UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Department of Industrial Engineering

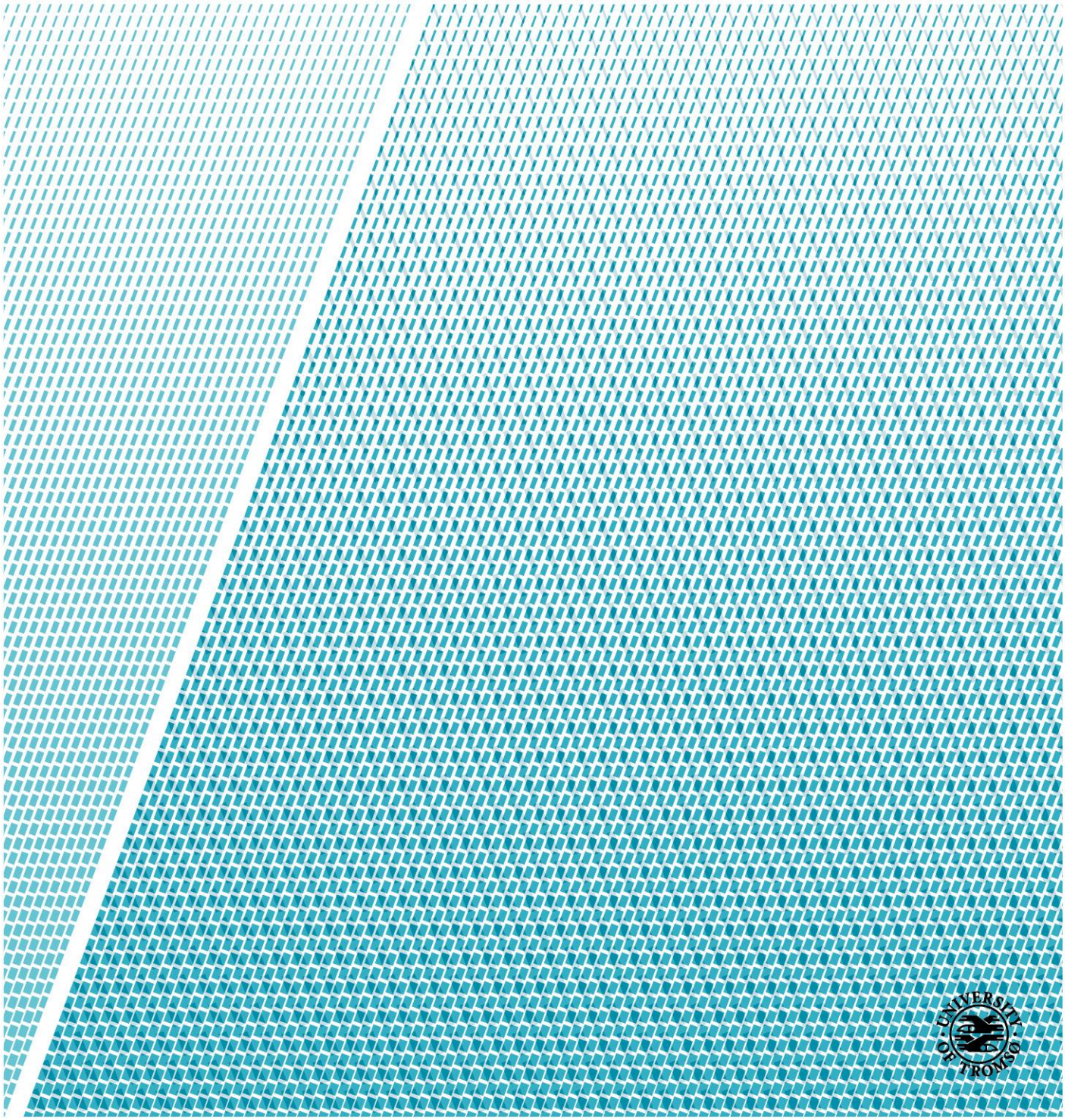
Master Thesis Part II-Pre study Report

Field data processing techniques

Faculty of Technology

Ivan Balachin

Master's thesis in SHO6266, Second of June 2017



Title: Field data processing techniques	*Date:31.01.2017*
	Classification:
Author(s):Ivan Balachin	*Number of Pages:*
	Number of Attachments:1
Subject Name: Master Thesis - Pre Study Report	*Subject Code:* SHO6266
Faculty: Technology	
Master Program: Industrial Engineering	
Supervisor: Geanette Polanco Pinerez	
Co-supervisor: Hu Qin	
External Organization/Company: Chongqing University ,China	
External Organization's/Company's Liaison:	
Keywords (max 10): basic statistics, summary statistics, graphical displays of data, regression analysis, cold weather effects on power lines.	

Abstract (max 150 words):

This paper is written in order to give overview of basic statistical concepts, summary statistics, graphical displays of data, regression analysis. This background will be used for finding regressions between environmental factors and electrical, mechanical response variables of experiment. Experimental data taken from overhead transmission line facility at Xuefeng Mountain belonged to Chongqing University, China.

Table of Contents

List of Tables

List of Figures

Symbols

a	additive constant for linear line equation.
A_2	constant for calculating of Shewhart control chart parameters.
b	multiplicative part of linear line equation.
B_3,D_4	constants for calculating of Shewhart control chart parameters.
f	data fraction.
i	index number of sample value.
j	index number of distance from straight line in y direction.
k	number of samples for average range method, average standard deviation and average moving range.
K	confidence interval constant.
M	sample median.
$\overline{MR}$	average moving range.
n	number of data points in sample.
N	number of data points in population.
q	sum of squared distances in developing of linear regression equation.
$Q_1, Q_2, Q_3, Q_4,$	quartiles.
r	correlation coefficient for sample.
R	range.
$\bar{R}$	average range.
S	variance of sample.
$\bar{S}$	average standard deviation.
S_{xy}	covariance of samples x and y.
SH_i	sum of high for cumulative sum plot.
SL_i	sum low for cumulative sum plot.
t	turning constant fro m-estimate.
w_i	weights of sample.
$y_1 \ldots . y_n$	initial sample values.
$\bar{y}$	sample mean.
y_{max}, y_{min}	max, min values in a sample.
α	significance level of hypothesis.
μ	mean in population.
$\hat{\mu}$	mean estimator for population.
$\tilde{\mu}$	median for population.
σ	std. Deviation fro population.
$\hat{\sigma}$	std. Deviation estimator.
σ_{xy}	covariance of population X and Y .
σ_x	variance of marginal distribution X.
σ_y	variance of marginal distribution Y.

Abbreviations

AMR	average moving range
CUSUM	cumulative sum
LCL	lower control limit
LOC	location of median and quantile
SPC	statistical process control
UCL	upper control limit

1 Introduction

Ice covering on overhead transmission lines can cause accidents such as, short-circuit, grounding, wire breakage, tower distraction or flashover (You-le Liu, 2008). This can lead to disturbances in electrical supply of consumers. In order to prevent these accidents icing monitoring on transmission lines is used. On line icing monitoring on transmission lines requires installing of monitoring devices. Xiang-jun Zenga, 2011, described methods of icing thickness monitoring based on reclouser transient travelling wave. In order to design transmission line should be collected some preliminary information like: electrical and geographical data. Geographical data including: maximum and minimum temperatures, maximum wind velocities with ice or without, radial thickness of ice expected on the conductors, existence of aggressive atmosphere's (Farr, 1980). Also it is important to collect and process data. Such databases are important for validation of experimental and theoretical simulations of the icing process (Poots, 1996). Also data is processed by using statistical principles.

2 Scope

The goals of the first phase of thesis are to give basic statistical concepts, definitions and fitting data techniques. Build the bases for the work of Phase II based on selected study case.

3 Statistics in Engineering and science

Role of statistics in engineering and science can not be overestimated. Quality and productivity are goals of industrial process (Robert L. Mason, 1989). Statistics used in monitoring of product quality and ensure that products are in specification limits. Best products are initiated in academic and industrial research, this requires collection of data. Statistics is the science of problem solving in the presence of variability.

For example statistics is used in studying of automobiles emissions, forces on pipes used in drilling oil wells, testing of commercial drugs, etc. All these processes involve some degree of

uncertainty. Figure 1 presents the relation of the different parameters involved when statistics in applied to science investigation.

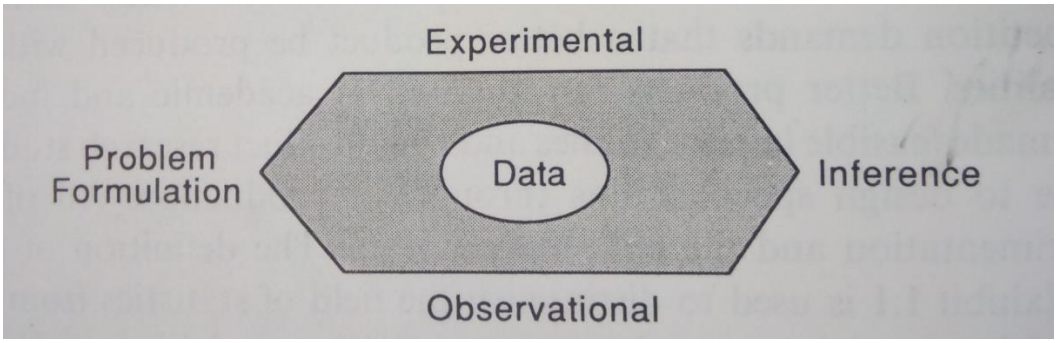


Figure 1 Statistics in scientific investigations

Data is the product of experimental and observational studies. Data is collected from different sources, which include variation in measurements. Variation exists because of changing in ambient conditions, errors in instrument readings or other unknown causes. In order to ensure that an experiment provides useful information, three conditions of experimental design and experiment analysis need to be satisfied.

First nature of data to be collected must be considered, what measurements to be taken and what factors might influence the variation of measurements. Secondary control limits should be selected, what variation is possible from known sources. And third feature is that a statistical analysis of experimental results, should allow to make conclusion how measurements and design factors are related.

Table 1 summarizes the role of statistics in engineering and science experimentation.

Table 1 Role of statistics in experimentation

Phase of experiment	Description phase
Planning	What is to be measured How large is likely variation What are the influential factors
Design	Control known source of variation Estimate size of the uncontrolled variation Investigate possible models
Analysis	Make inference on design factors Make next designs Suggest more suitable models

To guaranty a general understanding of the statistics, hereafter a definition of the object of study, samples, population and distributions, followed by the definition and description of the

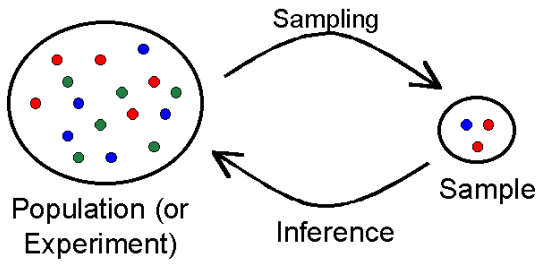
more used, statistical methods, data graphical displays, regression and correlation analysis techniques.

3.1 Basic statistical concepts

The three bases for understanding statistical inference are: distinguishing samples from populations, relating sample statistics to population parameters, deterministic and empirical modelling (Robert L. Mason, 1989).

First base is to distinguish samples from populations concept. Population is a group of possible items or units that determines an outcome of a well-defined experiment. Populations must be defined with respect to all known sources of variation in order to draw valid statistical inferences. Population can also represent processes.

Meanwhile, sample is a group of observations taken from population or a process. The use of samples obeys to economical and time constraints. Connection between sample and population shown on figure



Figur 1 Connection between sample and population

Both, population and sample are related in the way that the sample must be a representative part of the population, so there is no need to evaluate the whole population and using the information of the sample it is possible to elaborate conclusion about the population. Process is a repeatable characteristic or measurement. Measurements on a population of units can exhibit statistical differences based on the characteristics of interest in the experiment, know as variable.

