy, knowledge and wisdom are more "holistic". Knowledge and wisdom could only be expressed through systemic network patterns. Ackoff (1989) has similarities with Zeleny's (1987) hierarchy. In his conceptualization Ackoff (1989) presents a five-step hierarchy where data is at the bottom and information is placed on top of it. The higher level following information is knowledge. After knowledge there is an extra level, understanding, and wisdom is the highest level. In his context understanding is created through management support systems. These kinds of systems must be able to improve understanding, for instance by detecting errors, determining their causes and correcting them.

Many scholars including Lillrank and Forssén (1998), Tuomi (2000), Smith (2001), Rowley (2007), Harsh (2007), and Frické (2009) have correspondingly written

[7] T. S. Eliot's poem The Rock was originally published in 1934. In this dissertation the citation is from book T. S. Eliot Collected Poems 1909-1962 edited by John Dawson, Peter Holland and David McKitterick and published by Faber and Faber in 2002

about. DIKW-hierarchy (**Figure 2-1**). Other possible names used for this model are for instance the "Knowledge Hierarchy", the "Information Hierarchy" and the "Knowledge Pyramid" or the "DIKW pyramid". This hierarchy is usually used to contextualize data, information, knowledge and sometimes wisdom as well as the interrelationships between these terms.

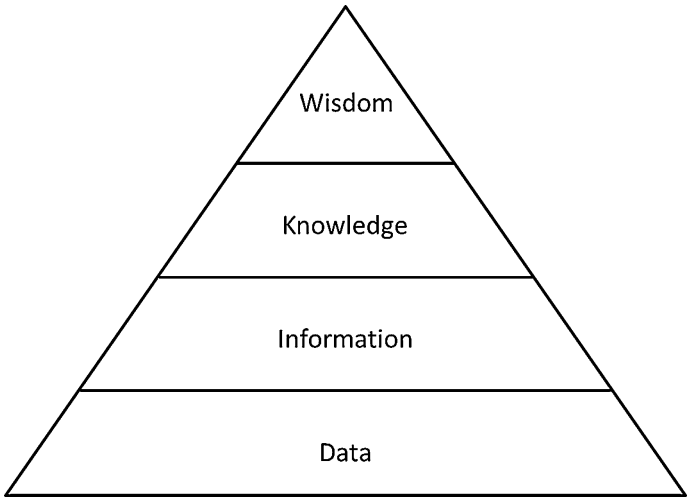

Figure 2-1: Data Information Knowledge Information Hierarchy (Rowley, 2007; Frické, 2009), adopted.

The basic assumption of the DIKW-hierarchy (**Figure 2-1**) is that data can be used to create information and that it is possible to use information to create knowledge and knowledge could be used to create wisdom. This DIKW-hierarchy is usually used to describe processes that transform something from a lower level of hierarchy into a higher level of hierarchy. (Rowley, 2007)

Lillrank and Forsen (1998) (**Figure 2-2**) have three levels in their hierarchy. Those levels are data, information and knowledge. The lowest level of this hierarchy is data and the highest is knowledge. In the context of this dissertation, this is a good description of the knowledge hierarchy. Data has symbolic representations and it is based on measurements. When the data is placed in context and value is added to it, information is created. On the other hand the metaphor "know-what" could be understood as what is happening in a stage of the process. Finally, knowledge is created when meanings and dynamics are added to the information – when the analysis results are interpreted. The metaphor "know-how" describes knowledge.

This could easily be placed into the industrial context: how a process or some stage of it is functioning.

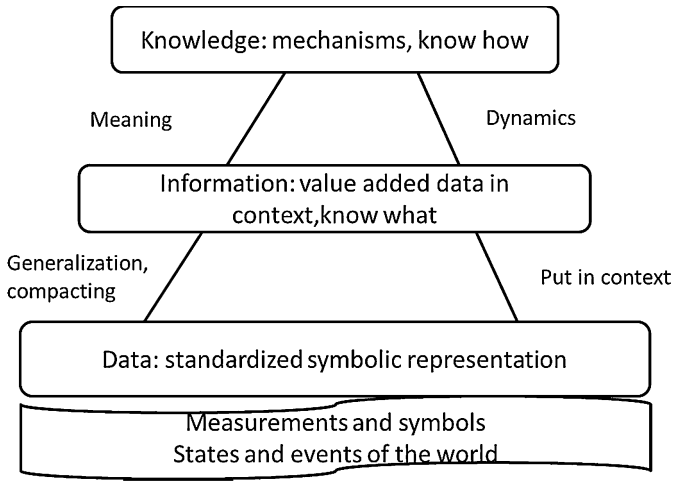

Figure 2-2: Knowledge hierarchy from Lillrank and Forsén (1998), adopted.

DIKW-hierarchy models have also been criticized. The hierarchy requires that it is built on a solid base. The data on which the pyramid is based should not be false and the data should not have been collected in the hope that someday it will be transformed into information. (Frické, 2009) In industrial applications these kinds of problems are reduced, through the use of the methods like the Gage R&R (Chapter 4.4.1), which ensure that it is possible to count on the data and that we are measuring the right object or phenomenon. It is, however, noteworthy that in industry, data collection is dominantly purpose driven and not done just for the purpose of gathering data. Since this classical DIKW-hierarchy has been criticized, several scholars have proposed their own structures. One suggestion is the reversed structure by Tuomi (2000). This reverse hierarchy (**Figure 2-3**) is interesting and some parts of it are adopted in the industrial knowledge process that is proposed in the following chapter (2.4.3).

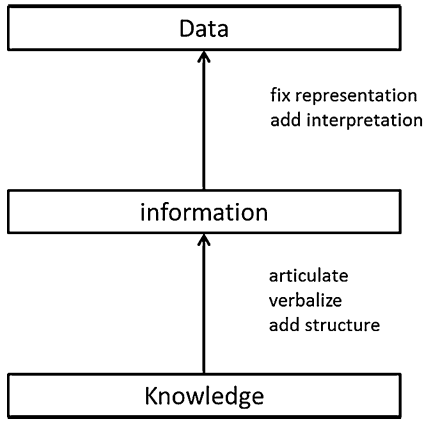

Figure 2-3: Knowledge hierarchy from Tuomi (2000), adopted.

The concept of Tuomi (2000) (**Figure 2-3**) is based on the assumption that data could appear only after knowledge and information are present. Data appears after the meaning and semantics of it are fixed and used to represent information, on the other hand there is no data before someone has created it using his or her knowledge. This structure is interesting due to its perception of how data is created. Knowledge is required when something is measured, to decide what is measured and how it should be measured. For instance in the context of this dissertation, when data is gathered using a survey, sufficient knowledge of survey creation is required.

2.4. Knowledge Creation Process

"Lack of knowledge … that is the problem"

Dr. W Edwards Deming[8]

The importance of knowledge can be directly gathered from this quote by Dr. Deming. It could be easily understood to mean that knowledge needs to be created.

[8] The citation is from http://www.brainyquote.com/quotes/authors/w/w_edwards_deming.html accessed on 28.03.2012

In the context of this dissertation the quote could be understood widely – data and information are not enough more is needed, knowledge has to be created.

Knowledge is created and organized based on the flow of information and it is attached to the beliefs and commitment of its owner. Knowledge includes many layers of meanings. (Nonaka, 1994) The classic view of knowledge conversion, the SECI model, is from Nonaka (1994) and Nonaka and Takeuchi (1995) (**Figure 2-4**), who suggested that knowledge is conversed through the interaction of tacit and explicit knowledge. This conversion could happen in four ways: from tacit knowledge to tacit knowledge (socialization), from tacit knowledge to explicit knowledge (externalization), from explicit knowledge to explicit knowledge (combination) and explicit knowledge to tacit knowledge (internalization).

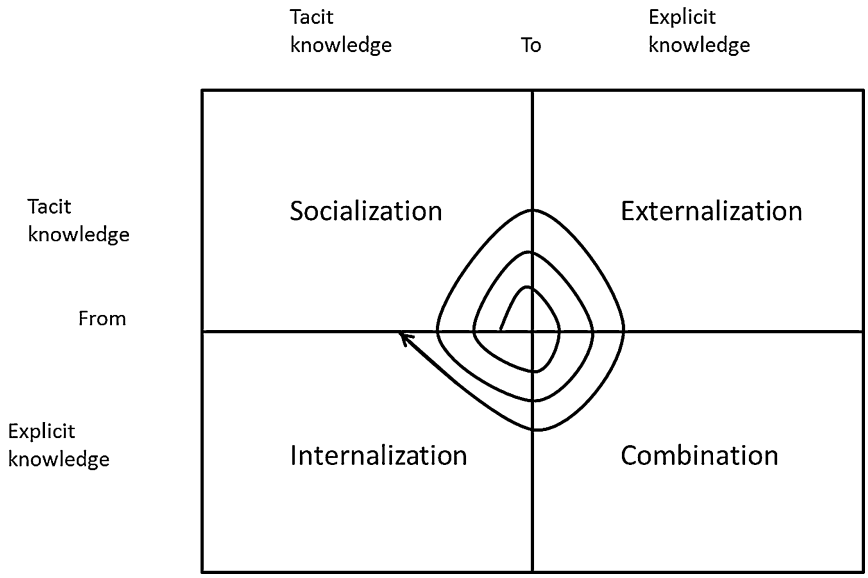

Figure 2-4: Knowledge conversion (SECI model) according to Nonaka (1994), adopted.

Figure 2-4 depicts graphically Nonaka's (1994) and Nonaka and Takeuchi (1995) idea on how knowledge could be conversed. Socialization, in their knowledge conversion model, refers to a situation where experience is shared as the result of tacit knowledge, in the way that mutual mental models and technical skills are formed. Explicit knowledge based on tacit knowledge is created through externalization. In this type of knowledge conversion the shapes of metaphors, analogies, concepts,

hypotheses or models are used for knowledge creation. Combination creates explicit knowledge based on explicit knowledge. This kind of knowledge creation contains different bodies of explicit knowledge. As an example, individuals share and exchange knowledge using different media. The internalization type of knowledge conversion is related to learning by doing. This kind of knowledge adaption is helped if knowledge is verbalized or documented. This helps individuals to internalize what they have experienced. (Nonaka & Takeuchi, 1995)

The previously described knowledge creation process, introduced by Nonaka and Takeuchi (1995), has received some critique. A spiral, as seen in the **Figure 2-4**, is usually added to their original conceptualization. The spiral describes the continuous social interaction between tacit and explicit knowledge. The spiral begins from socialization, goes through externalization and combination and ends at internalization, before beginning a new round. According to Gourlay (2006) there are problems related to this formulation, such as why does the process start from socialization and not from combination or externalization. The other challenges in Nonaka's model (**Figure 2-4**), as presented by Gourlay (2006) are too philosophical in nature to be discussed in the context of this dissertation. This dissertation aims to explain how statistical methods can be used for knowledge creation – not to describe the structure of the term itself. Because of this, the basic SECI model was adopted for this dissertation.

Based on the aforementioned SECI model (**Figure 2-4**) Nonaka and his co-workers (Nonaka et al., 1994; Nonaka & Takeuchi, 1995) have created a model for organizational knowledge creation (**Figure 2-5**). As this model is widely adopted in industrial knowledge creation, it is presented here as well. It is also noteworthy that this model is used as a part of the final statistical knowledge creation process as proposed in this dissertation.

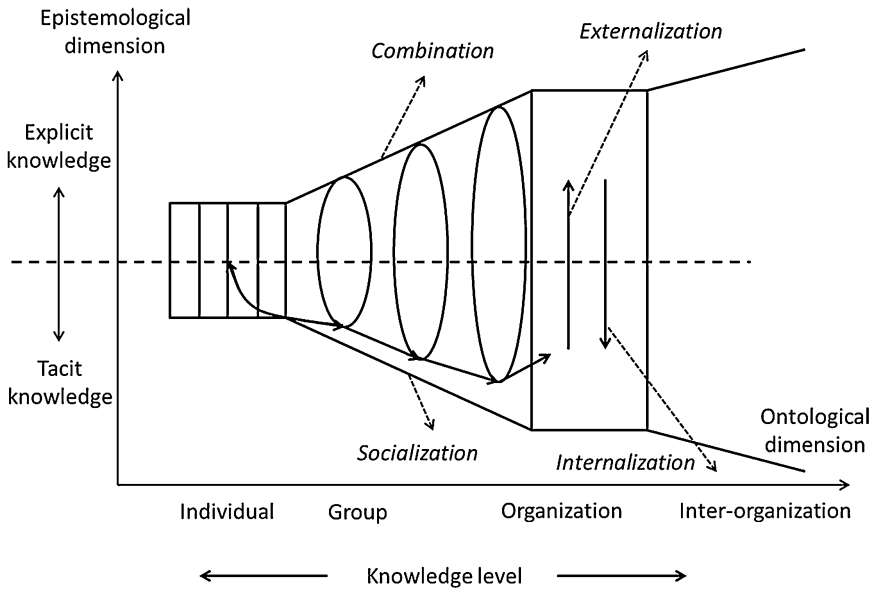

Figure 2-5: Organizational knowledge creation spiral according to Nonaka and Takeuchi (1995), adopted.

Nonaka and Takeuchi (1995) explain that their SECI model mainly concentrates on the epistemological parts of the industrial knowledge creation process. According to them, companies themselves cannot create knowledge, meaning that knowledge is created by individuals within the company. As a result, they have proposed a knowledge creation spiral (**Figure 2-5**). Individuals employed by companies have large amounts of tacit knowledge and this tacit knowledge is the basis of organizational knowledge creation. It is crucial to a company to activate this tacit knowledge and to amplify it through the four phases of the SECI model, finally crystallizing it at the higher ontological levels. This organizational knowledge creation spiral starts from the individual level and moves up through the shared level (group level) to the organizational level and sometimes even to the inter-organizational level. (Nonaka et al., 1994; Nonaka & Takeuchi, 1995)

2.4.1. Statistical Decision Making

There are two or sometimes three main approaches to statistical decision making: classical statistics, Bayesian statistics and decision theory. This discrimination is created based on how these approaches handle the data and prior information. (Barnett, 1982) These three main approaches are explained in the next chapters. There is also a lot of debate in this field. This discussion is about the right way to do statistical inference – classical statistics or Bayesian statistics. Efron (2005) claims that this conversation has been going on for 250 years[9]. Because there is so much disagreement and discussion among the theoretical statisticians in this field, the intention is not to take part in this debate. Here, these three approaches are only introduced because in the context of this dissertation they could be seen as a part of the knowledge process. This has similarities with Ackoff (1989) and Davenport and Prusak (2000) (Chapter 2.1.2), who explained that data is transformed into information by mathematical or statistical analysis. These methods describe how statistical analyses are made.

In the future work and limitations (Chapter 5.3) part of this dissertation some personal ideas and suggestions are presented. However, it is noticeable that all the analyses in the enclosed publications have been done following the classical statistical approach – not the Bayesian.

Classical statistics, sampling-theory, frequentist statistics, orthodox statistics or standard statistics is mainly based on the work of Sir R. A. Fisher, E. S. Pearson and J. Neyman. This approach is based on sampling theory and in this approach the information is limited to sample data. With this methodology statistical analysis is based on sample data which is evaluated through the frequency concept of probability. (Barnett, 1982; Cox, 2006)

The Bayesian approach is based on processing the sample data and some prior information. This methodology is based on the use Bayes theorem. The Bayes theorem[10] is based on the work of reverend Thomas Bayes (Barnard & Bayes, 1958).

[9] As a reference of this debate see for instance Efron (1986) including the discussion.

[10] Bayes' theorem is based on the work of reverend Thomas Bayes. His article was originally published posthumously in 1763.

This method uses inverse probability and in this approach prior information is modified through repeated use of sample data and Bayes theorem. (Barnett, 1982; Cox, 2006)

The decision theory is based on the work of Abraham Wald and it is based on decision rules. These rules are used for action under situations of uncertainty. This approach includes the assessment of different consequences of alternative actions, which is expressed as a mathematical theory with losses or los functions. The expected loss or risk of different decision rules proposing action on the basis of sample data or any other prior information is measured. The objective is to choose the alternative with "minimum risk". (Barnett, 1982)

2.4.2. Statistical Modeling as a Knowledge Creation Process

"Essentially, all models are wrong, but some are useful"

George E. P. Box[11]

The famous statistician, George E. P. Box cited at the beginning of this chapter states the fact about statistical models. Models are just models, an attempt to explain how nature or a process works. It can easily be understood that those are always faulty – some are better than others and that's how they are useful.

