Deggendorf Institute of Technology

Advanced Statistical Methods and Optimization

Prof. Dr. Tim Weber[1,*]

1. Deggendorf Institute of Technology
Applied Natural Sciences and Industrial Engineering (NUW)
Badstrasse 21, Germany

* *Correspondence:* Prof. Dr. Tim Weber tim.weber@th-deg.de

Table of contents

List of Figures

List of Tables

Preface

This is the script for the lecture "Advanced Statistical Methods and Optimization" at the DIT/Campus Cham. I do realize, that this body of knowledge has been repeated over and over, but have decided to do my own nonetheless so I can add my own flavor to the realms of statistics. This work is heavily inspired by (Wickham and Grolemund 2016). Please note that this material is copyrighted, you are not allowed to copy, at least ask for permission - you are likely to get it.

Tim Weber, Oct. 2024

Glossary

Text Abbreviations

ANOVA Analysis of Variance

CI Confidence Interval

CL Confidence Level

CDF cumulative function

CLT Central Limit Theorem

CTQ Critical To Quality

dof degree of freedom

DoE Design of Experiments

EDA Exploratory Data Analysis

FN false negative

FP false positive

gof goodness of fit

H0 Null Hypothesis

Ha Alternative Hypothesis

IQR Interquartile Range

KPI Key Performance Indicator

KS Kolmogorov Smirnov

LLN Law of Large Numbers

MLE Maximum Likelihood Estimation

MSA1 Measurement System Analysis Type I

UCL Upper Control Limit

LCL Lower Control Limit

UWL Upper Warning Limit

LWL Lower Warning Limit

PCC Pearson Correlation Coefficient

PDF Probability Density Function

PMF Probability Mass Function

PoI Parameter of Interest

p Population proportion

ppm parts per million

QC Quality Control

QQ Quantile-Quantie

SE Standard Error

TTF Time to failure

TN true negative

TP true positive

.w.r.t with respect to

Z Z-standardization

Symbol Abbreviations

α significance level

β false negative risk

ϵ residuals

μ_0 the true mean of a population

$\varphi(x)$ probability density function

$\phi(x)$ cumulative probability density function or cumulative distribution function

σ_0^2 the true variance of a population

σ_0 the true standard deviation of a population

C_g potential Measurement System Capability Index

C_{gk} Measurement Capability Index with systematic error

C_p potential process capability

C_{pk} actual process capability including centering

k number of predictors in a model

MSE mean squared errors

n number of data points/observations in the sample

N number of datapoints/observations in the population

P Probabilities

r^2 Coefficient of determination

$r^2_{adjusted}$ adjusted Coefficient of determination

sd the standard deviation of a dataset

SSE Sum of squared errors as calculated by

x_i the individual datapoints

$\bar{x}$ the mean value of the datas

X Predictor Variable

Y Response Variable

$\hat{y}$ predicted value

y_i true value

1 Basic Concepts

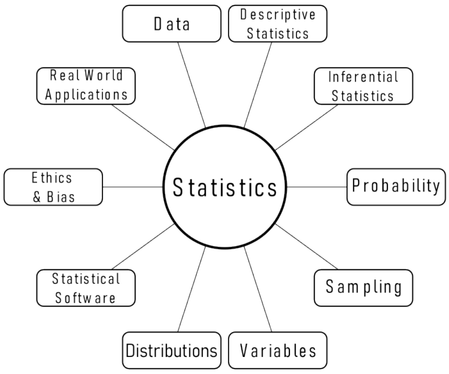


Figure 1.1: The necessary statistical ingredients.

Statistics is a fundamental field that plays a crucial role in various disciplines, from science and economics to social sciences and beyond. It's the science of collecting, organizing, analyzing, interpreting, and presenting data. In this introductory overview, we'll explore some key concepts and ideas that form the foundation of statistics:

1. **Data:** At the heart of statistics is data. Data can be anything from numbers and measurements to observations and information collected from experiments, surveys, or observations. In statistical analysis, we work with two main types of data: quantitative (numerical) and qualitative (categorical).

2. **Descriptive Statistics:** Descriptive statistics involve methods for summarizing and organizing data. These methods help us understand the basic characteristics of data, such as measures of central tendency (mean, median, mode) and measures of variability (range, variance, standard deviation).

3. **Inferential Statistics:** Inferential statistics is about making predictions, inferences, or decisions about a population based on a sample of data. This involves hypothesis testing, confidence intervals, and regression analysis, among other techniques.

4. **Probability:** Probability theory is the foundation of statistics. It deals with uncertainty and randomness. We use probability to describe the likelihood of

events occurring in various situations, which is essential for making statistical inferences.

5. **Sampling:** In most cases, it's impractical to collect data from an entire population. Instead, we often work with samples, which are smaller subsets of the population. The process of selecting and analyzing samples is a critical aspect of statistical analysis.

6. **Variables:** Variables are characteristics or attributes that can vary from one individual or item to another. They can be categorized as dependent (response) or independent (predictor) variables, depending on their role in a statistical analysis.

7. **Distributions:** A probability distribution describes the possible values of a variable and their associated probabilities. Common distributions include the normal distribution, binomial distribution, and Poisson distribution, among others.

8. **Statistical Software:** In the modern era, statistical analysis is often conducted using specialized software packages like R, Python (with libraries like NumPy and Pandas), SPSS, or Excel. These tools facilitate data manipulation, visualization, and complex statistical calculations.

9. **Ethics and Bias:** It's essential to consider ethical principles in statistical analysis, including issues related to data privacy, confidentiality, and the potential for bias in data collection and interpretation.

10. **Real-World Applications:** Statistics has a wide range of applications, from medical research to marketing, finance, and social sciences. It helps us make informed decisions and draw meaningful insights from data in various fields.

1.1 Probability

1.1.1 Overview

Probability theory is a fundamental concept in the field of statistics, serving as the foundation upon which many statistical methods and models are built.

1.1.2 What is Probability?

Probability is a mathematical concept that quantifies the uncertainty or randomness of events. It provides a way to measure the likelihood of different outcomes occurring in a given situation. In essence, probability is a numerical representation of our uncertainty.

1.1.3 Basic Probability Terminology

- **Experiment**: An experiment is any process or procedure that results in an outcome. For example, rolling a fair six-sided die is an experiment.
- **Outcome**: An outcome is a possible result of an experiment. When rolling a die, the outcomes are the numbers 1 through 6.
- **Sample Space (S)**: The sample space is the set of all possible outcomes of an experiment. For a fair six-sided die, the sample space is $\{1, 2, 3, 4, 5, 6\}$.
- **Event (E)**: An event is a specific subset of the sample space. It represents a particular set of outcomes that we are interested in. For instance, "rolling an even number" is an event for a six-sided die, which includes outcomes $\{2, 4, 6\}$.

1.1.4 Probability Notation

In probability theory, we use notation to represent various concepts:

- **P(E)**: Probability of event E occurring.
- **P(A and B)**: Probability of both events A and B occurring.
- **P(A or B)**: Probability of either event A or event B occurring.
- **P(E')**: Probability of the complement of event E, which is the probability of E not occurring.

1.1.5 The Fundamental Principles of Probability

There are two fundamental principles of probability:

- **The Addition Rule**: It states that the probability of either event A or event B occurring is given by the sum of their individual probabilities, provided that the events are mutually exclusive (i.e., they cannot both occur simultaneously).

$$P(A\ or\ B) = P(A) + P(B) \tag{1.1}$$

- **The Multiplication Rule**: It states that the probability of both event A and event B occurring is the product of their individual probabilities, provided that the events are independent (i.e., the occurrence of one event does not affect the occurrence of the other).

$$P(A\ and\ B) = P(A) * P(B) \tag{1.2}$$

1.1.6 Example: Rolling a Fair Six-Sided Die

Consider rolling a fair six-sided die.

- Sample Space (S): $\{1, 2, 3, 4, 5, 6\}$ (Figure 1.2)
- Event A: Rolling an even number = $\{2, 4, 6\}$ (Figure 1.2)
- Event B: Rolling a number greater than 3 = $\{4, 5, 6\}$ (Figure 1.2)

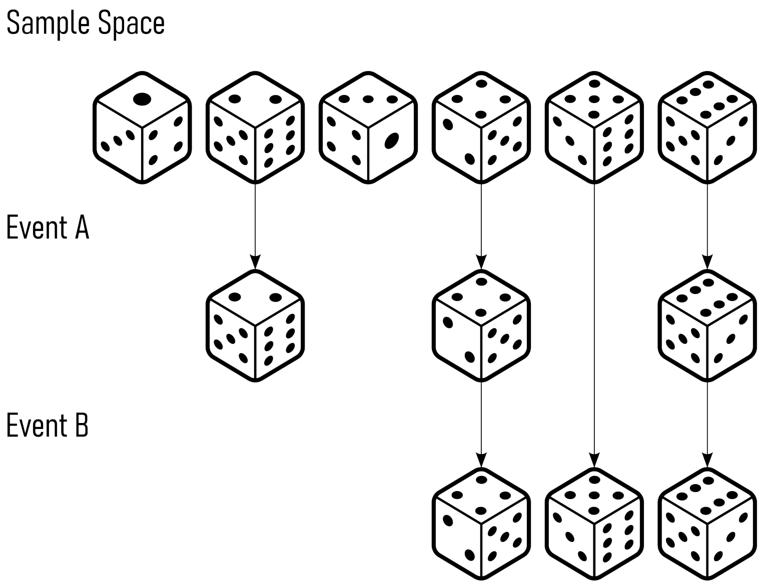


Figure 1.2: This example's sample space, as well as event A and event B.

1.1.7 Probability in action - The Galton Board

A Galton board, also known as a bean machine or a quincunx, is a mechanical device that demonstrates the principles of probability and the normal distribution. It was invented by Sir Francis Galton[1] in the late 19th century. The Galton board consists of a vertical board with a series of pegs or nails arranged in triangular or hexagonal patterns.

A Galton board, also known as a bean machine or a quincunx, is a mechanical device that demonstrates the principles of probability and the normal distribution. It was invented by Sir Francis Galton in the late 19th century. The Galton board consists of a vertical board with a series of pegs or nails arranged in triangular or hexagonal patterns.

1. **Initial Release**: At the top of the Galton board, a ball or particle is released. This ball can take one of two paths at each peg, either to the left or to the right.

[1] Sir Francis Galton (1822-1911): Influential English scientist, notable for his contributions to statistics and genetics.

The decision at each peg is determined by chance, such as the flip of a coin or the roll of a die. This represents a random event.

2. **Multiple Trials**: As the ball progresses downward, it encounters several pegs, each of which randomly directs it either left or right. The ball continues to bounce off pegs until it reaches the bottom.
3. **Accumulation**: Over multiple trials or runs of the Galton board, you will notice that the balls accumulate in a pattern at the bottom. This pattern forms a bell-shaped curve, which is the hallmark of a normal distribution.
4. **Normal Distribution**: The accumulation of balls at the bottom resembles the shape of a normal distribution curve. This means that the majority of balls will tend to accumulate in the center, forming the peak of the curve, while fewer balls will accumulate at the extreme left and right sides.

The Galton board is a visual representation of the central limit theorem, a fundamental concept in probability theory. It demonstrates how random events, when repeated many times, tend to follow a normal distribution. This distribution is commonly observed in various natural phenomena and is essential in statistical analysis.

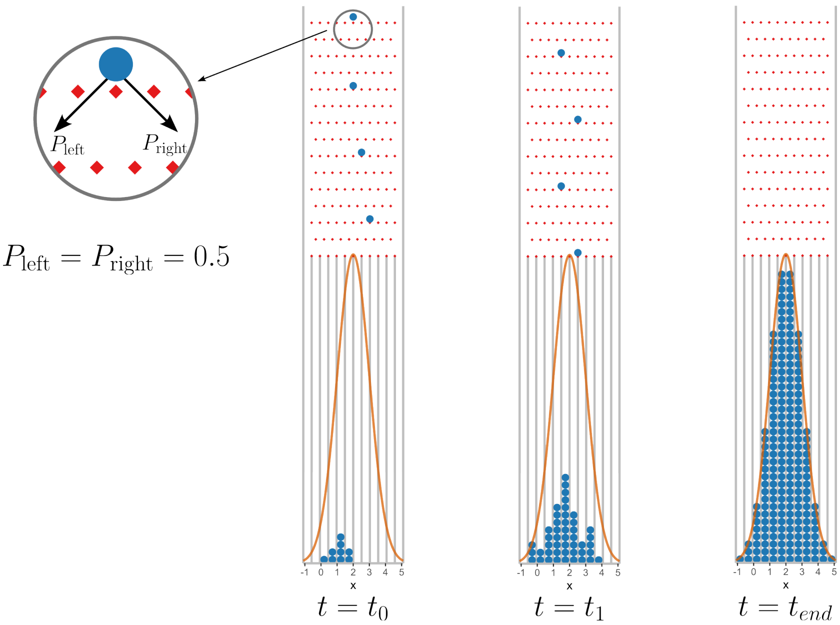


Figure 1.3: A Galton board in action.

1.1.7.1 Statistics and Probabbility

The Galton board is a nice example how statistics emerge from probability.

1.1.7.1.1 Define the problem

- The board has n rows of pegs (columns)
- Each ball has an equal probability of moving left or right (assuming no bias)
- The number of rightward moves determines the final position in the bins

1.1.7.1.2 Step 2: Binomial Probability Distribution

Each ball independently moves right (R) or left (L) with a probability of $p = 0.5$.

The number of rightwards moves follows a binomial distribution.

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.3}$$

n total number of columns (or pegs encountered)
k number of rightward moves
$\binom{n}{k}$ biomial coefficient, given by $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

with $p = 0.5$ this simplifies to

$$P(k) = \binom{n}{k} (\frac{1}{2})^n \tag{1.4}$$

1.1.7.1.3 Step 3: Position Mapping

The final position of a ball in a bin corresponds to the number of rightwards moves k. If the bins are indexed from 0 to n (where $k = 0$ means all left moves and $k = n$ means all right moves) the probability of landing in bin k is:

$$P(k) = \frac{n!}{k!(n-k)!} (\frac{1}{2})^n \tag{1.5}$$

1.2 Population

In statistics, a population is the complete set of individuals, items, or data points that are the subject of a study. Understanding populations and how to work with them is fundamental in statistical analysis, as it forms the basis for making meaningful inferences and drawing conclusions about the broader group being studied. It is the complete collection of all elements that share a common characteristic or feature and is of interest to the researcher. The population can vary widely depending on the research question

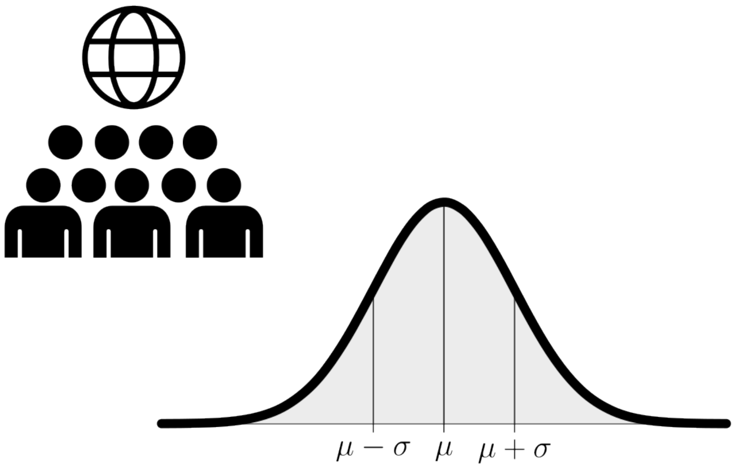


Figure 1.4: An example for a population.

or problem at hand. A populations *true mean* is depicted with μ_0 and the variance is depicted with σ_0^2.

1.3 Sample

The key principles behind a sample include its role as a manageable subset of data, which can be chosen randomly or purposefully. Ideally, it should be representative, reflecting the characteristics and diversity of the larger population. Statistical techniques are then applied to this sample to make inferences, estimate population parameters, or test hypotheses. The size of the sample matters, as a larger sample often leads to more precise estimates, but it should be determined based on research goals and available resources. Various sampling methods, such as random sampling, stratified sampling, or cluster sampling, can be employed depending on the research objectives and population characteristics. A samples *true mean* is depicted with $\bar{x}$ and the variance is depicted with *sd*.

1.4 Descriptive Statistics

Descriptive statistics are used to summarize and describe the main features of a data set. They provide a way to organize, present, and analyze data in a meaningful and concise manner. Descriptive statistics do not involve making inferences or drawing conclusions beyond the data that is being analyzed. Instead, they aim to provide a clear and

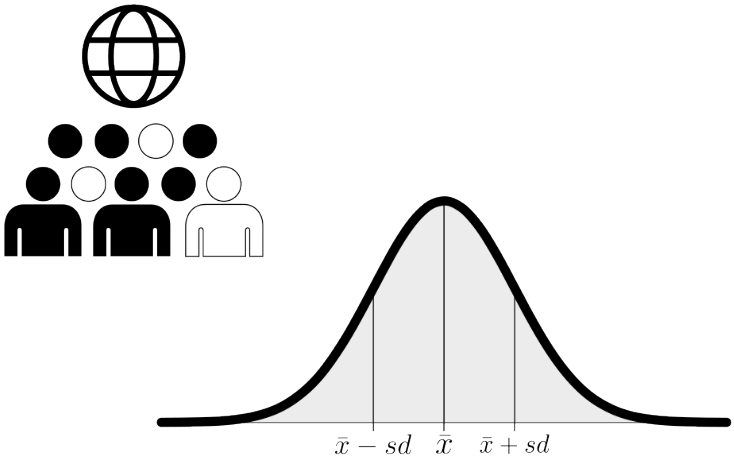


Figure 1.5: A sample drawn from the population.

accurate representation of the data set. Some common techniques and measures used in descriptive statistics include:

1.4.1 Example Data: The drive shaft exercise

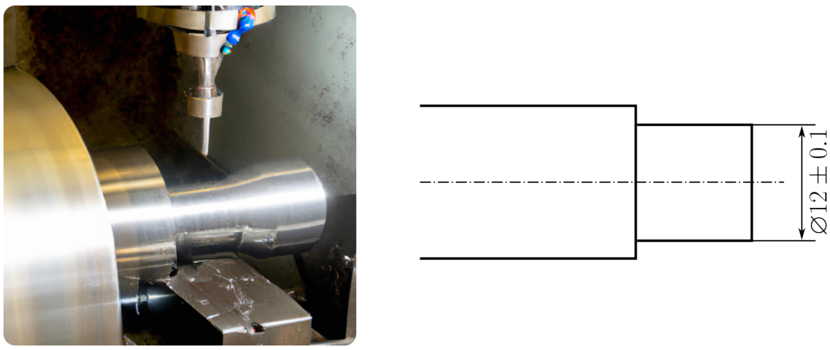


Figure 1.6: The drive shaft specification.