Variables can be divided on two categories: response variables and factors. A response variable can be defined using a probability model as function of one or several factors plus unknown constants. Factors are controllable experimental variables that influence on the observed values of response variable. For example in the study of ice load on wires: tension and torsion depending on environmental factors: pressure, temperature and speed. Power losses as a

function of environmental factors. Second important question what is it parameters and statistics.

Parameters are numerical characteristic of a population or a process. Statistics is a numerical characteristic that is computed from a sample of observations.

Examples of parameters are mean weight of cottage cheese pack at one plant, hardness of steel, maximum wear of stainless-steel ball bearings subjected to a prescribed wear-testing technique. Parameters denoted by Greek letters: μ and σ for standard deviation. Standard deviation is a measure of variability of the observations in a population. Population parameters are often used for defining specification limits or tolerances for manufactured products. Greek letters denote parameters, Latin letters variables. Denotation of parameters can be summarised in **Table 2** from (Michael L. George, 2005).

Table 2 Denotation of parameters in Population and Sample for standard deviation

Description	Population	Sampling
Number of data points	N	n
Mean	μ	$\bar{X}$
Mean estimator	$\hat{\mu}$	$\bar{X}$
Median	$\tilde{\mu}$	$\tilde{X}$
Std.Deviation	σ	s
Std.Devi.estimator	$\hat{\sigma}$	s

The most used is normal distribution that characterise populations and processes for many types of measurement. Likelihood of obtaining value represent the area under the curve and called density. For normal distribution needed mean μ and standard deviation in order completely specify probability model. The peak of the curve located above mean value μ, because probability density is highest around the mean, this shown on **Figure 2** .

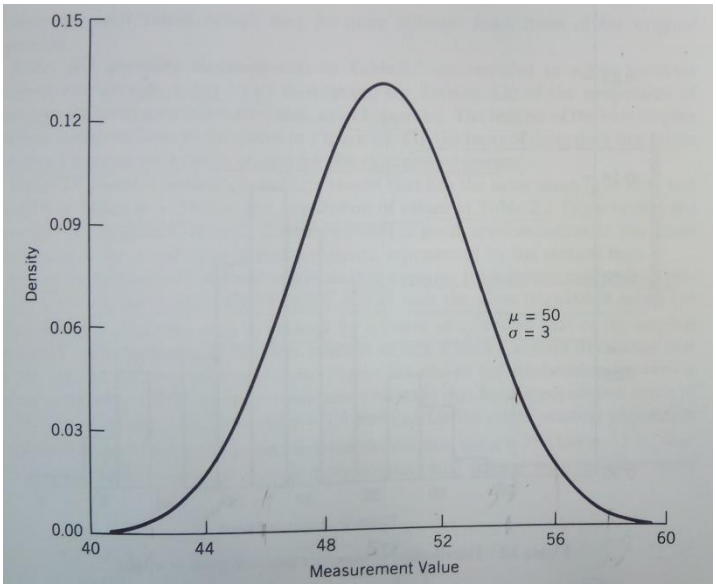


Figure 2 Normal distribution of measurement values *(Robert L. Mason, 1989)*

From normal distribution around 68% of measurement values lie between $\mu \pm \sigma$, 95% between $\mu \pm 2\sigma$ and 99%, between $\mu \pm 3\sigma$.

As was said statistics used sample values to estimate population parameters. For estimating mean of population used sample mean and population standard deviation can be obtained from standard deviation of sample. Several sample statistics can be used to estimate a population parameter.

The laws of statistics also happens in sampling distribution. It means that number of values taken according probability model can be determined by model of original population by the sampling procedure. It leads to definition of a sample distribution. Sample distribution is a theoretical model that describes the probability of obtaining the possible values of a sample statistics.

One of the most important quality of statistics is randomness.

Simple random sample is when every group of items of size n has an equal chance of being selected as the sample. Also (Kreyszig) gives overview of sampling techniques like: random sample with or without releasement, systematic random samples, stratified random samples, cluster sampling. Stratified random samples are based on dividing population into groups or strata of similar units and selecting simple random samples from each strata. It helps to check required observation in several groups in the sample.

Cluster sampling is dividing population on groups of units in such way that leads to randomly sample clusters and sample observations in each clusters. It is used as alternative to simple random sampling when process representing geographical location or lot of products.

This methods of sampling helps to make inferences about a population, process or phenomenon based on the information contained in a representative sample or collection of observations to exact distribution.

This proves one of the features of the normal model. Average from simple random sample of size n follow a normal probability model with the same population mean, but with a standard deviation that is reduced by a factor $\sqrt{n}$.

This feature shown on **Figure 3** , where the mean of probability model is $\mu = 35.4$ and standard deviation of $\sigma_2 = 2.5$ sample size 4 and for individual samples appropriate values: $\mu = 35.4$ and $\sigma = \frac{\sigma_2}{\sqrt{n}} = \frac{2,5}{\sqrt{4}} = 1.25$.

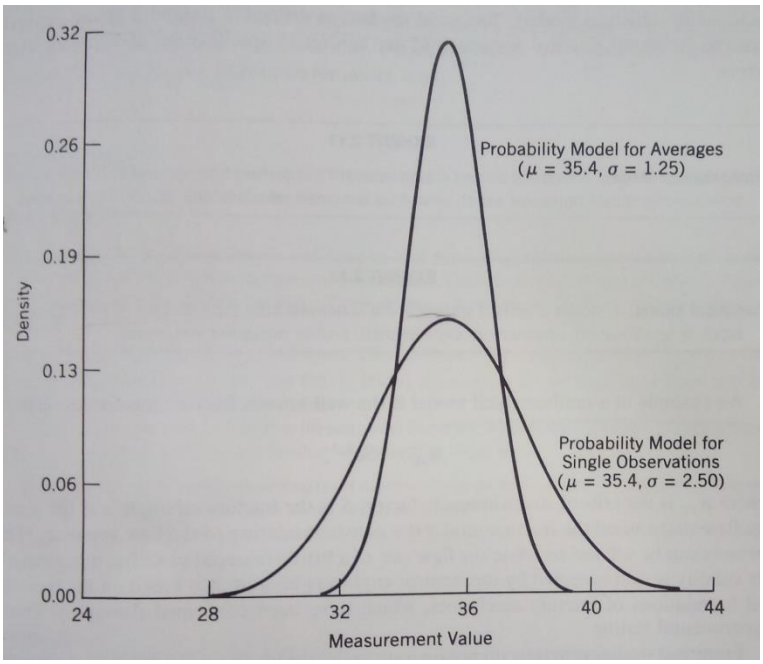


Figure 3 Comparison of normal distributions

It is achievable that the distribution of the averages is more concentrated around the population mean than distributions of individual observations.

It leads to conclusion that to obtain sample mean which is closer to population mean is easier than to obtain one observation which is close to population mean.

Third basic statistical concept is mathematical or statistical modelling. Models are common thing in engineering and physical sciences. Model is based on some knowledge about studying phenomenon. Experiments are conducted to prove or reject models.

Models build in order to characterize one or more response variables, through relationship with one or more factors. Models can be mathematical and statistical. Mathematical is model derived from the theoretical or mechanical considerations, that is based on assumed ideal (error-free relationships among the variables.

Statistical model is model derived from data that subjected to various types of specification, observation and measurement errors.

Example of a mathematical model is facture mechanics relation $k_{IC} = \gamma S a^{1/2}$ facture mechanism relation is based on theoretical foundations of fracture mechanics. Theoretical foundations were confirmed through extensive experimental testing.

But in reality is not always possible to make a mathematical model for mechanism being studied. Empirical studies do not made under the idealized conditions like mathematical model.

In this case statistical model is useful because it is include experimental error. Error can be additive or multiplicative. If to apply it to facture mechanics relation than it will take view like shown in formula (1),

$$\mathrm{K_{IC}} = \gamma \mathrm{Sa}^{1/2} + \mathrm{a} \quad \text{or} \quad \mathrm{K_{IC}} = \gamma \mathrm{Sa}^{1/2} + \mathrm{e} \qquad (1)$$

In formula (1) presence of error shows that model has uncontrolled source of variation. A mathematical model can be really proven with data. One of the best conclusions that experimental data is consistent with particular hypothesis model. Some typical mistakes when data collected over a very narrow range of variables. This make incorrect experimental data consistent with hypothesized model.

That why it is important to make proper experimental design and test mathematical model with experiment.

Statistical model should be based on mathematical model: law or relationship, than to be separate. In other words this type of model generalization can evolve to a theoretical model that adequately describes the studied phenomenon.

Block diagram of model generalization shown on **Figure 4**

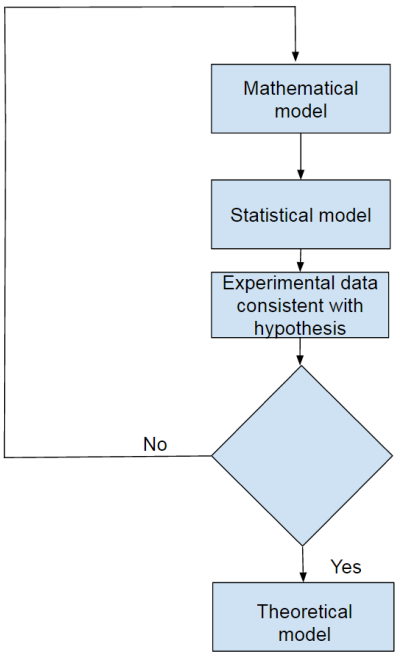


Figure 4 Model generalization loop

4 Statistical methods

Statistical methods divided on two categories descriptive statistics and inferential statistics (L.Jaech, 1985).

Descriptive statistics is some kind of data representation and it includes statistical graphs, charts, tables and indices.

Inferential statistics estimates behaviour of data sets based on behaviour of existing lower data set.

Inferential statistics uses same techniques as descriptive for getting intermediate results from the basis statistical statements about larger population of data.

This work will be focused on inferential statistics. Most important methods of statistical inference are estimation of parameters, determination of confidence intervals, hypothesis testing (application of quality control and acceptance sampling) regression and correlation analysis (Kreyszig).

Mathematical statistics makes conclusions about behaviour of populations by taking random variables, which are called samples, for example 20 parts from a total of 1000 part. Random

selection of the samples are required to obtain meaningful conclusions samples. Each of 1000 parts must have equal chance to be sampled.

Only than the sample mean $\bar{x}$ be a good approximation of the population mean μ. Accuracy of approximation increase with increasing n.

4.1 Descriptive statistics.

According to Rober L. Mason, 1989 descriptive statistics is divided on traditional summary statistics and summary statistics that is less sensitive to outliners in the data it is also called as robust summary statistics.

In traditional summary statistics sample mean is a value by which conclusions about typical behaviour of sample values and like following about behaviour of population values makes. Sample mean or average is a set of data values divided by number of observations,

$$\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n) \quad (2)$$

where $y_1 \ldots y_n$- data values, n-number of observations. However, it is not reliable only to take into consideration only sample mean, spreading of values around mean is important also.

Simplest measurement of spreading is taking into account maximum and minimum data values. This is made by parameter range

$$R = (y_{max} - y_{min}) \quad (3)$$

Values spreading is measured by the sample standard deviation S,

$$s = \sqrt{\frac{\sum(y_n - \bar{y})^2}{n - 1}} \quad (4)$$

, where $(y_n - \bar{y})^2$ squared deviations of sample values from mean. Deviations are squared because only amplitude value of the difference should be considered, square root is taken in order to get identical units with original observation.

Standard deviation is a measure of typical data values variation around sample mean. This parameter often used as measure a precision of a measurement process. Standard deviation gives useful information when compares with standard value such as specification limit or with

values obtained from similar measurements. For example several data sets from similar measurements and standard deviations are compaired. From magnitudes of them can be obtained conclusions about differences in variability of the process, from which were obtained data sets.

Traditional summary statistics uses measures of the centre and spread of a data set. It is allow to get few key statistics, which makes easier to understand large data sets. Sample mean value and standard deviation are properties of interest in the data analysis.

4.1.1 Robust summary statistics.

This type of statistics less sensitive to presence of outliners in the data. Outliners in data can occur when errors in experiment are made, reading mistakes or occasional large or small data occurs. Descriptive statistic methods help to spread few extreme observations.