Statistical modeling is one way to create knowledge. The steps of the traditional statistical modeling process are described in **Figure 2-6**. This statistical modeling process seeks a proper statistical model between the input variable or variables (x, independent variables) and the output variables (y, responses) (Breiman, 2001). As the model describes the relationship between the independent variables and the response, it will create information about how x predicts y, and this can be transformed into knowledge by interpreting the results. Statistical modeling is used in the publications included in this dissertation (Publication II and Publication III).

Barnard, G. A. and Bayes, T. "An Essay Towards Solving a Problem in the Doctrine of Changes." Philosophical Transactions (1763), 53, 370-418

[11] This famous quote is from the preface of Box's and Draper's book about response surface methodology (Box & Draper, 1987)

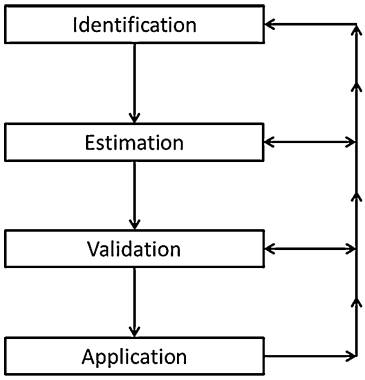

Figure 2-6: The process of statistical model building (Gilchrist, 1984), adopted.

The first step in statistical model building (**Figure 2-6**) is identification. At this stage, the appropriate model for the situation is recognized for and selected. There are two possible ways of doing this. In the "conceptual approach" the model is chosen based on knowledge about the situation, without a reference to any actual data. Another possibility is "empirical identification". This approach considers only the data, without to reference its meanings or the situation in which it arose. In the estimation step numerical values for model parameters are estimated. (Gilchrist, 1984)

Gilchrist (1984) explains the "validation" step in a very wide sense. At this stage the practical value and validity of the model in a given situation is considered. There are three phases for this step of the model building process. Those phases are development, testing and application. In the development phase, a fitted model is used to reveal the future aspects of data. In the testing phase data is gathered for future validation of the model. In the application phase monitoring processes are used to check whether this initially acceptable model remains valid in use. In this step, the linkage to the other sciences where statistical model are used is created. This means a connection with for example human sciences like psychology, where statistical model are applied. The last step of the statistical model building process is "iteration". The whole statistical model building process is iterative. The arrows in **Figure 2-6** illustrate the possibility of moving back and forth between the different stages in order to utilize additional information.

The traditional model building procedure, described above, has been criticized and an alternative method is proposed by Breiman (2001). In his work Breiman (2001) explains that instead of a data modeling culture an algorithmic modeling culture should be used. In this algorithmic modeling culture we try to find the function instead of statistical model that operates on x to predict y. As Breiman (2001) explains 98% of statisticians are using the data modeling culture and only 2% are using the algorithmic approach. Based on this, it could be argued that the state of the art method for statistical modeling is the traditional model building approach. This is the reason why in this dissertation the debate between these two approaches is not concentrated on and the algorithmic modeling is not explained.

Statistical data analysis starts with data. This assumes that data is being created by a black box. In this black box input variables (x, independent variables) are going in to the black box and output variables (y, responses) are coming out from that black box (**Figure 2-7**). (Breiman, 2001)

Figure 2-7: Nature related to the relationship between input variables (x) and response variables (y) (Breiman, 2001), adopted.

Traditionally there are two goals in data analysis. These are prediction and information. The goal of prediction is to try to foresee the response variables that might be input variables in the future. The purpose of information is to extract the information about the way that nature is related to the response variables and input variables. The analysis stage starts with the assumption that there is a stochastic data model, like a linear regression model or a logistic regression model or a Cox model inside the black box (**Figure 2-8**). (Breiman, 2001)

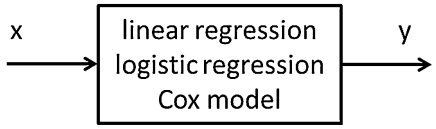

Figure 2-8: A Stochastic model related to the relationship between input variables (x) and response variables (y) (Breiman, 2001), adopted.

After the model is fitted between input variables (x) and response variables (y) it is validated. The validation or model selection is usually based on statistics. Validation is often based on using some goodness-of-fit tests or examining the model residuals. (Breiman, 2001)

This statistical modeling reveals the information that the data consists of. Using for example DoE methods (Chapter 4.3.1) it is possible to find the statistically significant input variables and the model between input variables and responses. But still going back to the Box's famous quote at the beginning of this chapter, models are just models – some are better than others and may provide some useful information which could be transformed into knowledge.

2.4.3. Knowledge Creation in Industry

Nowadays statistical methods are used in industry for knowledge creation according to the classical approach mentioned in the previous chapter (2.4.1). Data is used as a source of information and it is transformed into knowledge. As an example, statistical modeling and hypothesis testing are dealing with some goodness-of-fit tests and statistical significances (p-values).

It might be useful to expand this state of the art, classical approach. In many situations in industry, there is some prior information available. This prior information might be experience from process operators or something else. In these situations the use of Bayesian statistics, that take prior information into account, might be fruitful. Refer to scholars like Lindley (1956) who has written about Bayesian experimental designs and Chaloner and Verdinelli (1995), who have done a literature review on Bayesian statistical methods.

Although there are possible advantages to using Bayesian statistics, this dissertation follows the classical statistical approach (Chapter 2.4.1). This is done because the leading method in this area is the classical approach, on the other hand it should be mentioned that the number of the appliers of the Bayesian approach is rapidly increasing. The first step of industrial knowledge creation based on classical approach is data gathering. In this first step data is gathered according to a plan which requires knowledge about different statistical sampling methods. These sampling methods are for example random sample with or without return, sample by size, stratified sample, cluster sample and so on (Babbie, 1973; Sapsford, 2006). The

idea of sampling is based on the fact that in statistics it is possible to draw conclusions about the whole population based on a sample. For instance, the sample mean is non-biased estimator of population mean. This is presented in the **Figure 2-9**.

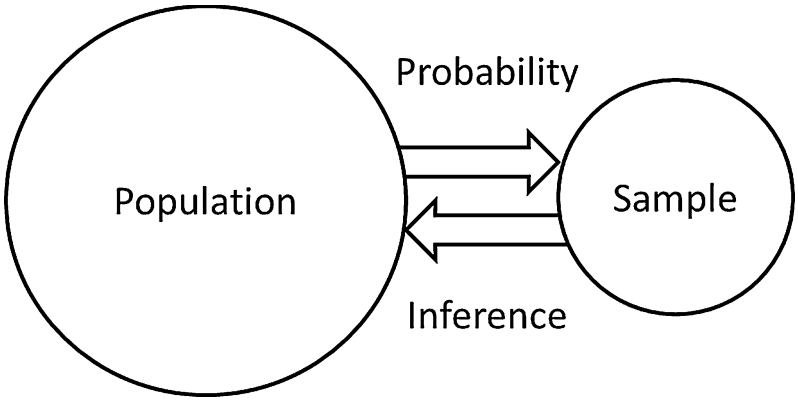

Figure 2-9: Inference about population based on sample (Devore, 2011), adopted

After the data has been gathered, the information based on it needs to generate. Now all possible and suitable statistical analysis methods are applicable.

The suitable method should be chosen based on the scale of data or the distribution of the data. It is possible to use all statistical methods could be used when analyzing data. But it is important to know that used analysis method has reasonable interpretation – what does the analysis results mean in real life. In this stage the statistical analyses are done, usually resulting in some p-values based on which interpretations are usually made. For this stage statistical analysis methods, such as regression analysis and χ^2 -test could be applied

The most important phase in the knowledge creation process is the stage where knowledge is generated based on information. In this stage, the capabilities of process specialists are needed, because the statistician is the specialist in statistical methods, but not a specialist in processes per se Vice versa, process specialists are not familiar with statistical methods and analyses; process specialist interprets with the help of statistician the statistical significances (p-values) and other statistics into common language and knowledge.

Following the conceptual framework of this dissertation, the DIKW-hierarchy (Chapter 2.3), the highest level of "knowledge creation" is wisdom and the next

phase in this knowledge creation process should be transforming knowledge into wisdom. However, in the context of this dissertation, this phase is not included in the proposed knowledge creation process.

In the next figure (**Figure 2-10**) a preliminary proposal for the knowledge creation process is presented. There are similarities with previously described models, but the conceptualization has been extended into the domain of industrial statistics.

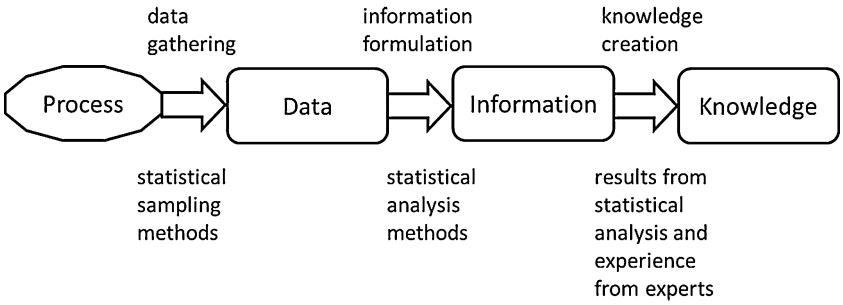

Figure 2-10: Knowledge creation process, first suggestion

Figure 2-10 shows the basic idea of the knowledge creation process. It is a straightforward generalization of previously mentioned DIKW-hierarchy. It has similarities with all previously mentioned knowledge hierarchies, like Eliot (1934), Cleveland (1982), Zeleny (1987) and Lillrank and Forsén (1998). At the lowest level there is a process where the data is gathered. Data is then analyzed to get the information, which is finally transformed into knowledge using statistical analysis results and experience.

This previously introduced first suggestion (**Figure 2-10**) for the knowledge creation process is not however sufficient to create a holistic view. In this approach, there is no conceptualization for the problem of how data is created. Usually data needs to be collected or measured. This might be described as a classical approach, because it has some similarities with decision making using classical statistics (Chapter 2.4.1). This classical approach is a straightforward generalization on DIKW-hierarchy in the industrial knowledge creation. In this approach, data is the sole source of information, which is finally transformed into knowledge.

Next the second iteration (**Figure 2-11**) for knowledge creation process is proposed. It has similarities with the previously described knowledge hierarchy based process (**Figure 2-10**). This suggestion is based on researcher's understanding that the traditional knowledge hierarchy is not sufficient – more is needed. In many real situations in industry, there are experts in many fields involved in knowledge creation process. As an example, in a Six Sigma (Chapter 4.4.1) project, the project group is formed in a way that there are process operators, process specialists and statistical specialists (Six Sigma specialists) working together to create knowledge. This means that the previously introduced classical approach (**Figure 2-10**) needs to be expanded. Data is not the only source of information: experience and expertise are also required, because they are used to create knowledge when information is formulated based on statistical analysis. In this approach, data is created based on the knowledge – how to measure something, for instance. This could be called Bayesian type of approach (**Figure 2-11**) and it is more suitable for knowledge creation. This approach has similarities with the reverse hierarchy of Tuomi (2000): data appears only after knowledge and information are present – knowledge is required for data creation.

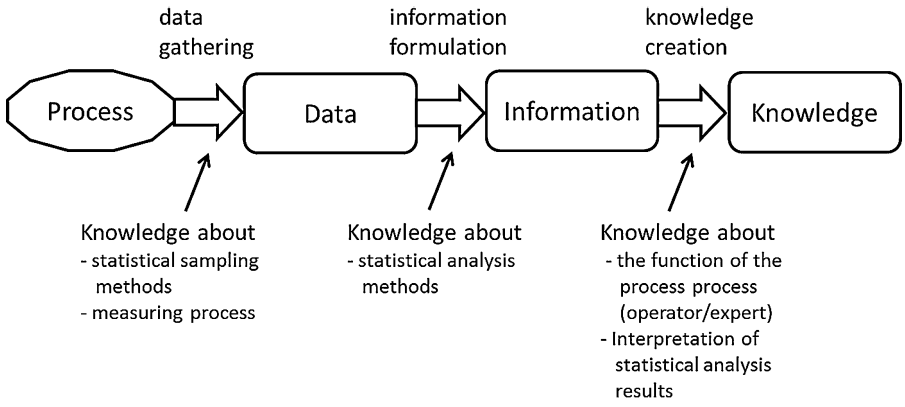

Figure 2-11: Knowledge creation process, second iteration

Bayesian type knowledge creation (**Figure 2-11**) starts with the data, which is gathered from some process. Already at this stage knowledge, regarding how to measure something, is mandatory. Statistical knowledge about how to create suitable a sampling plan, or a survey is required as well. As shown in (**Figure 2-11**) data consists of information. Next, the statistical knowledge is needed for trans-

forming the data into information. This statistical knowledge is formulated during the education of the statistician in the university and has similarities with previously mentioned (Chapter 2.1.2) definition of Ackoff (1989) that data could be transformed into information using statistical or mathematical analysis methods. The results for this stage are different statistics, such as χ^2-values, odds ratios and statistical significances.

Finally (**Figure 2-11**) the symbiosis of information and an expert's knowledge about the examined process gives form to the knowledge. In this stage, actual learning is not required. Learning has happened earlier, for example during education, when knowledge about statistics or knowledge about operating a process is achieved. In this dissertation, the second iteration of knowledge creation follows the Lee and Yang (2000) conceptualization. They argue that knowledge is created for example through interpretation (Chapter 2.1.3). Interpretation refers to the process of translating the results into everyday language; for instance, what do the results mean in the context of the examined process and how are significant factors interpreted. For example, by adjusting a particular factor it is possible to control the function of the analyzed process. The use of knowledge of statistical methods and experience from-data-to-information and from-information-to-knowledge phases could be considered to be a kind of externalization. Referring to Nonaka and Takeuchi's conceptualization, the tacit knowledge is used to create explicit knowledge.

This kind of knowledge creation has similarities with engineering projects. In engineering projects, there is knowledge that is not written in documents. This knowledge is only realized through the expertise and understanding of the project personnel. (Koskinen et al., 2003) Similarly, knowledge creation in **Figure 2-11** is based on the understanding and expertize of the from-data-to-information and from-information-to-knowledge phases of this process.

Expanding on these previously mentioned concepts of knowledge creation, a final process of industrial knowledge creation process based on statistics is proposed. For this purpose, the previously described organizational knowledge creation spiral Figure 2-5 and Bayesian type knowledge creation process (**Figure 2-11**) are needed, although augmented with the organizational knowledge creation cycle proposed by Bhatt (2000) (**Figure 2-12**).

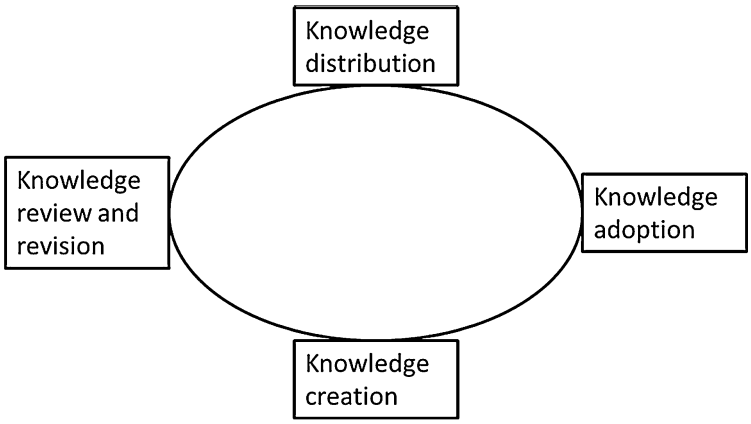

Figure 2-12: Organizational knowledge creation process (Bhatt, 2000), adopted

The organizational knowledge creation process introduced by Bhatt (2000) (**Figure 2-12**) shares similarities with the ideas of Nonaka and his co-workers; namely that the individual knowledge moves up from the individual level to the group level and the organizational level. In this knowledge creation process, knowledge is created through understanding and interpreting within an organizational context – an organization generates knowledge through its individuals. In knowledge adoption, a company adopts knowledge from other sources and uses it, which is usually considered to be a strategy that saves costs. This is useful in situations where a firm can create knowledge based on other projects etc. Because knowledge is a key organizational resource, it has to be distributed and shared through the organization. Finally, in the knowledge review and revision stage, knowledge needs to be taken into use - if not, it can easily be ignored and forgotten. (Bhatt, 2000)

The proposed industrial knowledge creation process is based on ideas expressed in Nonaka's and Takeuchi's (1995) organizational knowledge creation spiral (**Figure 2-5**), Bayesian type knowledge creation (**Figure 2-11**) and Bhatt's (2000) organizational knowledge creation process (**Figure 2-12**) and synthesized by the author. In this model (**Figure 2-13**), the way that the statistically created knowledge is transferred from the individual level to the organizational level is described.