1.4.2 Measures of Central Tendency

Measures of central tendency are essential in statistics because they provide a single value that summarizes or represents the center point or typical value of a dataset. The main reasons for using these measures include:

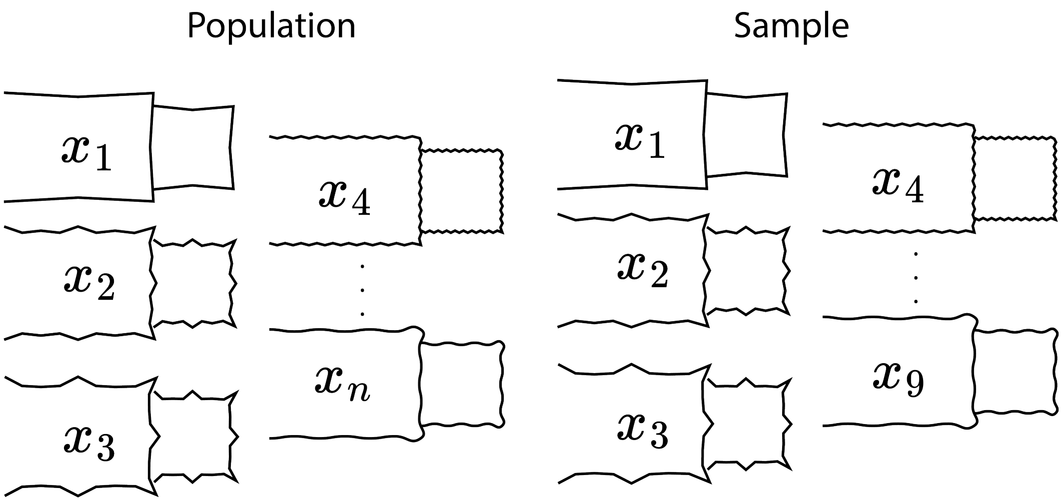


Figure 1.7: Difference between the population of ALL drive shafts and a sample of drive shafts.

- Simplification of Data: They condense large sets of data into a single representative value, making the data easier to understand and interpret.
- Comparison Across Datasets: They allow for straightforward comparison between different groups or datasets by providing a common reference point.
- Foundation for Further Analysis: Many statistical techniques and models rely on an understanding of central tendency as a starting point, such as in regression analysis or hypothesis testing.
- Decision-Making: In fields such as economics, education, and public policy, central tendency helps inform decisions based on typical outcomes or behaviors (e.g., average income, median test scores).
- Identification of Patterns: They help identify patterns and trends over time, especially in time-series data or longitudinal studies.

1.4.2.1 Mean

$$\text{population: } \mu = \frac{1}{N}\sum_{i}^{N} x_i \tag{1.6}$$

$$\text{sample: } \bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \tag{1.7}$$

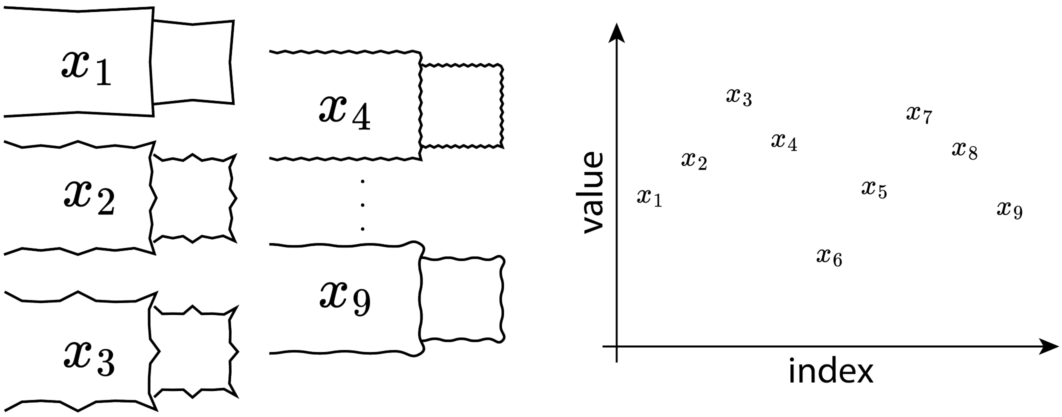


Figure 1.8: Some drive shaft sample data in a 2D plot of sample index vs. variable value

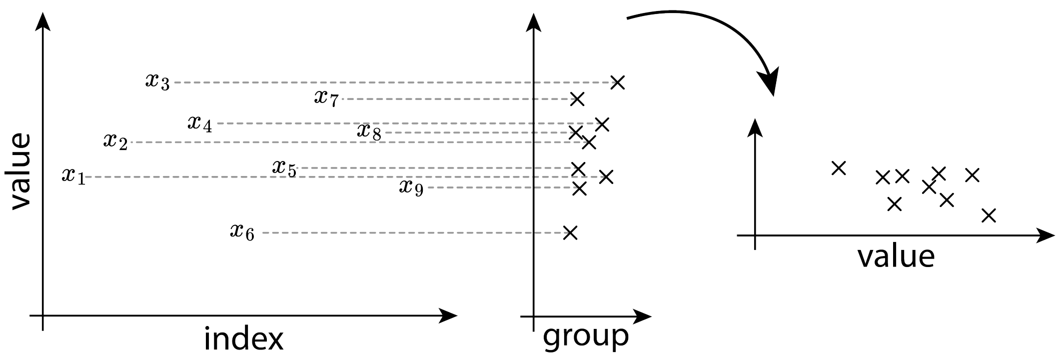


Figure 1.9: Some drive shaft sample data in a 2D plot of sample index vs. variable value

1.4.2.2 Median

$$\text{population: } m = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even} \end{cases} \tag{1.8}$$

$$\text{sample: } \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases} \tag{1.9}$$

1.4.3 Measures of Spread

Measures of spread (also called measures of dispersion or variability) are essential in statistics to provide information about the distribution of data — specifically, how much the data values differ from each other and from the central tendency.

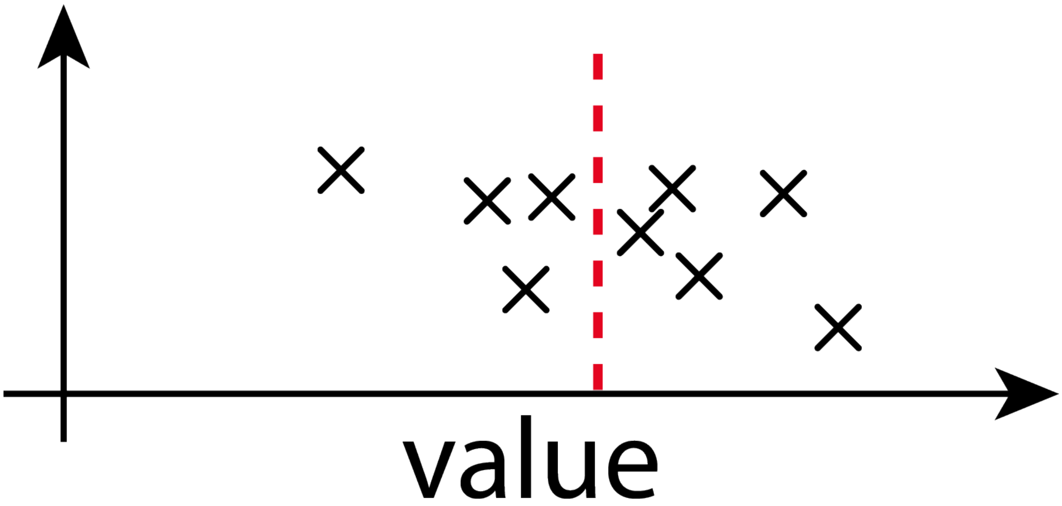


Figure 1.10: A graphical depiction of the mean

- Contextualizing Central Tendency: The mean or median alone does not give a complete picture of the data. Two datasets can have the same mean but very different spreads.
- Understanding Data Consistency: Measures of spread indicate how consistent or reliable the data are. A small spread suggests the values are closely clustered around the mean, while a large spread indicates greater variability and less predictability.
- Identifying Outliers: Large measures of spread may indicate the presence of outliers — values that are significantly different from others in the dataset. This can be important in quality control, risk assessment, and anomaly detection.
- Comparing Distributions: Spread allows for meaningful comparison between different datasets.
- Informing Statistical Models: Many statistical methods, such as regression, hypothesis testing, and confidence intervals, rely on measures of spread (like variance or standard deviation) to estimate error, assess significance, or make predictions.

1.4.3.1 Range

$$\text{Range} = x_{\text{max}} - x_{\text{min}} \tag{1.10}$$

There is no difference in computing the range for the population or the sample

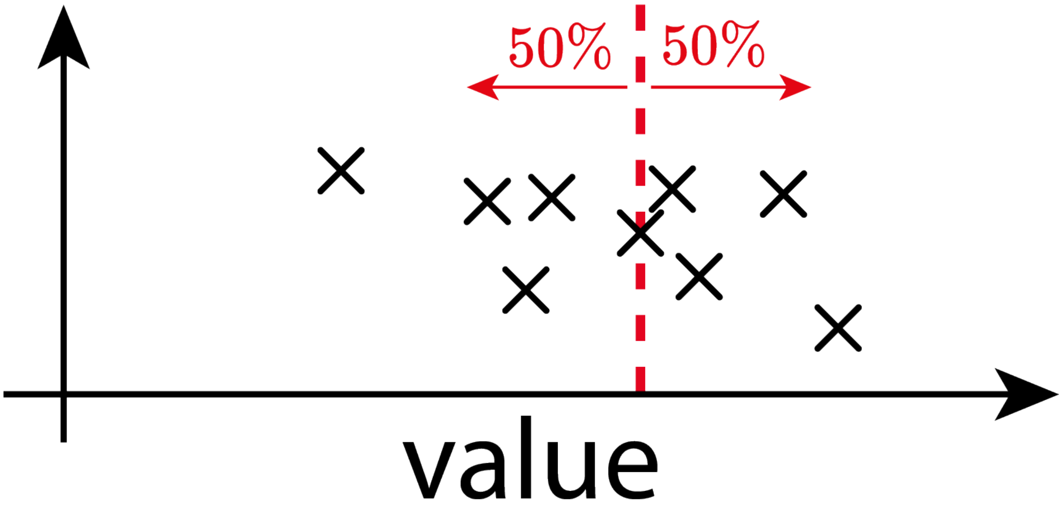


Figure 1.11: A graphical depiction of the median

1.4.3.2 Variance

$$\text{population: } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{1.11}$$

$$\text{sample: } sd^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.12}$$

1.4.3.2.1 The Bessel's correction

The variance calculated from a sample has one *degree of freedom* less, then the population variance.

Imagine you have 5 candies, and you want to give them to 5 friends — one candy to each. You decide how to give the first candy, then the second, third, and fourth. But when you get to the last candy, you have no choice — you have to give it to the last friend, so everyone gets one.

That's kind of like degrees of freedom in statistics. It means how many things you're free to choose before something has to be a certain way.

So if you're working with 5 numbers, and they all have to add up to a certain total (like a mean), you can choose 4 of them freely, but the last one has to be whatever makes the total come out right. That's why we say there are 4 degrees of freedom — 4 numbers you can choose any way you want.

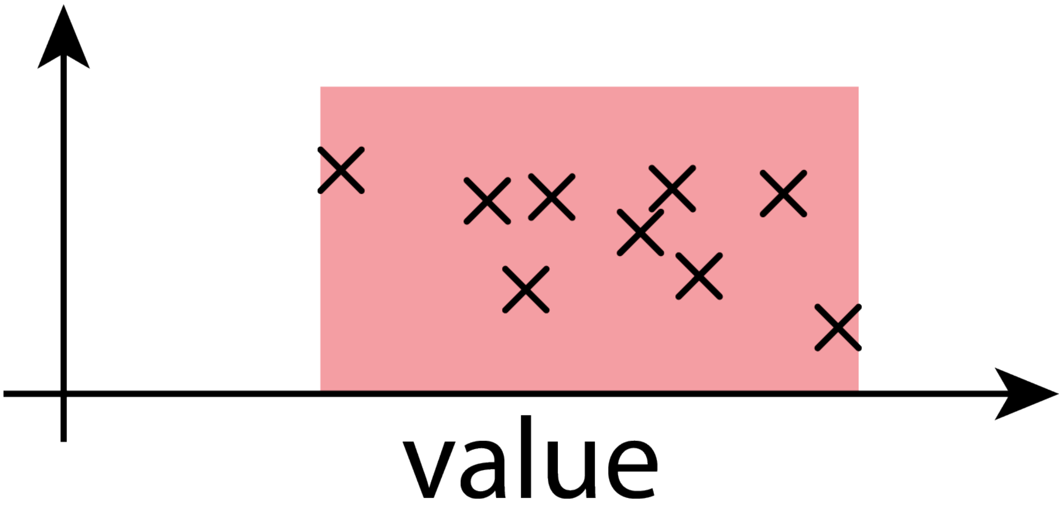


Figure 1.12: Spread, Dispersion, Variance … many names for measuring variability of data

$$\{2, 4, 6\}$$

- Mean: $\bar{x} = \frac{2+4+6}{3} = 4$
- Deviations: $-2, 0, 2$
- Squared Deviations: $4, 0, 4$
- Sum of squared deviations: 8

with Bessel's correction: $sd^2 = \frac{8}{3-1} = 4$

without Bessel's correction: $sd^2 = \frac{8}{3} \approx 2.67$ (biased, underestimates variance)

When computing the variance from a sample, we need to calculate $\bar{x}$, which *uses up* one degree of freedom and biases our estimate

1.4.3.2.2 Bessel's correction with increasing sample size

1.4.3.3 Standard Deviation

1.4.3.4 Percentiles, quantiles

1.4.4 Histogram

An example for descriptive statistics is shown in Figure 1.17 as a histogram. It shows data from a company that produces pharmaceutical syringes, taken from Ramalho (2021).

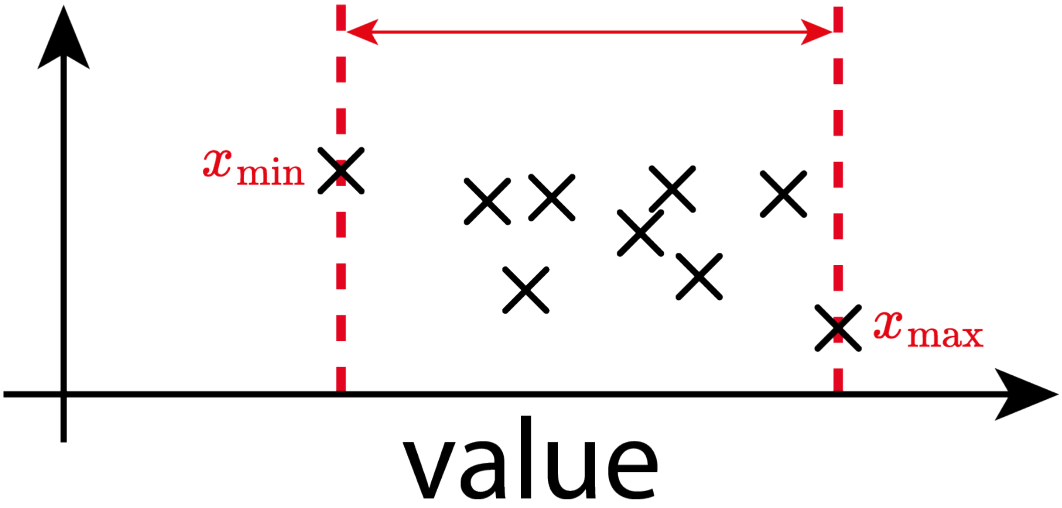


Figure 1.13: A graphical depiction of the range

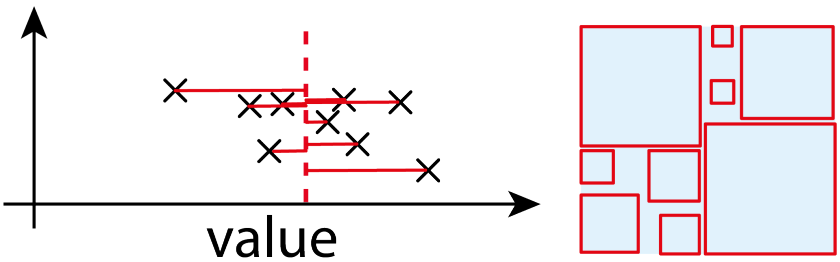


Figure 1.14: A graphical depiction of the variance

During the production of those syringes, the so called *barrel diameter* is a critical parameter to the function of the syringe and therefore of special interest for the Quality Control.

A histogram as shown in Figure 1.17 shows the data of 150 measurements during the QC. On the `x-axis` the *barrel diameter* is shown, while the count of each *binned* diameter is shown on the `y-axis`. The binning and of data is a crucial parameter for such a plot, because it already changes the appearance and width of the bars. Binning is a trade-off between visibility and readability.

1.4.5 Density plot

Density plots are another way of displaying the statistical distribution of an underlying dataset. The biggest strength of those plots is, that no binning is necessary in order to show the data. The limitation of this kind of plot is the interpretability. An example of

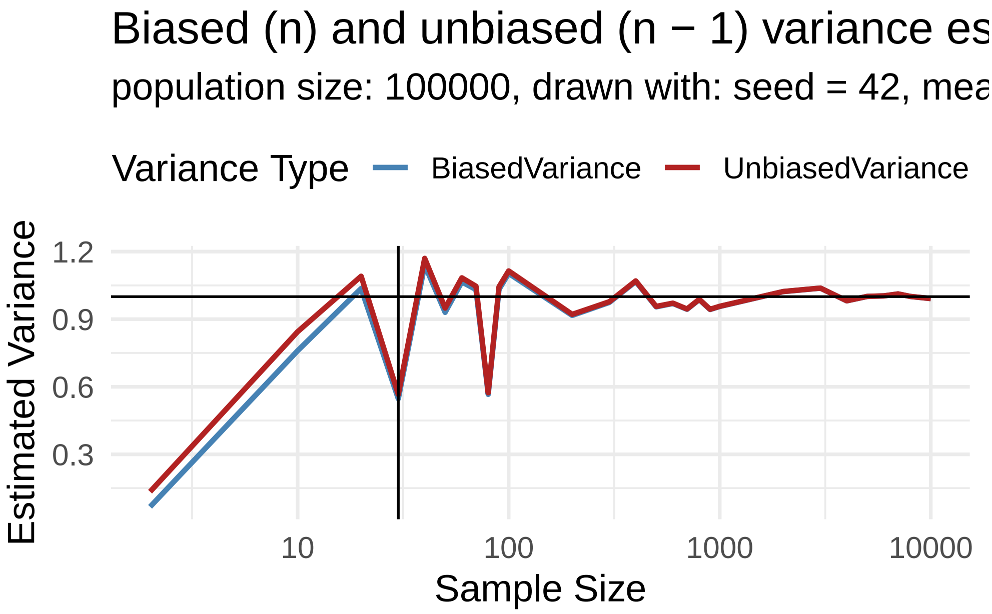


Figure 1.15: The biased (n) variance is not as prcises as the unbiased (n-1) variance estimate. This effect decreases with increasing sample size.

a density plot for the syringe data is shown in Figure 1.18. On the `x-axis` the syringe barrel diameter is shown (as in a histogram). The `y-axis` in contrast does not display the count of a binned category, but rather the Probability Density Function for the specific diameter. The grey area under the density curve depicts the probability of a syringe diameter to appear in the data. The complete area under the curve equals to 1 meaning that a certain diameter is sure to appear in the data.