In this field exists to alternative to sample mean as a measure of centre of data values: sample median and the m-estimator.

Sample median is a number that divides ordered data values into two groups of equal size and determines as follows in **Table 3**

Table 3 Determination of sample median

Number of step	Description of step
1	Order the data from the smallest to the largest values, $y_1 \leq y_2 \leq \cdots \leq y_n$
2	Determine the median as: $M = y_{(q)}$ if n is odd, where q= (n+1)/2 Otherwise $M = (y_{(q)} + y_{(q+1)})/2$ if n is even, where q= n/2

From steps in **Table 3** it is achievable that only 50% of data used in determining sample median.

Next method for determining centre of data set values is m-estimators. M-estimators are weighted averages of data values. For not extreme data values weights are equal to one and for extreme less than one. Formula for determining M-estimator is:

$$m = \sum w_i y_i / \sum w_i \qquad (5)$$

Where w_i – weights of sample values determined by logical expression,

$$w_i = \begin{cases} -\dfrac{tv}{y_i - m} \text{ if } y_i < m < tv, \\ 1 \text{ if } m - tv < y_i < m + tv, \\ +\dfrac{tv}{y_i - m} \text{if } m + tv < y_i. \end{cases} \quad (6)$$

Where t is tuning constant and depends from how strong influence of extreme observations should be covered. Usually taken between 1.345 and 1.5. Robust measurement of experiment v is determined as an absolute median deviation (MAD). MAD determined by formula:

$$MAD = \frac{\text{median}(|y_i - M|)}{0{,}6745} \quad (7)$$

After selecting t and v observation weights of values starting to be assigned. Algorithm for calculating MAD is given in **Table 4** (Robert L. Mason, 1989).

Table 4 Steps for calculation of MAD

Number of step	Description of step
1	Determine the sample median M
2	Calculate the deviations from the median, $y_i - M$
3	Take the absolute values of the deviations
4	Rewrite absolute values of the data from smallest to largest
5	Find median of the ordered absolute values of the deviations, $median(\|y_i - M\|)$
6	Divide $median(\|y_i - M\|)$ by 0,6745

After calculation of MAD m-estimator iterative procedure is summarised in **Figure 4**.

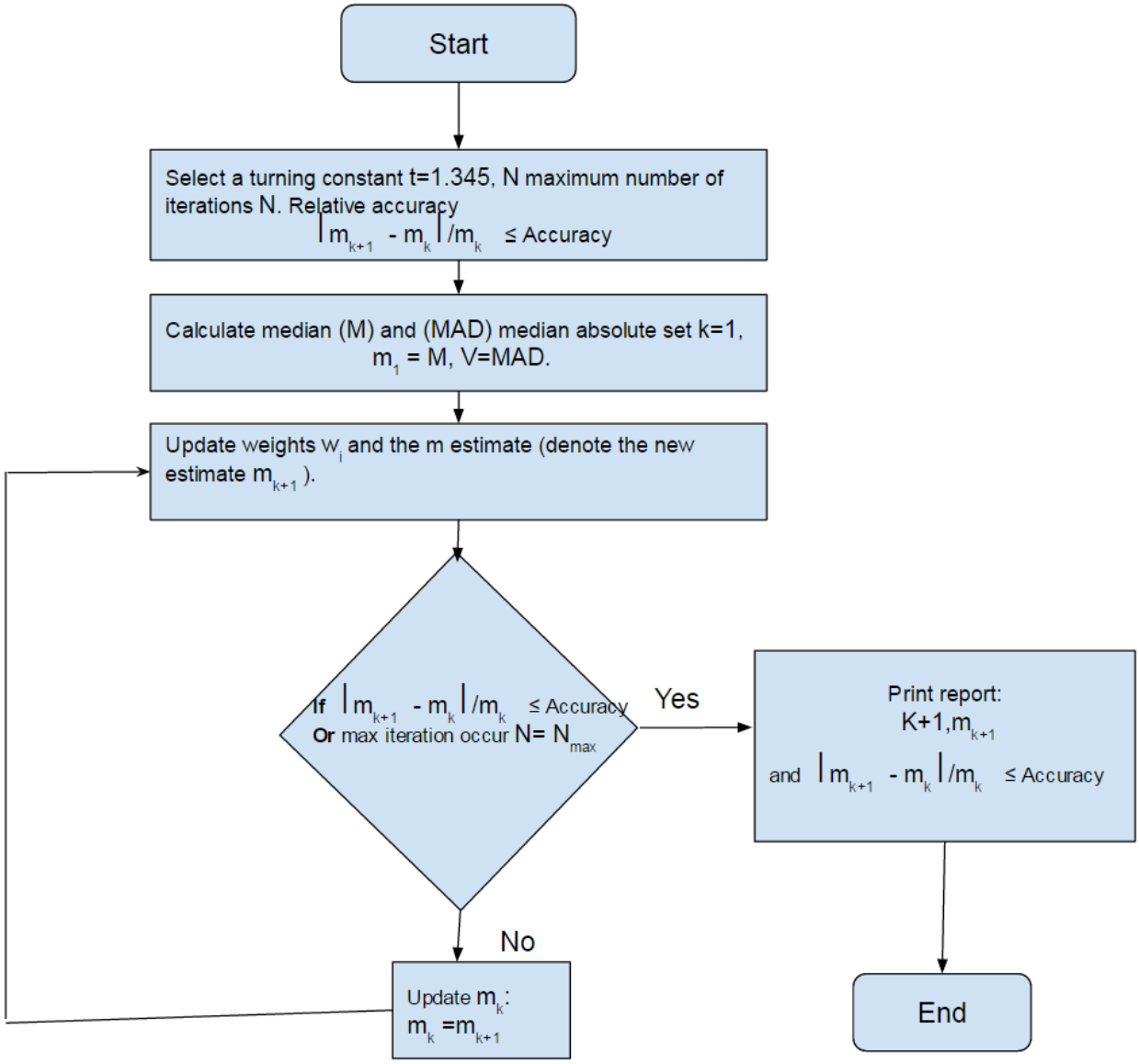


Figure 5 m-estimator iterative procedure

M-estimates method can be used not only for calculating mean of sample but for identification of influential observations. This is achieved by given appropriate weight to the value w_i, weights significantly less than 1 show extreme observations.

Another method of measuring of data values spreading is quartiles. Advantage of them that they are usually unaffected by a few extreme measurements. Quartiles Q_1, Q_2 and Q_3 are numerical values that divide a sample of observations into groups so that each group on 25% less than each previous one. The second quartile Q_2 is the sample median M. After the quartiles selecting procedure calculates semi-interquartile range (SIQR). Steps in order to obtain SIQR described in **Table 5**.

Table 5 Steps for calculating

Number of step	Description of step
1	Order the data values $y_{(1)} \leq y_{(2)} \leq \ldots y_{(n)}$
2	If n is odd then q=(n+1)/2 and if n is even then q=n/2.Then $Q_2 = M = y_{(q)}$ if n is odd, $(y_{(q)} + y_{(q+1)})/2$ if n is even.
3	If q is odd, then r=(q+1)/2, after $Q_1 = y_{(r)}$and $Q_3 = y_{(n+1-r)}$ If q is even, then r=q/2, after $Q_1 = \frac{y_{(r)}+y_{(r+1)}}{2}$ and $Q_3 = \frac{y_{(n+1-r)}+y_{(n-r)}}{2}$
4	Calculate semi-interquartile range (SIQR) is SIQR=$\frac{Q_3+Q_1}{2}$

Advantages of semi-interquartile range method in judging spreading of data around sample mean are: quick to compute in comparing with m-estimation and like m-estimation method less affected by extremes in the data than sample standard deviation method.

Than less SIQR is than less spreader sample values. Weights for the m-estimate shows dispersion of observations around mean, than closer weight to one than it is less dispersed. This observations allows to get information about the factors which influenced on response of experiment.

5.Graphical displays of data

According to Robert L. Mason, 1989 graphical displays of data can be classified on: raw data displays, tabulation and graphical summaries, graphical displays used in statistical process control.

If to look on outcome of experiments data gives information about response and factors of experiment. Raw data display methods makes easier understanding and presenting conclusions from experiment.

5.1 Raw data display methods

Raw data display methods are: point plots, scatter plots and sequence plots. In point plots horizontal axis covering the range of data values, vertical axis shows frequency of data values repeating. This plots allows to see what data values repeated more often. By using only summary statistics: average and standard deviation, it is not possible to see repetition of data values, as shown in figure 6 (Robert L. Mason, 1989).

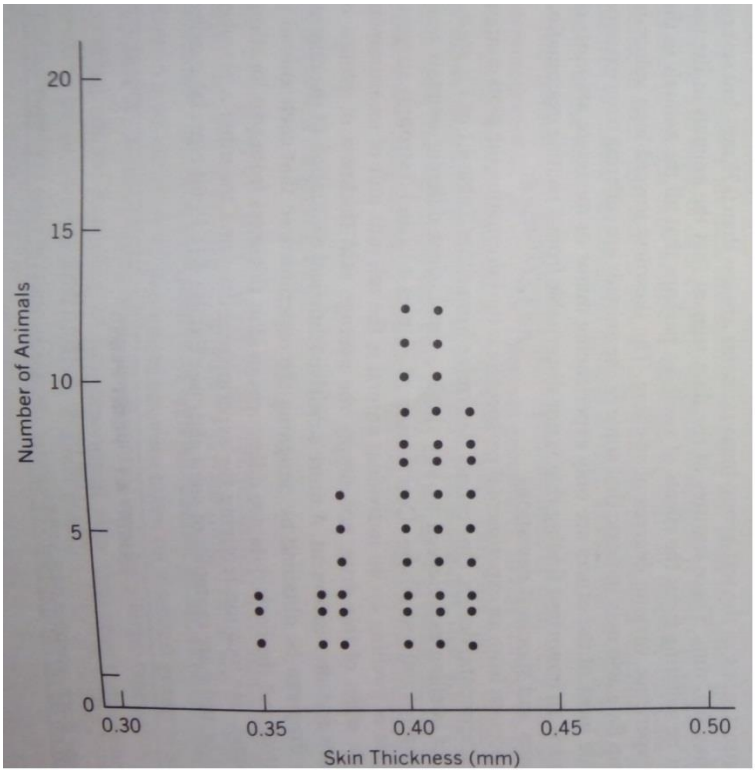


Figure 6 Point plot

For example two groups of values repeated more often than others, it means that distributions of them are different and need to be studied.

Another important raw data display method is scatterplot. Scatterplot depicts horizontal and vertical axes that cover the ranges of the two variables and plot (x_i, y_i) points for each experimental value. It is allows to see extreme values by visual inspection with upper control limits (UCL) and lower control limits (LCL).

Another deviation of scatter plot is sequence plot. Sequence plot is a scatter plot where value numbers placed in chronological ordering. In sequence, plot line which connects two points should connect only successful point, no two points which is out of control limits. Sequence plot view shown on a **Figure 7** (Robert L. Mason, 1989).

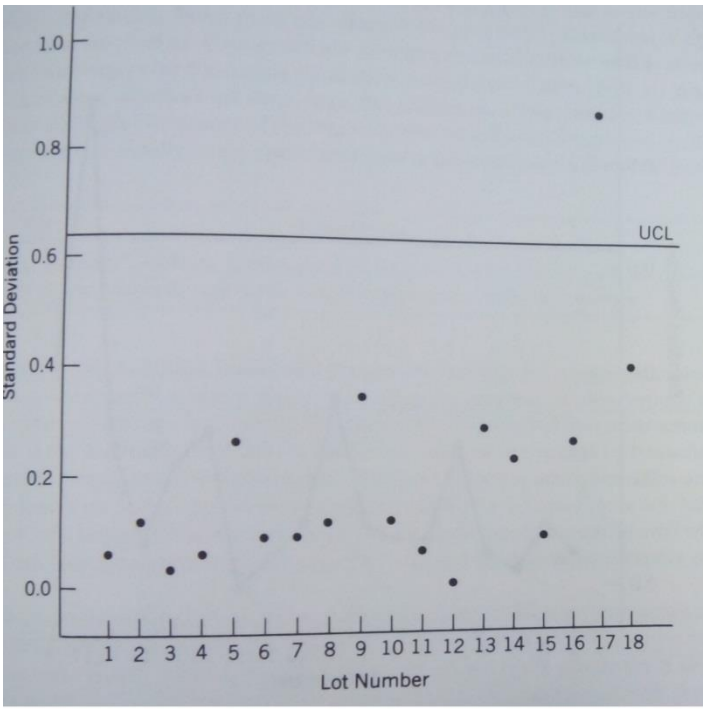


Figure 7 Sequence Plot

All these plots can be modified by labelling of the points, this help to avoid increasing number of axis's and graphs. Point plots are good for displaying a small or moderate size data values.