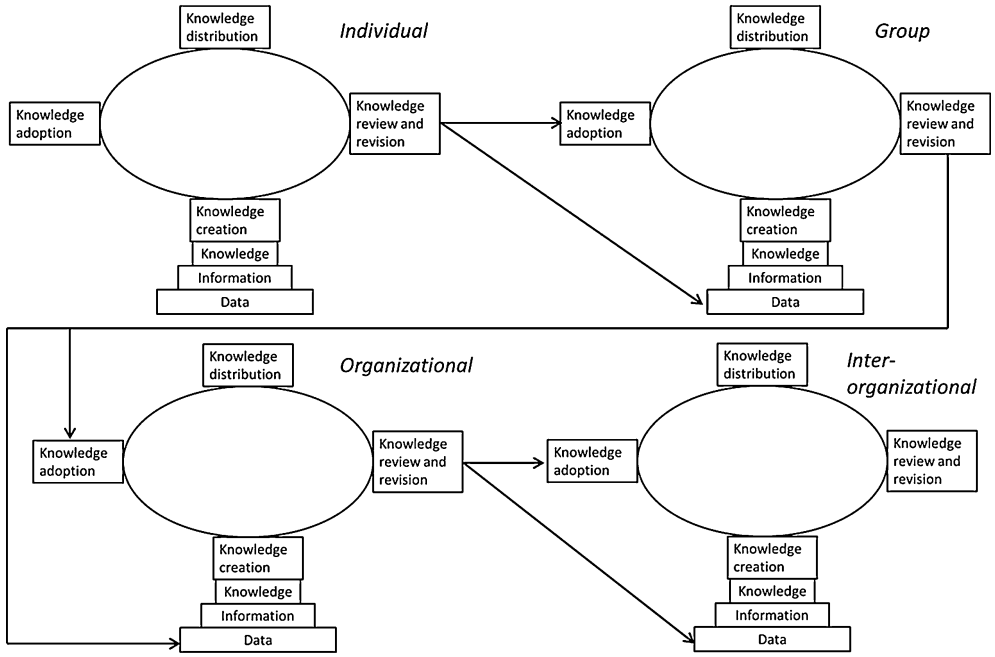

Figure 2-13: Statistical analysis based industrial knowledge creation process

The statistical analysis based industrial knowledge creation process (**Figure 2-13**) starts with Bayesian type knowledge creation (**Figure 2-11**). In the knowledge adoption stage previously created knowledge is adopted. It means that previously created knowledge is understood in the context of each examined situation (for example process. After this phase, the knowledge is distributed. This refers to a situation where a statistician and a project specialist work together and share their knowledge to a wider project group. After that, the project group reviews and revises this knowledge before it moves on to the group level, for instance manufacturing. At this stage, the knowledge could be used as a source of new data. For instance, we now know that "something happens", so it is necessary to gather data that can be transformed into knowledge about "why it happens" (back from the group level to the individual level). It is also possible to adopt this "something happens" type of knowledge to other similar processes in the company.

As an example, after the group level the "why it happens" type of knowledge moves up to the organizational level. At this stage, knowledge can be used as a source of new data. We now know what kind of data is required to generate, for

example, "why it is happening in all our processes" kind of knowledge. It is also possible to adopt knowledge from other similar processes.

The inter-organizational step of this process may be achieved in situations where a company publishes or presents its knowledge to a wider audience, rather than just inside the company itself. Typical examples are, for instance, situations where knowledge is presented in an academic journal or at a scientific conference. These journals and conferences are important because based on them; a company can adopt new knowledge or start creating new knowledge based on other companies' findings.

This industrial knowledge creation process (**Figure 2-13**) was followed when the included publications were written. For example in publication IV, the knowledge was first generated using the Bayesian type knowledge creation process (**Figure 2-11**), after which the knowledge was distributed when an employee told his supervisor about our findings. In the next phase, the knowledge that there is no need for sample collection was reviewed and revised and then applied when collecting data to generate knowledge about why something occurred. Hopefully this knowledge reached the organizational level and was thus adopted across the entire organization. The inter-organizational level was achieved when the work was published, allowing other companies to adopt the information presented in the publication. According to Lee and Yang (2000), knowledge is generated when information is read, interpreted or understood.

3. Data in Industry

The amount of data in industry is expanding and simultaneously the data processing capacity of computers is increasing as well. This data needs to be transformed into knowledge. An appropriate method for analyzing data is the use of statistics. The use of a suitable statistical method depends on the type of data.

This chapter concentrates on one possible categorization of data in industry. This categorization will differ from the traditionally used continuous/discrete or qualitative/quantitative classifications. This chapter aims to categorize data based on how it is gathered: process data, survey data or big data. There are other possibilities to collect data such as automatic optical inspection (AOI). According to own expressions AOI is widely used in electronics industry. Based on the aforementioned, another goal of this chapter is to present some suggestions about suitable statistical analysis methods based on this categorization.

3.1. Data as a Source of Knowledge in Industry

As described earlier, data and specialist knowledge are the sources of information and further knowledge, during the process where statistical analyses are done and

the results are interpreted. It is possible to divide data into qualitative data and quantitative data (**Figure 3-1**). Qualitative or nonmetric data could be defined as categorical properties for instance occupation (physician, attorney, professor). (Hair et al., 1998)

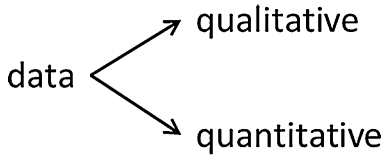

Figure 3-1: Qualitative and quantitative data

Psychologist S. S. Stevens (1946) created a well-known classification of measurement data. He classified measurement data into four different types of scales (**Figure 3-2**). Those scales are: nominal scale, ordinal scale, interval scale and ratio scale. Stevens (1946) also proposed that the analysis method for data depends on these scales. (Gaito, 1980; Khurshid & Sahai, 1993; Velleman & Wilkinson, 1993).

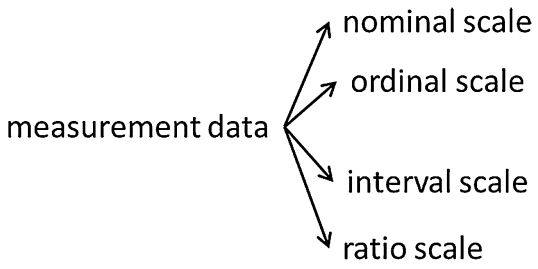

Figure 3-2: Scales of measurement data (Stevens, 1946), adopted.

Measurement data that lack natural ordering are called nominal scale variables. This kind of data could be, for example, a person's favorite type of music (classical, country, folk, jazz and rock) or type of residence (apartment, condominium, house and other) (Agresti, 2002; Agresti, 2010) According to Stevens (1946) nominal scale data is in a "primitive form" and it can be divided into two subtypes. These types could be characterized as follows: a) the numbers used to identify football players and b) the numbers used to denote a class, where each member of the class gets the same number. The only statistics that is relevant for nominal data is the number of cases, which in this case would mean counting the players. Type b,

nominal scale data, can be used to calculate the mode (the most commonly occurring value in the data). When using a nominal scale of measurement, the resulting categories of variables differ only by name and only tell the name of the category. (Stevens, 1946; Khurshid & Sahai, 1993).

Measurement data which have ordered categories are called ordinal scale variables. An ordinal type of data could, for example, describe the size of an automobile (subcompact, compact, midsize or large), social class (worker or manager), patient condition (good, fair, serious or critical). There might also be a scale where the middle category is neutral. That kind of a scale is usually called a Likert scale. It is not possible to use mean and standard deviation with ordinal data. These are faulty statistics, because the differences between the successive values on these scales are not of an equal in size. For example percentile measures may be applicable or mode or median as a measure of central tendency. (Agresti, 2002; Stevens, 1946)

The next level in measurement data classification is the interval scale. Common statistics such as mean and standard deviation are applicable to an interval scale of data. Typical example of an interval scale type of data is temperature measured in Centigrades and Fahrenheits, because in both there is an arbitrary zero, not absolute. In contrast to the ordinal scale, the differences in interval rankings are stated in a way that has the same absolute value over the whole range of observations. Based on this it is possible to state how much the former is greater than the latter. (Stevens, 1946; Siegel, 1957; Khurshid & Sahai, 1993)

The highest level of measurement data is the ratio scale. In ratio scales there is always an absolute zero involved, like in the Kelvin temperature scale. In ratio scale type of data there are quantities: equality, rank-order, equality of intervals and equality of ratios. All kinds of statistics could be applied to ratio scale data. (Stevens, 1946)

Stevens (1946) presented ideas about the way that suitable statistical analysis methods could be chosen based on the type of scale used in the measurement data. As explained earlier, for instance mean is not a suitable measure of central tendency or standard deviation is not a proper measure of dispersion when working with ordinal data. This kind of procedure is criticized by many scholars. For instance Lord (1953) claimed that measurement issues should be understood independently from statistical issues and Velleman and Wilkinson (1993) also argued that this

does not work because real data does not follow the requirements of data scales. They also argued as well that good data analyses are impossible if the data is asserted to have a scale type that outlaws something. It is not a good approach to select the probable hypotheses or the methods that are applicable based on the scale type. Khurshid and Sahai (1993) studied several scholars who have criticized Stevens (1946) classification and concluded that the proper statistical analysis method depends on the population distribution rather than Stevens' scales.

As described earlier, all methods should be considered when analyzing data, but it is important to understand the assumptions for the used method. This is important because it will have an effect on interpretations. It is possible to use an analysis method that requires continuous, normally distributed data when dealing with discreet data. The problem might occur during the interpretation of this kind of analysis results. For example, using mean as a measure of central tendency in Likert scale data does not provide correct information – a suitable measure of central tendency in this case.

3.2. Process Data

There are many types of processes. Some examples of processes that are not producing, distributing or selling anything are managerial processes like budgeting, business planning and process reporting. In addition, there are manufacturing processes, which are easy to see and follow. Inputs of the process might be "inherent process inputs" like raw materials, "controlled variables" like process temperature or "uncontrolled noise variables" like raw material lots. (Breyfogle, 2003; Breyfogle, 2008) The typical process functioning could be characterized with three terms input-process-output (**Figure 3-3**).

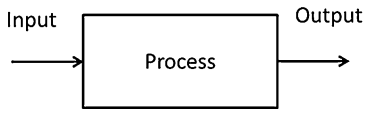

Figure 3-3: Input-process-output (Breyfogle, 2003; Breyfogle, 2008), adopted.

Usually there are consecutive steps in processes. These steps can be modeled using tools like SIPOC-chart (Suppliers, Input, Process, Output and Customer) (**Figure 3-4**).

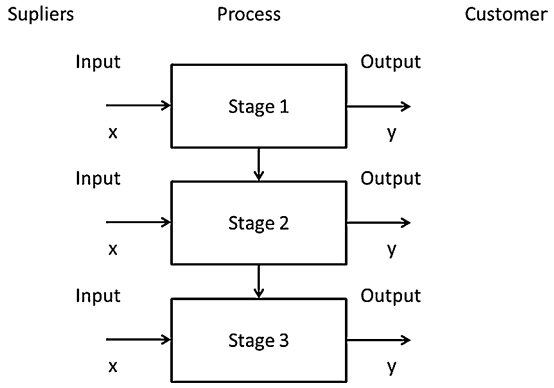

Figure 3-4: SIPOC (Breyfogle, 2003), adopted.

Examples of the stages in **Figure 3-4** could be drilling, cutting and painting in a manufacturing process. More examples of graphical process modeling could be found in Breyfogle (2003). Process data, like time spent, could be collected, measured or gathered from each process step. This fits the earlier definition of data used in this dissertation (Chapter 2.2). Data characterizes the function of the process. Traditional methods of industrial statistics like SPC (Chapter 4.3.2) could be used to monitor the stability of the examined process and DoE (Chapter 4.3.1) methods are applicable when analyzing which factors have a significant effect on the process output. (Breyfogle, 2003; Box et al., 2005; Montgomery, 2008)

3.3. Survey Data

Surveys are traditionally widely used in social sciences and IT. Different types of surveys might be censuses of population, public opinion polls, market research studies, academic studies of prejudice and epidemiological studies. Survey is a research method where systematic observation or interviewing is used to describe a natural population. Three main objectives for surveys are: description, explanation and exploration. (Babbie, 1973; Sapsford, 2006)

Because surveys are tools for sampling they have a close relationship to different sampling methods. Typical sampling methods are: simple random sampling (with or without return), systematic sampling, stratified sampling, cluster sampling and sample by size. (Babbie, 1973)

Basic types of surveys are cross-sectional surveys and longitudinal surveys. In a cross-sectional type of survey, data is collected at a certain time. This sample will characterize the population at that time. In longitudinal surveys, data is collected over time. In this type of studies it is possible to report changes in descriptions and explanations. There are three main types of longitudinal designs: trend studies, cohort studies and panel studies. In trend studies the general population will be sampled and studied at different points in time and the samples consists of different persons. While trend studies are based on the general population, cohort studies focus on same, specific population each time that data is collected. It is still possible that the samples studied in cohort studies might be different. In panel studies data is collected over time from the same sample of respondents. (Babbie, 1973)

The questions in a survey can be classified in different ways: they can be either direct or indirect. Direct questions could be formulated like "how many children you have" or "what is your age". When dealing with opinions, attitudes or beliefs it is recommended to use less direct questions. An important part of surveys is measurement scales. In some questions there is no point in asking about someone's level of "clinical depression" using direct questions. Possible formats for this kind of questions are Visual Analogue Scaling or Likert scales. Visual Analogue Scale is a continuous line between two adjectives and respondents are asked to mark on the line where their opinion is. Likert scales are for example scales from 1 to 5 or strongly disagree to strongly agree.

Likert scales are explained more thoroughly here because a Likert scale based questionnaire is used for data collection in one of the publications included in this dissertation (Publication I). Likert scales are commonly used as a format for surveys. In these questionnaires, respondents are asked to rank quality from high to low or best to worst using five or seven levels. Five-grade and seven-grade Likert scales are the most popular. These grades, scores or degrees are organized in an ascending order of agreement or approval of the individual with respect to the val-

ue statement. Likert scales were developed in 1932 by Rensis Likert. (Allen & Seaman, 2007; Göb et al., 2007; Clason & Dormody, 1994)

Likert scales produce ordinal data. This means that for example mean and standard deviation cannot be calculated, because they are inappropriate for ordinal data. So instead of them, median or mode should be used as a measure of central tendency. In many cases, these differences are not understood and it has been a common practice to consider Likert scale data as interval-level measurements. The problem here is the fact that it cannot be assumed that the difference between the feelings "strongly disagree" and "disagree" is similar to the differences between the other consecutive categories on the Likert scale. (Jamieson, 2004)

Because of the previously mentioned facts, analysis methods suitable for continuous data cannot be used with Likert scales. For example, the commonly applied multivariate methods Factor Analysis (FA) and Principal Component Analysis (PCA) could not be used. Instead, the counterparts suitable for categorical data analysis need to be used. For instance Korhonen and Siljamäki (1998) have examined PCA suitable for ordinal multivariate data and Meulman et al. (2004) and Linting et al (2007) have studied nonlinear PCA in their articles and for instance Batholomew (1980) has written about FA for categorical data.

3.4. Big Data

Nowadays there are different databanks consisting of large dataset that have not been gathered for analysis. Because the data is not collected specially for analysis purposes, it is not sampled from a pre-defined population and might be insufficient to meet analysis requirements. This kind of data might be collected, for example, on customers and their behavior. Another type of a large dataset is bibliometrics and especially evaluative bibliometrics. These datasets contain data such as the number of publications, citations or patents. Bibliometric analysis is based on the quantitative and statistical analysis of publications. (Narin et al., 1994; Feelders et al., 2000; Georghiou et al., 2008)

When describing the structure of large datasets, the co-occurrences of terms, authors, references or institutions can be seen as examples of bibliometric data. Kess-

ler (1963a; 1963b) and Small (1973) where two of the earliest academics in this field. They started to examine the connections between two scientific papers. Leydesdorf and Vaughan (2006) argued that co-occurrence matrices like co-citation and co-word provide useful information about the underlying document sets. These co-occurrence matrices might be symmetrical or asymmetrical. A symmetrical co-citation matrix is symmetrical in relation to its diagonal. The same objects appear in both the rows and columns and there are same amount of rows and columns. Asymmetric co-citation matrix does not have these properties.