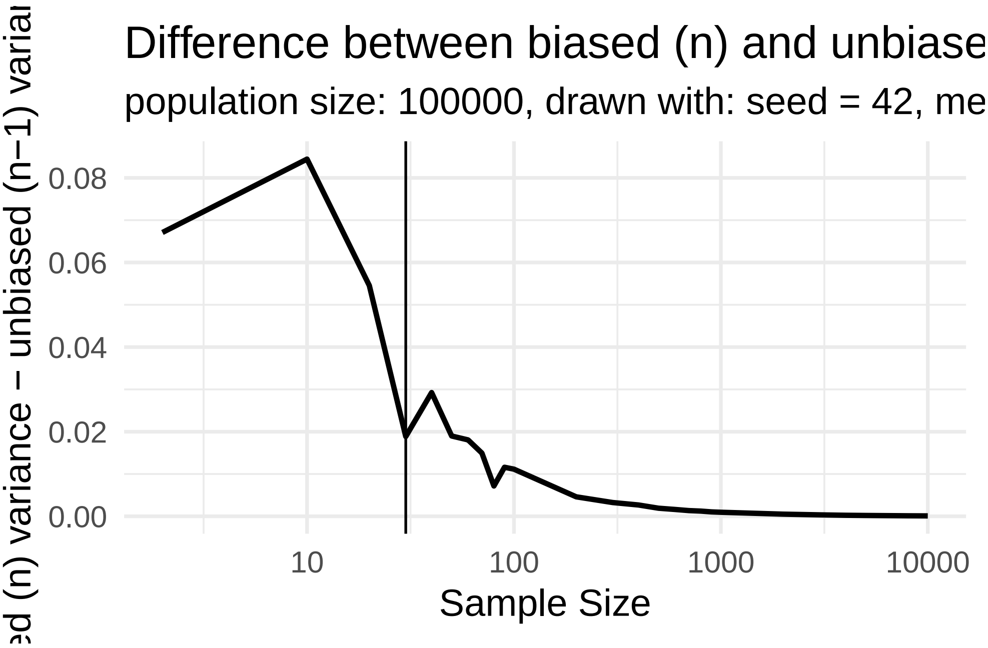


Figure 1.16: The difference between the biased (n) and unbiased (n-1) shows, that the Bessel's correction increases variation while decreasing bias

1.4.6 Boxplot

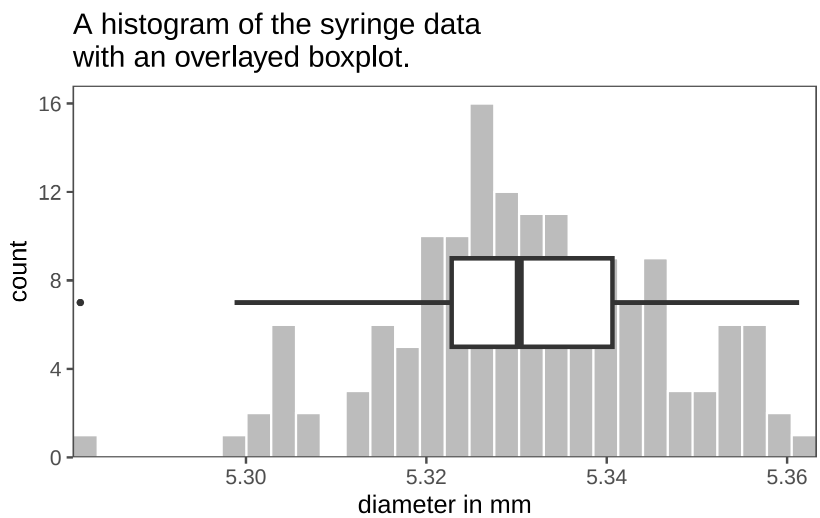


Figure 1.19: A boxplot of the same syringe data combined with the according histogram.

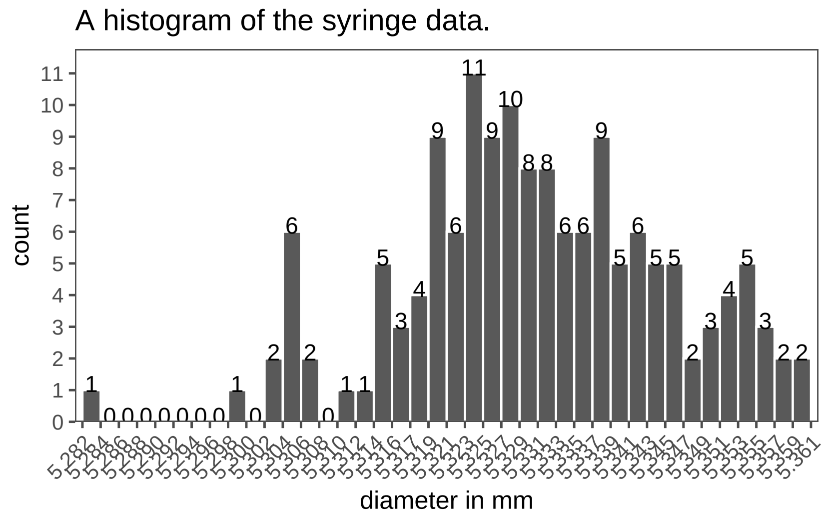


Figure 1.17: An example for descriptive statistics (histogramm)

It is very common to include and inspect measures of central tendency in the graphical depiction of data. A `boxplot`, also known as a box-and-whisker plot, is a very common way of doing this. A `boxplot` is a graphical representation of a dataset's distribution. It displays the following key statistics:

1. Median (middle value).
2. Quartiles (25^{th} and 75^{th} percentiles), forming a box.
3. Minimum and maximum values (whiskers).
4. Outliers (data points significantly different from the rest).

The syringe data in boxplot form is shown in Figure 1.19 as an overlay of the histogram plot before. Boxplots are useful for quickly understanding the central tendency, spread, and presence of outliers in a dataset, making them a valuable tool in data analysis and visualization.

1.4.7 Average, Standard deviation and Range

Very popular measures of central tendency include the *average* (mean) and the *standard deviation* (variance) of a dataset. The computed mean from an actual dataset is depicted with $\bar{x}$ and calculated via (1.13).

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.13}$$

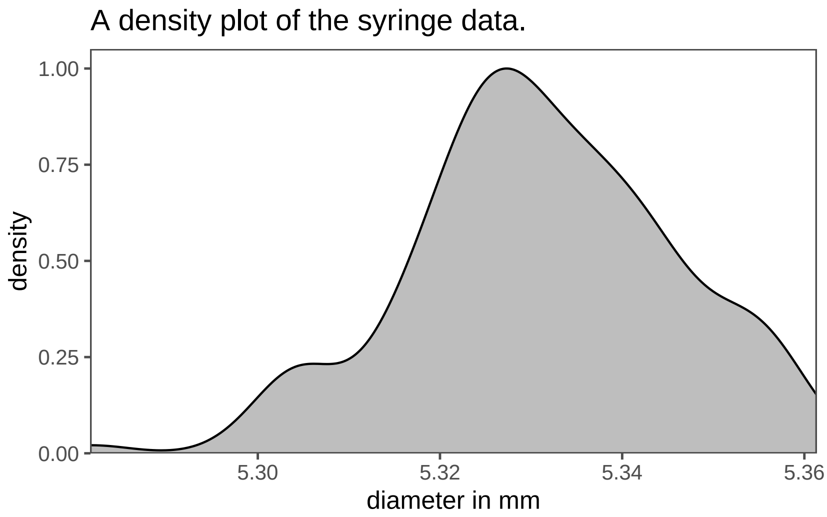


Figure 1.18: An example for a density plot for the syringe data (barrel diameter).

With n being the number of datapoints and x_i being the datapoints. The *mean* is therefore the sum of all datapoints divided by the total number n of all datapoints. It is not to be confused with the true mean μ_0 of a population.

The computed *standard deviation* from an actual dataset is depicted with sd and calculated via (1.14).

$$sd = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.14}$$

The *standard deviation* can therefore be explained as the square root of the sum of all differences of each individual datapoints to the mean of a dataset divided by the number of datapoints. It is not to be confused with the true variance σ_0^2 of a population. The variance of a dataset can be calculatd via (1.15).

$$\sigma = sd^2 \tag{1.15}$$

The *range* from an actual dataset is depicted with r and calculated via (1.16).

$$r = \max(x_i) - \min(x_i) \tag{1.16}$$

The *range* can therefore be interpreted as the range from minimum to maximum in a dataset.

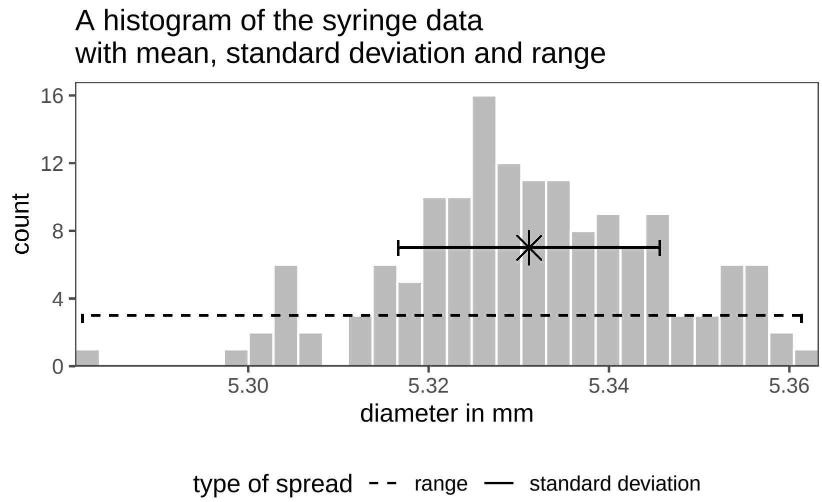


Figure 1.20: A histogram of the syringe data with mean, standard deviation and range.

1.5 Visualizing Groups

1.5.1 Boxplots

The methods described above are especially useful when it comes to visualizing groups in data. The data is discretized and the information density is increased. As with every discretization comes also a loss of information. It is therefore strongly advised to choose the right tool for the job.

If the underlying distribution of the data is unknown, a good start to visualize groups within data is usually a boxplot as shown in Figure 1.21. The syringe data from Ramalho (2021) contains six different groups, one for every sample drawn. Each sample consists of 25 observations in total. On the `x-axis` the *diameter* in mm is shown, the `y-axis` depicts the sample number. The `boxplots` are then drawn as described above (median, 25^{th} and 75^{th} percentile box, 5^{th} and 95^{th} whisker). The 25^{th} and 75^{th} percentile box is also known as the Interquartile Range

1.5.2 Mean and standard deviation plots

If the data follows a normal distribution, showing the mean and standard deviation for each group is also very common. For the syringe dataset, this is shown in Figure 1.22. The plot follows the same logic as for the boxplots (`x-axis-data`, `y-axis-data`), but the data itself shows the mean with a $\times$-symbol, as the length of the horizontal errorbars accords to $\bar{x} \pm sd(x)$.

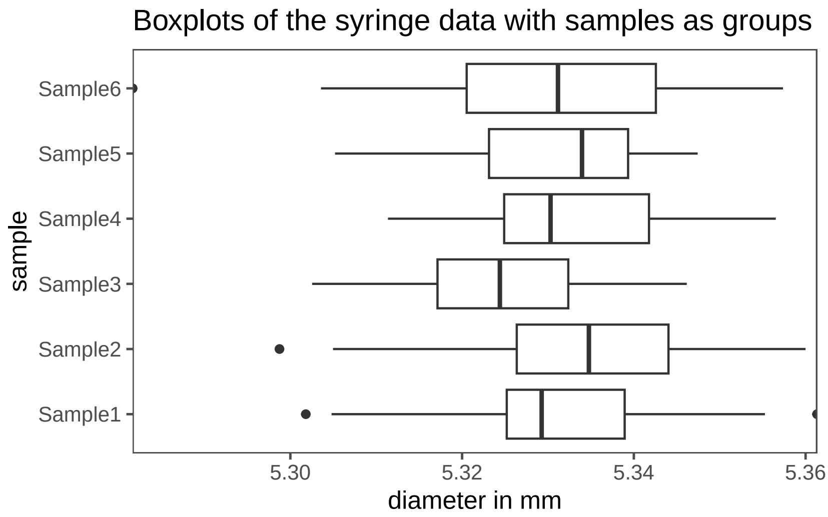


Figure 1.21: Boxplots of the syringe data with the samples as groups.

1.5.3 Half-half plots

Boxplots and mean-and-standard-deviation plots sometimes hide some details within the data, that may be of interest or simply important. Half-half plots, as shown in shown in Figure 1.23, incorporate different plot mechanisms. The left half shows a violin plot, which outlines the underlying distribution of the data using the PDF. This is very similar to a density plot. The right half shows the original data points and give the user a visible clue about the sample size in the data size. Note that the `y`-position of the points is jittered to counter *overplotting*. Details can be found in Tiedemann (2022).

1.5.4 Ridgeline plots

Figure 1.24 shows so called *ridgeline* plots as explained in Wilke (2022). They are in essence density plots that use the `y-axis` to differentiate between the groups. On the `x-axis` the density of the underlying dataset is shown. More info on the creation of these plots and graphics is available in Wickham (2016) as well as "The R Graph Gallery – Help and Inspiration for r Charts" (2022).

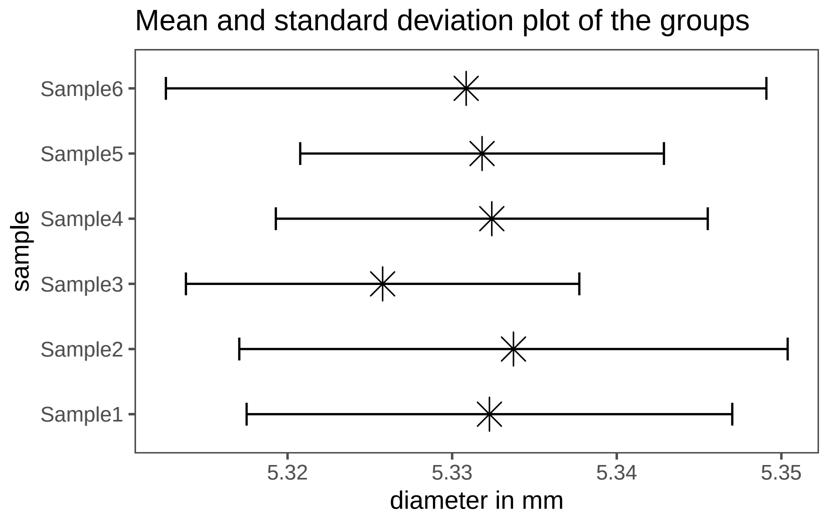


Figure 1.22: Mean and standard deviation plots of the groups in the dataset.

1.6 The drive shaft exercise

1.6.1 Introduction

A drive shaft is a mechanical component used in various vehicles and machinery to transmit rotational power or torque from an engine or motor to the wheels or other driven components. It serves as a linkage between the power source and the driven part, allowing the transfer of energy to propel the vehicle or operate the machinery.

1. Material Selection: Quality steel or aluminum alloys are chosen based on the specific application and requirements.

2. Cutting and Machining: The selected material is cut and machined to achieve the desired shape and size. Precision machining is crucial for balance and performance.

3. Welding or Assembly: Multiple sections may be welded or assembled to achieve the required length. Proper welding techniques are used to maintain structural integrity.

4. Balancing: Balancing is critical to minimize vibrations and ensure smooth operation. Counterweights are added or mass distribution is adjusted.

5. Surface Treatment: Drive shafts are often coated or treated for corrosion resistance and durability. Common treatments include painting, plating, or applying protective coatings.

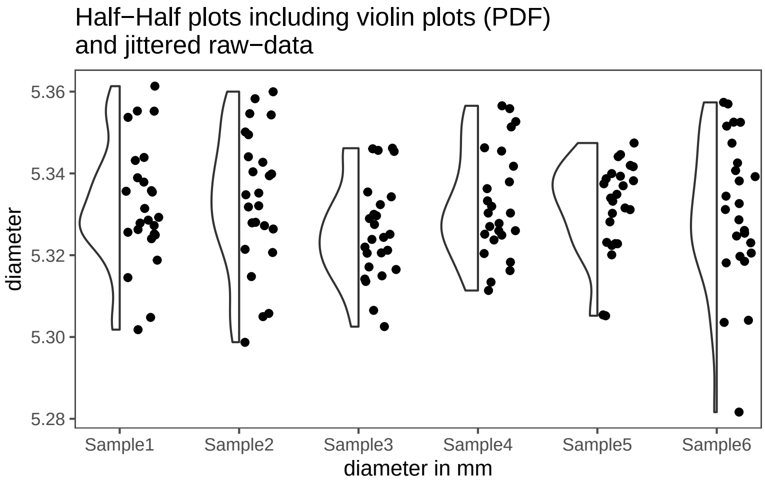


Figure 1.23: Half-half plots that incooperate different types of plots

6. Quality Control: Rigorous quality control measures are employed to meet specific standards and tolerances. This includes dimensional checks, material testing, and defect inspections.
7. Packaging and Distribution: Once quality control is passed, drive shafts are packaged and prepared for distribution to manufacturers of vehicles or machinery.

The end diameter of a drive shaft is primarily determined by its torque capacity, length, and material selection. It needs to be designed to handle the maximum torque while maintaining structural integrity and flexibility as required by the specific application. For efficient load transfer, there are ball bearings mounted on the end diameter. Ball bearings at the end diameter of a drive shaft support its rotation, reducing friction. They handle axial and radial loads, need lubrication for longevity, and may include seals for protection. Proper alignment and maintenance are crucial for their performance and customization is possible to match specific requirements.

The end diameter of the drive shaft shall be $\emptyset 12 \pm 0.1 mm$ (see Figure 1.25). This example will haunt us the rest of this lecture.

1.6.2 Visualizing all the Data

First, some descriptive statistics of $N = 500$ produced drive shafts are shown in Table 1.1 ($\bar{x}(sd), median(IQR)$). This first table does not tell us an awful lot about the sample, apart from the classic statistical measures of central tendency and spread.

Table 1.1: The summary table of the drive shaft data

Variable	**N = 500**[1]
diameter	12.17 (0.51), 12.03 (0.58)

[1]Mean (SD), Median (IQR)

In Figure 1.26 the data and the distribution thereof is visualized using different modalities. The complete `drive shaft data` is shown as a histogram (Figure 1.26a) and as a density plot (Figure 1.26b). A single boxplot is plotted over the histogram data in Figure 1.26a, providing a link to Table 1.1 (median and IQR). One important conclusion may be draw from those plots already: There may be more than one dataset hidden inside the data. We will explore this possibility further.

1.6.3 Visualizing groups within the data

Fortunately for us, the groups that may be hidden within the data are marked in the orginal dataset and denoted as `group0x`. Unfortunately for us, it is not known (purely from the data) how these groups come about. Because we did get the dataset from a colleague, we need to investigate the *creation* of the dataset even further. This is an important point, for without knowledge about the history of the data, it is *impossible* or at least *unadvisable* to make valid statements about the data. We will go on with a table of summary statistics, see Table 1.2. Surprisingly, there are five groups hidden within the data, something we would no be able to spot from the raw data alone.

Table 1.2: The group summary table of the drive shaft data

Variable	**N = 100**[1]
group01	12.02 (0.11), 12.02 (0.16)
group02	12.36 (0.19), 12.34 (0.25)
group03	13.00 (0.10), 13.01 (0.13)
group04	11.49 (0.09), 11.49 (0.12)
group05	12.001 (0.026), 12.000 (0.030)

[1]Mean (SD), Median (IQR)

Again, the table is good to have, but not as engagingi for ourself and our co-workers to look at. In order to make the data more approachable, we will use some techniques shown in Section 1.5.