Next graphical representation of data is tabulating and displaying of distributions.

5.2 Tabulating and displaying of distributions

A distribution is a measure of occurrence data in a population, process or sample. Collection of data distributions is known as a histogram. They are very useful when large amount of a data should be proceed. Histograms are made by dividing the range of data on several intervals, counting the number of observations in each interval and making bar charts of the counts. Number of intervals varies between 8 and 20 depending from number of observations, view of histogram shown on a **Figure 8** (Robert L. Mason, 1989).

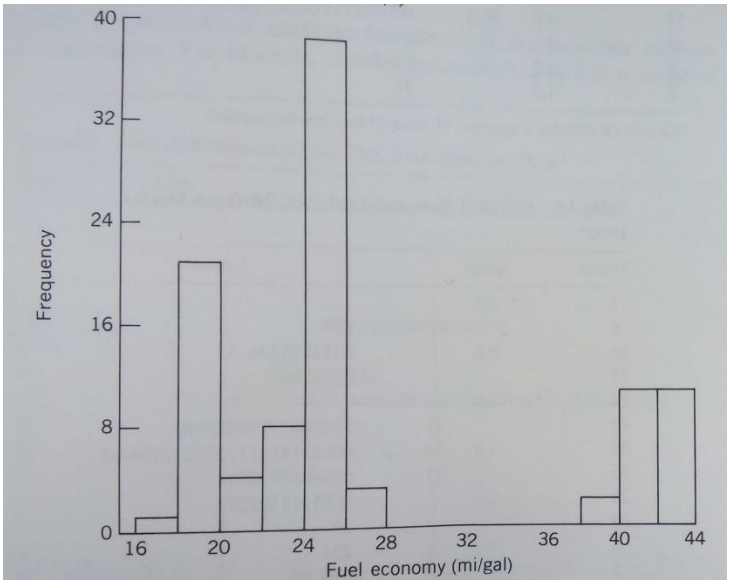


Figure 8 Histogram

When sample size is large it can be useful to make relative-frequency histograms. Another plot that remind histogram is stem and leaf. It represents data range, distribution shape and density. View of steam and leaf histogram shown on Figure 9 (Robert L. Mason, 1989).

Depth	Stem		Leaves
0	1		
0			
0			
1			7
22			88899999999999999999
26	2	Q	0000
34			23333333
39		M	44444444444444444444444455555555555555
27		Q	677
24			
24	3		
24			
24			
24			
24			99
22	4		00000001111
11			22222223333

*Q: interval includes a quartile; M: interval includes the median.

Figure 9 Stem and leaf Plot

Digits are divided into stems and leaves. Stems are leading digits and leaves are trailing digits. The depths of the steam is a represent repetition of values from the closer end of the data.

Depth values shows medians quartiles. Number of leafs can show median value. Location of median and quartiles can be found by formula,

$$LOC = (n + 1)(\frac{P}{100}) \quad (8)$$

Where p – percentage of data covering by quartile, n- number of data values which is the same as the number of leafs. For first quartile Q_1 data covering percentage is 25%, for second it is 50%. Also, second quartile is median of data sample. For third quartile data covering percentage p equal 75%.

In conclusion can be said that steam and leaf plot make data set compact depicted. It uses all advantages of histogram without missing data. Not all digits need to be illustrated. Allows to get fast median and quartiles using stem depth values. Lengths of stem gives impression about number of observation in row. Shape of steam and leaf plot depicts horizontal histogram by which can be made conclusion about type of distribution.

Data from experiment can be represented by box plot, it provides big amount of information about data set.

View of box plot (Robert L. Mason, 1989) is shown in **Figure 10** .

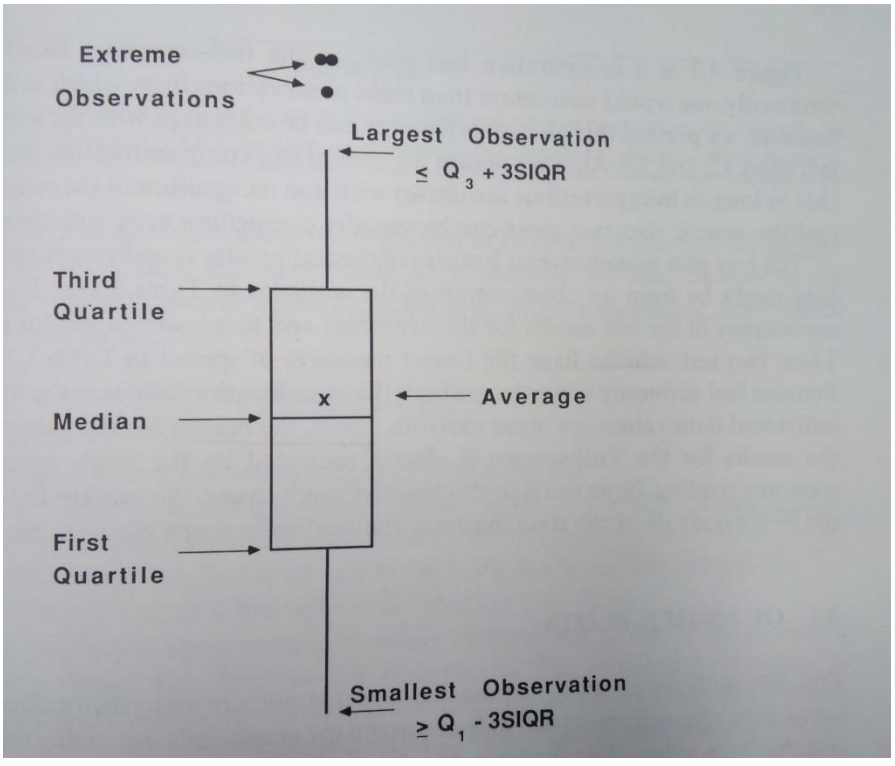


Figure 10 Box plot

Summary of box plot construction is given in **Table 6**.

Table 6 Box plot construction steps

Step	Description
1	Calculate the averages and quartiles of the data sets
2	Calculate the semi-interquartile range, SIQR= (Q_3-Q_1)/2
3	Draw a rectangle with upper and lower boundaries at the third and first quartile.
4	Horizontal line identifying the median an x symbol depicts median
5	Draw lines from the centre of each edge of the rectangle to extreme data positioned no more than 3SIQR from each edge
6	Plot points that lying after 3SIQR border.

Quantile plots takes big part in data representing. These plots display distributional features of data set. They show easily repeated data. Then more dense data then more horizontal will be graph. Quantile plots allows to find median and all quantiles from the plot. View of Quantile plot (Robert L. Mason, 1989) is shown on **Figure 11**.

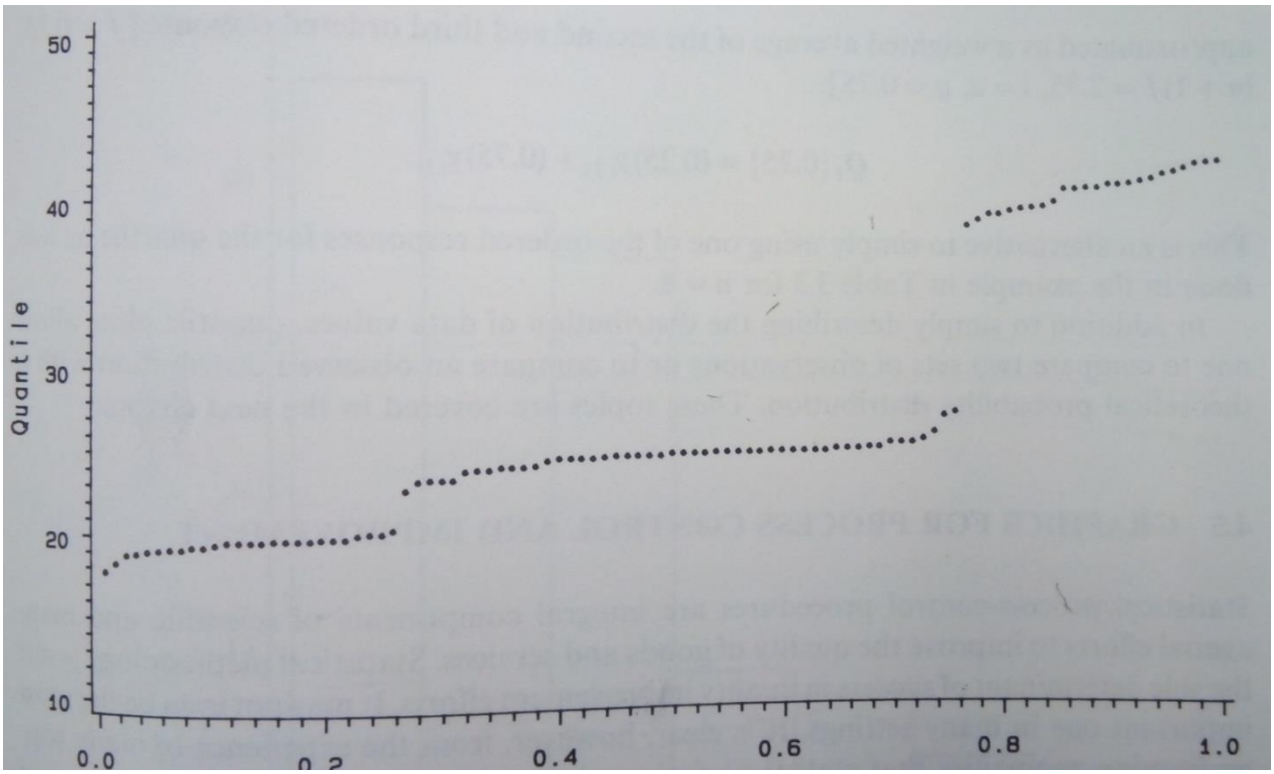


Figure 11 Quantiles plot

Very nice feature is that quantile can be interpolated by formula

$$Q_I\{f\} = (1 - g)Q_I\{f_i\} + gQ_I\{f_{i+1}\} \quad (9)$$

Formula is valid for intervals of data fractions f from (1/n+1) to n/(n+1).

5.3 Graphics for process control and improvement

Managerial decisions are made from quality information obtained from process. Information from process data collected from it can be summarised by graphical and analytical statistical methods.

In order to monitor factors that influence process very important tool is Pareto diagram, it shows problems according to their frequency from lowest to highest. Pareto principle says that three types of defects account over 90% of the visual defects, Pareto chart is shown on Figure 12 (Robert L. Mason, 1989).