Data mining methods are used to efficiently obtain summaries and to identify interesting structures and relationships within large datasets such as the previously mentioned consumer databanks and bibliometrics. More generally these data mining methods are used to extract knowledge from these datasets. Data mining is a field at the intersection of statistics, machine learning, database management and data visualization. Nevertheless, data mining is usually considered to be a part of the information sciences and there has been some debate about the position of data mining amongst scientific disciplines. A characterization of the differences between statistics and data mining is presented in **Table 3-1**. (Friedman, 1997; Hosking et al., 1997; Feelders et al., 2000)

Statisticians' issues	Data miners' issues
Model specification	Accuracy
Parameter estimation	Generalizability
Diagnostic checks	Model complexcity
Model comparison	Computational complexity
Asymptotics	Speed of computation

Table 3-1: Statisticians' and data miners' issues in data analysis (Hosking et al., 1997), adopted.

As seen from **Table 3-1**, both statistics and data mining are concerned with drawing inferences from data, but the major difference between these two is the amount of the analyzed data. In data mining, there are extremely large datasets compared to the more controlled volumes of data in statistics. Statisticians are used to working with models where the speed of computation is not an issue, while the speed of computation is important to data miners. (Hosking et al., 1997). Both statistics and data mining are focused on learning from data and transforming data into information (Glymour et al., 1997).

There are many valuable insights that statisticians and data miners could gain from each other. It would be beneficial for a data miner to study the problems that an outlier might cause in data analysis. It would also be important for a data miner to understand the use of diagnostics in model accuracy checking. A statistician might underestimate the significance of the asymptotic accuracy of the model based estimates. For example, it is possible to use statistical models like cluster analysis, discriminant analysis and nonparametric regression to explain relationships within datasets to make predictions. (Hosking et al., 1997)

4. Industrial Applications of Statistics

> *"Statistics is a science of collection, analysis, and presentation of data. Statisticians contribute to scientific enquiry by applying their knowledge to the design of surveys and experiments; the collection, processing, and analysis of data; and the interpretation of the results."*[12]

This definition of statistics describes an important fact about statistics and statisticians. Statistics mostly deals with data. The knowledge of the statistician concerns gathering data for an analysis and then analyzing this data. Finally we need to interpret the results into an everyday language, for example what do statistical significances (p-values) mean.

[12] Definition of Statistics from the webpage of American Statistical Association, accessed on 09.03.2012 (http://www.amstat.org/careers/whatisstatistics.cfm)

In statistics two sciences – mathematics and philosophy – are acting together. Statisticians are dealing with the greatest philosophical challenge in science: how information can be translated into knowledge. Statistics has the tools to translate data into decisions, to recognize changes and to tell when changes are needed. (Senn, 2003; Smith, 2001)

Statistics has a very wide area of applicability. It is possible to apply statistics into biostatistics, including medical science. As an example, it is possible to analyze the effectiveness of a new drug or treatment compared to the present one. In economics statistical methods like time series modeling[13] and econometrics[14] are widely applied. In physics, probability theory could be used to solve physical problems. In social sciences, for instance psychology[15], statistical methods such as multidimensional scaling (MDS) could be used. In actuarial sciences, like insurance, statistical methods are used in different kinds of risk evaluations, for example. In engineering sciences, such as manufacturing, industrial statistics plays an important role. The list of the areas of science where statistics is applicable is almost endless. This is based on the definition of statistics at the beginning of this chapter: statistics is a science for collecting, analyzing and presenting data. As a generalization, when data is present, statistics is applicable.

4.1. Statistical Methods in Industry

The roots of industrial statistics are in the beginning of the 20th century. These methods were developed because statisticians were dealing with real, not theoretical problems. Most of these problems were from industry, the chemical and physical sciences and from engineering sciences. As an example Dr. Walter A. Shewhart

[13] The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2003 was divided equally between Robert F. Engle III *"for methods of analyzing economic time series with time-varying volatility (ARCH)"* and Clive W.J. Granger *"for methods of analyzing economic time series with common trends (cointegration)"*.
http://www.nobelprize.org/nobel_prizes/economics/laureates/2003/ accessed on 09.03.2012

developed Statistical Process Control (SPC, Chapter 4.3.2) when dealing with telephone manufacturing. (Montgomery, 2000; Montgomery, 2001)

4.2. Industrial Statistics

Industrial statistics is one area of the previously mentioned applied statistics. The term "industrial statistics" might be misleading, because in many cases it might refer to some statistical notes on some area of industry – for example the amount of sales in a particular year. Industrial statistics cannot be thought of as a branch of mathematics. The relationship between mathematics and industrial statistics is more complex. The way that statistics is applied is rather far from mathematical reasoning. Here, the objectives of the reasoning are inferences about real systems. There is somewhat similar relationship between mathematics and physics in addition to mathematics and industrial statistics. Both, physics and industrial statistics use mathematics as an analysis tool. (De Mast & Does, 2006)

The term *industrial statistics* and the greatest challenges within this discipline were analyzed using a questionnaire sent to experts of this field. These experts were chosen from publications in this field as well as based on the author's own impressions. The questionnaire was sent to 26 experts, 7 of whom answered during January and February 2012. Based on expert opinion, the simplest way to determine industrial statistics is the following: industrial statistics is statistics used in industry. There are no fundamental differences between industrial statistics and other areas where statistics is applied.

> *"Industrial Statistics is Statistics used in industry. What else?"*[16]

> *"Industrial Statistics is the field of applying statistical thinking and statistical methods toward the ultimate success of the industry served."*[17]

[16] Anonymous respondent "Expert opinion study on Industrial Statistics"
[17] Anonymous respondent "Expert opinion study on Industrial Statistics"

Here, the term industry primarily involves engineering as opposed to healthcare, government finance and other industries. If pharmaceutical companies are considered a part of industry, then many biostatistical methods become relevant. As one of the respondents says

> *"If one includes pharmaceutical companies as industry then many biostatistical methods are included."*[18]

The role of statistics and industrial statisticians in industry is considered to be important among the industrial statisticians. There has been a lot of discussion in the United States about industrial statistics, industrial statisticians and their role in industry. As an example in 2008 a panel discussion[19] was organized where industrial statistician discussed about their discipline. A respondent of our own survey describes the role of an industrial statistician as follows.

> *"…successful industrial statistics requires the statistician to take leadership role to help address the organization's major issues. In this role he/she must promote statistical thinking throughout the organization."*[20]

Conventionally industrial statistics has included DoE, (Chapter 4.3.1), SPC (Chapter 4.3.2), Capability Analysis (Chapter 4.3.3), Reliability (Chapter 4.3.4) and Acceptance Sampling (Chapter 4.3.5).[21] In the next chapter, the traditional methods of industrial statistics are introduced, since this dissertation uses the conventional definition for industrial statistics. Industrial statistics is statistical methods and thinking used in industry toward the success of the industry served.

[18] Anonymous respondent "Expert opinion study on Industrial Statistics"
[19] See Steinberg (2008)
[20] Anonymous respondent "Expert opinion study on Industrial Statistics"
[21] Anonymous respondents "Expert opinion study on Industrial Statistics"

4.3. Traditional Methods of Industrial Statistics

In the following chapters the traditional methods of industrial statistics: DoE (Chapter 4.3.1), SPC (Chapter 4.3.2), Capability Analysis (Chapter 4.3.3), Reliability (Chapter 4.3.4) and Acceptance Sampling (Chapter 4.3.5) are introduced. This is done to give a comprehensive view for the broad applicability of statistics in industry. DoE and SPC are more thoroughly explained, because those two are the most effective tools to create information about how the examined process is functioning and those methods could be used to find an answer to the following questions: Is the process under control? What are the significant factors that affect on the output of the process?

4.3.1. Design of Experiment

The creation of the new knowledge can be expensive, take time and pose many challenges. Knowledge is power and a fundamental when dealing with innovation and profit. By using statistical methods and especially DoE[22] it is possible to increase the efficiency of knowledge creation. (Box et al., 2005)

The history of statistical experimental design could be divided into four eras. The first era, or the agricultural era started in the 1920s and early 1930s with the pioneering work of Sir Ronald A. Fisher. Those principles are randomization, replication and blocking as well as experimental methods like factorial designs and the analysis of variance (ANOVA). (Montgomery, 2008)

The second or industrial era started with the development of response surface methodology (RSM) in 1951 by Box and Wilson (1951). They found out that many industrial experiments were fundamentally different from their agricultural counterparts. Two main differences were 1) the response variable can usually be observed (nearly) immediately and 2) the experimenter can quickly learn the crucial information from a small number of runs that can be used to plan the next experiment. (Montgomery, 2008)

The third era of statistical experiments started in the late 1970s. There was increasing interest in the Western industry towards quality improvement, which started

[22] Design of experiment is a part of DMAIC road map but is developed much earlier

this era. In this era Taguchi methods were introduced: the statistical design of controlled variables and statistical design about noise variables and crossing these two. These Taguchi methods have been criticized because there are some important problems with the experimental strategy and the data analysis methods. Despite the criticism Taguchi's work has had some positive impact, such as the wider use of designed experiments. (Box et al., 2005; Montgomery, 2008)

In the fourth era, general interest towards DoE methods increased among researchers and practitioners and many new and useful approaches to experimental problems in the industrial world were developed. The education about these methods also became a part of engineering programs in many universities. (Montgomery, 2008)

There are several DoE methods, as Box et al. (2005) and Montgomery (2008) describe: factorial designs, randomized blocks designs, latin squares, RSM methods, Taguchi methods and split plot designs. Though this area of industrial statistics is worthy of many studies and dissertations, here we focus on statistically designed 2^k factorial designs as an example of these methods. Theoretical facts concerning this and other DoE methods could be studied, for instance, from Box et al. (2005) and Montgomery (2008)

Randomization, replication and blocking are the basic principles of experimental designs. The use of statistical methods in designed experiments is based on randomization. Randomization means that the runs in an experiment are made in a random order. The use of statistical methods requires that observations are independently distributed variables. Randomization usually makes this valid. Replication refers to repeating the factor combinations. Blocking is usually used to handle factors that might influence the experiment, but which the researchers are not directly interested in. (Montgomery, 2008)

In many experiments there are more than two factors and the goal is to study the effect of those factors. In these kinds of situations the use of factorial designs is the most appropriate option. In factorial designs all possible combinations of levels of factors are examined. In these kinds of experiments it is possible to analyze the main effects and interactions of the examined factors. (Montgomery, 2008) The following figures (**Figure 4-1** and **Figure 4-2**) give an illustration of the situation, were there are two factors and those factors have two levels.

Factor	
A	B
1	-1
-1	-1
1	1
-1	1

Figure 4-1: 2^k factorial design for two factors (Montgomery, 2008), adopted.

The **Figure 4-1** shows the illustration of a 2^k factorial design for two factors (k=2) and two levels per factor. In this figure, all of the combinations of these two factors and their levels are shown. In this design there are no center points, which would be coded as "0" and are checking for the linearity of the relationship between factor A and B. When the amount of factors increases it will cause an increase in the amount of runs (5 factors =>2^5 runs).

Randomization could be seen from the **Figure 4-1** as well. The combinations or runs are in a random order. Replication in this example would mean that all four runs are done again. Blocking can be applied if, for instance, the experiment is done on Monday and Tuesday. In this, the blocking variable will record the weekday. (Box et al., 2005; Montgomery, 2008)

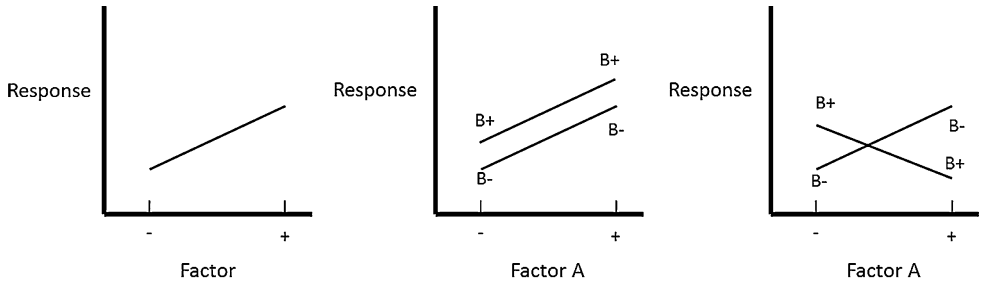

Figure 4-2: Graphical Illustration of main effect (left) and without integration (in the middle) and with interaction (right) (Montgomery, 2008), adopted.

The **Figure 4-2** graphically shows the main effects and interactions in 2^k factorial experiments. Significant interaction means that the effect of A is different in each level of B. Interactions are the most important reasons that statistically designed experiments should be used. When dealing with more than one factor, interactions

must be examined. One-factor-at-the-time (OFAT) or one-variable-at-the-time (OVAT) methods do not handle interactions. Analyzing the interactions produces important information. (Box, 1989)

4.3.2. Statistical Process Control

Dr. Walter A. Shewhart (1930) developed the original concept of SPC in the 1920's. A few years later he worked at Bell Laboratories and introduced the concept of the univariate control chart. (Montgomery, 2009) During the Second World War, SPC techniques were used in the military industry. The results in the war time industry were excellent because SPC made it possible to reduce the amount of reworking and faulty products. After the war, the use of SPC diminished in western countries because energy and materials were cheap and there was high demand for products on markets. However Dr. Deming's ideas, such as the use of SPC had a great impact on Japan in the 1950's when he was lecturing there. Western countries adopted SPC methods again in the 1970's. (Wheeler, 1995; Wheeler, 1992)

It is possible to divide this area into two groups, these groups are univariate control charts and multivariate control charts. The univariate control charts are not thoroughly explained, because facts about univariate charts can easily be studied from different authors like Breyfogle (2003), Montgomery (2009), Ryan (1989) and Wheeler (1992). Ideas of multivariate SPC are more thoroughly introduced than univariate charts, because many industrial processes are multivariate and the methods are not as widely applied. Using several univariate control charts in these kinds of circumstances might cause situations where special causes are not noticed. A more detailed view on multivariate SPC can be found from several scholars like Ryan (1989), Stoumbos et al. (2000) and Bersimis et al. (2007) .