First in Figure 1.27a the raw data points are shown as points with overlayed boxplots. On the `x-axis` the groups are depicted, while the Parameter of Interest (in this case the *end diameter* of the drive shaft) is shown on the `y-axis`. Because we are interested how the manufactured drive shafts behave with respect to the specification limit, the

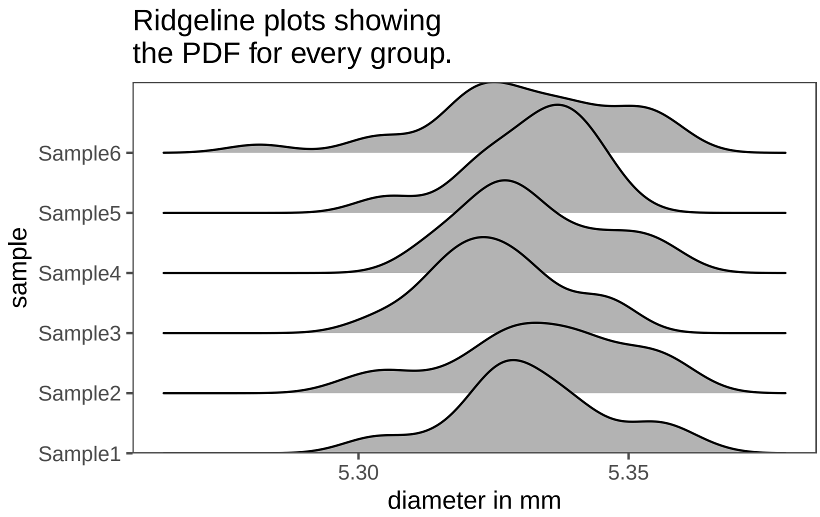
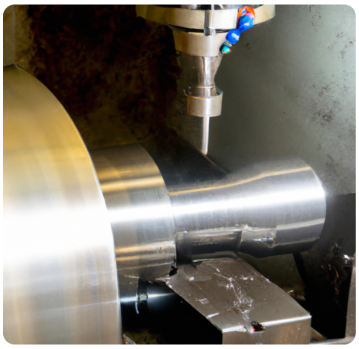


Figure 1.24: Ridgeline plots for distributions within groups.



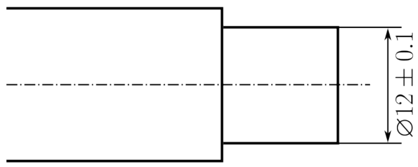
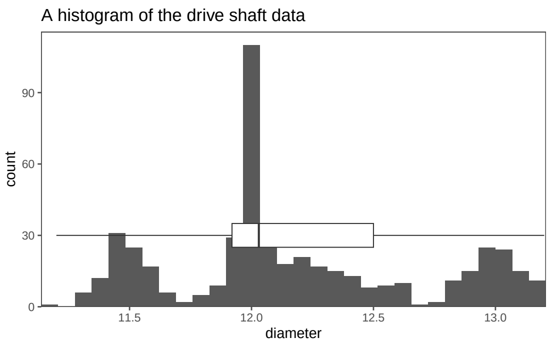


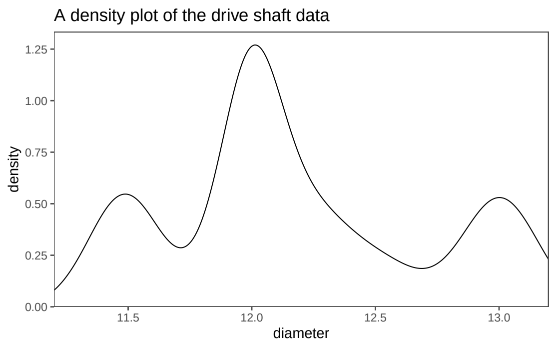
Figure 1.25: The drive shaft specification.

`nominal` value as well as the `uppper` and the `lower` specification limit is also shown in the plot as horizontal lines.

In Figure 1.27b the data is shown as ridgeline density plots. On the `x-axis` the diameter is depiected, while the `y-axis` shows two types of data. First, the groups 1 ... 5 are shown. For the individual groups, the probability is depicted as a line, therefore indicating which values are most probable in the given group. Again, because we are interested how the manufactured drive shafts behave .w.r.t the specification limit, the `nominal` value as well as the `uppper` and the `lower` specification limit is also shown in the plot as vertical lines.

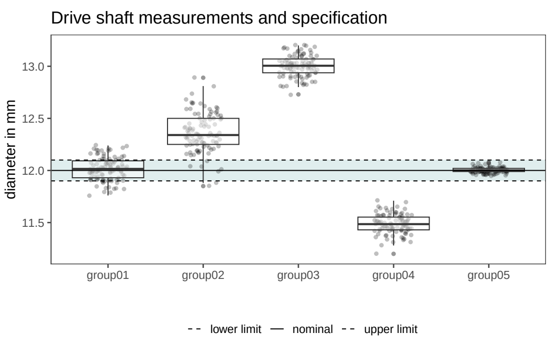


(a) The drive shaft data shown in a histogram.

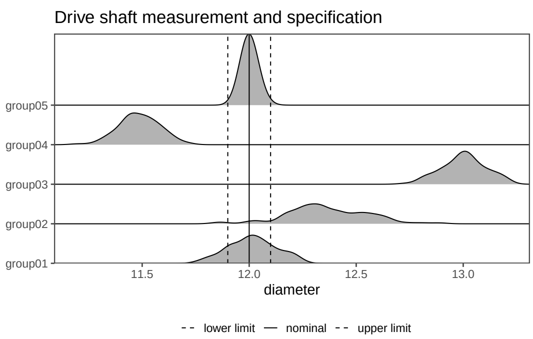


(b) The drive shaft data shown in a density plot.

Figure 1.26: The raw data of the measured drive shaft diameter.



(a) The groups visualized as boxplots (including the specification)



(b) The groups visualized as ridgeline plots

Figure 1.27: The raw data of the measured drive shaft diameter.

2 Statistical Distributions

2.1 Types of data

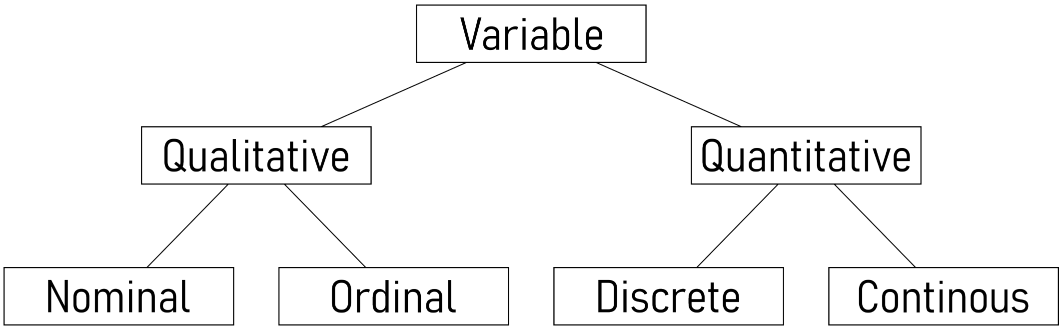


Figure 2.1: Data can be classified as different types.

1. **Nominal Data:**
   - Description: Nominal data represents categories with no inherent order or ranking.
   - Examples: Colors, gender, or types of fruits.
   - Characteristics: Categories are distinct, but there is no meaningful numerical value associated.
2. **Ordinal Data:**
   - Description: Ordinal data has categories with a meaningful order or ranking, but the intervals between them are not consistent or measurable.
   - Examples: Educational levels (e.g., high school, bachelor's, master's), customer satisfaction ratings (e.g., low, medium, high).
   - Characteristics: The order is significant, but the differences between categories are not precisely quantifiable.
3. **Discrete Data:**
   - Description: Discrete data consists of separate, distinct values, often counted in whole numbers and with no intermediate values between them.
   - Examples: Number of students in a class, number of cars in a parking lot.

- Characteristics: The data points are distinct and separate; they do not have infinite possible values within a given range.

4. **Continuous Data:**

- Description: Continuous data can take any value within a given range and can be measured with precision.
- Examples: Height, weight, temperature.
- Characteristics: Values can be any real number within a range, and there are theoretically infinite possible values within that range.

2.1.1 Nominal Data

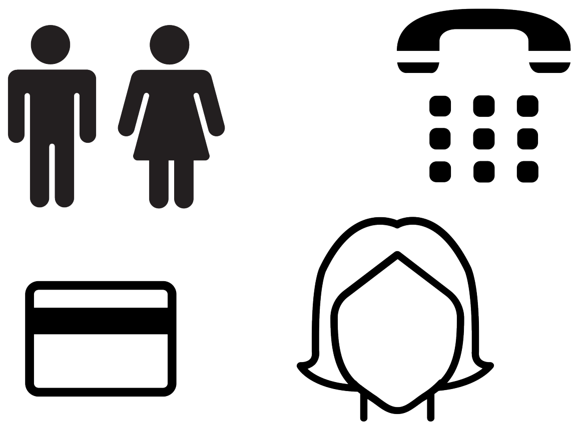


Figure 2.2: Some example for nominal data.

Nominal data is a type of data that represents categories or labels without any specific order or ranking. These categories are distinct and non-numeric. For example, colors, types of fruits, or gender (male, female, other) are nominal data. Nominal data can be used for classification and grouping, but mathematical operations like addition or subtraction do not make sense in this context.

2.1.2 Ordinal Data

Ordinal data represents categories that have a specific order or ranking. While the categories themselves may not have a consistent numeric difference between them, they can be arranged in a meaningful sequence. A common example of ordinal data is survey responses with options like “strongly agree,” “agree,” “neutral,” “disagree,” and “strongly disagree.” These categories indicate a level of agreement, but the differences between them may not be uniform or measurable.

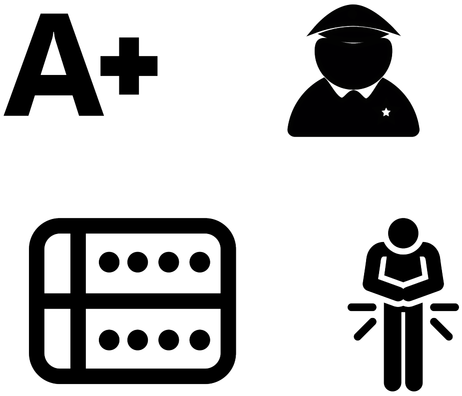


Figure 2.3: Some example for ordinal data.

2.1.3 Discrete Data

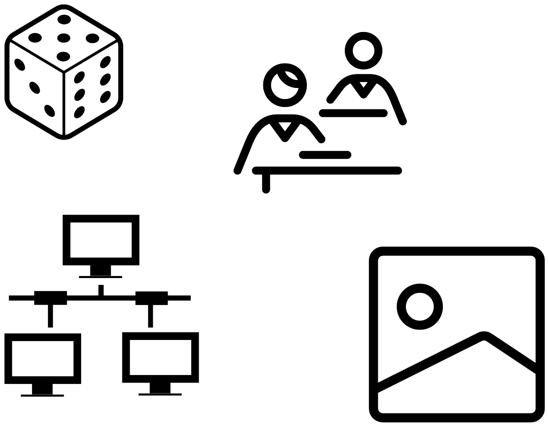


Figure 2.4: Some example for discrete data.

Discrete data consists of distinct, separate values that can be counted and usually come in whole numbers. These values can be finite or infinite, but they are not continuous. Examples include the number of students in a class, the count of cars in a parking lot, or the quantity of books in a library. Discrete data is often used in counting and can be represented as integers.

One quote in the literature about discrete data, shows how difficult the classification of data types can become (J. Bibby (1980)): "... All actual sample spaces are discrete, and all observable random variables have discrete distributions. The continuous distribution is a mathematical construction, suitable for mathematical treatment, but not practically observable. ...”

2.1.4 Continous Data

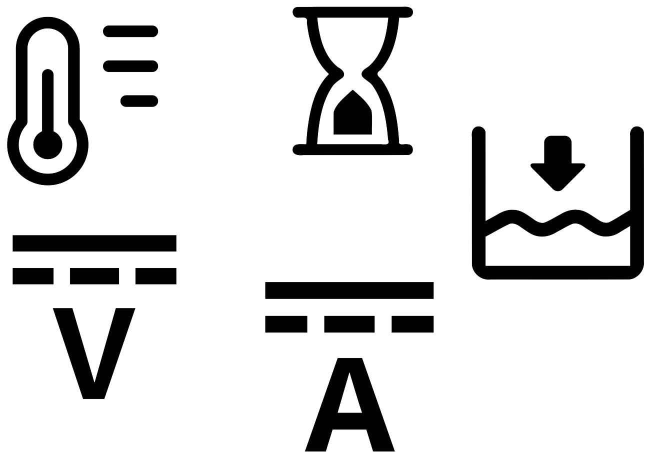


Figure 2.5: Some example for continous data.

Continuous data encompasses a wide range of values within a given interval and can take on any real number. There are infinite possibilities between any two points in a continuous dataset, making it suitable for measurements with high precision. Examples of continuous data include temperature, height, weight, and time. It is important to note that continuous data can be measured with decimals or fractions and is not limited to whole numbers.

2.2 Bionmimal Distribution

The binomial distribution is a **discrete** probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. A Bernoulli trial, named after Swiss mathematician Jacob Bernoulli[1], is a random experiment or trial with two possible outcomes: success and failure. These outcomes are typically labeled as 1 for success and 0 for failure. The key characteristics of a Bernoulli trial are:

[1] Jacob Bernoulli (1654-1705): Notable Swiss mathematician, known for Bernoulli’s principle and significant contributions to calculus and probability theory.

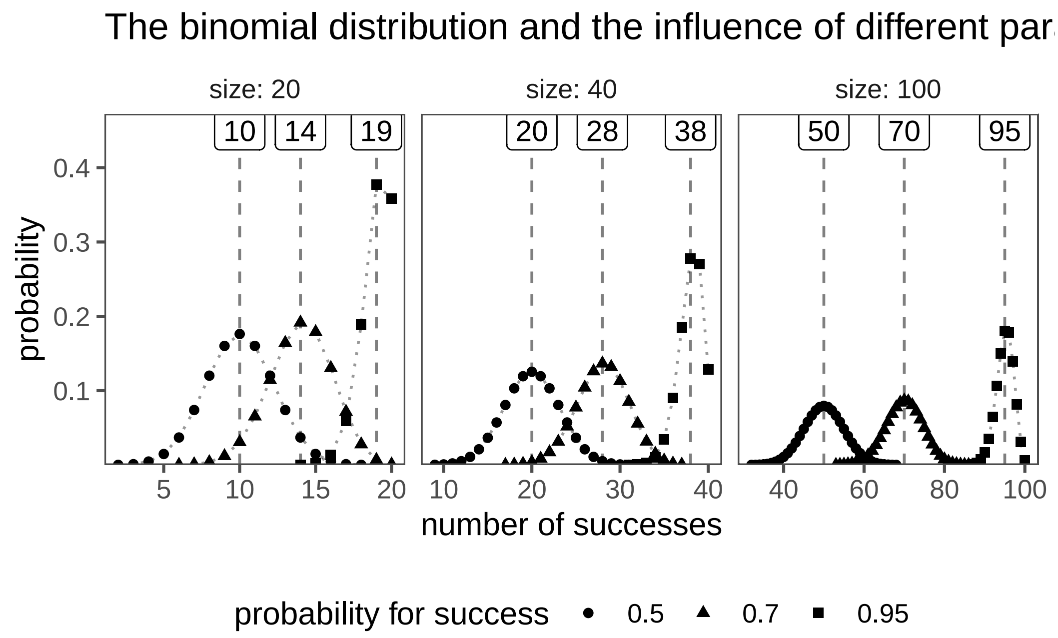


Figure 2.6: The binomial distribution

1. **Two Outcomes:** There are only two possible outcomes in each trial, and they are mutually exclusive. For example, in a coin toss, the outcomes could be heads (success, represented as 1) or tails (failure, represented as 0).
2. **Constant Probability:** The probability of success remains the same for each trial. This means that the likelihood of success and failure is consistent from one trial to the next.
3. **Independence:** Each trial is independent of others, meaning that the outcome of one trial does not influence the outcome of subsequent trials. For instance, the result of one coin toss doesn't affect the result of the next coin toss.

Examples of Bernoulli trials include:

- Flipping a coin (heads as success, tails as failure).
- Rolling a die and checking if a specific number appears (the number as success, others as failure).
- Testing whether a manufactured product is defective or non-defective (defective as success, non-defective as failure).

The Bernoulli trial is the fundamental building block for many other probability distributions, including the binomial distribution, which models the number of successes in a fixed number of Bernoulli trials.

2.2.1 Probability Mass Function (PMF)

The probability mass function (PMF), also known as the discrete probability density function, is a fundamental concept in probability and statistics.

- Definition: The PMF describes the probability distribution of a discrete random variable. It gives the probability that the random variable takes on a specific value. In other words, the PMF assigns probabilities to each possible outcome of the random variable.
- Formal Representation: For a discrete random variable X, the PMF is denoted as P(X = x), where x represents a specific value. Mathematically, the PMF is defined as: $P(X = x) = \text{probability that } X \text{ takes the value } x$
- Properties: The probabilities associated with all hypothetical values must be non-negative and sum up to 1. Thinking of probability as "mass" helps avoid mistakes, as the total probability for all possible outcomes is conserved (similar to how physical mass is conserved).
- Comparison with Probability Density Function (PDF): A PMF is specific to *discrete* random variables, while a PDF is associated with continuous random variables. Unlike a PDF, which requires integration over an interval, the PMF **directly** provides probabilities for individual values.
- Mode: The value of the random variable with the largest probability mass is called the mode.
- Measure-Theoretic Formulation: The PMF can be seen as a special case of more general measure-theoretic constructions. It relates to the distribution of a random variable and the probability density function with respect to the counting measure.

The PMF for the binomial distribution is given in (2.1)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.1}$$

2.2.2 The drive shaft exercise - Binomial Distribution

In the context of a drive shaft, you can think of it as a model for the number of defective drive shafts in a production batch. Each drive shaft is either good (success) or defective (failure).

Let's say you have a batch of 100 drive shafts, and the probability of any single drive shaft being defective is 0.05(5%). You want to find the probability of having a certain number of defective drive shafts in this batch.

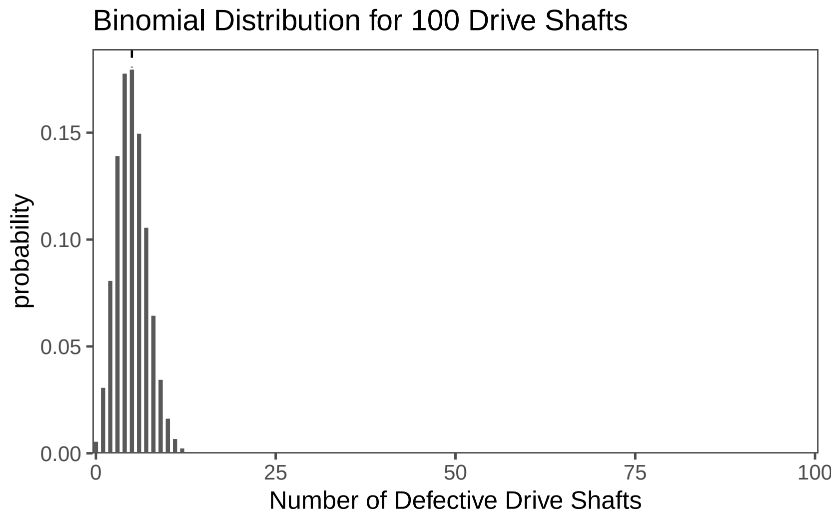


Figure 2.7: The binomial disitribution and the drive shaft exercise.