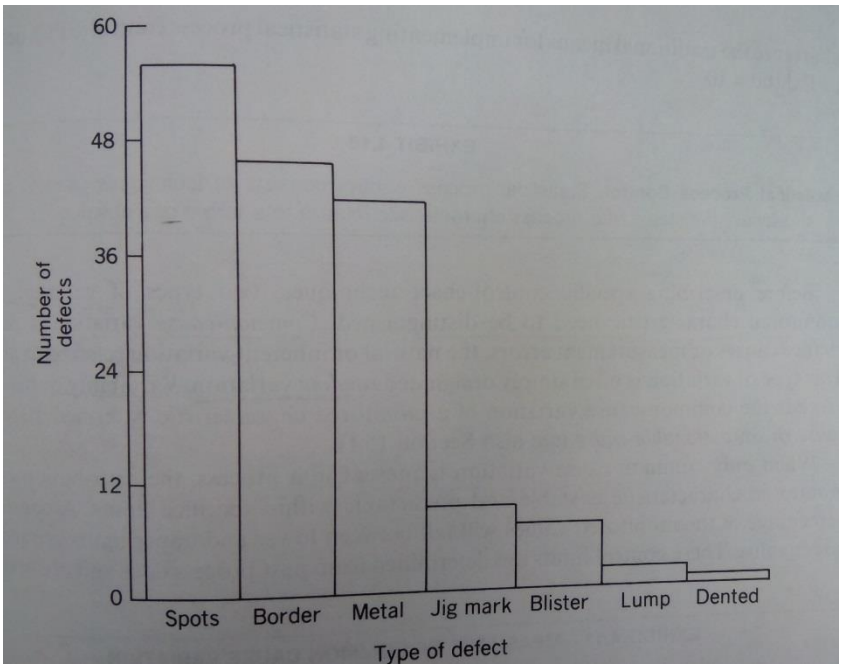


Figure 12 Pareto chart

According to this information manager should discuss and determine why main defects occur (determine factors) in order to arrange continuous procedure for monitoring of process control.

For this purpose control charts are used. Control charts also help to minimize overcontrol and undercontrol of process.

Overcontrol of a process occurs when measurements are made too often and control charts will give information about temporary variations in a process.

Undercontrol occurs when measurements are taken too seldom and process is not controlled in some period of time, as a result periods of off-target operation and increased product variability.

Control chart is sequence plot, with time on horizontal axis and control limits. Control chart is one of the main tools in statistical process control (SPC).

SPC is techniques used to get status of a process characteristic according to target or aim value. Each control chart contain limits, this limits determined by common case variation or random variation. This type variation is determined by quality of machines, mechanism, tools and other factors. When it is present in a process, than it is predictable within specification limit around the target value.

Common case variation or control limits obtained from past process data, which is collected and updated for example every 4 months or when significant changes in a process were made.

Common requirements for monitoring control charts are: a sampling plan, a target value and forecasting of the random variability based on ranges or standard deviations. Sample plans are connected with simple random sample or systematic random sample. Methods for determining random variation described in **Table 7**.

Table 7 Random variation determination

Method	Description
Average-Range Method	Observations for $n > 1$, takes for samples at the i th time period (i=1,2,..k). For each sample calculates range R_i and takes average: $\bar{R} = \frac{R_1 + R_2 + \cdots + R_k}{k}$
Average-Standard Deviation Method	Takes samples with $n > 1$ observations and calculate S_i- sample standard deviation is: $\bar{S} = \frac{s_1 + s_2 + \cdots + s_k}{k}$
Average-Moving range method	In each sample takes only one measurement calculates moving range for the i th sample as: $\overline{MR_i} = \lvert y_i - y_{i-1} \rvert$, i = 2,3….. k. $AMR = \frac{MR_2 + MR_3 + \cdots + MR_k}{k-1}$ Determines estimated standard deviation $S_{AMR} = \frac{AMR}{1.128}$

For having appropriate data, which measure random-variation satisfactorily rules should be satisfied data sampling should cover long-enough time intervals. Sample size should contain at least 30 and 40 samples over two-three months periods.

Most common in industry are Shewhart and cumulative (CUSUM) charts. View of this type of charts shown on **Figure 13**.

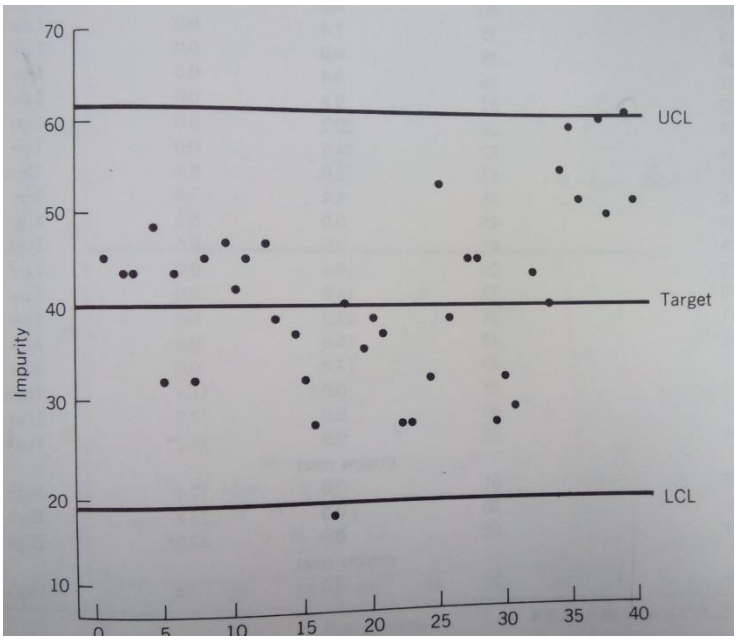


Figure 13 Shewhart Plot

Construction steps for Shewhart control chart is given in **Table 8** constants can be found in (Robert L. Mason, 1989) Appendixes.

Table 8 Algorithm of building Shewhart control charts

Step	Description
1	Take samples according to sample plan, each of k time periods
2	Calculate means ranges and average moving range as described in **Table 7**
3	Calculate UCL and LCL for averages using constants from (Robert L. Mason, 1989) table A2 of the Appendix If $n > 1$: $LCL = target - A_2\bar{R}$, $UCL = target + A_2\bar{R}$, $n = 1$: $LCL = target - 3S_{AMR}$, $UCL = target + 3S_{AMR}$.
4	Upper and lower control limits for dispersion, using the constants in tables A2, A3 of the appendix : Range: $LCL = D_3\bar{R}$, $UCL = D_4\bar{R}$, S.D.: $LCL = B_3\bar{S}$, $= B_4\bar{S}$.
5	Plot, averages, or standard deviation charts on the both. If values outside control limits exists, the process should be studied on reasons of this.

In Shewhart control charts average is used as a target value and centre line. As alternative to average-standard deviation in Stewart chart can be used "pooled" estimation as the measure of random variation, described in formula (10)

$$s_p = \left(\frac{s_1^2 + s_2^2 \ldots + s_k^2}{k}\right)^{1/2} \quad (10)$$

Appropriate appear and lower control limits then calculated as in formula (11)

$$\text{LCL} = \text{target} - \frac{3S_p}{\sqrt{n}} \text{ , UCL= target} + \frac{3S_p}{\sqrt{n}} \quad (11)$$

In comparing with standard deviation "root mean squared" standard deviation makes Shewhart control chart more sensitive to values out of control limits, in order to see this advantage observations in sample should be: $n > 5$, if $n < 5$ then simpler to use average standard deviation in control chart.

Like was said before if some data values goes out of control limits than process should be adjusted to return them back to target.

Another more sensitive to little changes in a process is a cumulative sum charts (CUSUM) charts. In this chart monitored characteristics is an average or dispersion of sample. Cumulative sum (CUSUM) control chart building steps described in **Table 9**

Table 9 Algorithm of building CUSUM control charts

Step	Description
1.	Obtain n random samples
2.	Calculate sample average range using the $\bar{y}_i$; For random variation chart calculate $ln\, S_i$,for each sample and the moving range using the $ln\, S_i$.
3.	Calculate the "Sum high" (SH) and "Sum low" (SL) statistics for the i th sample; where Z_i is interested characteristic ($\bar{y}$ or S_i). $SH_i = SH_{i-1} + [Z_i - (target + 0.5S_{AMR})]$, $SL_i = SL_{i-1} + [(target - 0.5S_{AMR}) - Z_i]$, Note that $SH_0 = SL_0 = 0$ if at any time SH_i or SL_i is negative, it is set to zero.
4.	Calculate UCL and LCL: $SH_i > 4S_{AMR}$ or $SL_i > 4\, S_{AMR}$ If then investigation for case is conducted and appropriate correction in a process conducted, than reset cumulative sum. It can be making zero SH_i or SL_i consequence of correction actions will be seen in next sample observations.
5.	Plot max $\{SH_i; SL_i\}$ with labels H and L for sums and upper control limit at 4 S_{AMR}.

View of cumulative sum diagram shown on **Figure 14** (Robert L. Mason, 1989).

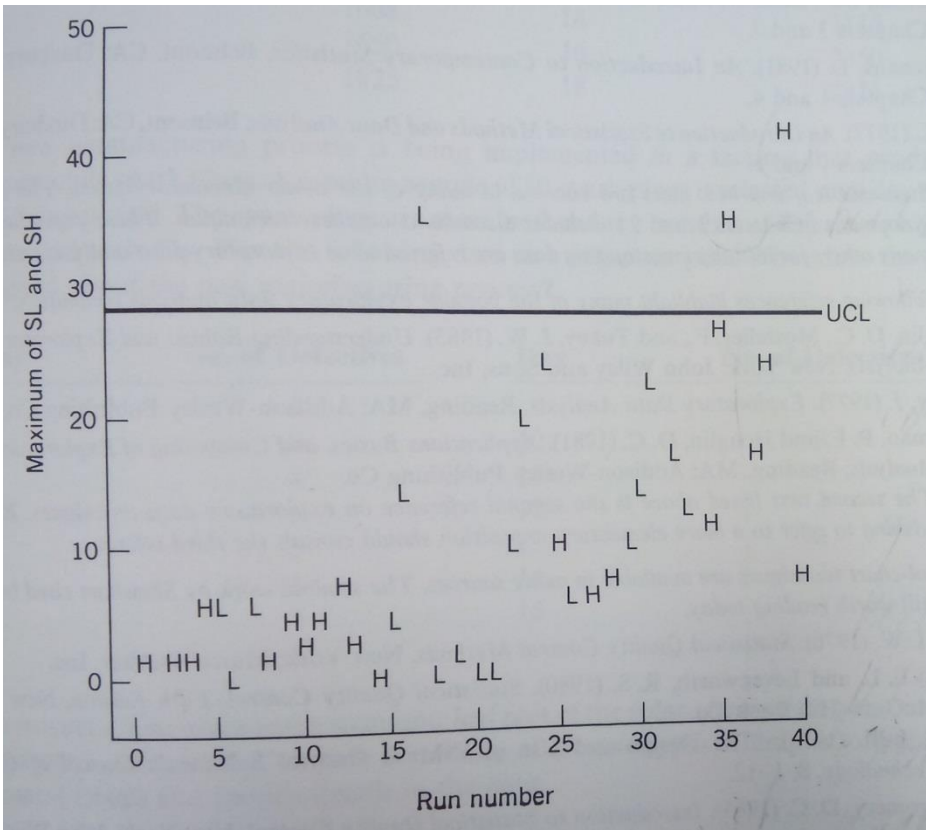


Figure 14 CUSUM Plot

The main feature of CUSUM control charts is that they show successive deviations of the process characteristics from point value - $target \pm 0.5S_{AMR}$. Like was mentioned before difference in the Shewhart and the CUSUM charts is that last one is more sensitive to random variation. Also do not need to be updated because of reference point step and values after correction. Shewhart control charts are easier in construction than CUSUM, but needs updating of control limits.

5.4 Graphical Comparison of distributions

In this chapter will be raised up questions of: box plot comparison of two sample distributions, quantile plot comparison of two sample distributions, comparison of sample distribution with a theoretical reference.

For analysing experimental results by graphs can be used box and quantile plots.

5.4.1 Comparison by using box plots

Some times response from experiment has influence of several factors on output of experiment. It is difficult to separate influence of several factors on output of experiment. By another words response of experiment will be function of several factors. In this case comparative box plots will be very useful. In order to show application of box plots will be used example of Chemical measurements from an infrared scanning instrument from (Robert L. Mason, 1989) table 5.1.

In this experiment instrument for measuring chemical properties of industrial liquids was tested. Were determined factors that are influence on measurements of chemical response by this device. This factors are summarised in Table 10, this factors are potential sources of variability.