All processes have variation. This variation could be divided into two types. These types of variation are common cause variation and special cause variation. These two types of variation need to be separated, because they have different effects on the output of the process. The tool for separation are Dr. Shewharts SPC methods. There are different charts for attribute type of data and measurement type of data. Attribute data could be for example the amount of defects in a product or defective products. (Ryan, 1989; Wheeler, 1992)

Some examples of charts that are suitable for attribute data are the p-chart and c-chart. When using a p-chart, data is assumed to be binomially distributed. The p-chart is suitable for situations where nonconforming units in each sample are studied. It is also possible to use the p-chart for plotting nonconforming units per sample if the sample size is constant. In the case of a c-chart, data is assumed to be Poisson distributed. The c-chart is suitable for controlling the number of nonconformities per each sampling unit. Charts that are suitable for measurement data are for instance the R-chart, s-chart, Xbar-chart, CUSUM-chart and EWMA-chart. It is assumed that when using charts for measurement data, the distribution must be normal. (Montgomery, 2009)

R-chart or Range-chart is constructed for controlling of the process variation and data is assumed to be normally distributed. Standard deviation or s-chart is also used for monitoring the process variation and the data is assumed to be normally distributed. The use of the s-chart instead of the R-chart is recommended due to the way it is calculated[23]. The Xbar-chart is used to control the process mean and the distribution is assumed to be normal. Usually Xbar and s as well as Xbar and R charts are done as a pair (Xbar-s and Xbar-R). (Ryan, 1989)

In multivariate industrial processes it is possible that there are correlated variables. In this kind of a situation, the process variables are not independent of each other. This means that there is a multivariate property that defines the quality of the product – several variables define it together. This may cause a situation where a change in one process variable will cause a ripple effect to other variables. (Mason & Young, 2004)

The Hotelling T^2 chart is based on the Hotelling T^2 statistics, which Harold Hotelling introduced in 1947 (Lowry & Montgomery, 1995). The idea in this chart is a formula that is capable of combining different types of measurements into one single measure of excellence. This kind of a measure is the generalization of Student t. This generalization, T was introduced by Harold Hotelling in 1931. The T^2 statistic calculates generalized distance, which in this situation is the distance between p-dimensional sample points to the mean vector. In this situation there is the

[23] R=max-min

$$s = \sqrt{\frac{\sum_1^5 (y_i - \bar{y})}{n-1}}$$

same assumption than in traditional Student t test: the data has to be normally distributed and here the data has to be multivariate normally distributed. The ideal value of a T^2 chart is zero, which also is the smallest possible value. T^2 reflects distance and it is not possible to have negative distance. Zero also means that an observation is located at the center of the process. Small T^2 values are good and large T^2 values signal that the process is not functioning as it should. (Hotelling, 1947; Murphy, 1987; Mason et al., 1997; Mason & Young, 2000)

The MEWMA-chart was introduced by Lowry, Woodall, Champ and Rigdon in 1992 (Rigdon, 1995). It is the multivariate counterpart of the univariate EWMA-chart. Here the examined data has to follow multivariate normal distribution as well. This chart has similarities with the Hotelling T^2 chart, for instance the examined data has to be multivariate normally distributed. The ideal value of the chart is zero, because this chart analyzes the distance from the sample point to the mean vector. There is one difference between the MEWMA-chart and the Hotelling T^2 chart – MEWMA chart is more sensitive to small changes. (Lowry et al., 1992; Stoumbos et al., 2000; Bersimis et al., 2007)

4.3.3. Capability Analysis

A process Capability Analysis can only be performed when the process is under statistical control. A process is in statistical control when it is not affected by special causes and process is predictable. It is possible to use SPC charts for process capability and performance checking, but the actual Capability Analysis is based on calculating and the interpretation of capability indexes. The basic process capability indexes are C_p, , C_{pm} and C_{pk} . These indexes could be interpreted based on some cutoff values. This means that when the indexes are greater than the cutoff values, process could be stated to be capable or world class. For instance many companies use $C_{pk}>1.33$ as a definition of a capable process. (Pyzdek, 1999; Kotz & Johnson, 2002; Breyfogle, 2003; Montgomery, 2009)

However, it is possible that a process is under statistical control but still has poor capability. In this situation, the mean of the process is not correctly in target and data is not centered between control limits. (Pyzdek, 1999; Breyfogle, 2003) Spiring et al. (2003) have done a bibliographical analysis of process capability papers between 1990-2002. This reviews the great deal of sources for this specific area.

4.3.4. Reliability

The scope of Survival Analysis is on failure times of group or groups of individuals. Sometimes the interest is in the distribution of failure times in a single group. More often the failure times of two or more groups are compared, whether the failure times in one group are systematically longer than in another group (new drug/old drug). Generally survival data measures time to some event. Typically, in medical research, this event is death. In industrial applications, this is usually time to failure or in economics this might be time to acceptance of a job offer. In many other cases the event is a transition from one state to another. (Cox & Oakes, 1984; Hougaard, 1999)

Survival Analysis is not part of the standard statistical analysis procedures. There are some reasons for this. Firstly, this kind of data is typically not symmetrically distributed, and is (usually positively) skewed. Because of this, it is not possible to assume that data is normally distributed. Secondly, survival data might be censored. The survival time is said to be censored, when the end point of interest is not observed. (Collett, 2003; Clark et al., 2003; Singh & Mukhopadhyay, 2011)

4.3.5. Acceptance Sampling

Acceptance Sampling is a quality control technique. It could be applied to discreet lots of products or batches. This method was once the main quality control procedure. Nowadays it is understood that Acceptance Sampling should never be used for processes in quality control, because it is detection methodology. Quality control methods should be preventative methods such as SPC. According to the results of the inspection of a sample, it is possible to accept or reject the lot or batch. There are three main elements in Acceptance Sampling. The first one is the sampling plan, which refers to the number of units to be inspected and to the acceptance criteria. The second is the action taken with the current lot or batch (sort, scarp, return to vendor, etc.). The third one is actions taken for the future such as switching to reduced or tightened sampling, 100% sampling or maybe shutting down the process. (Pyzdek, 1999)

Acceptance Sampling is mainly operated by different types of sampling plans. There are specific sampling plans for attributes (quality characteristics measured in go, no-go basis) and variables (quality characteristics measured in numerical scale).

There are standards[24] that are created to assist creating different sampling plans. (Pyzdek, 1999; Montgomery, 2009)

4.4. Methods Used for Knowledge Creation

In this chapter the statistical methods used for knowledge creation in the publications included in this thesis are explained. This clarification included because typically the clarification in scientific publications is very straightforward and strict; data, methods and results. A more detailed explanation will enable a deeper understanding of the way in which these methods can be applied and how the methods create information. As a part of knowledge creation process proposed in this dissertation is applied in the attached publications, expert knowledge is an important part of the final transformation from information into knowledge.

In many cases in industry, the examined data might be univariate or multivariate. It is important to know what kind of data is gathered because it will determine the applicable statistical analysis methods. Based on the type of data, univariate methods are needed when dealing with univariate data and multivariate methods when data is of a multivariate type.

This part of the dissertation starts by explaining a method that is widely applied in industry for knowledge creation. This method, Six Sigma, is mainly used by companies to improve their understanding and knowledge about their customers and processes. The use of Six Sigma is highly recommended, because it is a structured roadmap from data to knowledge.

4.4.1. Six Sigma

Here the method called Six Sigma is presented as a specific routine to create knowledge from processes. The use of this procedure follows specific Define, Measure, Analyze, Improve and Control- roadmap (DMAIC). The roots of this method are in Motorola. Their senior engineer, Bill Smith, was studying the correlation between a product's field life and how often the product had been repaired

[24] ANSI/ASQC Z1.4, MIL-STD-105, MIL-STD-414

during the manufacturing process. Smith presented his findings in 1985. His conclusions were that if a product was found to be defective and corrected during the production process, other flaws were bound to be hidden and found later during the early use of the product by the customer. Smith also concluded that if the product was assembled error free, it seldom failed during early use by the customer. He created the original statistics and formulas which were the beginning of Six Sigma culture

During the history of Six Sigma three different generations are recognized. The first generation of Six Sigma focused on the elimination of defects as well as basic variability reduction, mainly in manufacturing. In the second generation of Six Sigma, the defect elimination and variability reduction remained but there was also a strong tendency to connect these with projects and activities that improved business performance through improved product design and cost reduction. The third generation of Six Sigma focuses on creating value throughout organization and for its stakeholders. (Montgomery & Woodall, 2008)

Traditionally the Greek letter sigma (σ) describes variability in statistics. In Six Sigma it is a quality level that tells how often defects are likely to occur. If the sigma level is high, it indicates that process is less likely produce defects than a process with a low sigma level. The definition of a Six Sigma quality level indicates 3.4 defects per million opportunities. The best products get the value of Six Sigma as a measure of excellence that is world class (Breyfogle, 2003; Harry, 1998)

The backbone of the Six Sigma is statistics, because statisticians have created the tools of the data analysis and industrial designs used in Six Sigma. Consequently, Six Sigma is a field of applied statistics. (Hahn et al., 1999) Six Sigma is not a new way of thinking and it does not provide a new set of quality tools. Instead, it could be defined as a combination of many earlier developed quality methods.

The DMAIC-roadmap forms the base for Six Sigma. It is a systematic procedure for finding significant causes (x's) among all the causes. During each step of this procedure, the amount of possible significant x's is diminished.

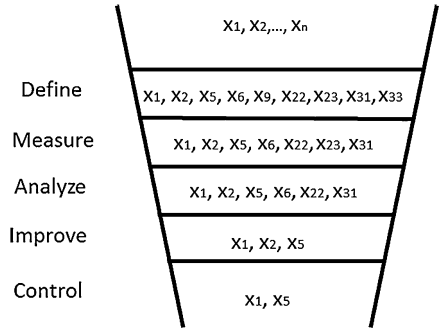

Figure 4-3: DMAIC-roadmap as a funnel, idea adopted from George (2002) and Breyfogle (2003).

The idea of DMAIC-roadmap as a funnel is presented in **Figure 4-3**. This figure shows the idea that at the beginning there are all possible x's and the amount of x's diminishes during a Six Sigma project. Finally, at the control phase there are only few statistically significant x's to control.

In the Define stage, the problem is stated and a project timetable is defined. It is also important to outline the goals of the project. Otherwise there is a risk that if project has no time table, it will never be complete. It is important to also note that it is not possible to solve the whole world with one project. In this stage, process description tools like SIPOC are used for process description. (Pyzdek, 2003)

In the Measure stage, the measurements are performed. The whole stage starts with a Gage R&R or measurement system analysis. The purpose of these procedures is to check the capability of the measurement system. This means that the source of variation in data must be the actual process, not the measurement process. (Breyfogle, 2003).

In the Analyze stage, the data collected during the previous stages is analyzed. It is important to notice that data is worth nothing without proper analysis. This analysis reveals the most important thing about the problem stated before – is it real or caused by an accident. Analysis reveals the statistically significant x's among the all x's. (Bass & Lawton, 2009).

In the Improve stage, attempts are made to solve the previously located real problem. After the solution is found, it is tested. Testing is important because it shows how the solution is functioning with the real process variables. (Bass & Lawton, 2009)

The final stage of DMAIC-roadmap is Control. In this stage, the goal should be to create a system that ensures that the improved situation is maintained after the improvement process has ended. In this stage some measurement procedures might be created for process monitoring to ensure the stability of the current situation. The most important step in this stage is to educate the staff about the lessons learned. (Breyfogle, 2003) Refer to Breyfogle (2003) for a more detailed view on DMAIC.

Six Sigma is a method that does not consist of any redesigning or modifying the structure of original process. It is just a tool that attempts to find solutions to eliminate the root causes of problems within process performance and performance variation. Six Sigma leaves the process unchanged. A method called Design for Six Sigma (DFSS) allows the changing and redesigning of the structure of the underlying process. DFSS is closely related to the product development process and it can significantly improve the product development process, for example in terms of innovation, product design quality, reliability and quality. So it is possible to use DFSS as a tool for technology development and product design. (Creveling et al., 2003; Yang & El-Haik, 2008)

Six Sigma and DFSS share a lot of statistical methods such as DoE and SPC. In DFSS, mainly SPC charts for continuous data are used to check whether the process under statistical control (Creveling et al., 2003). A Similar approach was chosen in Publication 3. In this publication the stability of new fuel cell manufacturing process were studied using SPC.

4.4.2. General Linear Model

Linear models are the foundation of statistical modeling and can be divided into two categories. These two categories are general linear models and generalized linear models. The history of general linear models is based on the work of Gauss and Legendre, but the foundations of generalized linear models are quite new. (Monahan, 2008; McCullagh & Nelder, 1989)

Widely applied statistical analysis methods known as regression models and analysis of variance (ANOVA) models are based on general linear model (Monahan, 2008). The form of general linear model is presented in the following equation (4-1).

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \tag{4-1}$$

This equation (4-1) shows the well-known form of the general linear model. In this model, **y** (1×N) is vector of observed responses**. X** (N×p) is a matrix consisting of the fixed constants and **b** (p×1) is a vector of unknown parameters that are fixed. Vector **e** (N×1) contains of the unobserved errors and it is assumed to have zero mean. (Monahan, 2008)

In the regression model, the response variable (y_i) and the independent variable (x_i) are related to each other. The important thing about the relationship between these variables is the fact that it is linear. The situation where there is only one independent variable and one response variable is called simple linear regression. There is a possibility that there is only one response variable and several independent variables. This type of regression model is a multiple regression model. In some situations there are several response variables related to several independent variables. Regression models of this type are multivariate multiple regression models. In linear regression models, it is assumed that the distributions of the response or responses as well as the predictors are normal. Suitable tests for normality checking are for instance the Anderson-Darling (1952) normality test or the Shapiro-Wilk (1965) normality test. (Rencher, 2002; Monahan, 2008)

In many situations, such as DoE, the comparison between *k* treatments would be interesting. In this kinds of situations the one-way ANOVA is a suitable statistical method. (Monahan, 2008) There are assumptions for using the one-way ANOVA. It should be assumed that the treatment data are normally distributed and have equal variances. This simple one-way analysis method can easily be generalized for comparing more than one treatment. The resulting methods are called two-, three-, etc. higher-way methods. (Scheffé, 1963) A suitable test for checking the normality assumption would be Anderson-Darling (1952) normality test or Shapiro-Wilk (1965) normality test. The equality of variances could be inspected using Bartlett's (1937) (for normally distributed data) test or Levene's (1960) test (for any continuous distribution).

In situations where the previously mentioned assumption for a normal distribution is not fulfilled, the Kruskall-Wallis non-parametric one-way ANOVA should be used (Kruskall & Wallis, 1952). The assumptions for the use of this test are only

that all observations are independent and that the populations are approximately the same shape. The interpretation of the Kruskall-Wallis non-parametric one-way ANOVA test is similar to the interpretation of the one-way ANOVA. It can be concluded that a significant value from this test signifies a difference between populations, but not necessarily the means. (Kruskall & Wallis, 1952)

The Mood's median test is another possibility for non-parametric ANOVA. In this test the equality of medians is examined. In this test it is assumed that random samples taken from different populations have the same continuous distribution in shape. This test is more robust with relation to the outliers than Kruskal-Wallis one-way ANOVA, but it is less powerful in situations where there are batches of data with several different distributions. (Breyfogle, 2003)

4.4.3. Generalized Linear Model

Nelder and Wedderburn (1972) introduced the idea of generalized linear models. As McCullagh and Nelder (1989) describe, these models are a natural generalization of the classical linear models. Agresti (2002) explains that generalized linear models are an extension of the ordinary regression model for situations where a non-normally distributed response is examined and the functions of its mean are modeled.

The logistic regression model is the most important model for categorical response data. This type of a statistical model has a wide applicability in many fields such as social sciences, biomedical studies and marketing. Logistic regression model is calculated using the proportional odds model. McCullagh (1980) explains that in the proportional odds model, there are *k* ordered categories of the response. Similarly each of these categories has its own probabilities, $\pi_1(\mathbf{x})$, $\pi_2(\mathbf{x})$, …, $\pi_k(\mathbf{x})$. In these probabilities vector **x** represents explanatory factor or covariates. According to McCullagh (1980), a logistic regression model could be defined using the following equation (4-2).

$$log\left[\frac{\gamma_{j\,(x)}}{\{1-\gamma_{j\,(x)}\}}\right] = \theta_j - \boldsymbol{\beta}^T x\,, (1 \le j \le k) \qquad (4\text{-}2)$$

Log-linear models are for Poisson-distributed counted data, not proportions. Typically, the response in this type of a model might be for example radiation count or

particles per second. These models are analogous with ANOVA type models and linear regression models. Log-linear models are usually used to analyze the statistical independence for r×c contingency tables. These models should be used when two variables are the response variables. The statistically dependent row variable X and column variable Y satisfy the following log-linear model, equation (4-3). (McCullagh & Nelder, 1989; Agresti, 2002)

$$log\ y_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \qquad (4\text{-}3)$$

In this model (4-3) λ_ι^X and λ_j^Y are the row effect in this r×c contingency table and as well as λ_{ij}^{XY} are presents the interaction between X and Y.

4.4.4. Multivariate Statistical Methods

Multivariate statistical analysis contains methods that are suitable for analyzing multivariate data. These methods could be applied when there are several measurements from every individual or object in one more samples. Any simultaneous analysis of two or more variables can be, in a broad sense, seen as a multivariate analysis. Multivariate methods are based on univariate and bivariate statistics. Univariate statistics deal with data where there is only one response (independent variable, y) and one or more independent variables (x). (Hair et al., 1998; Rencher, 2002)

There are a wide range of multivariate methods in statistics, such as multivariate analysis of variance (MANOVA), discriminant analysis, multivariate regression, canonical correlation, cluster analysis and FA (Rencher, 2002). In the context of this dissertation, it is not relevant or even possible to describe all these methods. Here, only the methods that have been used in the enclosed publications are explained. These methods are PCA (for categorical data) (Publication I) and MDS (Publication III).

The main idea of PCA is from Karl Pearson. In his paper (Pearson, 1901) he invented an idea to represent a system of points on a line or plane. A few years later Hotelling (1933a; 1933b) investigated similar of representations.

PCA is a dimension reduction method. An important issue while calculating PCA is how many principal components there are in the final solution. The maximum amount is the same as the amount of variables. Jolliffe (1986) introduced several

rules for that purpose. One of the most widely used among them is "Cumulative Percentage of Total Variation". When using this rule, so many principal components are included that they explain around 80 per cent or 90 per cent of the data variation.