2.3 The Normal Distribution

The normal distribution is a fundamental statistical concept that holds immense significance in the realms of engineering and production. It is often referred to as the Gaussian distribution or the bell curve, is a mathematical model that describes the distribution of data in various natural and human-made phenomena, see Johnson (1994). It forms a symmetrical curve when plotted, is centered around a mean (μ_0) and balanced on both sides (Figure 2.8). The spread or dispersion of the data points is characterized by σ_0^2. Those two parameters completley define the normal distribution. A remarkable property of the normal distribution is the empirical rule, which states that approximately 68% of the data falls within one standard deviation from the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations (Figure 2.8). The existence of the normal distribution in the real world is a result of the combination of several factors, including the principles of statistics and probability, the Central Limit Theorem, and the behavior of random processes in nature and society.

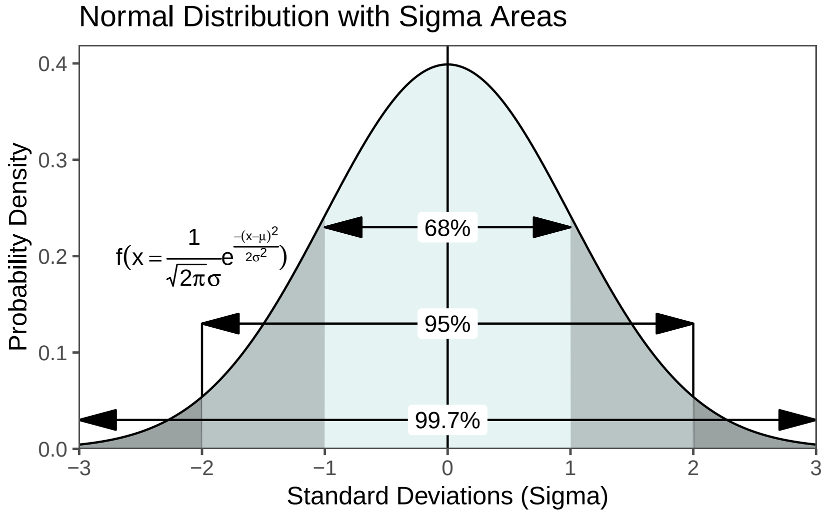
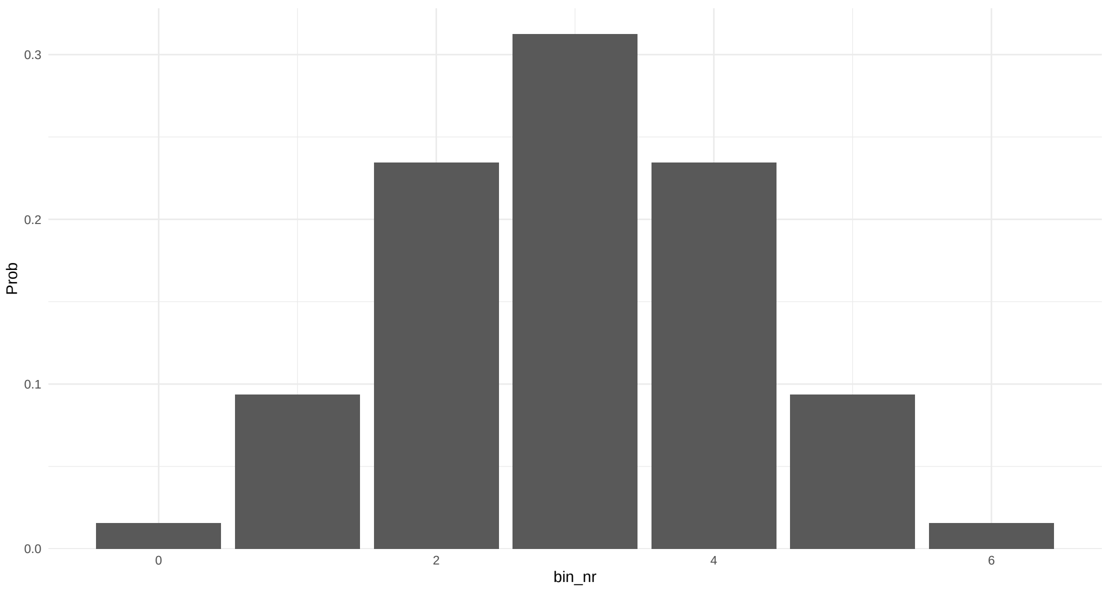
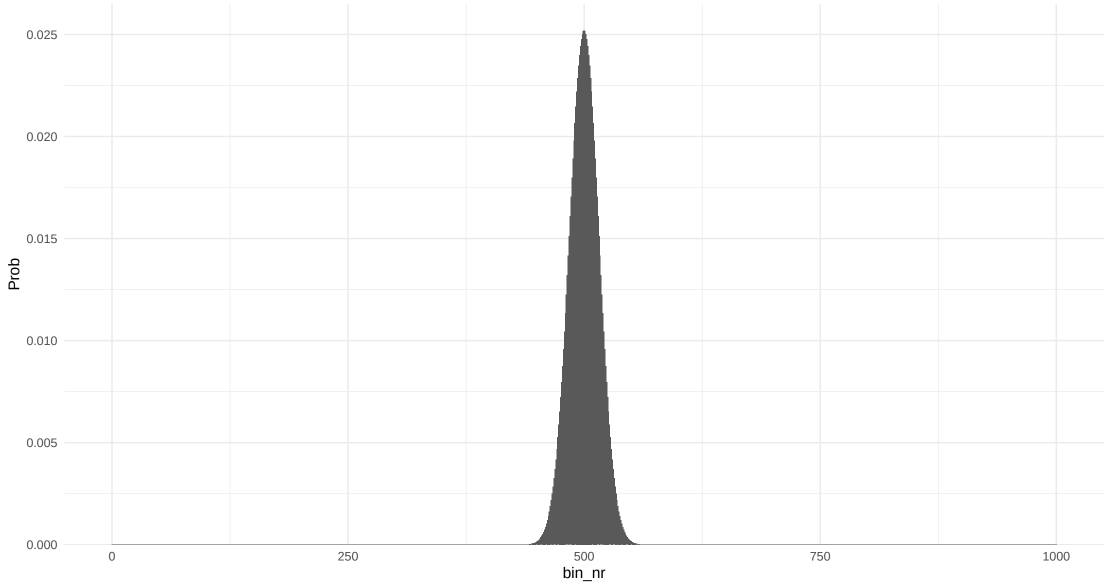


Figure 2.8: The standarized normal distribution

2.3.1 Emergence





2.3.1.1 The math behind

$$P(k) = \binom{n}{k} \cdot \frac{1}{2}^{n}$$

k	$n!$	$k!$	$(n-k)!$	$\binom{n}{k} = \binom{6}{k}$	$P(k) = \binom{6}{k} \cdot \frac{1}{2}^{n=6}$
0	720	1	$(6-0)! = 6! = 720$	$\frac{720}{1 \cdot 720} = 1$	$P(0) = 1 \cdot \frac{1}{64} = 0.015624$
1	720	1	$(6-1)! = 5! = 120$	$\frac{720}{1 \cdot 120} = 6$	$P(1) = 6 \cdot \frac{1}{64} = 0.093750$
2	720	2	$(6-2)! = 4! = 24$	$\frac{720}{2 \cdot 24} = 15$	$P(2) = 15 \cdot \frac{1}{64} = 0.234375$
3	720	6	$(6-3)! = 3! = 6$	$\frac{720}{6 \cdot 6} = 20$	$P(3) = 20 \cdot \frac{1}{64} = 0.312500$
4	720	24	$(6-4)! = 2! = 2$	$\frac{720}{24 \cdot 2} = 15$	$P(2) = 15 \cdot \frac{1}{64} = 0.234375$
5	720	120	$(6-5)! = 1! = 1$	$\frac{720}{120 \cdot 1} = 6$	$P(5) = 6 \cdot \frac{1}{64} = 0.093750$
6	720	720	$(6-6)! = 0! = 1$	$\frac{720}{720 \cdot 1} = 1$	$P(6) = 1 \cdot \frac{1}{64} = 0.015624$

2.3.1.2 Mean and Standard Deviation from the Galton board

- What is the expected value?

$$\mu = E[X] = np$$

$$p = 0.5 \rightarrow \mu = n \times 0.5 = \frac{1}{2}n$$

On *average* a ball will land in $k = \frac{n}{2}$

- What is the spread

$$\sigma = \sqrt{np(1-p)}$$

$Var(X) = E[(X-\mu)^2]$ (Variance in general)

$Var(X) = np(1-p)$ (number of right moves)

2.3.1.3 Understanding spread

$$\sigma^2 = np(1-p)$$

- Every step is independent, for a single step ($p = 0.5$):

$$Var(X) = p(1-p)$$

- Therefore for n steps:

$$Var(X) = np(1-p) \rightarrow \sigma = \frac{\sqrt{n}}{2}$$

- For $n = 100$

$\sigma = \frac{\sqrt{100}}{2}$ so most balls will land between 45 and 55

2.4 Z - Standardization

The Z-standardization, also known as standard score or z-score, is a common statistical technique used to transform data into a standard normal distribution with a mean of 0 and a standard deviation of 1 (Taboga 2017). This transformation is useful for comparing and analyzing data that have different scales and units (2.2).

$$Z = \frac{x_i - \bar{x}}{sd} \tag{2.2}$$

How the z-score can be applied is shown in Figure 2.9 and Figure 2.10. The data for group `X` and group `Y` may be measured in different units (Figure 2.9). To answer the question, which of the values $x_i (i = 1 \dots 5)$ is more probable, the single data points are transformed to the respective z-score using (2.2). In Figure 2.10, the z-scores for both groups are plotted against each other. The perfect correlation of the datapoints shows, that for every x_i the same probability applies. Thus, the datapoints are comparable.

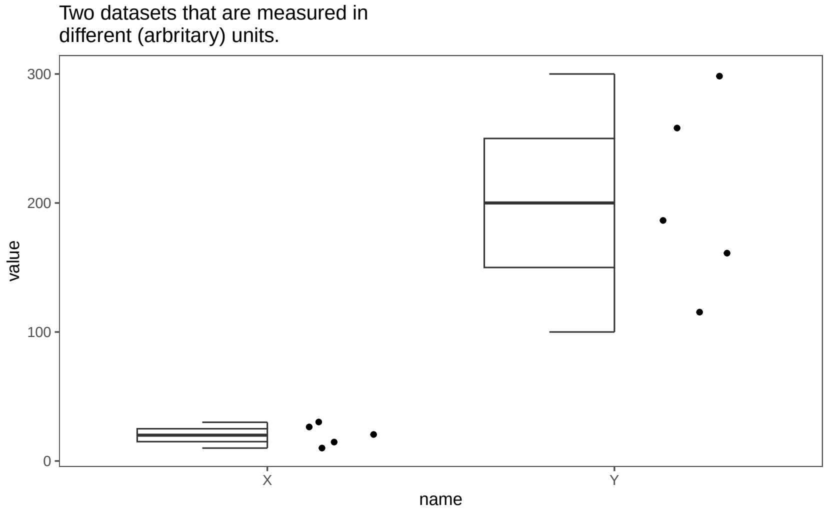


Figure 2.9: The original data of group X and group Y

2.4.1 The drive shaft exercise - Z-Standardization

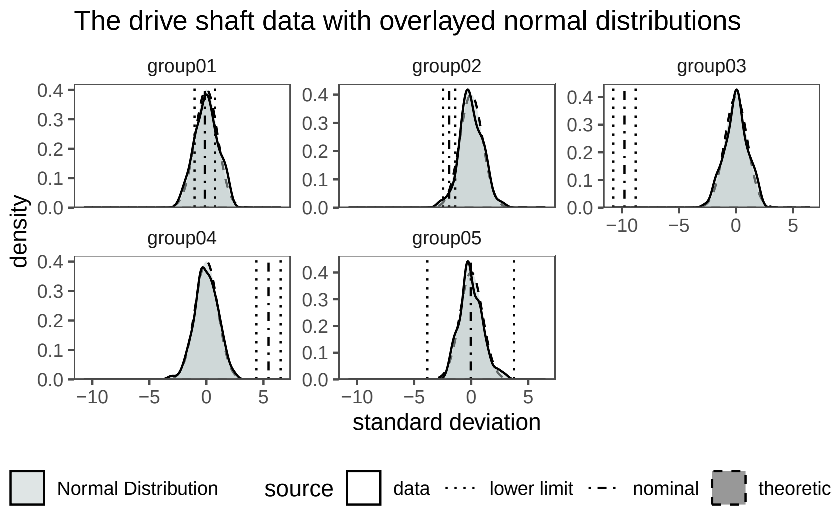


Figure 2.11: The standardized data of the drive shaft data.

In Figure 2.11 the standardized drive shaft data is shown. The mean of the data ($\bar{x}$) is now centered at 0 and the standard deviation is 1. For this case, the specification limits have also been transferred to the respective z-score (even though they can not be interpreted as such anymore). For every x_i the probability to be within a normal distribution is now known. When comparing this to the transferred specification limits, it is clear to see that for `group01` "most" of the data points are within the limits in

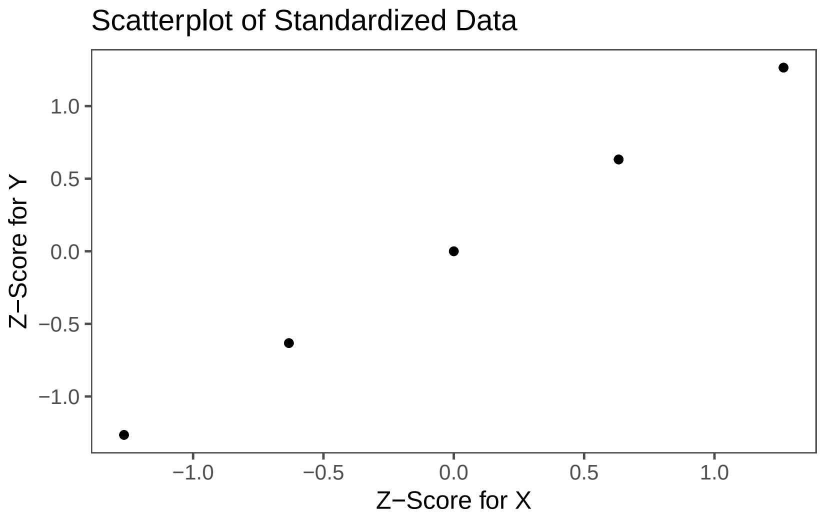


Figure 2.10: The correlation of the z-score shows, that every point x_i is equally probable

contrast to `group03` where none of the data points lies within the specification limits. When looking at `group03` we see, that the *nominal* specification limit is -9.78 standard deviations away from the centered mean of the datapoints. The probability of a data point being located there is $6.8605273 \times 10^{-23}$ which does not sound an awful lot. We will dwelve more into such investigation in another chapter, but this is a first step in the direction of inferential statistics.

2.4.2 Central Limit Theorem (CLT)

The primary reason for the existence of the normal distribution in many real-world datasets is the Central Limit Theorem (Taboga 2017). The CLT states that when you take a large enough number of random samples from any population, the distribution of the sample means will tend to follow a normal distribution, even if the original population distribution is not normal. This means that the normal distribution emerges as a statistical consequence of aggregating random data points. This is shown in Figure 2.12.

From $n = 10000$ uniformly disitrbuted data points (the *population*) ($min = 1, max = 100$) either $2, 10, 50$ or 200 samples are taken randomly (the *samples*). For each of the samples the mean is calculated, resulting in 1000 mean values for each $(2, 10, 50$ or $200)$ sample size. In Figure 2.12 the results from this numerical study are shown. The larger the sample size, the closer the mean calculated $\bar{x}$is to the population mean (μ_0). The effect is especially large on the standard deviation, resulting in a smaller standard deviation the larger the sample size is.

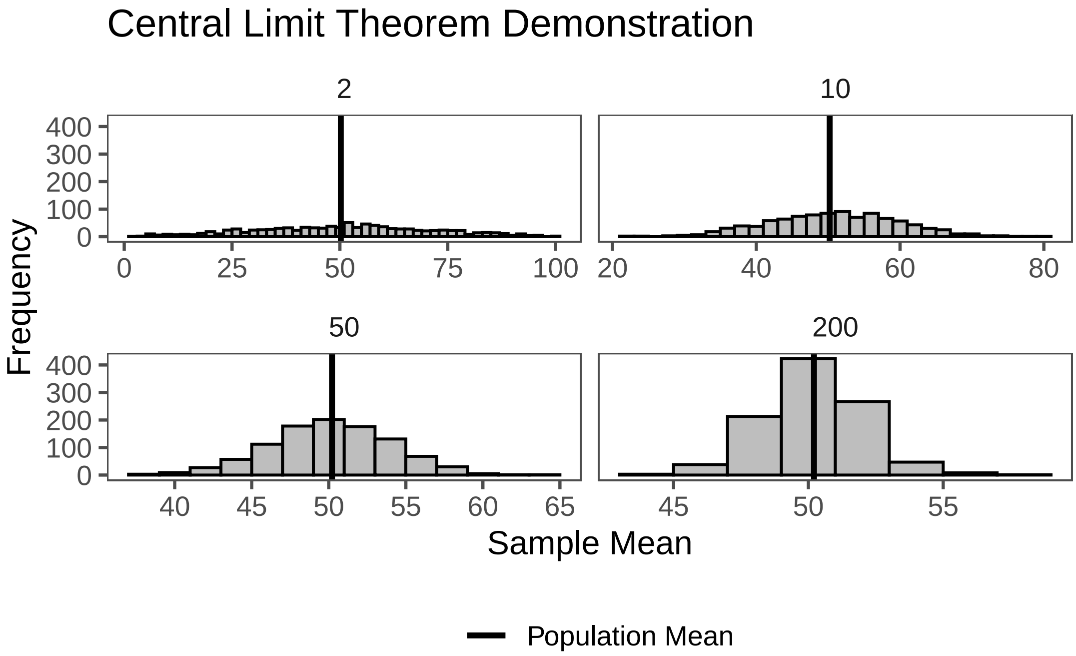


Figure 2.12: The central limit theorem in action.