Table 10 Factors of interest

Number	Factor
1	Chemical concentration (32%,36%)
2	Sample preparations (samples were prepared for each concentration)
3	Amount of scans per observation
4	Repeatable measurements for each sample mix

From examination of table 5.1 from (Robert L. Mason, 1989) main conclusion that, than higher concentration of chemical than higher response of measurement value. Samples and amount of scans gives some variability, but not significant. Because of concentration effect and number of scans was made box plot which is shown on Figure 15 .

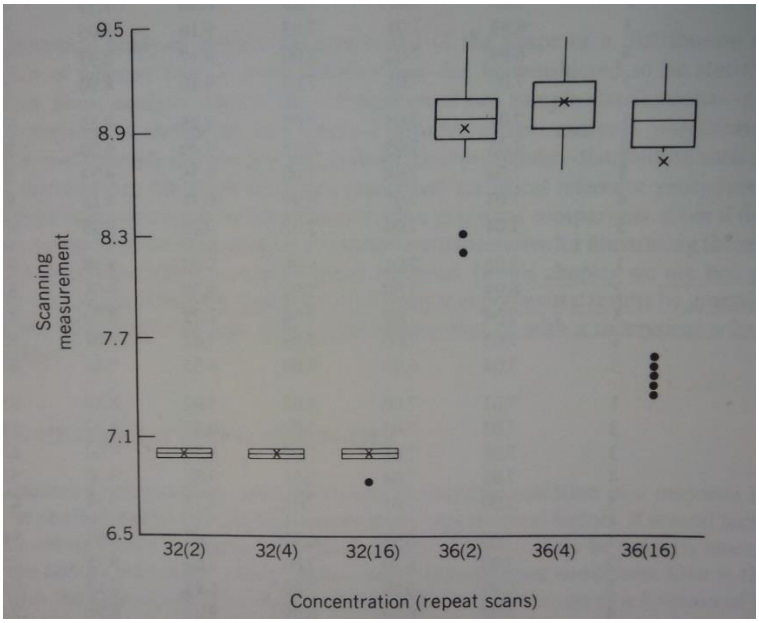


Figure 15 Box plot of measurements *(Robert L. Mason, 1989)*

From graph can be seen fluctuation of data by size of box and that average value of scanning measurement not depend from repetition of scans.

From graph achievable that for 32% and 36% concentration measurement groups average and median approximately same.

In 36% concentration measurement group exist extreme observations for amount of scans 2 and 16.

For determining factors that influence on extreme values, ware made another two graphs.

One graph it is dependence of average value from number of repeat scans **Figure 16** other is graph of standard deviation from number of scans **Figure 17**. First flashing values in sample for standard deviation of third value for number of repeated scans 2 and 4 for 36 % concentration.

By checking plots on figures: **Figure 16** and **Figure 17** can be made conclusion that this extreme values are sudden and not connected with factors of experiment. It is proved by low variability of average values for 2 and 4 number of repeated scans on **Figure 16**, and low standard deviation on **Figure 17**. So, can be made conclusion that such extreme values are sudden and not because of factors from **Table 10**.

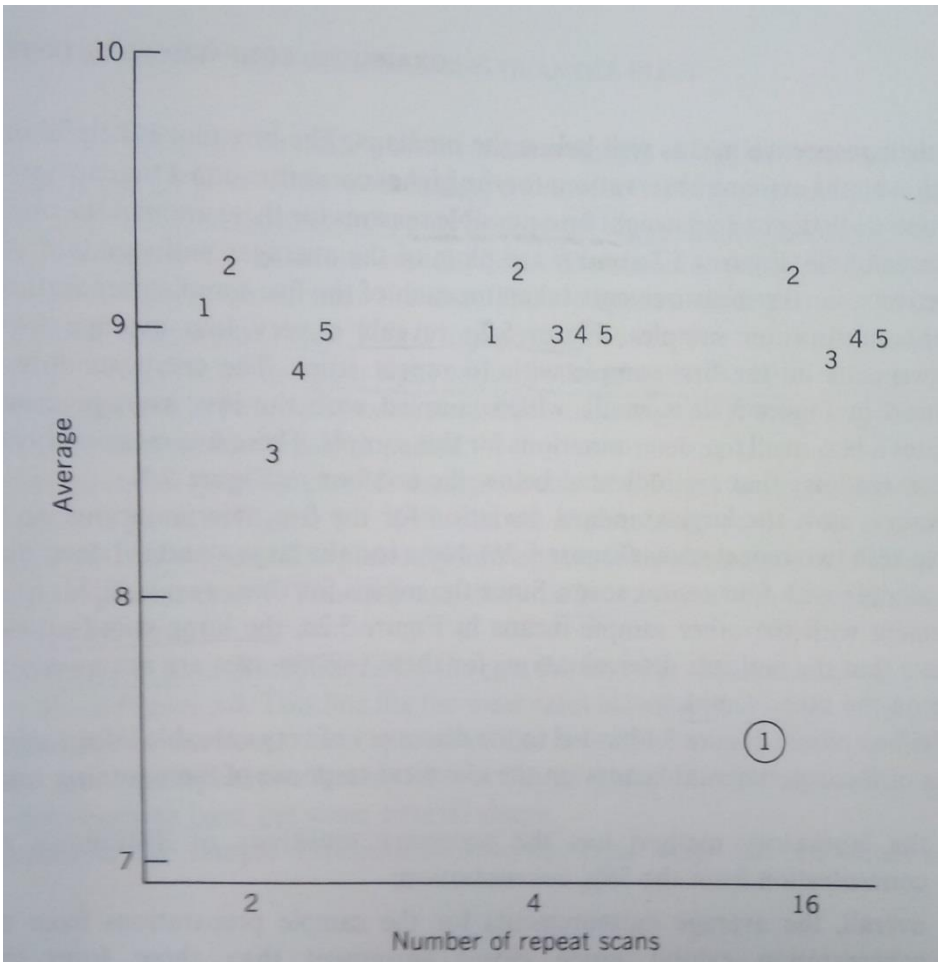


Figure 16 Average values versus number of repeat scans *(Robert L. Mason, 1989)*

In conclusion can be said that: box plot is very useful for discovering information about the effect of experimental factors on response of instrument.

It showed that device has good sensitivity in order to differ 32% concentration from 36% concentration. Average measurements for 32% concentration is closer than 36% concentration to mean. It means that precision is affected by concentration.

Variation in the five measurements for 3 rd mix a for 2 and 4 scans significantly exceeds the variation than in all other mixtures **Figure 17**.

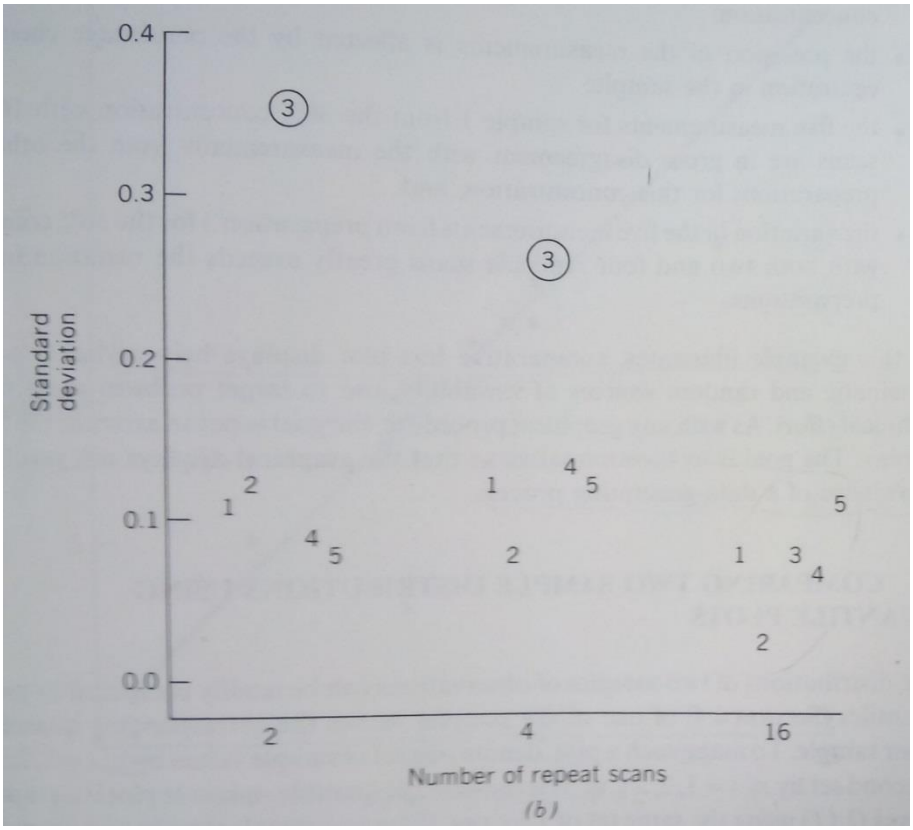


Figure 17 Standard deviation versus number of repeat scans *(Robert L. Mason, 1989)*

However, checking average values for same numbers of repeated scans, can be made conclusion that such extreme values are sudden and not because of factors from **Table 10**. This is proved by low variability in average values for number of repeated scans.

5.4.2 Comparison of two sample distributions using Quantile plot

By plotting quantiles of one sample versus corresponding quantile of another sample to quantiles distributions can be compared. If two sets of sample values obtained: set n with y_i and $i = 1,2 \dots n$ and m with x_i, $i = 1,2 \dots m$, than two-sample quantile-quantile plot is graph of Q_y to $Q_x(f)$ using same set of f –values.

If m=n, sizes of samples equal than a plot will be from pairs of observations $(x_{(i)}, y_{(i)})$. Example of such plot shown on Figure 18 from (Robert L. Mason, 1989).

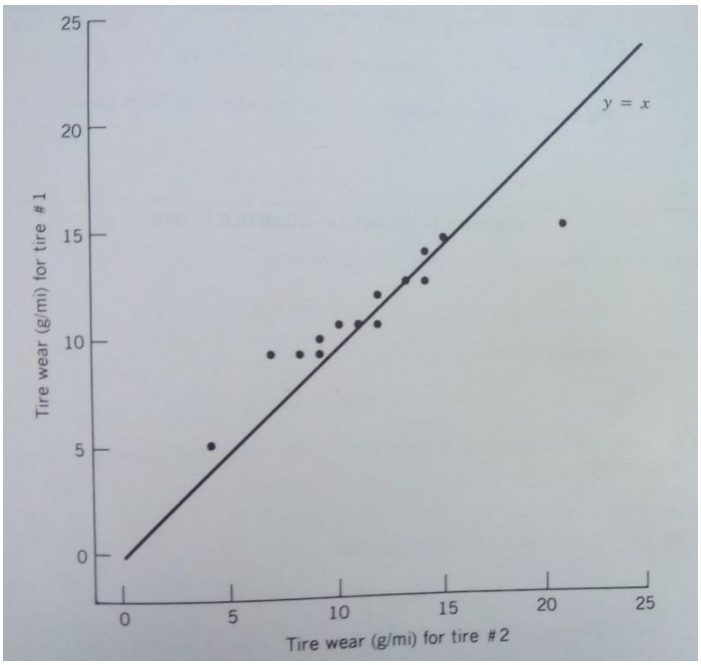


Figure 18 Quantile-quantile plot for wear tiers of same brand

On Figure 18 shown tire wear of two tiers from one manufacturer. The line y=x fit them, except of one point. One point can be extreme because of unexpectedly sharp break by driver.

A measurement of tire wear plot from two manufacturer given on **Figure 19** (Robert L. Mason, 1989). By dashed line denoted two tire brands line and by line y=x one brand tiers.