PCA tries to seek a linear combination of variables that maximizes the variance. The first principal component has the maximal variance and the second principal component has the maximal variance in an orthogonal direction in relation to the first principal component. The orthogonality between principal components explains the fact that principal components are uncorrelated with each other. Principal components can be calculated based on a data covariance matrix or a correlation matrix. (Rencher, 2002) A principal component could be easily presented according to following equation (4-4) .

$$\begin{aligned} Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{4-4}$$

In these equations (4-4) the constant (a) tells the weight of each variable (X) in each principal component. It is noticeable that the sum of these constants in each principal component is 1. (Tabachnick & Fidell, 2006; Rencher, 2002)

PCA assumes that variables are normally distributed or at least continuous. If these assumptions are violated the calculation of the linear correlation is not correct. (Tabachnick & Fidell, 2006) In situations where the analyzed data is, for example, from a Likert-scale the previously mentioned assumptions are not fulfilled. Due to this, linear PCA is not suitable for this kind of situations – a modification should be used instead. Non-linear PCA is suitable for these kinds of situations.

According to Ellis et al. (2006) the calculation of non-linear PCA is based on the work of Gifi (1991) and more recently Michaelidis and de Leeuw (1998). Non-linear PCA is calculated using a categorical method. This procedure follows the

method described by Meulman et al[25] (2004). There is a relationship between Cronbach's (1961) coefficient α and the eigenvalue (the total sum of squared component loadings in each dimension). (Meulman et al., 2004) Based on the previously explained relationship between Cronbach's (1961) coefficient α, and the eigenvalue, the amount of required principal components can be determined according to the following rule. A cut-off value for Cronbach's coefficient α is chosen and all the principal components where α is bigger than the cut-off is included in the further analysis

MDS consists of mathematical methods that enable a researcher to find hidden structures of data. The purpose of MDS is to represent the measurements of similarity or dissimilarity between pairs of objects as distances between points in a low dimensional space. The original idea of MDS was proposed by Torgerson[26]. (Kruskal & Wish, 1981; Cox & Cox, 2001; Borg & Groenen, 2005) Borg and Groenen (2005) present four purposes for MDS. They argue that MDS is a method that represents dissimilarity or similarity of data as distances in a low-dimensional space. The idea here is to make data accessible for visual examination and exploration. MDS is also described as a technique that allows the examination of whether and in what way the empirical differences between certain objects are related to criteria in question. MDS could also be seen as an analysis method that helps to discover the dimensions where the judgments of dissimilarity or similarity are based on. As Borg and Groenen (2005) explain, MDS can be used as a psychological model that explains decisions where similarities and dissimilarities are are a part of a rule that minimizes a particular type of a distance function.

In metric MDS, it is assumed that there are *n* objects with dissimilarities. In metric MDS it is attempted to find a set of points in a space where points represent the objects and the distance is defined as a function of the dissimilarities. Similarity and dissimilarity are two possible measures of proximity. Usually it is assumed that proximities are ratio-scaled values. Metric MDS analyzes objects with dissimilarities (δ_{rs}) and tries to find a set of points in space were one point represents one

[25] This method were chosen because it was provided by SPSS19, which was possible to use for analysis

[26] Torgerson (1952; 1962)

object. The distances (d_{rs}) between points could be defined according to the following equation (4-5). (Cox & Cox, 2001)

$$d_{rs} \approx f(\delta_{rs}) \tag{4-5}$$

According to Cox and Cox (2001) the function in the previous equation is a continuous parametric monotonic function. It is possible that the function may be a transformation function or a transformation function that transforms dissimilarities to a distance like form. Scholars like Gower and Legendre (1986) and Jackson et al. (1989) have discussed about similarity and dissimilarity measures for quantitative data.

In some cases, for instance in classification, only the rank-order of the proximities is valid. In these cases, the use of nonmetric MDS such as ordinal MDS is appropriate. In ordinal MDS, the variables are only values of ranking numbers. Nonmetric MDS models represent only the ordinal properties of the data. Nonmetric MDS is the most important MDS method in practice. (Cox & Cox, 2001; Borg & Groenen, 2005)

4.5. Future Challenges for Industrial Statistics

As industry is evolving, so do the needs for industrial statistics. These needs have evolved and expanded beyond engineering and manufacturing applications, into areas containing more service applications. In these areas, almost always some type of financial analysis and risk analysis is applied. Similarly, the amount of data is increasing rapidly. One major data related problem that has always been there and will remain in the future is the fact that for a non-statistician it is hard to understand variability and its effect to observations of the world and conclusions made based on these observations. Another problem is that some people might think that mastering statistical software makes them statisticians. Keller-McNulty (Steinberg, 2008) argues that statisticians are always required in industry; statistical software and tools will not replace them. Two respondents of the expert survey had similar opinions.

> *"The same as always. Most non-statisticians – that includes scientists, engineers, accountants, lawyers, managers, hair-dressers, plumbers, movie stars – you name it! – do not understand variability and how it enters their observations of the world and the conclusions they make from them"*[27]

> *"Industrial Statistics runs the risk of being taken over by those who believe that if they possess some statistical software, they will be able to do everything they need to make sensible decisions statistically"*[28]

It is pointed out that a challenge for professionals would be to develop and encourage a culture in organizations, where statistical thinking is valued and utilized. The value of industrial statisticians is nowadays too rarely recognized. This is quite harshly described by two respondents.

> *"Except, of course, to note that it does not help the poor statisticians trying to do their jobs when those who they might potentially work with think statistics is a plague visited on humanity by a malicious deity"* [29]

> *"It is important that statisticians demonstrate repeatedly their worth to the organization through frequent and large contributions to the organizational bottom line."* [30]

There are some problems related to the education of industrial statistics in the universities today. Mainly, this is due to the fact that many statistical departments are changing their focus towards something else than the traditional methods of industrial statistics. An opinion from one of the respondents describes the current situation.

[27] Anonymous respondent "Expert opinion study on Industrial Statistics"
[28] Anonymous respondent "Expert opinion study on Industrial Statistics"
[29] Anonymous respondent "Expert opinion study on Industrial Statistics"
[30] Anonymous respondent "Expert opinion study on Industrial Statistics"

> *"As more and more statistics departments move toward biostatistics there might be a danger that fewer universities will teach industrial statistics. This might be a problem. "*[31]

Based on the earlier chapters there is wide applicability of statistics in industry. It is also important that many companies are using statisticians for data analysis. This is done mainly based on the current situation – amount data of in industry is increasing and statistics is the discipline for data analyzing. The increased capabilities of computers for data analysis have also assisted the current development.

The practical applications of statistics are becoming more and more important to industry. As Google's chief economist, Hal Varian states this.

> *"I keep saying that the sexy job in the next ten years will be statisticians. People think I'm joking, but who would guessed that computer engineers would've been the sexy job of the 1990s"*[32]

But still, there is a concern that the traditional methods of industrial statistics are not taught in the universities. In many universities there are biostatistics programs and theoretical statistics programs. The development has to be toward a mix of statistical programs where applied statistics are part of either the statistical or engineering study programs.

The term statistical engineering is proposed for study programs where statistics are a part of engineering studies. This kind of a development is highly recommended. As an example, some universities have Six Sigma trainings as a part of their engineering study programs. In industry, more and more people are participating in different levels of Six Sigma trainings to become a specialist of this area in their companies.

[31] Anonymous respondent "Expert opinion study on Industrial Statistics"
[32] Hal Varian, The McKinsey Quaterly, January 2009 http://flowingdata.com/2009/02/25/googles-chief-economist-hal-varian-on-statistics-and-data/ accessed on 22.03.2012

5. Conclusions

This section of the dissertation will draw conclusions based on the previous sections and the enclosed publications. The section will also explain the relevance of the publications included in this dissertation. Through this, the final statement about the way that knowledge is created in industry is defined. Some own suggestions for industrial knowledge process are also argued and finally the limitations are told and some future research possibilities are suggested.

The structure of this final section is built in the following way. First part of the section summarizes the included publications and explains their relevance in the context of this thesis. Each of these publications will illustrate the wide applicability of statistical methods in the industrial knowledge process. This is the reason that a different type of a data has been selected in each publication. The next section will describe the scientific contribution of this dissertation and the way that this work is interconnected with the scientific discipline. The discussion part does the overall summary of the work and final part will suggest possible new research topics based on this dissertation.

5.1. Summary of the Publications

Following section is a summary of each included publication. This part of the dissertation presents the argumentation for including these publications in this dissertation. This section also outlines the contribution of the researcher for each of these publications.

Data Type	Publication	Analysis Method
Big Data	V	Multidimensional Scaling
Measurement Data	III	Statistical Process Control
Survey Data	I	Non-linear Principal Component Analysis
	II	Ordinal Logistic Regression
	II	Log-linear Model
	IV	Frequency Table
	IV	Binary Logistic Regression
	IV	Kruskall-Wallis One Way Analysis Of Variance
	IV	Mood's Median Test

Table 5-1: The structure of the dissertation.

The structure of the dissertation is presented in **Table 5-1**. This table shows that the data that is analyzed in the enclosed publications can be categorized into three groups. This discrimination is based on the way that how the data is gathered. In the middle column, the publication in which the method has been used for analysis is listed. The last column in the table illustrates the analysis methods used in each enclosed publication.

5.1.1. Publication I

Publication I is a case study where the mobile phone use of young people is studied. This study is approached using questionnaire with a number of open questions as well as five point Likert-scale questions. The focus of this paper was to study mobile phone and mobile service adoption among young people. In this study, self-collected dataset and previously published study on the same topic by Wilska

(2003) was used. Wilska's (2003) original study was analysed using FA, which we found unsuitable for the situation and replaced with a suitable method: non-linear PCA.

The problem in this publication was the selection of a proper analysis method. In this type of studies the researchers usually have a lack of statistical knowledge and have chosen the wrong analysis method, like Wilska has done. This might resemble the situation described in the expert opinion on page 69 (footnote 28) – mastering a statistical analysis software might lead a researcher to think that he or she masters statistical methods as well. Mastering an appropriate level of statistical knowledge the author of this dissertation has chosen a method suitable for the analysis of Liker scale data. This kind of data is discrete and should be treated as such (Jamieson, 2004; Allen & Seaman, 2007). Based on the previously mentioned facts, an analysis method suitable for discrete data needs to be chosen.

The analysis part of the study consists of some descriptive statistics and categorical PCA for both, our new data and Wilska's (2003) data. Because the Likert-scale data is discrete and ordinal, analysis methods must be chosen accordingly (Jamieson, 2004; Allen & Seaman, 2007). Based on the categorical PCA some comparisons between the datasets might be performed. The comparison is complicated by the fact that Wilska's (2003) dataset includes both sexes, while ours only includes males and is also considerably smaller. Possible changes between 2003 and 2012 would be interesting to analyse in a whole new study.

As a result of the analysis using non-linear principal components of both data sets, only one principal component is sufficient to explain the analysed group. This is based on the value of Cronbach's (1961) alpha coefficient. The interpretation of these categorical principal components is based on component loadings. Based on the component loadings, we have presented some changes in consumption styles. According to our study, the most important thing for young people in their mobile phone use is to receive calls and text messages. It is also important that the mobile phone represents the latest technology. Our two categorical principal components have the same interpretation as two out of six Wilska's (2003) six factors.

In this publication survey data is transformed into information using descriptive statistics and a dimension reduction method called categorical PCA. These statistical methods provide simple tools for transforming data into information. The in-

formation is mainly composed of descriptive statistics and the values of categorical principal component loadings. The knowledge is created when these analysis results are interpreted. The created knowledge is the understanding about which are the most important factors for young people in their mobile phone use.

5.1.2. Publication II

Publication II is a case study. In this publication the perceived quality of mobile terminals in the mobile industry is analysed. This study is based on the collection of empirical data from the field using a questionnaire. Over 160 different service instances were studied using statistical methods. Statistical methods were used because these methods reduce the information that data consists into simple analysis results.

The selection of suitable statistical analysis methods and the proper use of these methods was a problem in this examination. The author of this dissertation brought statistical knowledge into this project and helped to choose the proper analysis method for each of the questions about the data. After the author had analysed the data, knowledge was created together by interpreting the analysis results.

The analysis in this publication focuses on the quality of the corrective actions and the amount of time spent resolving the issues, which were collected together with the perceived quality of issue description and additional information from resolvers. The data is analysed using (Chapter 4.4.3) ordinal logistic regression and the log-linear model. Ordinal logistic regression provides information about predictor, predictive variable interrelationship and the association between those variables. The log-linear model is used to find out possible significant interactions between the terms.

Based on statistical analysis, there is a significant difference between the levels in our response. The regression equations for the change from score 1 to score 2 and from score 1 to score 3 were determined. According to the association measures, Goodman and Kruskall's γ (1954) and Somers' D (1962) and Kendall's τ-a (1938) there is some positive association between the response and predictors. This interpretation is based on Goodman and Kruskall (1954).

Statistical analysis in this publication created information about the data. The statistician and the process specialist transformed these results into knowledge. As a

result of this publication, knowledge about different time-factors is achieved. The perceived amount of time spent on resolving issues must be reduced. It means that the effects of absolute time and perceived time are totally different. Perceived time is the more important one – for someone a day is a long time and for someone else, a week is a short time.

5.1.3. Publication III

The study in publication III was a part of a bigger fuel cell project funded by TEKES[33] and executed in the electronics productization research group. In this fuel cell project there were three milestones. The first one was to evaluate the capabilities of the fuel cell technology and to solve the issues crucial to the research group. As the second milestone, the focus was to develop a concept and products based on the first milestone. The third milestone focused on demonstrating the commercial potential of the technology by developing the planned product concepts.

The challenge for the author was to use a statistical method to check the stability of the new manufacturing process for the newly designed fuel cell. As the product was totally new, there was no previous data available. Another challenge was the fact that small changes in manufacturing process would "break" the product. The author of the dissertation provided statistical knowledge for choosing the proper method for process stability checking. The solution for the checking the performance and stability of this fuel cell manufacturing process was the use of one of the basic methods of industrial statistics, SPC.

It is possible to use SPC charts when data is normally distributed, and process capability checking could be done using SPC charts (Montgomery, 2009). The normality of the examined data is checked using a normal probability plot (**Figure 5-1**). In normal probability plot the distribution of the data is analyzed using the Anderson-Darling normality test (Anderson & Darling, 1952) at the 95% confidence interval. When the statistical significance (p-value) in this test is <0.05, it means that result there is no statistically significant difference between normal distribution and the examined data.

[33] TEKES is a Finish agency for funding technology research projects

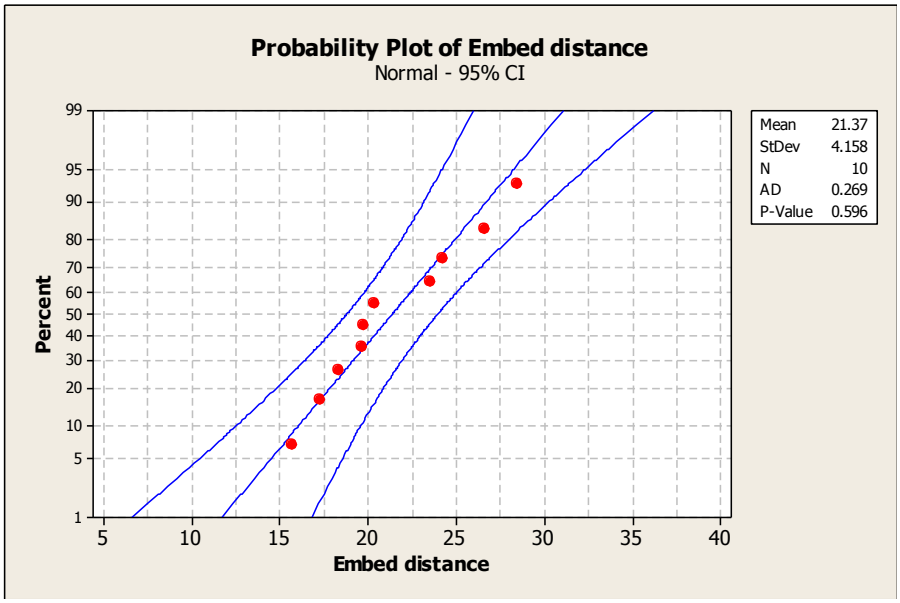

Figure 5-1: Normal Probability Plot for Embed distance.