2.4.3 Law of Large Numbers

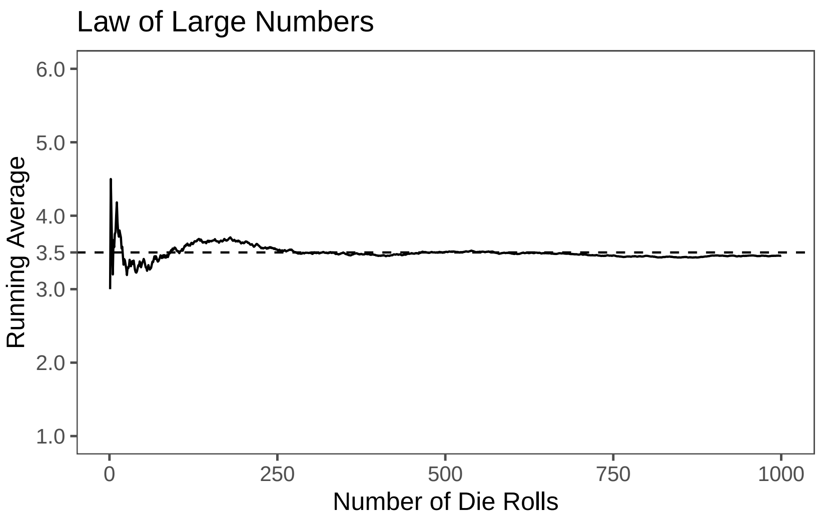
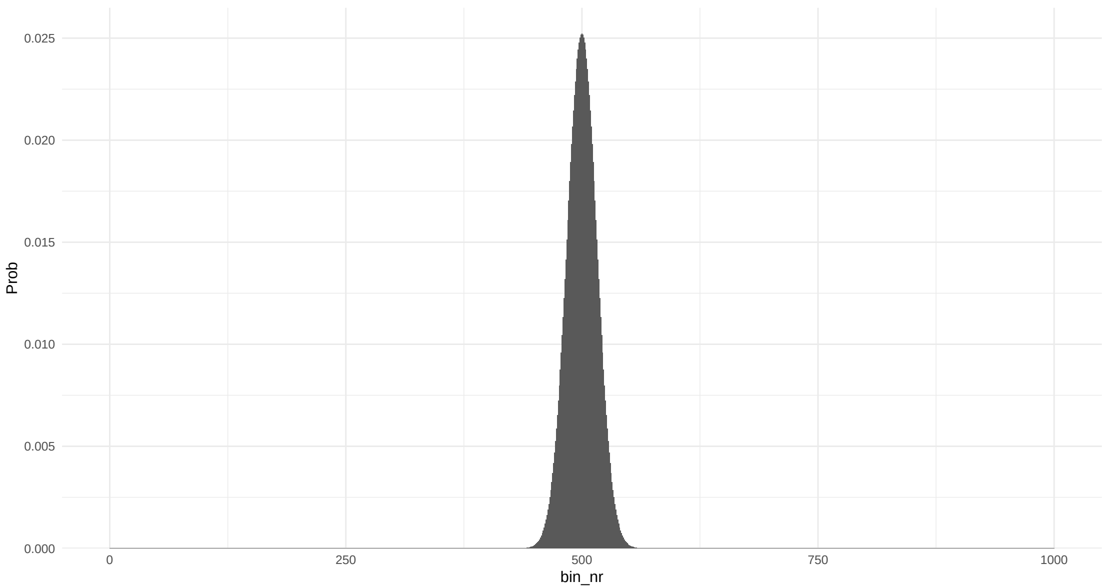


Figure 2.13: The Law of Large Numbers in Action with die rolls as an example.

The Law of Large Numbers states that as the size of a random sample increases, the sample average converges to the population mean. This law, along with the CLT, explains why the normal distribution frequently arises. When you take many small, independent, and identically distributed measurements and compute their averages, these averages tend to cluster around the true population mean, forming a normal distribution Johnson (1994).

The LLN ar work is shown in Figure 2.13. A fair six-sided die is rolled 1000 times and the running average of the roll results after each roll is calculated. The resulting line plot shows how the running average approaches the expected value of 3.5, which is the average of all possible outcomes of the die. The line in the plot represents the running average It fluctuates at the beginning but gradually converges toward the expected value of 3.5. To emphasize this convergence, a dashed line indicating the theoretical expected value which is essentially the expected value applied to each roll. This visualization demonstrates the Law of Large Numbers, which states that as the number of trials or rolls increases, the *sample mean* (running average in this case) approaches the *population mean* (expected value) with greater accuracy, showing the predictability and stability of random processes over a large number of observations.

2.4.4 The Z-transform and the Galton Board



2.4.4.1 Applying the Z-transform

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{X - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

$$\lim_{n\to\infty} P(a \le Z \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}\, dz$$

2.4.4.2 Converting the bionmial Formula to a Normal Form

Stirling appoximation: $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$

Appprox: $\binom{n}{k} \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi n}\left(\frac{k}{e}\right)^k \cdot \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}}$

simplifies to: $\binom{n}{k} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}$

substituting $p = 0.5$: $P(X = k) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}}$

Which is the **Probability Density Function**

2.4.5 The drive shaft exercise - Normal Distribution

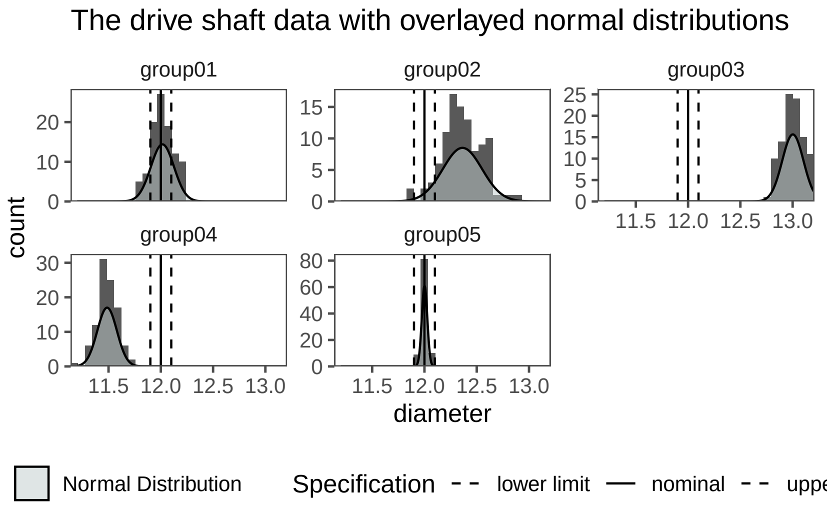


Figure 2.14: The drive shaft data with the respective normal distributions.

In Figure 2.14 the `drive shaft data` is shown for each group in a histogram. As an overlay, the respective *normal distribution* (with the groups $\bar{x}, sd$) is overlayed. If the data is normally distributed, is a different question.

2.5 Probability Density Function (PDF)

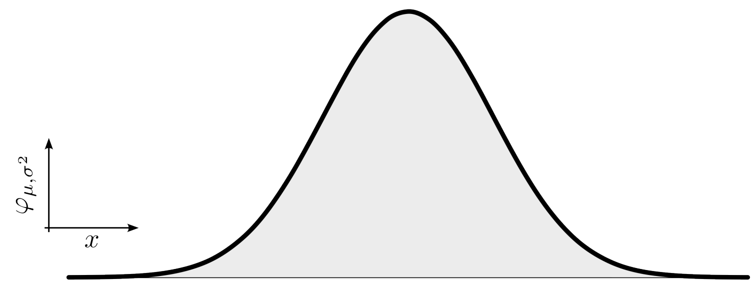


Figure 2.15: A visual represenstation of the PDF for the normal distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.3}$$

A probability density function (PDF) is a mathematical function that describes the *likelihood* of a continuous random variable taking on a particular value. Unlike discrete probability distributions, which assign probabilities to specific values of a discrete random variable, a PDF describes the relative likelihood of the variable falling within a particular range of values. The total area under the curve of a PDF over its entire range is equal to 1, indicating that the variable must take on some value within that range. In other words, the integral of the PDF over its entire domain equals 1. The probability of a continuous random variable falling within a specific interval is given by the integral of the PDF over that interval.

2.6 Cumulative Density Function (CDF)

A cumulative density function (CDF), also known as a cumulative distribution function, describes the probability that a random variable will take on a value less than or equal to a given point. It is the integral of the PDF from negative infinity to a certain value. The CDF provides a comprehensive view of the probability distribution of a random variable by showing how the probability accumulates as the value of the random variable increases. Unlike the PDF, which gives the probability density at a particular point, the CDF gives the cumulative probability up to that point.

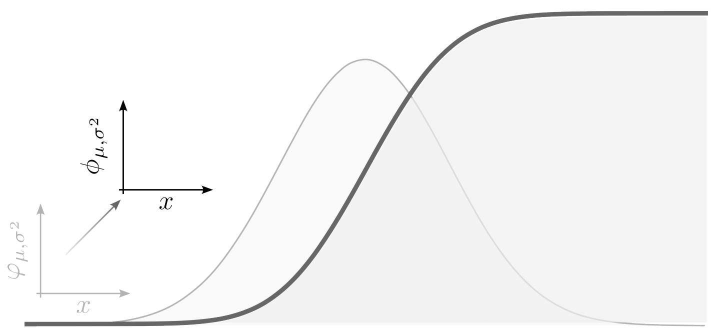


Figure 2.16: A visual represenstation of the CDF for the normal distribution.

$$z = \frac{x-\mu}{\sigma}$$

$$\varphi(x) = \frac{1}{2\pi} e^{\frac{-z^2}{2}} \tag{2.4}$$

$$\phi(x) = \int \frac{1}{2\pi} e^{\frac{-x^2}{2}}\, dx \tag{2.5}$$

$$\lim_{x\to\infty} \phi(x) = 1$$

$$\lim_{x\to-\infty} \phi(x) = 0$$

2.7 Likelihood and Probability

Likelihood refers to the chance or plausibility of a particular event occurring given certain evidence or assumptions. It is often used in statistical inference, where it indicates how well a particular set of parameters (or hypotheses) explain the observed data. Likelihood is a measure of how compatible the observed data are with a specific hypothesis or model.

Probability represents the measure of the likelihood that an event will occur. It is a quantification of uncertainty and ranges from 0 (indicating impossibility) to 1 (indicating certainty). Probability is commonly used to assess the chances of different outcomes in various scenarios.

In summary, while both likelihood and probability deal with the chance of events occurring, likelihood is often used in the context of comparing different *hypotheses or models* based on *observed data*, while probability is more broadly used to quantify the chances of *events happening* in *general*.

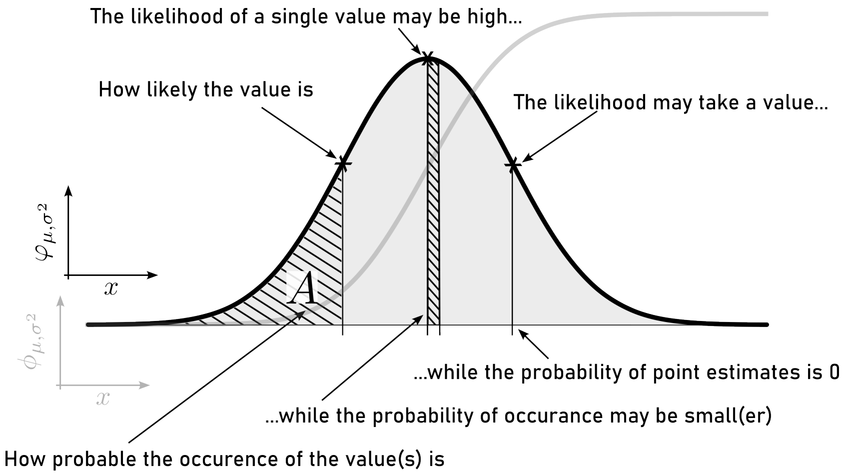
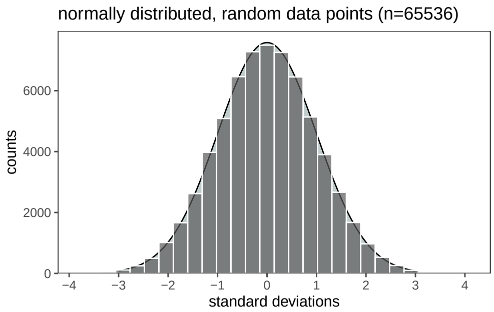
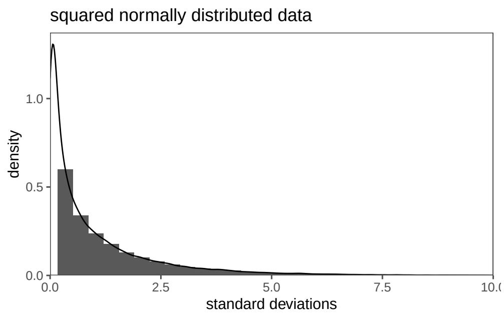


Figure 2.17: The subtle difference between likelihood and probability.

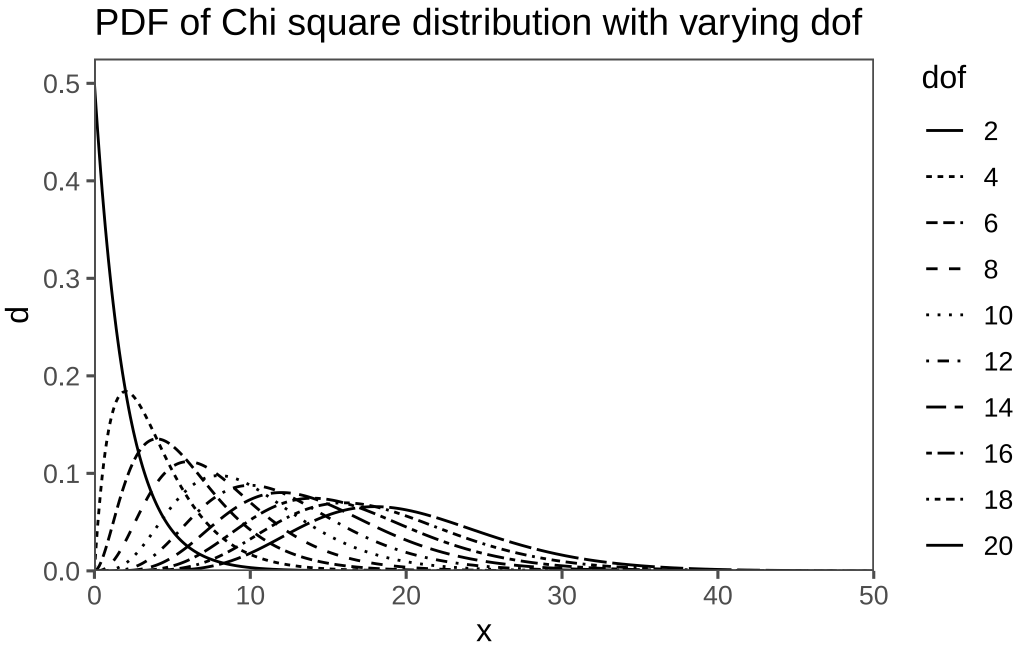
2.8 Chi² - Distribution



(a) a normal distribution



(b) the standard normal variable square ($dof = 1$)



(c) the χ^2 distributions with varying degrees of freedom

Figure 2.18: What a χ^2 distribution reprepresents and how it relates to a the normal distribution.

The χ^2 distribution is a continuous probability distribution that is widely used in statistics (Taboga 2017). It is often used to test hypotheses about the independence of categorical variables.

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{2.6}$$

The connection between the chi-squared distribution and sample variance holds significant importance in statistics.

1. **Distribution of Sample Variance:** When calculating the sample variance from a dataset, it follows a chi-squared distribution. Specifically, for a random sample from a normally distributed population with mean μ_0 and variance σ_0^2, the sample variance (adjusted for bias) divided by σ_0^2 follows a χ^2 distribution with $n-1$ degrees of freedom, where n is the sample size.
2. **Hypothesis Testing:** In statistical analysis, hypothesis testing is a common technique for making inferences about populations using sample data. The χ^2 distribution plays a crucial role in hypothesis testing, especially when comparing variances between samples.
    - **χ^2 Test for Variance:** The χ^2 distribution is used to test whether the variance of a sample matches a hypothesized variance. This is applicable in various scenarios, such as quality control, to assess the consistency of a manufacturing process.
3. **Confidence Intervals:** When estimating population parameters like population variance, it's essential to establish confidence intervals. The χ^2 distribution aids in constructing these intervals, allowing researchers to quantify the uncertainty associated with their parameter estimates.
4. **Model Assessment:** In regression analysis, the χ^2 distribution is related to the F-statistic, which assesses the overall significance of a regression model. It helps determine whether the regression model is a good fit for the data.

In summary, the link between the chi-squared distribution and sample variance is fundamental in statistical analysis. It empowers statisticians and analysts to make informed decisions about population parameters based on sample data and evaluate the validity of statistical models. Understanding this relationship is essential for those working with data and conducting statistical investigations.

2.8.1 The drive shaft exercise - Chi^2 Distribution

In Figure 2.19 the squared standad deviation for every datapoint (from the stanardized data) is shown as a histogram for every group with an overlayed (and scaled) density plot. In the background of every group the theoretical χ^2-distribution with $dof = 1$ is plotted to visually compare the empirical distribution of the datapoints to the theorectial.

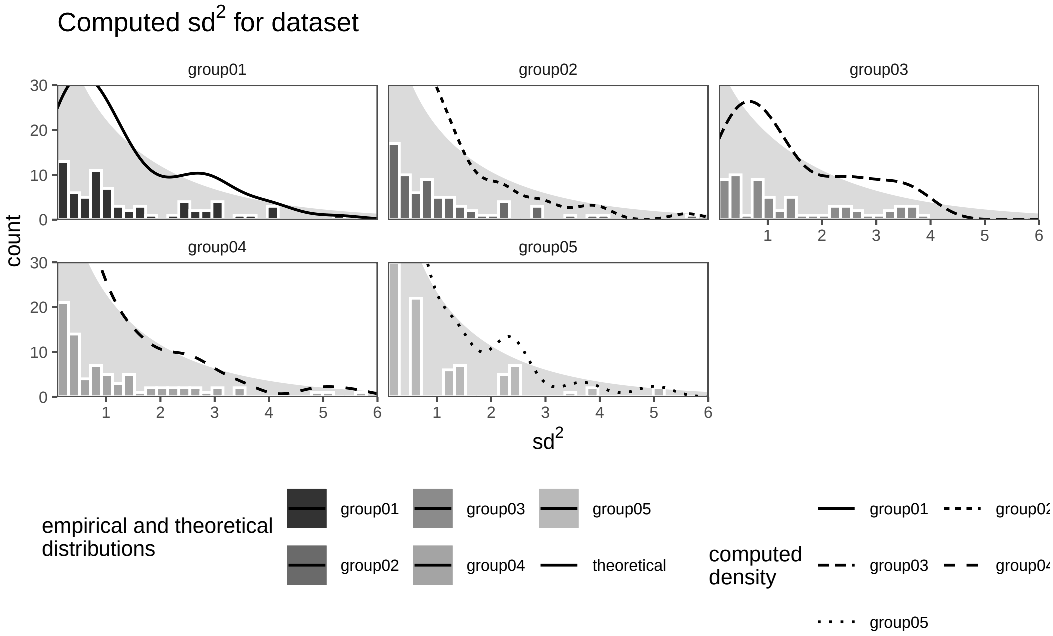


Figure 2.19: The χ^2 disitribution of the drive shaft data.

2.9 t - Distribution

The t-distribution, also known as the Student's t-distribution (Student 1908), is a probability distribution that plays a significant role in statistics[2]. It is a symmetric distribution with a bell-shaped curve, similar to the normal distribution, but with heavier tails. The key significance of the t-distribution lies in its application to inferential statistics, particularly in hypothesis testing and confidence interval estimation.

1. **Small Sample Sizes:** When dealing with small sample sizes (typically less than 30), the t-distribution is used to make inferences about population parameters, such as the mean. This is crucial because the normal distribution assumptions are often violated with small samples.

2. **Accounting for Variability:** The t-distribution accounts for the variability inherent in small samples. It provides wider confidence intervals and more conservative hypothesis tests compared to the normal distribution, making it more suitable for situations where sample size is limited.

3. **Degrees of Freedom:** The shape of the t-distribution is determined by a parameter called degrees of freedom (df). As the df increases, the t-distribution approaches

[2] William Sealy Gosset (June 13, 1876 - October 16, 1937) was a pioneering statistician known for developing the t-distribution, a key tool in modern statistical analysis.