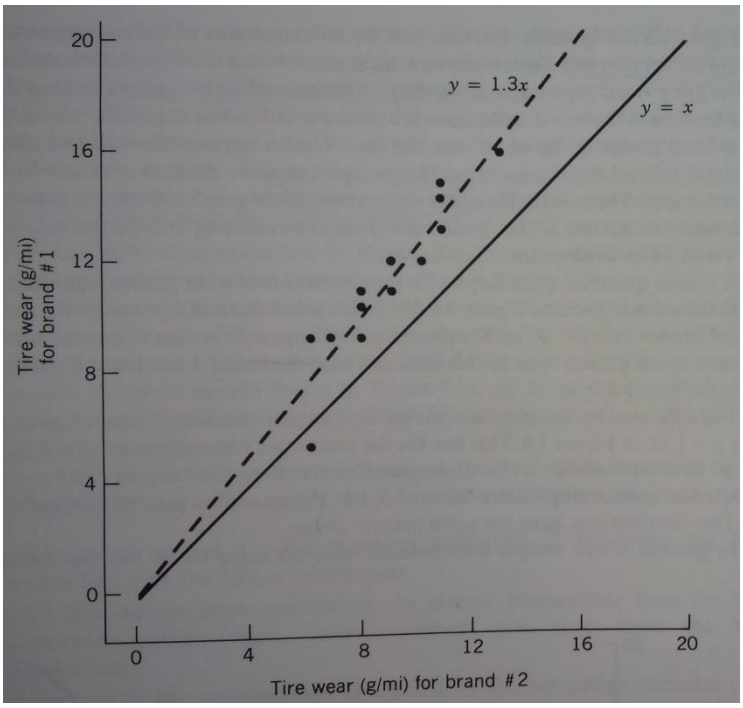


Figure 19 Quantile-quantile plot for wear tiers of different brand

It is achievable that two distributions has same shape nut estimates of the center spread are different on 1.3 multiplicative factor. Quantile-quantile plot gives conclusion that two sample distributions has same shape, but differ by addition or multiplicative constants. It can be described by straight line equation,

$$y_{(i)} = a + bx_{(i)} \quad (12)$$

For quartiles view equation can be rewrite by

$$Q_y\{f\} = a + bQ_{(x)}\{f\} \quad (13)$$

If two distributions do not have the same shape, then quantile-quantile plot is not a linear-line equation. If two data sets are not equal size, than data points of smaller sample plot versus bigger sample. The procedure for making quantile-quantile plots shown in a Table 11 from (Robert L. Mason, 1989).

Table 11 Algorithm of building Quantile-quantile plot

Step	Description
1	Mark the smaller sample values by $y_{(n)}$ and order then in: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$, order. Larger sample mark by: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(m)}$
2	Find interpolated quantile $x`(f_i)$ for the data fraction n of smaller sample n, $f_i = i/n$; If n=m, $x`(f_i) = x_{(i)}$; If n<m, set h = (m + 1) f_i , then $x`(f_i) = (1-g)x_{(k)} + gx_{(k+1)}$,where k is the integer portion of h, g = h − k. $[If\ k \geq m,\ x`(f_i) = x(m).]$
3	Plot $Q_y\{f_i\} = y_{(i)}$, versus $Q_x\{f_i\} = x`(f_i)$, $i = 1,2,\ldots,n$.

In conclusion can be said that distribution quantile-quantile plots gives visual information about distribution character of two data sets, shifts in locations can be determined from coefficients of linear equation. Also quantile-quantile plots can be used for comparing observed distribution with a theoretical reference distribution.

5.4.3 Comparison with a reference distribution

Like was said before quantile plots can be used in comparing a sample distributions. Robert L. Mason, 1989 described method of using quantile-quantile plot for comparing of it with

reference distribution. Than closer quantile-quantile plot to straight line, than closer sample distribution to reference distribution.

Normal distribution has higher density around mean values and standard deviation σ shows spread of the values. Normal probability distribution described by density function:

$$f(y) = {}^{-1/2}\sqrt{(2\pi\sigma^2)}\exp\{-(y-\mu)^2/2\sigma^2\} \quad (14)$$

Where y is measurement value, μ – population or process mean, σ – population or process standard deviation. View of it shown on **Figure 20** (Robert L. Mason, 1989).

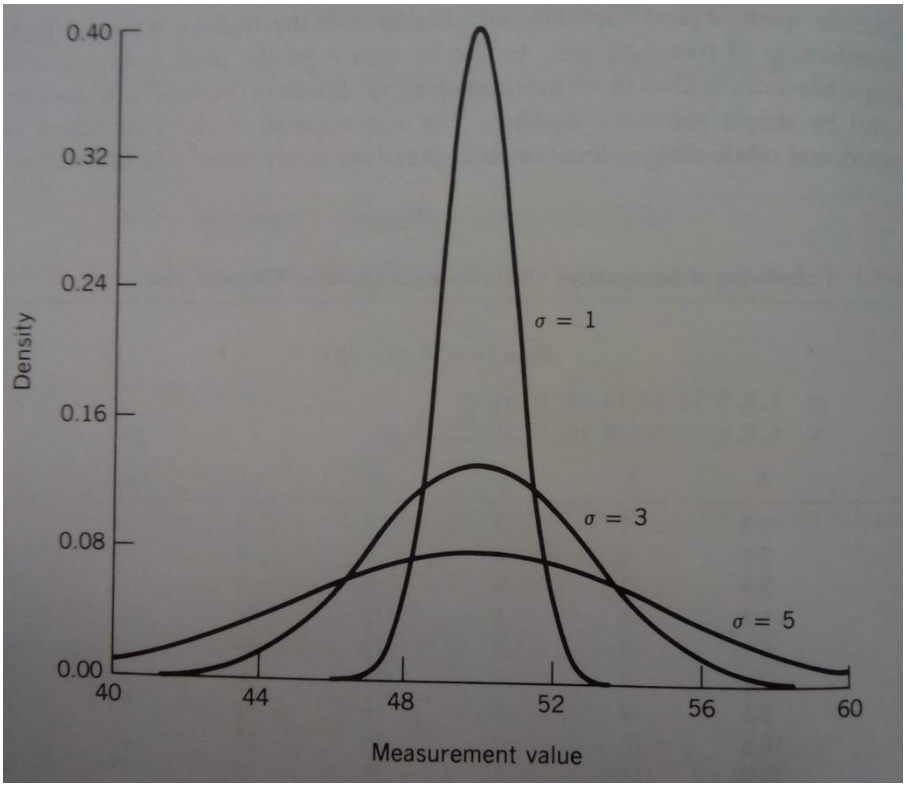


Figure 20 Normal distributions with means 50 and diff. standard deviations.

With approximation that $\mu = 0$ and $\sigma = 1$ normal quantile function has view:

$$Q_{SN}\{f\} = 4{,}91\ \{f^{0.14} - (1-f)^{0.14}\} \quad (15)$$

In order to obtain quantile plot for any mean μ and standard deviation σ can be used formula

$$Q_N\{f\} = \mu + \sigma Q_{SN}\{f\} \quad (16)$$

The algorithm of plotting a quantiles of sample observations against theoretical normal quantiles described in a table:

Table 12 Algorithm of making normal quantile-quantile plot

Description	Description
1	Order data values: $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$
2	Set $Q_y\{f\} = y_{(i)}, i = 1,2, \ldots, n$
3	Calculate $Q_{SN}\{f\}$ for $f_{(i)} = (i - 3/8)/(n + \frac{1}{4})$ which is $Q_{SN}\{f\} = 4{,}91[f_i^{0.14} - (1 - f_i)^{0.14}]$
4	Plot $Q_y\{f\}$ against $Q_{SN}\{f_i\}$ If quantile depicted linearly than the data consistent with a normal distribution.

As example let take **Figure 21** (Robert L. Mason, 1989).

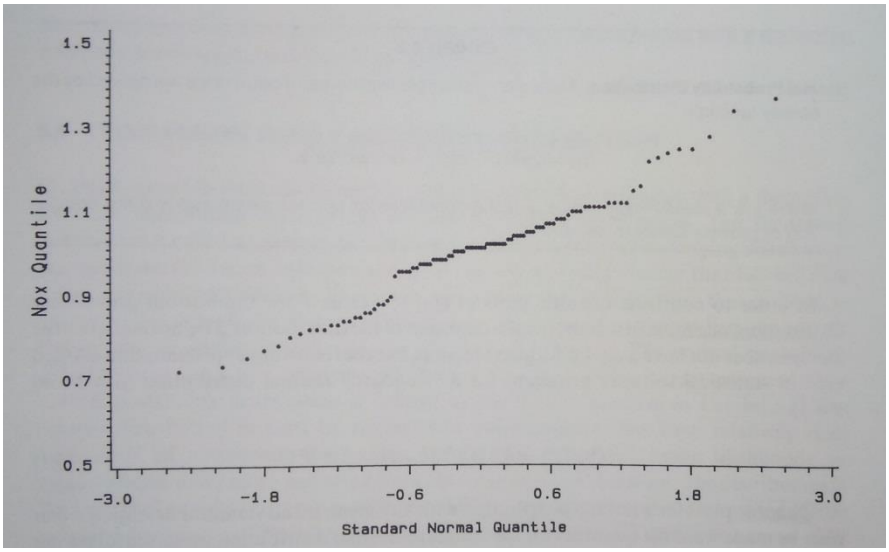


Figure 21 Normal quantile-quantile plot example 1

From first view can be made wrong conclusion that quantiles from sample consistent with a normal distribution, but it is not like this, it is three normal distributions for different types of cars. It contains two breaks in plot.

Figure 22 shows that quantiles of samples do not follow linearity like a normal distribution.

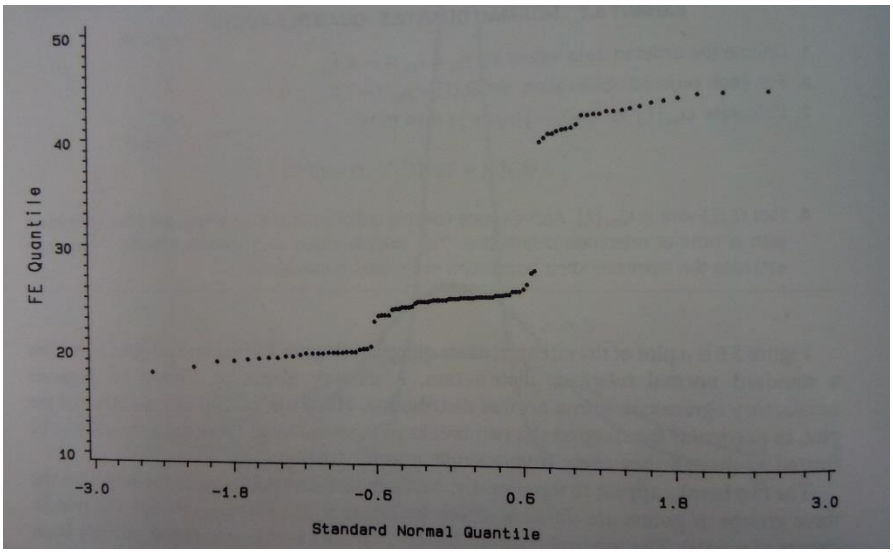


Figure 22 Normal quantile-quantile plot example 2

In conclusion can be said that similar procedure with normal quantile plot can be applied for other reference distributions.

This method do not requires normal probability paper.

5 Regression and correlation analysis

From (Kreyszig) regression analysis is case when one of two variables can be considered as an ordinary X and another variable is random and interest is dependence Y from X. This analysis applied when X can be measured with low error. Examples of regression are dependence of blood pressure Y on the edge of a person X, gain of weight Y from daily ration of food X.

In correlations, both quantities X and Y are random variables and the goal is to find relation between them. Examples of correlations are: X and Y wear of the cars front tyres, hardness of steel X in the centre and hardness Y near the edges of the plate.

5.1 Regression analysis technique

For quality regression model analysis exist different phases, they are described in (Robert L. Mason, 1989) they are: investigate, specify, estimate, asses, select (ISEAS). Summary of regression analysis given in Table 13

Table 13 Strategy of regression analysis

Step	Description
Investigate	Data searched for outliners (negative values, decimal point in wrong position).This will decrease effect of outliners. Calculated summary statistics, plotted variables.
Specify	Format regression model(polynomial, linear, exponential). Graphs can help in it. Formulate initial model. Reexpressing variables if needed.
Estimate	Estimate model parameters using software, calculate statistics which summarise adequacy of the fit.
Asses	Check are assumptions correct. If model fit data can errors be considered as normal distributed with zero means and constant standard deviations.
Select	Select statistically significant predictor variables.