The p-value for the Anderson-Darling normality test in **Figure 5-1** is greater than 0.05 so the embed distance data could be interpreted as normally distributed. As this assumption is satisfied it is possible to use SPC charts for process capability checking. The used chart was I-MR chart. In this chart there are two parts: the individual value (I) chart and the moving range (MR) chart. In the I chart there are the observation values and in the MR part, there are values calculated y_{t+1}-y_t (formula for moving range). Naturally the Lower Control Limit (LCL) for the moving range is 0. It is a measure of range, so it cannot be less than 0. This chart was chosen because the number of observations was so low.

Usually the UCL (Upper Control Limit) and the LCL (Lower Control Limit) in SPC charts was chosen based on the mean±3*standard deviation (3σ limits). In the I-MR chart this basic rule does not provide useful information about the function of the examined process. Instead of traditional 3σ limits the use of 4σ limits is recommended. This is based on the following equation (5-1).

$$I - MR\ chart$$

$$n = 1,\ \ \mu = the\ value\ of\ each\ observation, \sigma = 0$$
$$\Rightarrow \sigma^2 = 0 \tag{5-1}$$
$$if\ x \sim N(\mu, \sigma^2)$$

$$then\ according\ to\ CLT\ \bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

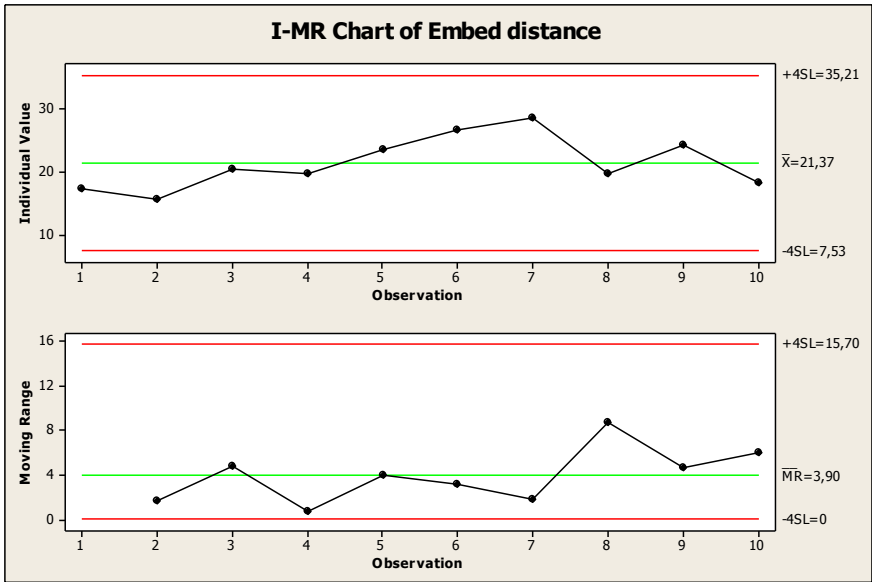

Figure 5-2: I-MR chart for new fuel cell design stability.

Figure 5-2 shows the I-MR chart for new fuel cell design. It can be observed that based on the I-MR chart, the process can be considered stable. There are no values outside the control limits and there are not so many consecutive points at the same side of mean.

The use of the SPC chart is an easy way to conclude whether a process is stable or not. The SPC charts were created specifically for this purpose. (Chapter 4.3.2). The data is transformed into information and further into knowledge with the use of SPC chart. Looking the SPC chart provides information about the examined process. The dotted line (observations) alternating up and down between control lim-

its. This information was transformed into knowledge that the examined process is stable and under control.

5.1.4. Publication IV

The publication IV is a case study were survey data was studied. The association between a sample and the perceived quality of corrective actions was analyzed. The effect of the quality of the samples on the issue resolution time was also analyzed. Finally, calculations were performed to determine whether there were statistically significant differences in the sample collection times between different sales units or in issue resolution times between different software platforms.

A problem in the study reported here in publication IV was the selection of suitable statistical analysis methods and the proper use of these methods. In this case study, the researcher had a set of data and some questions about the data. The author of this dissertation brought the statistical knowledge into this project and helped to choose the proper analysis method for each of the questions based on the data. Based on this statistical knowledge, the transformation from data to knowledge was achieved. After the author had analysed the data, knowledge were created together by interpreting the analysis results.

Frequency tables were used to analyze the association between customer expectations and the sample. A similar kind of analysis is performed using logistic regression. According to the association measures analyzed according to Goodman and Kruskall (1954), there was no association. The dissimilarities in sample collection time in different sales units and the possible statistically significant difference in issue resolution times between different software platforms, can be easily analyzed using different kind of variance analysis methods (mainly parametric or non-parametric).

The datasets were first studied graphically and the equally assumed variances were tested. Because the analyzed datasets were not normally distributed, the differences were mainly studied using Kruskall and Wallis one way ANOVA and Mood's median test. Both are non-parametric alternatives for parametric ANOVA.

In this publication statistical analysis results provided information based on data. This information was transformed into knowledge when the statistician and the process specialist were interpreting the analysis results. Important knowledge based

on this publication was generated when the association measures were interpreted. Because the association is not statistically significant the interpretation says that there is no need for sample collection. It means that there is no difference in issue resolution time. Knowledge about differences between different software platforms and sales unit is important, as it provides new research questions. An additional study examining the causes would be worth performing.

5.1.5. Publication V

Publication V analyzes a big dataset, especially bibliometric data. This type of data is not primarily collected for analysis. This kind of data might be databanks about customer behavior or publication databanks such as Elsevier, Science Direct or Pubmed in medical sciences. Some examples of data that can be gained from different databanks are co-occurrence data and more specific co-citation data. This kind of data describes how two scholars have cited each other.

The problem in this publication was the intention to take more theoretical point of view about the statistical methods applied in this kind of bibliometric studies. There is also a general problem which is examined using morphological analysis and which is therefore outside the scope of this dissertation. There are two statistical problems in this publication. Those are 1) which standardization method for co-occurrence data should be used and 2) what is the appropriate MDS method for this situation. Scholars in this field do not usually mention what type of MDS they have used (van Eck et al., 2010). The co-citation data is ordinal scale data and remains so after the transformation. The author of this dissertation provided the statistical knowledge to the study and solved these statistical issues based on the research of other academics in this field. The author has chosen an approach for statistical analysis that follows mainly van Eck et al. (2010). The use of association strength for normalizing the co-occurrence data is an appropriate method. This kind of transformed data could be treated as ratio scale measurement data, but if the proximities are similarities there is no use in calculating ratio MDS. (van Eck et al., 2010)

In this publication, the co-citation data is collected from ISI Web of Science by using a search algorithm of “fuel cell” or “fuel cells” mentioned in the title or topic. The resulting data forms the co-citation matrix of the study. Leydesdorff and Vaughan (2006) have argued that the co-citation matrix could be directly used as a

proximity matrix. On the other hand, many scholars have stated their opinion about the co-citation matrix and proximities. Borg and Croenen (2005) agree about the complications related to the direct use of co-occurrence data for proximities. The direct use depends on the definition of direct. In some cases the criterion should have been clear to the respondents so the direct use is acceptable. In some cases the characterization of co-occurrences as proximities is based on the interpretation of the researcher so the direct use is not acceptable. Waltman and van Eck (2007), van Eck and Waltman (2009) and van Eck et al. (2010) have argued that the direct use of the co-citation matrix as a proximity matrix should be avoided and co-occurrence data should be normalized before the MDS analysis instead.

The analysis of this publication was mainly based on van Eck et. al (2010), who propose that the proper method for transforming co-occurrence frequencies is the association strength, equation (5-2).

$$AS_{ij} = \frac{c_{ij}}{c_i c_j} \quad for\ i \neq j \tag{5-2}$$

Where c_{ij} indicates the number of items in which scholar i and j both occur and c_i represents the number of items in which scholar i occurs. (van Eck & Waltman, 2007a)

Similarities that are calculated based on association strength can be treated as measurements on a ratio scale. On the other hand it makes no sense to use ratio MDS when the proximities in MDS are similarities. (van Eck et al., 2010) The use of ordinal or interval MDS in these cases might pose some problems. Instead of that the use of so-called visualization of similarities (VOS) method proposed by van Eck and Waltman (2007b) is more appropriate. (van Eck & Waltman, 2007a; van Eck et al., 2010) Instead of this, the traditional visualization method MDS was chosen for the analysis. To be more exact, the method chosen was interval MDS which is metric MDS.

Statistical analysis in this publication created information about the data. The data in this examination was co-citation data and the study was a bibliometric study in the field of fuel cells. Statistician and process specialist transformed analysis results into knowledge. The statistical analysis was used as a validating tool for the expert opinion analysis. As a statistical result, this publication showed the similari-

ties between the analyzed objects in a two dimensional space. Small distances in the analysis denote a high level of similarity between the analyzed objects. Large distances denote high dissimilarities between the analyzed objects. We discovered that the statistical results were able to identify the most visible group in the dataset.

5.2. Contribution and Conclusion

This dissertation is written with a clear understanding that is not possible to solve all the problems in the world with one single dissertation. The purpose of the thesis is to show a glimpse of the wide applicability of statistical methods in relation to different types of data in industry. It is also examined how data is transformed into information and further to knowledge using statistical methods. This was done based on the conceptual framework of this dissertation, the widely known DIKW-hierarchy. In this part of the dissertation a structure consisting of data, information and knowledge in industrial processes was proposed. The first proposition is named as classical approach (Chapter 2.4.3). This classical model is a straightforward generalization of the DIKW-hierarchy into an industrial context. This structure was found insufficient for the examined situation. Because of that second iteration were made. This proposed own structure is based on DIKW-hierarchy as well and it is named Bayesian approach (Chapter 2.4.3). This Bayesian type model is based on the fact that in each step of the process some additional knowledge is required. In the first step of the process data is measured. This is not possible to do without the knowledge about measuring – how to measure. Similarly when data is transformed into information using statistical analysis knowledge about statistical analysis methods is required. At the final step when knowledge is created, knowledge about process is required. This knowledge from process specialist interprets the statistical analysis results into knowledge about the process. Based on organizational knowledge creation spiral (Nonaka & Takeuchi, 1995) and organizational knowledge creation process (Bhatt, 2000) own statistical analysis based industrial knowledge creation process were proposed.

The proposed model for industrial knowledge creation was used in each enclosed publications. During the research of each publication the project group was created in a way that there was a specialist about data gathering present. This stage re-

quired knowledge about process and measurement as well as statistics. Knowledge about the process and measurement process were need to measure things right. Statistical knowledge was required for estimation of suitable amount of samples for statistical analysis. After that statistician, who have the knowledge about statistics calculated the proper analysis. In final stage person with statistical knowledge and process knowledge interpreted analysis results to knowledge. Eventually this knowledge goes up from individual level to organizational and inter-organizational levels.

The use of statistical methods for knowledge creation in industry is strongly recommended. In some references the term "statistical engineering" is mentioned. This kind of educations should be added to the academic education of engineers. Adding applied statistics and industrial statistics as a part of technical education will have positive effect in industry. In this kind of situation technical experts would solve the problems in their process by themselves. In some universities there are departments of industrial statistics. This as well helps companies to recruit statistically oriented people to analyse their data and create knowledge to help their operation in markets.

5.3. Future Work and Limitations

In this dissertation statistical methods for knowledge creation in industrial processes were studied. In the future it would be interesting to examine the other possible methods for this purpose – as the amount of data increases and increases, is there any proper alternatives for statistics. If there are, how efficient those are comparing to statistics.

Statistics is a powerful tool for transforming data into information and knowledge. The use of statistics in different companies must be increased. As argued at the beginning of this dissertation the amount of data will be increasing more and more in the future. Because of that the knowledge of methods like statistics and data mining will be advantage in the future. Studies were these two methods are combined will have interesting and important result. To be competitive in the hard markets requires that the process from data to knowledge must be shortened.

Knowledge must be created rapidly and must be available similarly with the data, otherwise it is hard to survive.

References

Ackoff, R.L., 1989. From Data to Wisdom. *Journal of Applied System Analysis*, 16, pp.3--6.

Agresti, A., 2002. *Categorical Data Analysis*. Wiley-Interscience.

Agresti, A., 2010. *Analysis of Ordinal Categorical Data*. Wiley.

Allen, I.E. & Seaman, C.A., 2007. Likert Scales and Data Analyses. *Quality Progress*, 40, pp.64--65.

Anderson, T.W. & Darling, D.A., 1952. Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23, pp.193--212.

Argyris, C. & Schon, D.A., 1978. *Organizational Learning: A Theory of Action Perspective*. MA: Addison-Wesley Publishing.

Babbie, E.R., 1973. *Survey Research Methods*. Belmont: Wadsworth Publishing Company, Inc.

Barnard, G.A. & Bayes, T., 1958. Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Changes. *Biometrika*, pp.293--315.

Barnett, V., 1982. *Comparative Statistical Inference*. Chichester: John Wiley & Sons.

Bartholomew, D.J., 1980. Factor Analysis for Categorical Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, pp.293--321.

Bartlett, M.S., 1937. Properties of Sufficiency and Statistical Tests. In S. Kotz, ed. *Breakthroughs in Statistics Vol.1*. Springer Verlag 1992.

Bass, I. & Lawton, B., 2009. *Lean Six Sigma Using SigmaXL and Minitab*. McGraw-Hill Professional.

Bersimis, S., Psarakis, S. & Panaretos, J., 2007. Multivariate Statistical Process Control Charts: an Overview. *Quality and Reliability Engineering International*, 23, pp.517--543.

Bhatt, G.D., 2000. Organizing Knowledge in the Knowledge Development Cycle. *Journal of Knowledge Management*, 4(1), pp.15-26.

Borg, I. & Groenen, P.J., 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer.

Botha, E.M., 1989. Theory Developmenting Perspective: the Role of Conceptual Frameworks and Models in Theory Development. *Journal of Advanced Nurcing*, 14(1), pp.49--55.

Box, G.E.P., 1989. *Do Interactions Matter*. Report No. 46. Madison, Wisconsin: University of Wisconsin-Madison Center for Quality and Productivity Improvement.

Box, G.E.P. & Draper, N.R., 1987. *Empirical Model-Building and Response Surfaces*. Wiley.

Box, G.E.P., Hunter, J.S. & Hunter, W.G., 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience.

Box, G.E.P. & Wilson, K.B., 1951. On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13, pp.1--45.

Breiman, L., 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16, pp.199--215.

Breyfogle, F.W., 2003. *Implementing Six Sigma: Smarter Solutions Using Statistical Methods*. John Wiley & Sons.

Breyfogle, F.W., 2008. *Integrated Enterprise Excellence, Vol. III Improvement Project Execution: A Management and Black Belt Guide for Going Beyond Lean Six Sigma and the Balanced Scorecard*. Bridgeway Books.

Brown, J.S. & Duguid, P., 2002. *The Social Life of Information*. Harvard Business School Press.

Bryman, A., 1988. *Quantity and Quality in Social Research*. London: Unwin Hyman.

Chaloner, K. & Verdinelli, I., 1995. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3), pp.273--304.

Clark, T.G., Bradburn, M.J., Love, S.B. & Altman, D.G., 2003. Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, 89, pp.232--238.

Clason, D.L. & Dormody, T.J., 1994. Analyzing Data Measured By Individual Likert-Type Items. *Journal of Agricultural Education*, 35, pp.31--35.

Cleveland, H., 1982. Information as a Resource. *The Futurist*, pp.34--39.

Collett, D., 2003. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC.

Cox, D.R., 2006. *Principles of Statistical Inference*. Cambridge: Cambirdge University Press.

Cox, T.F. & Cox, M.A., 2001. *Multidimensional Scaling*. Chapman and Hall/CRC.

Cox, D.R. & Oakes, D., 1984. *Analysis of Survival Data*. Chapman and Hall/CRC.

Creveling, C.M., Slutsky, J.L. & Antis, D., 2003. *Design for Six Sigma*. New Jersey: Pearson Education, Inc.

Cronbach, L.J., 1961. Coefficient Alpha and the Internal Structure of Tests. *Psychometrica*, 16(3), pp.297--334.

Davenport, T.H. & Prusak, L., 2000. *Working Knowledge*. Harvard Business Review Press.

De Mast, J. & Does, R.J., 2006. Industrial Statistics: a Discipline with Opportunities and Challenges. *Statistica Neerlandica*, 60, pp.270--282.

Devore, J.L., 2011. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Dretske, F.I., 1981. *Knowledge and the Flow of Information*. Blackwell Publishers.

Efron, B., 1986. Why Isnt Everyone a Bayesian? *The American Statistician*, 40, pp.1--11.