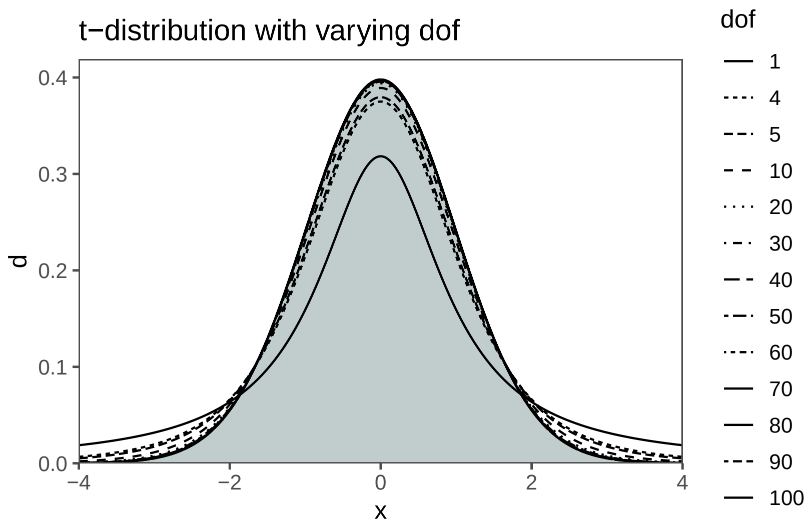


Figure 2.20: PDF of t-distribution with varying dof

> the normal distribution. When df is small, the tails of the t-distribution are fatter, allowing for greater uncertainty in estimates.

Statisticians found that if they took samples of a constant size from a normal population, computed a statistic called a *t-score* for each sample, and put those into a relative frequency distribution, the distribution would be the same for samples of the same size drawn from any normal population. The shape of this sampling distribution of t's varies somewhat as sample size varies, but for any n, it is always the same. For example, for samples of 5, 90% of the samples have t-scores between -1.943 and $+1.943$, while for samples of 15, 90% have t-scores between ± 1.761. The bigger the samples, the narrower the range of scores that covers any particular proportion of the samples (2.10) (Note the similarity to (2.2)). Since the *t-score* is computed for every x_i the resulting sampling distribution is called the *t-disitribution*.

$$t_i = \frac{x_i - \mu_o}{sd/\sqrt{n}} \tag{2.7}$$

In Figure 2.20 it is shown, that with increasing dof (in this case *sample size*), the *t-distribution* approximates a normal distribution (gray area). Figure 2.20 also shows an example of the *t-distribution* in action. Of all possible samples with 9 dof 0.025 ($2\frac{1}{2}$%) of those samples would have t-scores greater than 2.262, and .975 (97.5%) would have t-scores less than 2.262. The advantage of the *t-score* and *t-distribution* is clearly visible. All these values can be computed from sampled data, the population can remain *estimated* (2.10).

2.9.1 The drive shaft exercise - t-Distribution

The t-score computation and the z-standardization look very familiar. While the z-score calculation needs some population parameters, the t-score calculation does not need such. It therefore allows us, to estimate population parameters based on a sample - a very frequent use case in statistics.

Suppose we have some data (maybe the drive shaft exercise?) with which calculations can be done. First, the mean $\bar{x}$ and sd is calculated according to (1.13) and (1.14). After this, the confidence level (we will get to this later in more detail) is specified. A value of 95% is a common choice of cl.

$$ci = 0.95 \quad \text{(for a 95\% confidence level)} \tag{2.8}$$

Then the Standard Error (SE) is calculated using (2.9), which takes the sd and n of a sample into account (notice, how we did not use any population estimation?).

$$SE = \frac{sd}{\sqrt{n}} \tag{2.9}$$

In the next step, the critical *t-score* is calculated using the cl as shown in (2.10). qt in this case returns the value of the inverse cumulative function of the t-distribution given a certain random variable (or datapoint x_i) and $n-1$ dof. Think of it as an automated look up in long statistical tables.

$$t_{score} = qt\left(\frac{1-ci}{2}, df = n-1\right) \tag{2.10}$$

With this, the *margin of error* can be calculated using the SE and the *t-score* as shown in (2.11).

$$margin\ of\ error = t_{score} \times SE \tag{2.11}$$

In the last step the Confidence Interval is calculated for the `lower` and the `upper` bound with (2.12) and (2.13).

$$lo = \bar{x} - margin\ of\ error \tag{2.12}$$
$$hi = \bar{x} + margin\ of\ error \tag{2.13}$$

It all looks and feels very similar to using the normal disitrbution. Why this is the case, is shown in Figure 2.21. In **?@fig-ds-t-1** the raw dataset is shown with the underlayed specification limits for the manufacturing of the drive shaft. For some groups the judgement if the drive shaft is wihtin specification is quite clear (`group 1`, `group 2` and `group 5`). For the other groups, this can not be done so easily. For the drive shaft data, we of course now some population data, therefore the *normal distribution* can be compared to the *t-distribution*. This is done in **?@fig-ds-t-2**. On the `x-axis` the diameter is shown, the `y-axis` depicts the groups (as before). The distribution on top of the estimated parameters is the population (normal distribution), the distribution on the bottom follow a *t-distribution*. With $n > 30$ (as for this dataset), the difference between disitrbution is very small, further showcasing the use of the *t-distribution* (also see Figure 2.20 for comparison).

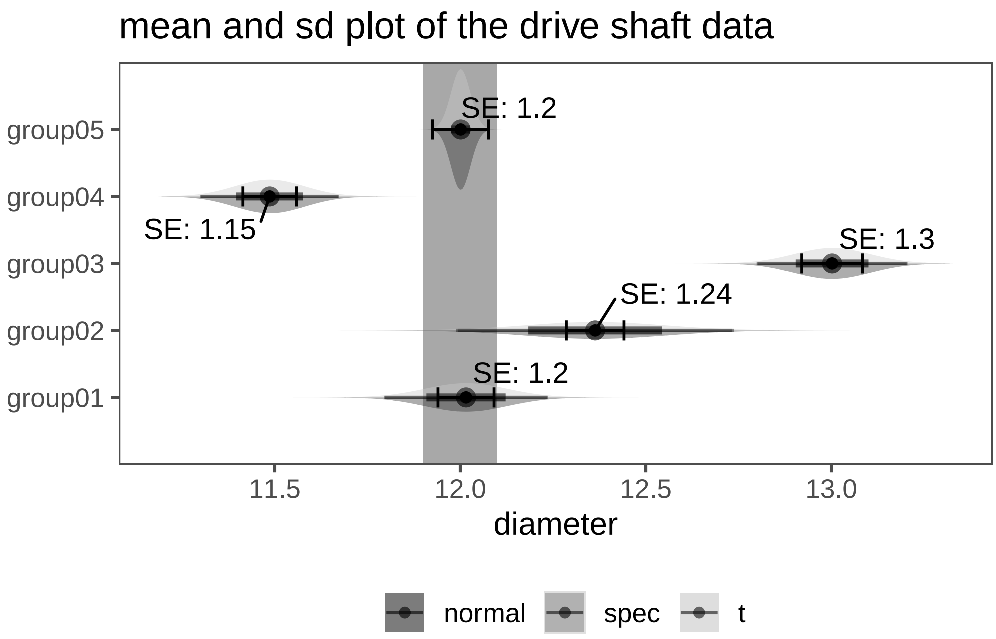
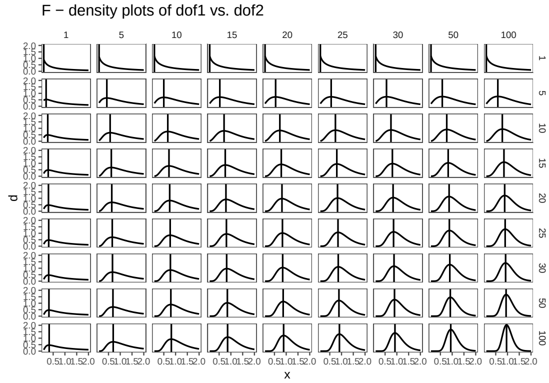


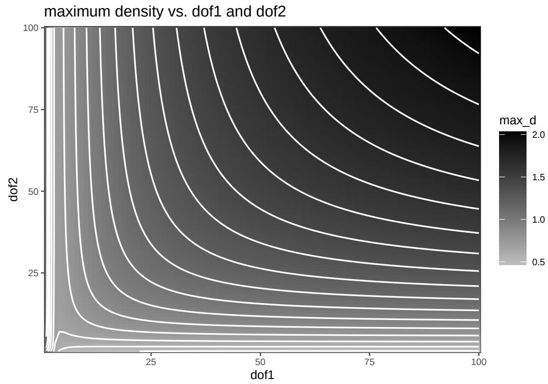
Figure 2.21: The drive shaft data with normal disitribution, t-distribution and confidence intervalls using the t-distribution

2.10 F - Statistics

F-statistics, also known as the *F-test* or *F-ratio*, is a statistical measure used in analysis of variance and regression analysis (Taboga 2017). It assesses the ratio of two variances, indicating the extent to which the variability between groups or models is greater than the variability within those groups or models. The *F-statistic* plays a crucial role in hypothesis testing and model comparison.



(a) F-distribution for dof_1 on the horizontal and dof_2 on the vertical axis



(b) the maximum density as a function of dof_1 and dof_2 in a continous parameter space

Figure 2.22: The influence of dof_1 and dof_2 on the density in the F-disitribution

Significance of F-statistics: The significance of the F-statistic lies in its ability to help researchers determine whether the differences between group means or the goodness-of-fit of a regression model are statistically significant. In ANOVA, a high F-statistic suggests that at least one group mean differs significantly from the others, while in regression analysis, it indicates whether the regression model as a whole is a good fit for the data.

Applications of F-statistics: 1. **Analysis of Variance (ANOVA):** F-statistics are extensively used in ANOVA to compare means across two or more groups. It helps determine whether there are significant differences among the means of these groups. For example, an ANOVA might be used to compare the mean test scores of students taught using different teaching methods.

2. **Regression Analysis:** F-statistics are used in regression analysis to assess the overall significance of a regression model. Specifically, in multiple linear regression, it helps determine whether the model, which includes multiple predictor variables, is better at explaining the variance in the response variable compared to a model

with no predictors. It tests the null hypothesis that all coefficients of the model are equal to zero.

The degrees of freedom in an *F-distribution* refer to the two sets of numbers that determine the shape and properties of the distribution (Figure 2.22).

Numerator Degrees of Freedom (dof_1): The numerator degrees of freedom, often denoted as dof_1, is associated with the variability between groups or models in statistical analyses (Figure 2.22a - horizontal axis). In the context of ANOVA, it represents the dof associated with the differences among group means. In regression analysis, it is related to the number of predictors or coefficients being tested simultaneously.

Denominator Degrees of Freedom (dof_2): The denominator degrees of freedom, often denoted as dof_2, is associated with the variability within groups or models (Figure 2.22b - vertical axis). In ANOVA, it represents the degrees of freedom associated with the variability within each group. In regression analysis, it is related to the error or residual degrees of freedom, indicating the remaining variability not explained by the model.

The F-distribution is used to compare two variances: one from the numerator and the other from the denominator. The F-statistic, calculated as the ratio of these variances, follows an F-distribution (2.14).

$$f(x; dof_1, dof_2) = \frac{\Gamma\left(\frac{dof_1+dof_2}{2}\right)}{\Gamma\left(\frac{dof_1}{2}\right)\Gamma\left(\frac{dof_2}{2}\right)} \left(\frac{dof_1}{dof_2}\right)^{\frac{dof_1}{2}} \frac{x^{\frac{dof_1}{2}-1}}{\left(1+\frac{dof_1}{dof_2}x\right)^{\frac{dof_1+dof_2}{2}}} \tag{2.14}$$

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n} \tag{2.15}$$

In practical terms: A higher numerator degrees of freedom (dof_1) suggests that there are more groups or predictors being compared, which may result in larger F-statistic values. A higher denominator degrees of freedom (dof_2) implies that there is more data within each group or model, which may lead to smaller F-statistic values. The F-distribution is right-skewed and always positive. It has different shapes depending on the values of dof_1 and dof_2 (Figure 2.22b). The exact shape is determined by these degrees of freedom and cannot be altered by changing sample sizes or data values (Figure 2.22b). Researchers use F-distributions to conduct hypothesis tests, such as F-tests in ANOVA and F-tests in regression, to determine if there are significant differences between groups or if a regression model is statistically significant.

In summary, degrees of freedom in the F-distribution are critical in hypothesis testing and model comparisons. They help quantify the variability between and within groups or models, allowing statisticians to assess the significance of observed differences and make informed statistical decisions.

2.11 Interconnections

1. Normal Distribution The **Normal Distribution** is characterized by its mean (μ) and standard deviation (σ), see Figure 2.23. It serves as the foundation for many statistical analyses.

2. Standardized Normal Distribution The **Standardized Normal Distribution**, denoted as $Z \sim N(0, 1)$, is a special case of the normal distribution. It has a mean (μ) of 0 and a standard deviation (σ) of 1. It is obtained by standardizing a normal distribution variable X: $Z = \frac{X-\mu}{\sigma}$ (Figure 2.23).

3. t Distribution The **t Distribution** is related to the normal distribution and depends on degrees of freedom. As dof increases, the t-distribution approaches the standard normal distribution (Figure 2.23).

4. Chi-Square Distribution The **Chi-Square Distribution** is indirectly connected to the normal distribution through the concept of "sum of squared standard normals." When standard normal random variables (Z) are squared and summed, the resulting distribution follows a chi-square distribution.

5. F Distribution The **F Distribution** arises from the ratio of two independent chi-square distributed random variables. It is used for comparing variances between groups in statistical tests like ANOVA.

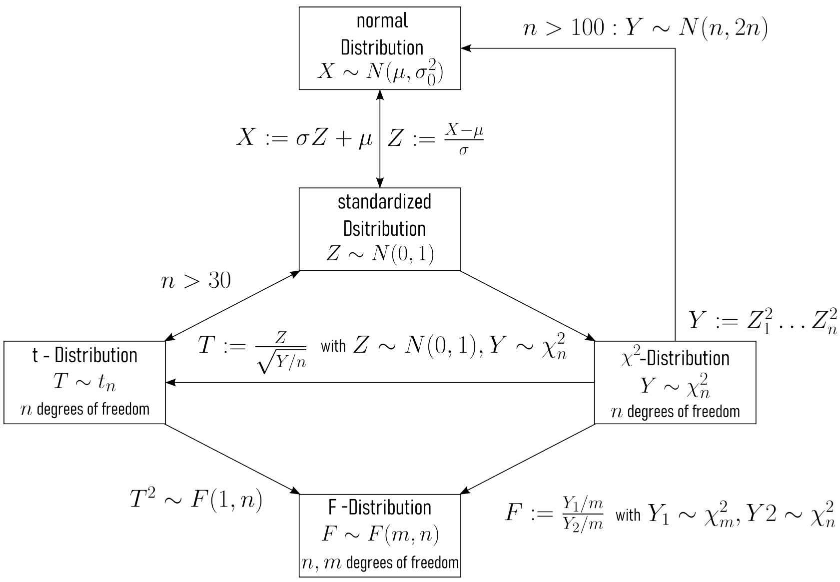


Figure 2.23: The distributions are interconnected in several different ways.

2.12 Weibull - Distribution

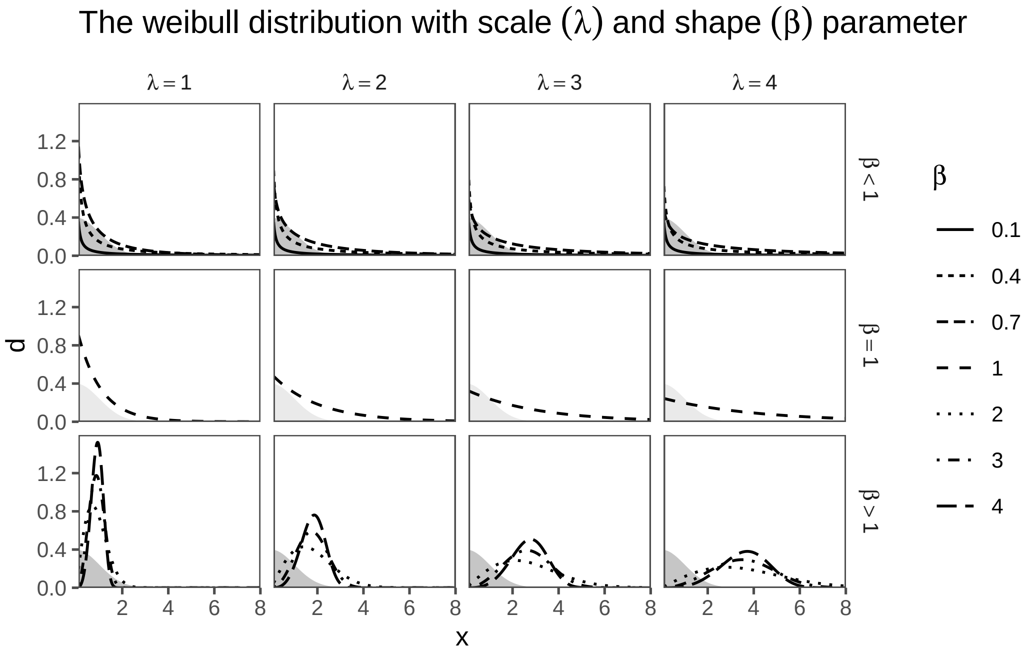


Figure 2.24: The weibull distribution and the influence of β and λ

The Weibull distribution is a probability distribution frequently used in statistics and reliability engineering to model the time until an event, particularly failures or lifetimes. It is named after Wallodi Weibull[3], who developed it in the mid-20th century (Weibull 1951).

The Weibull distribution is characterized by two parameters:

Shape Parameter (β): This parameter determines the shape of the distribution curve and can take on values greater than 0. Depending on the value of β, the Weibull distribution can exhibit different behaviors:

If $\beta < 1$, the distribution has a decreasing failure rate, indicating that the probability of an event occurring decreases over time. This is often associated with "infant mortality" or early-life failures. If $\beta = 1$, the distribution follows an exponential distribution with a constant failure rate over time. If $\beta > 1$, the distribution has an increasing failure rate, suggesting that the event becomes more likely as time progresses. This is often associated with "wear-out" failures.

[3] Waloddi Weibull (1887–1979) was a Swedish engineer and statistician known for his work on the Weibull distribution, which is widely used in reliability engineering and other fields.

Scale Parameter (λ): This parameter represents a characteristic scale or location on the time axis. It influences the position of the distribution on the time axis. A larger λ indicates that events are more likely to occur at later times.

Applications: - Reliability Engineering: The Weibull distribution is extensively used in reliability engineering to assess the lifetime and failure characteristics of components and systems. Engineers can estimate the distribution parameters from data to predict product reliability, set warranty periods, and plan maintenance schedules.

- Survival Analysis: In medical research and epidemiology, the Weibull distribution is employed to analyze survival data, such as time until the occurrence of a disease or death. It helps in modeling and understanding the progression of diseases and the effectiveness of treatments.
- Economics and Finance: The Weibull distribution is used in finance to model the time between financial events, like market crashes or loan defaults. It can provide insights into risk assessment and portfolio management.