During investigation stage response variables gives not linear trend. It means that in Specify step we need to make appropriate regression model. Box and steam and leaf plot will give good information about outliners in a single variables. In assessment step for testing of normal distribution of errors uses: quantile-quantile plot. But graphical assessment not so precise like Sapiro-Wilk test for normality. It is used for sample sizes below fifty. If sample size higher than fifty than should be used Kolmogorov-Smirnov test and Anderson-Darling test.

5.2 Linear Regression analysis

The sample regression line formula,

(17)

$$y - \bar{y} = k_1(x - \bar{x})$$

Where k_1 regression coefficient of the sample and determined by,

(18)

$$k_1 = \frac{S_{xy}}{S_x^2}$$

Where sample covariance's S_{xy} and S_x^2 are given by

(19)

$$S_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)\left(y_j - \bar{y}\right)$$

and

(20)

$$S_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right]$$

Formula (17) can be obtained by applying Least Square principle and assumption that x - values in sample are not equal to a straight-line equation:

(21)

$$y = k_0 + k_1 x$$

straight line must go through point such way that sum of squared distances should be minimum in vertical direction (Y). Like following the sum squares of distances is determined by,

(22)

$$q = \sum_{j=1}^{n}\left(y_j - k_0 - k_1 x_j\right)^2$$

Than by derivation (22) with respect to k_0 and k_1 and equalling derivatives to zero gives:

(23)

$$k_0 n + k_1 \sum x_j = \sum y_j$$

$$k_0 \sum x_j \quad + k_1 \sum x_j{}^2 = \sum x_j y_j$$

Dividing first equation of (23) by n and using formula of mean value got $k_0 = \bar{y} - k_1 \bar{x}$ substituting it in (17) gives (18). By solving equation (23) with Cramer's rule can be obtained formulas (18),(19),(20).

5.3 Confidence Intervals in Regression Analysis

From (Kreyszig) in order to get confidence interval for regression should be made assumptions about of distribution of random variable Y. For this assume that for each fixed x the random variable Y is normal with mean $\mu(x) = k_0 + k_1 x$ and independence of sampling.

Under this, assumptions can be obtained confidence interval for k_1. Determination of it shown in Table 14.

Table 14 Determination of confidence interval for linear regression

Number of Step	Description of step
1	Choosing of a confidence level γ (95%,99% or others)
2	Determine the solution c of the equation $F(c) = 1/2(1+\gamma)$, from the table of the t-distribution with n-2 degrees of freedom.
3	Using a sample $(x_1, y_1), \ldots, (x_n, y_n)$ compute $(n-1)\,S_x{}^2$ from (10) $(n-1)\,S_{xy}$ from (9), k_1 from (8) $$(n-1)s_y^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2$$ and $$q_0 = (n-1)\left(s_y^2 - k_1^2 y s_x^2\right).$$
4	Compute $K = \sqrt[c]{\frac{q_0}{(n-2)(n-1)s_x^2}}$. The confidence interval is $CONF_\gamma\{k_1 - k \le k_1 \le k_1 + k \quad\}$

5.4 Correlation analysis

Like was said before correlation analysis is connected with the relation between X and Y in a two-dimensional random variable (X,Y). A sample consists of n ordered pairs of values (x_1, y_1), (x_n, y_n). The interrelation between x and y values in the sample is measured by the sample covariance Sxy or by the sample correlation coefficient:

(24)

$$r = \frac{Sxy}{SxSy}$$

Sx and Sy – given by formula (10) and (9), sample correlation coefficient Sxy (9).

Sample correlation coefficient is positioned in interval $-1 \leq r \leq 1$, $r = \pm 1$ if sample values lies on the straight line. Examples of samples with different correlation coefficient given on Figure 23.

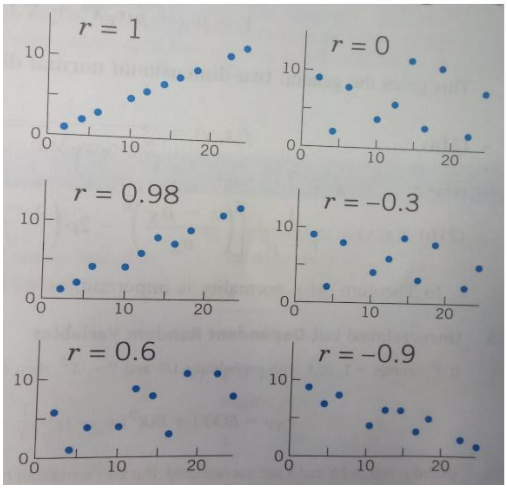


Figure 23 Examples of sample correlations from *(Kreyszig)*

Than for population, correlation coefficient denoted by ρ will be

(25)

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

, where $\mu_x = E(X)$, $\mu_y = E(Y)$, means of marginal distribution and $\sigma_x = E([X - \mu_x]^2)$ and $\sigma_y = E\left([X - \mu_y]^2\right)$ variances of the marginal distribution of X and Y, $\sigma_x \sigma_y$ covariance of X and Y given by

(26)

$$\sigma_x \sigma_y = E\big([X - \mu_x][Y - \mu_y]\big) = E(XY) - E(X)E(Y)$$

From theorem about correlation coefficient for population, the correlation coefficient ρ satisfies $-1 \leq \rho \leq 1$, and $\rho = \pm 1$ if and only if X and Y are linearly related, that is $Y = \gamma X + \delta, X = \gamma^* Y + \sigma^*$.

If X and Y are uncorrelated than $\rho = 0$.

Another theorem says that: independent X and Y are uncorrelated and If (X,Y) is normal then uncorrelated X and Y are independent.

For testing of the Correlation Coefficient of population (Kreyszig) gives algorithm which contained several steps and described in Table 15. Where t is an observed value of random variable that has t distribution with n-2 degrees of freedom.

Table 15 Test of hypothesis $\rho = 0$ with alternative $\rho > 0$ in the case of Two-Dimensional Normal Distribution

Number of steps	Description of step
Step 1	Choose a significance level α (5%, 1%)
Step 2	Determine the solution c of the equation $P(T \leq c)$=1- α
Step 3	Compute r from (14) using a sample $(x_1, y_1), \ldots, (x_n, y_n)$
Step 4	Compute $t = r\sqrt{\frac{n-2}{1-r^2}}$ If $t \leq c$, accept the hypothesis others $t > c$, reject hypothesis.

Problem definition for thesis phase number 2

In part 2 of master Thetis will be applied regression analysis techniques to experimental data from overhead transmission line facility at Xuefeng Mountain belonged to Chongqing University, China. Regressions between environmental factors and electrical, mechanical response variables of experiment will be founded. Will be made appropriate graphical display of the data and developed regressions. Project Gantt chart shown in Attachment 1.

References

ABB. (n.d.). *DC motors type DMI.* ABB.

Bhaskar K, S. S. (2012). AWNN assisted wind power forecasting using feed forward neural network. *IEEE trans Sustain Energy*, 306.

Ekelund, T. (2000). Yaw control for reduction of structural dynamics loads in wind turbines. *Journal of wind Engineering and Industrial aerodynamics*, 241-262.

Farr, H. H. (1980). *Transmission Line Design Manual.* Denver,Colorado: United States Department of the interior Whater and Power Resource service.

Kreyszig, E. (n.d.). *Advanced engineering mathematics.* Wiley international edition .

L.Jaech, J. (1985). *Statistical Analysis of measurement errors.* An Exxon Monograph .

Liebherr. (2017). *Components for wind turbines.*

Michael L. George, D. R. (2005). *The Lean Six Sigma Pocket Toolbook.* New York: McGraw-Hill.

Nordkraft AS. (n.d.). *Nordkraft*. Retrieved from http://www.nordkraft.no/kraftverk/nygardsfjellet-vindpark-article368-110.html

Poots, G. (1996). *Ice and snow accertion on structures.* New York: John Wiley & Sons Inc.

Rice, J. A. (1987). *Mathematical statistics and data analysis.* California: Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove.

Robert L. Mason, R. F. (1989). *Statistical Design and analysis of experiments.*

S. Stubkier, H. J. (2014). Analysis of load reduction possibilities using a hydraulic soft yaw system for a 5 MW turbine and its sensitivity to yaw-bearing friction. *Engineering Structures*, 123-124.

S. Stubkier, H. P. (2014). Analysis of load reduction possibilities using a hydraulic soft yaw system. *Engineering Structures*.

Simens. (2015). *Wind Turbine SWT-2.3-93.* Hamburg: Simens.

Simens. (2015). *Wind Turbine SWT-2.3-93.* Hamburg.

Tinghui Ouyang, A. K. (2017). Predictive model of yaw error in a wind turbine. *Energy*, 119-130.

Wolfson, R. (2012). *Essential University physics.* Pearson.

Xiang-jun Zenga, X.-l. L.-z.-t. (2011). A novel thickness detection method of ice covering on overhead transmission line. *International Conference on Advances in Energy Engineering.*

You-le Liu, B. C.-n. (2008). *Automation of electric power systems.*

Attachment 1

Field data processing techniques

UIT campus Narvik

Today's Date: 23.01.2017 mandag (vertical red line)

Project Lead: Ivan Balachin

Start Date: 03.10.2016 mandag

First Day of Week (Mon=2): 2

WBS	Tasks	Task Lead	Start	End	Duration (Days)	% Complete	Working Days	Days Complete	Days Remaining
1	**Writing report**		2.12.15	4.17.17	796	0 %	##	0	##
1.1	Prestudy report part 2		10.3.16	12.4.16	63	0 %	45	0	63
1.1.1	Literature review		10.3.16	12.2.16	61	0 %	45	0	61
1.1.2	Preperation for presentation		2.12.15	2.16.15	5	0 %	3	0	5
1.1.3	Presentation Part 1		1.10.17	1.10.17	1	0 %	1	0	1
1.1.4	Delivery of prestudy report part 2		1.27.17	1.27.17	1	0 %	1	0	1
2	**Prject part 2**		1.27.17	4.17.17	81	0 %	57	0	81
2.1	Calculating of summary statistics of data		1.27.17	2.16.17	21	0 %	15	0	21
2.1.1	Graphical representation of data		2.16.17	3.8.17	21	0 %	15	0	21
2.1.2	Finding regressions		3.8.17	3.28.17	21	0 %	15	0	21
2.1.3	Graphical representaition of regressions		3.28.17	4.17.17	21	0 %	15	0	21
2.1.4	Preperations for presentation		4.17.17	4.21.17	5	0 %	5	0	5
2.2	Printing and preparing for delivery		4.21.17	4.21.17	1	0 %	1	0	1
2.3	Delivery of master thesis		4.21.17	4.21.17	1	0 %	1	0	1
2.4	Prodject dedline		5.26.17	5.26.17	1	0 %	1	0	1

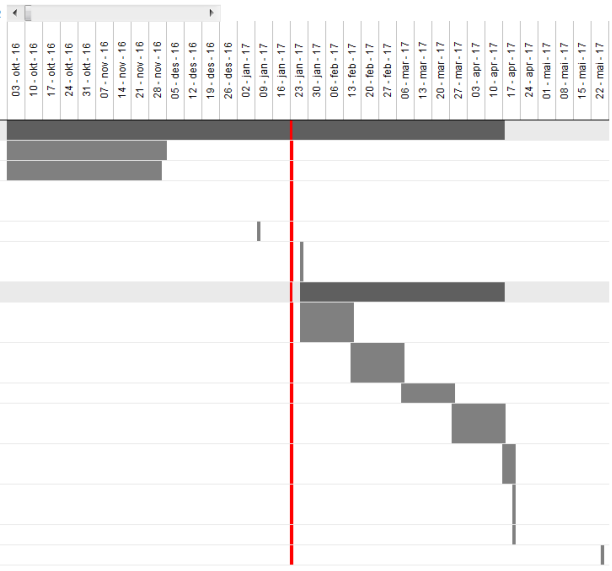


Figure 2 Project Gantt chart