Efron, B., 2005. Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association*, 100, pp.1--5.

Eliot, T.S., 2002. *T. S. Eliot Collected Poems 1909-1962*. Faber and Faber.

Ellis, R.N., Kroonenberg, P.M., Harch, B.D. & Basford, K.E., 2006. Non-linear Principal Components Analysis: an Alternative Method for Finding Patterns in Environmental Data. *Environmetrics*, 17, pp.1--11.

Feelders, A., Daniels, H. & Holsheimer, M., 2000. Methodological and Practical Aspects of Data Mining. *Information & Management*, 37, pp.271--281.

Fisher, R.A., 1925. Theory of Statistical Estimation. *Proceedings of the Cambrige Philosophical Society*, 22, pp.700--725.

Frické, M., 2009. The Knowledge Pyramid: a Critique of the DIKW Hierarchy. *Journal of Information Science*, 35, pp.131--142.

Friedman, J.H., 1997. *Data Mining and Statistics: What's the Connection?* Houston. Keynote Speech of the 29th Symbosium on the Interface: ComputingScience and Statistics.

Gaito, J., 1980. Measurement Scales and Statistics: Resurgence of an Old Misconception. *Psychological Bulletin*, 87(3), pp.564--567.

George, M.L., 2002. *Lean Six Sigma Combining Six Sigma Quality with Lean Speed*. New York: McGraw-Hill.

Georghiou, L. et al., eds., 2008. *The Handbook of Technology Foresight*. Edward Elgar Publishing Limited.

Gettier, E., 1963. Is Justified True Belief Knowledge. *Analysis*, 23(6), pp.121--123.

Gifi, A., 1991. *Nonlinear Multivariate Analysis*. Chichester: Wiley.

Gilchrist, W., 1984. *Statistical Modelling*. John Wiley & Sons Ltd.

Glymour, C., Madigan, D., Pregibon, D. & Smyth, P., 1997. Statistical Themes and Lessons for Data Mining. *Data Mining and Knowledge Discovery*, 1, pp.11--28.

Göb, R., McCollin, C. & Ramalhoto, M.F., 2007. Ordinal Methodology in the Analysis of Likert Scales. *Quality & Quantity*, 41, pp.601--626.

Goodman, L.A. & Kruskall, W.H., 1954. Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, 49, pp.732--764.

Gourlay, S., 2006. Conceptualizing Knowledge Creation: A Critique of Nonaka's Theory. *Journal of Management Studies*, 43(7), pp.1415--1436.

Gower, J.C. & Legendre, P., 1986. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, pp.5--48.

Hahn, G.J., Hill, W.J., Hoerl, R.W. & Zinkgraf, S.A., 1999. The Impact of Six Sigma Improvement--A Glimpse into the Future of Statistics. *The American Statistician*, 53, pp.208--215.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W.C., 1998. *Multivariate Data Analysis*. New Jersey: Prentice-Hall PTR.

Halfpenny, P., 1979. The Analysis of Qualitative Data. *Sociological Review*, 27(4), pp.799--825.

Harry, M.J., 1998. Six Sigma: A Breakthrough Strategy for Profitability. *Quality Progress*, 31, pp.60--64.

Harsh, O.K., 2007. Data, Information and Knowledge & Reuse Management Techniques. In *Proceedings of the World Congress on Engineering*. London, 2007. WCE.

Helenius, H., 1995. *Tilastollisten Menetelmien Perustiedot*. Tampere: Painomainos Oy.

Hosking, J.R.M., Pednault, E.P.D. & Sudan, M., 1997. A Statistical Perspective on Data Mining. *Future Generation Computer Systems*, 13, pp.117--134.

Hotelling, H., 1933b. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6), pp.417-41.

Hotelling, H., 1933a. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(7), pp.498-520.

Hotelling, H., 1947. Multivariate Quality Control, Illustrated by the Air Testing of Sample Bomsights. In E. Churchill, M.W. Hastay & W.A. Wallis, eds. *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Management Engineering*. McGraw-Hill. pp.111-84.

Hougaard, P., 1999. Fundamentals of Survival Data. *Biometrics*, 55, pp.13--22.

Jackson, D.A., Somers, K.M. & Harvey, H.H., 1989. Similarity Coefficients: Meaures of Co-occurence and Association or Simply Measures of Occurence. *The American Naturalist*, 133(3), pp.436--453.

Jamieson, S., 2004. Likert Scales: How to (Ab)use Them. *Medical Education*, 38(12), pp.1217-18.

Johnson, B., Lorenz, E. & Lundvall, B.-Å., 2002. Why all this Fuss about Codified and Tacit Knowledge. *Industrial and Corporate Change*, 11(2), pp.245--262.

Jolliffe, I.T., 1986. *Principal Component Analysis*. New York: Springer-Verlag.

Kendall, M.G., 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1), pp.81--93.

Kessler, M.M., 1963a. Bibliographic Coupling Between Scientific Papers. *American Documentation*, pp.10--25.

Kessler, M.M., 1963b. An Experimental Study of Bibliographic Coupling Between Technical Papers. *IEEE Transactions on Information Theory*, pp.49--51.

Khurshid, A. & Sahai, H., 1993. Scales of Measurements: An Introduction and a Selected Bibliography. *Quality and Quantity*, 27, pp.303--324.

Korhonen, P. & Siljamäki, A., 1998. Ordinal Principal Component Analysis Theory and an Application. *Computational Statistics & Data Analysis*, 26, pp.411--424.

Koskinen, K.U., Pihlanto, P. & Vanharanta, H., 2003. Tacit Knowledge Acquisition and Sharing in a Project Work Context. *International Journal of Project Management*, 21, pp.281--290.

Kotz, S. & Johnson, N.L., 2002. Process Capability Indices-a Review, 1992-2000 / Discussion / Response. *Journal of Quality Technology*, 34(1), p.2.

Kruskall, W.H. & Wallis, W.A., 1952. The Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), pp.583--621.

Kruskal, J.B. & Wish, M., 1981. *Multidimensional Scaling*. Beverly Hills: Sage Publications, Inc. Qualitative Applications in the Social Sciences.

Lee, C.C. & Yang, J., 2000. Knowledge Value Chain. *Journal of Management Development*, 19, pp.783--794.

Lehrer, K., 1990. *Theory of Knowledge*. Routledge.

Levene, H., 1960. Properties of Sufficiency and Statistical Tests. In I. Olkin, ed. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press.

Levitin, A.V. & Redman, T.C., 1998. Data as a Resource: Properties, Implications, and Prescriptions. *Sloan Management Review*, 40, pp.89--101.

Leydesdorff, L. & Vaughan, L., 2006. Co-occurence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment. *Journal of the American Society for Information Science and Technology*, 57(12), pp.1616--1628.

Lillrank, P.M. & Forssèn, M., 1998. *Managing for Knowledge: Perspectives and Prospects*. Helsinki University of Technology.

Lindley, D.V., 1956. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27, pp.986--1005.

Linting, M., Meulman, J.J., Groenen, P.J. & van der Koojj, A.J., 2007. Nonlinear Principal Components Analysis: Introduction and Application. *Psychological Methods*, 12, pp.336--358.

Lord, F.M., 1953. On the Statistical Treatment of Football Numbers. *American Psychologist*, 8, pp.750--751.

Lowry, C.A. & Montgomery, D.C., 1995. A Rewiev of Multivariate Control Charts. *IIE Transactions*, 27, pp.800-10.

Lowry, C.A., Woodal, W.H., Champ, C.W. & Rigdon, S.E., 1992. A Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics*, 34(1), pp.46-53.

Mahoney, J. & Goertz, G., 2006. A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *Political Analysis*, 14, pp.227--249.

Marshall, C. & Rossman, G.B., 2006. *Designing Qualitative Research*. Sage Publications.

Mason, R.L., Tracy, N.D. & Young, J.C., 1997. A Practical Approach for Interpreting Multivariate T2 Control Chart Signals. *Journal of Quality Technology*, 29, pp.396--406.

Mason, R.L. & Young, J.C., 2000. Interpretive Features of a T(2) Chart in Multivariate SPC. *Quality Progress*, 33, pp.84--89.

Mason, R.L. & Young, J.C., 2004. Multivariate Thinking. *Quality Progress*, 37, pp.89--91.

McCullagh, P., 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, pp.109--142.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models*. Chapman and Hall/CRC.

Meulman, J.J., van der Kooij, A.J. & Heiser, W.J., 2004. Principal Component Analysis with Nonlinear Optimum Scaling Transformations for Ordinal and Nominal Data. In Kaplan, D. *Handbook of Quantitative Methodology for Social Sciences*. London: Sage. pp.49-70.

Michaelidis, G. & de Leeuw, J., 1998. The Gifi Sysytem of Descriptive Multivariate Analysis. *Statistical Science*, 13(4), pp.307--336.

Monahan, J.F., 2008. *A Primer on Linear Models*. Chapman & Hall/CRC.

Montgomery, D.C., 2000. The Future of Industrial Statistics. *Orion*, 16(1), pp.1-21.

Montgomery, D.C., 2001. Opportunities and Challenges for Industrial Statisticians. *Journal of Applied Statistics*, 28(3&4), pp.427--439.

Montgomery, D.C., 2008. *Design and Analysis of Experiments*. Wiley.

Montgomery, D.C., 2009. *Introduction to Statistical Quality Control*. 6th ed. John Wiley & Sons.

Montgomery, D.C. & Woodall, W.H., 2008. An Overview of Six Sigma. *International Statistical Review*, 76, pp.329--346.

Murphy, B.J., 1987. Selecting Out of Control Variables With the Multivariate Quality Control Procedure. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 36, pp.571--581.

Narin, F., Olivastro, D. & Stevens, K.A., 1994. Bibliometrics/Theory, Practice and Problems. *Evaluation Review*, 18(1), pp.65--76.

Nelder, J.A. & Wedderburn, R.W.M., 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), pp.370--384.

Nonaka, I., 1994. A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, 5, pp.14--37.

Nonaka, I., Byosiere, P., Borucki, C.C. & Konno, N., 1994. Organizational Knowledge Creation Theory: A First Comprehensive Test. *International Business Review*, 3(4), pp.337-51.

Nonaka, I. & Takeuchi, H., 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, USA.

Nonaka, I. & von Krogh, G., 2009. Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organization Knowledge Creation Theory. *OrganizationScience*, pp.635--652.

Pearson, K., 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2, pp.559-72.

Polanyi, M., 1966. *The Tacit Dimension*. New York: Doubleday & Co.

Polanyi, M., 1969. *Knowing and Being: Essays by Michael Polanyi*. University Of Chicago Press.

Pyzdek, T., 1999. *Quality Engineering Handbook*. Quality Publishing, LLC.

Pyzdek, T., 2003. *The Six Sigma Project Planner : A Step-by-Step Guide to Leading a Six Sigma Project Through DMAIC*. McGraw-Hill.

Rencher, A.C., 2002. *Methods of Multivariate Analysis*. Wiley-Interscience.

Rigdon, S.E., 1995. A Double-integral Equation for the Average Run Length of a Multivariate Exponentially Weighted Moving Average Control Chart. *Statistics & Probability Letters*, 24, pp.365--373.

Robson, C., 2002. *Real World Research*. Blackwell Publishing.

Rowley, J., 2006. Where Is the Wisdom That We Have Lost in Knowledge? *Journal of Documentation*, pp.251--270.

Rowley, J., 2007. The Wisdom Hierarchy: Representations of the DIKW Hierarchy. *Journal of Information Science*, 33, pp.163--180.

Ryan, T.P., 1989. *Statistical Methods for Quality Improvement*. John Wiley & Sons, Inc.

Sapsford, R., 2006. *Survey Research*. Sage Publications Ltd.

Scheffé, H., 1963. *The Analysis of Variance*. John Wiley & Sons.

Senn, S., 2003. *Dicing with Death: Change, Risk and Healt*. Cambridge: Cambridge University Press.

Shannon, C.E., 1948a. A Mathematical Theory of Communication, Part I - Part II. *The Bell System Technical Journal*, XXVII(3), pp.379--423.

Shannon, C.E., 1948b. A Mathematical Theory of Communication, Part III - Part IV. *The Bell System Technical Journal*, XXVII(3), pp.623--656.

Shapiro, S.S. & Wilk, M.B., 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, pp.591--611.

Shewhart, W., 1930. *Economic Quality Control of Manufactured Products*. Bell Telephone Laboratories.

Siegel, S., 1957. Nonparametric Statistics. *The American Statistician*, 11(3), pp.13--19.

Singh, R. & Mukhopadhyay, K., 2011. Survival Analysis in Clinical Trials: Basics and Must Know Areas. *Perspectives in Clinical Research*, 2, p.145.

Small, H., 1973. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of American Society for Information Science*, pp.256--269.

Smith, E.A., 2001. The Role of Tacit and Explicit Knowledge in the Workplace. *Journal of Knowledge Management*, 5, pp.311--321.

Somers, R.H., 1962. A New Symmetrical Measure of Association for Ordinal Variables. *American Sociological Review*, 27, pp.73--80.

Spiring, F., Leung, B., Cheng, S. & Yeung, A., 2003. A Bibliography of Process Capability Papers. *Quality and Reliability Engineering International*, 19, pp.445--460.

Steinberg, D.M., 2008. The Future of Industrial Statistics: A Panel Discussion. *Technometrics*, 50(2), pp.103--127.

Sternberg, R.J., 2003. *Wisdom, Intelligence, and Creativity Synthesized*. Cambridge University Press.

Stevens, S.S., 1946. On the Theory of Scales of Measurement. *Science*, 103, pp.677 --680.

Stoumbos, Z.G., Marion R. Reynolds, J., Ryan, T.P. & Woodall, W.H., 2000. The State of Statistical Process Control as We Proceed into the 21st Century. *Journal of the American Statistical Association*, 95, pp.992--998.

Tabachnick, B.G. & Fidell, L.S., 2006. *Using Multivariate Statistics*. Allyn & Bacon.

Torgerson, W.S., 1952. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17(4), pp.401-19.

Torgerson, W.S., 1962. *Theory and Methods of Scaling*. John Wiley & Sons, Inc.

Tuomi, I., 2000. Data is More Than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. *Journal of Management Information Systems.*, 13(3), pp.103--117.

van Eck, N.J. & Waltman, L., 2007a. Bibliometric Mapping of the Computational Inteligence Field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5), pp.625--645.

van Eck, N.J. & Waltman, L., 2007b. VOS: a New Method for Visualizing Similarities Between Objects. In Lenz, H.-J. & Decker, R., eds. *Proceedings of the 30th Annual Conference of German Classification Society, Studies in Classification, Data Analysis and Knowledge Organizations*., 2007b. Springer.

van Eck, N.J. & Waltman, L., 2009. How to Normalize Co-Occurence Data? An Analysis of Some Well-Known Similarity Measures. *Journal of the American Society for Information Science and Technology*, 60(8), pp.1635--1651.

van Eck, N.J., Waltman, L., Dekker, R. & van den Berg, J., 2010. A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12), pp.2405--2416.

Velleman, P.F. & Wilkinson, L., 1993. Nominal, Ordinal, Interval and Ratio Typologies are Misleading. *The American Statistician*, 47(1), pp.65--72.

Walpole, R.E. et al., 2002. *Probability and Statistics for Engineers and Scientists*. Prentice Hall.

Waltman, L. & van Eck, N.J., 2007. Some Comments on the Question Whether Co-Occurence Data Should Be Normalized. *Journal of the American Society for Information Science and Technology*, 58(11), pp.1701--1703.

Wheeler, D.J., 1992. *D. S. Chamberss D. J. Wheelers Understanding Statistical Process Control*. SPC Press, Inc.

Wheeler, D., 1995. *Advanced Topics in Statistical Process Controll.* SPC Press, Inc.

Wilska, T.-A., 2003. Mobile Phone Use as Part of Young People's Consumption Styles. *Journal of Consumer Policy*, 26, pp.441--463.

Wonnacott, T.H. & Wonnacott, R.J., 1990. *Introductory Statistics*. Wiley.

Yang, K. & El-Haik, B., 2008. *Design for Six Sigma: A Roadmap for Product Development*. The McGraw-Hill Companies.

Zeleny, M., 1987. Management Support Systems: Towards Integrated Knowledge Management. *Human Systems Management*, 7(1), pp.59--70.

Zins, C., 2007. Conceptual Approaches for Defining Data, Information and Knowledge. *Journal of the American Society for Information Science and Technology*, 58, pp.479--493.