2.12.1 The drive shaft exercise - Weibull distribution

The weibull distribution can be applied to estimate the probability of a part to fail after a given time. Suppose there have been $n = 100$ drive shafts produced. In order to assure that the assembled drive shaft would last during their service time, they have been tested in a test-stand that mimics the mission profile[4] of the product. This process is called *qualification* and a big part of any product development (Meyna 2023). The measured hours are shown in Figure 2.25 in a histogram of the data. On the `x-axis` the `Time to failure`is shown, while the `y-axis` shows the number of parts that failed within the time. They histogram plot is overlayed with an empirical density plot as a solid line, as well as the theoretical distribution as a dotted line (Luckily, we know the distribution parameters).

2.13 Poisson - Distribution

The Poisson distribution is a probability distribution commonly used in statistics to model the number of events that occur within a fixed interval of time or space, given a known average rate of occurrence. It is named after the French mathematician Siméon Denis Poisson[5].

[4] A mission profile for parts is a detailed plan specifying how specific components in a system should perform, considering factors like environment, performance, safety, and compliance.

[5] Siméon Denis Poisson (1781-1840) was a notable French mathematician, renowned for his work in probability theory and mathematical physics.

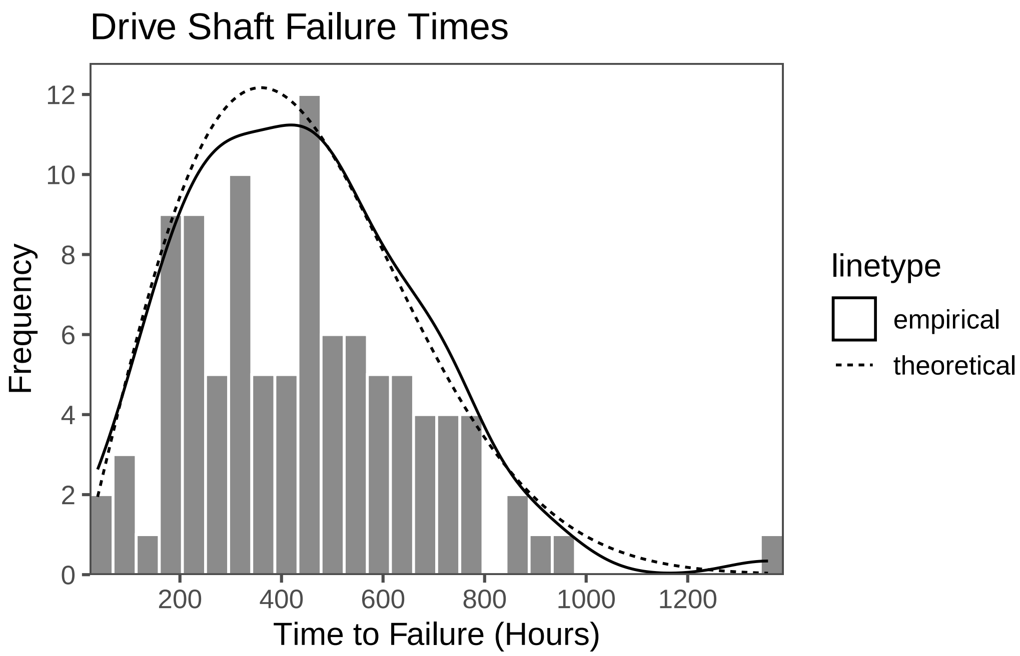


Figure 2.25: The measured hours how long the drive shafts lasted in the test stand.

The Poisson distribution is an applicable probability model in such situations under specific conditions:

1. Independence: Events should occur independently of each other within the specified interval of time or space. This means that the occurrence of one event should not affect the likelihood of another event happening.

2. Constant Rate: The average rate (*lambda*, denoted as λ) at which events occur should be constant over the entire interval. In other words, the probability of an event occurring should be the same at any point in the interval.

3. Discreteness: The events being counted must be discrete in nature. This means that they should be countable and should not take on continuous values.

4. Rare Events: The Poisson distribution is most appropriate when the events are rare, meaning that the probability of more than one event occurring in an infinitesimally small interval is negligible. This assumption helps ensure that the distribution models infrequent events.

5. Fixed Interval: The interval of time or space in which events are counted should be fixed and well-defined. It should not vary or be open-ended.

6. Memorylessness: The Poisson distribution assumes that the probability of an event occurring in the future is independent of past events. In other words, it does not take into account the history of events beyond the current interval.

7. Count Data: The Poisson distribution is most suitable for count data, where you are interested in the number of events that occur in a given interval.

In the context of a Poisson distribution, the parameter lambda (λ) represents the average rate of events occurring in a fixed interval of time or space. It is a crucial parameter that helps define the shape and characteristics of the Poisson distribution.

Average Rate: λ is a positive real number that represents the average or expected number of events that occur in the specified interval. It tells you, on average, how many events you would expect to observe in that interval.

Rate of Occurrence: λ quantifies the rate at which events happen. A higher value of λ indicates a higher rate of occurrence, while a lower value of λ indicates a lower rate.

Shape of the Distribution: The value of λ determines the shape of the Poisson distribution. Specifically:

When λ is small, the distribution is skewed to the right and is more concentrated toward zero (Figure 2.26). When λ is moderate, the distribution approaches a symmetric bell shape (Figure 2.26). When λ is large, the distribution becomes increasingly similar to a normal distribution(Figure 2.26).

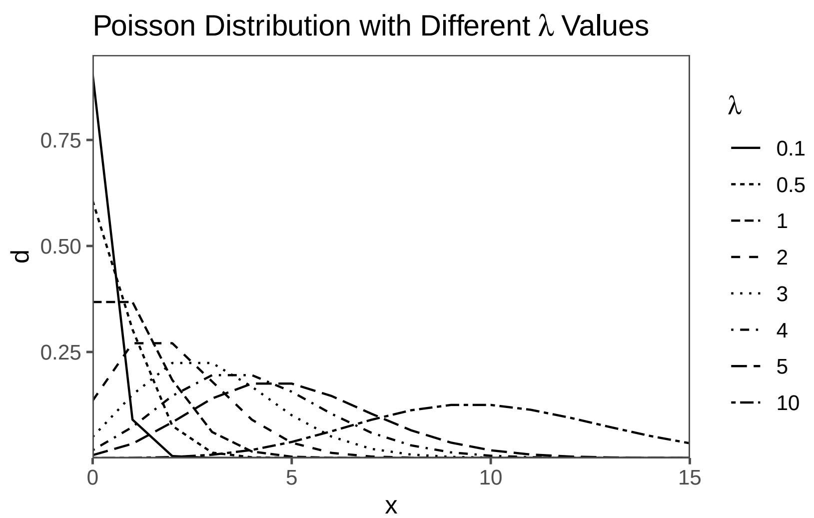


Figure 2.26: The poisson distribution with different λ values.

2.14 Gamma - Distribution

The gamma distribution is a probability distribution that is often used in statistics to model the waiting time until a Poisson process reaches a certain number of events. It is a continuous probability distribution with two parameters, typically denoted as α (shape parameter) and β (rate parameter).

Key points about the gamma distribution:

1. It is often used to model the waiting times for events that occur at a constant rate, such as the time between arrivals in a Poisson process.
2. The exponential distribution is a special case of the gamma distribution when $\alpha = 1$ (Figure 2.27).
3. The gamma distribution is right-skewed for $\alpha > 1$ and left-skewed for $0 < \alpha < 1$ (Figure 2.27).
4. The mean of the gamma distribution is $\frac{\alpha}{\beta}$, and its variance is $\frac{\alpha}{\beta^2}$ (Figure 2.27).

It is widely used in various fields, including reliability analysis, queuing theory, and finance.

The connection to other distributions:

Exponential Distribution: The exponential distribution is a special case of the gamma distribution with $\alpha = 1$.

χ^2: When α is an integer, the gamma distribution with shape parameter α is equivalent to the chi-squared distribution with 2α degrees of freedom.

Erlang Distribution: The Erlang distribution is a specific case of the gamma distribution where α is an integer, representing the sum of α exponentially distributed random variables.

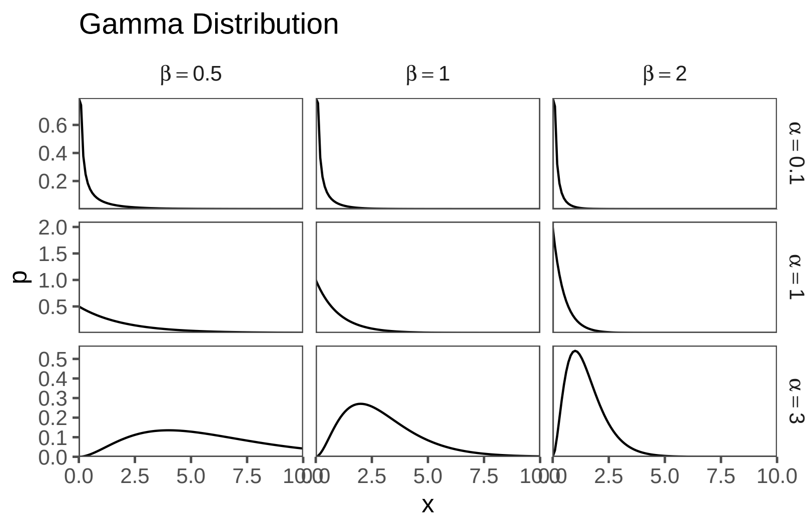
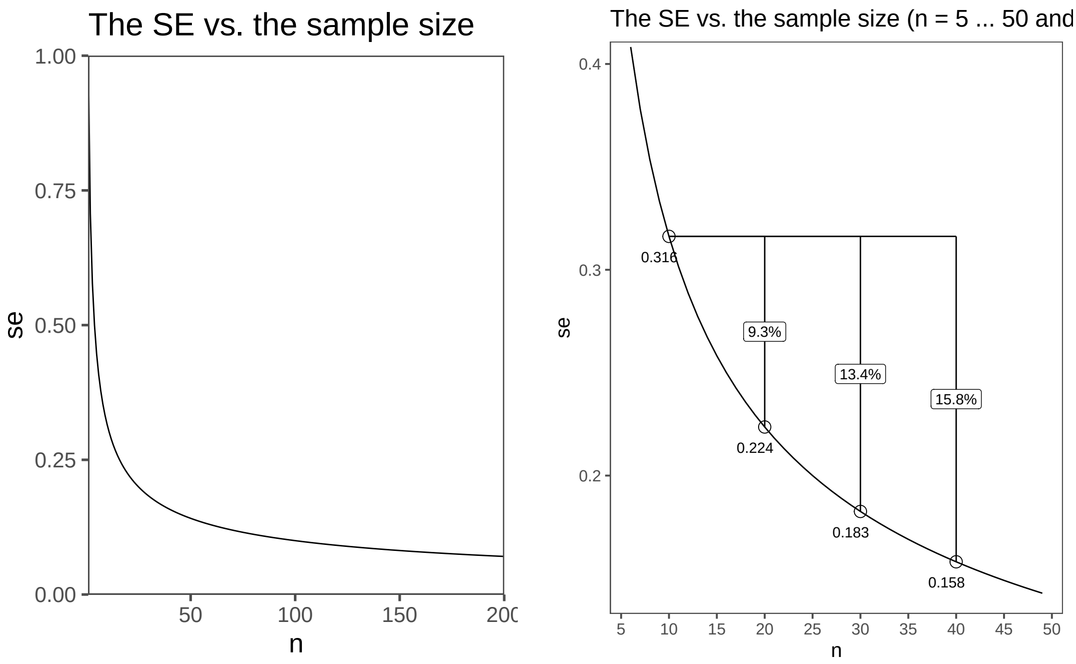


Figure 2.27: The Gamma distribution with varying α (shape) and β (scale)

3 Sampling Methods

3.1 Sample Size

3.1.1 Standard Error



(a) maximum sample size $n = 200$ (b) sample size for $n = 5 \ldots 50$

Figure 3.1: The SE for varying sample sizes n

Standard error is a statistical measure that quantifies the variation or uncertainty in sample statistics, particularly the mean (average). It is a valuable tool in inferential statistics and provides an estimate of how much the sample mean is expected to vary from the true population mean.

$$SE = \frac{sd}{\sqrt{n}} \tag{3.1}$$

A smaller standard error indicates that the sample mean is likely very close to the population mean, while a larger standard error suggests greater variability and less precision in estimating the population mean. Standard error is crucial when constructing confidence intervals and performing hypothesis tests, as it helps in assessing the reliability of sample statistics as estimates of population parameters.

Variance vs. Standard Deviation: The standard error formula is based on the standard deviation of the sample, not the variance. The standard deviation is the square root of the variance.

Scaling of Variability: The purpose of the standard error is to measure the variability or spread of sample means. The square root of the sample size reflects how that variability decreases as the sample size increases. When the sample size is larger, the sample mean is expected to be closer to the population mean, and the standard error becomes smaller to reflect this reduced variability.

Central Limit Theorem: The inclusion of $\sqrt{n}$ in the standard error formula is closely tied to the Central Limit Theorem, which states that the distribution of sample means approaches a normal distribution as the sample size increases. $\sqrt{n}$ helps in this context to ensure that the standard error appropriately reflects the distribution's properties.

3.2 Random Sampling

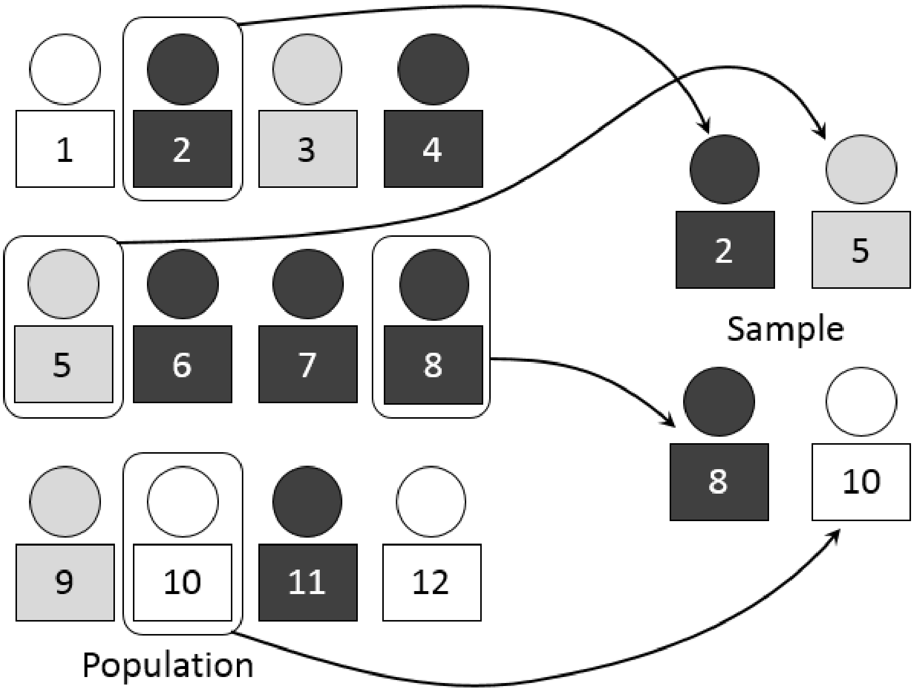


Figure 3.2: The idea of random sampling (Dan Kernler).

- **Definition:** Selecting a sample from a population in a purely random manner, where every individual has an equal chance of being chosen.
- **Advantages:**
  - Eliminates bias in selection.
  - Results are often representative of the population.
- **Disadvantages:**
  - Possibility of unequal representation of subgroups.
  - Time-consuming and may not be practical for large populations.

3.3 Stratified Sampling

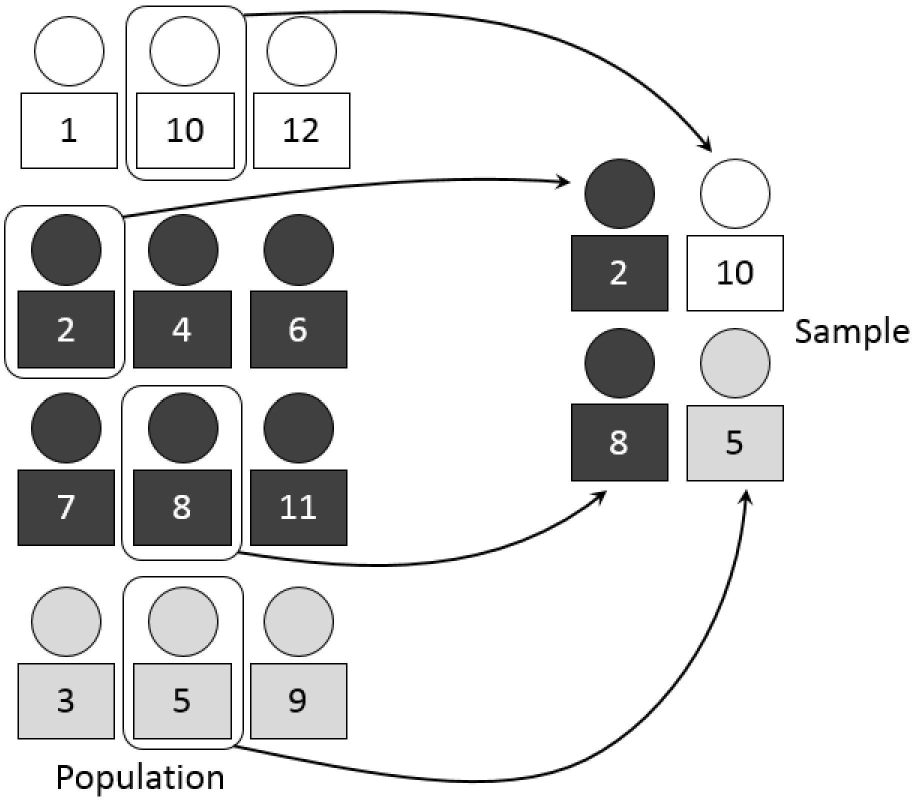


Figure 3.3: The idea of stratified sampling (Dan Kernler)

- **Definition:** Dividing the population into subgroups or strata based on certain characteristics and then randomly sampling from each stratum.
- **Advantages:**
  - Ensures representation from all relevant subgroups.
  - Increased precision in estimating population parameters.
- **Disadvantages:**
  - Requires accurate classification of the population into strata.
  - Complexity in implementation and analysis.

3.4 Systematic Sampling

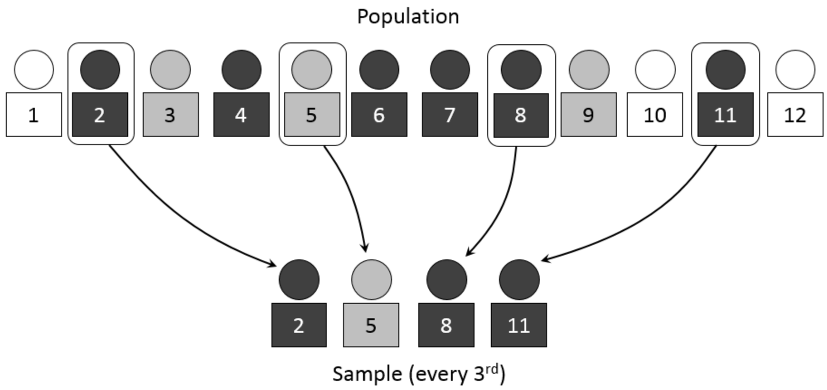


Figure 3.4: The idea of systematic sampling (Dan Kernler)

- **Definition:** Choosing every kth individual from a list after selecting a random starting point.
- **Advantages:**
  - Simplicity in execution compared to random sampling.
  - Suitable for large populations.
- **Disadvantages:**
  - Susceptible to periodic patterns in the population.
  - If the periodicity aligns with the sampling interval, it can introduce bias.

3.5 Cluster Sampling

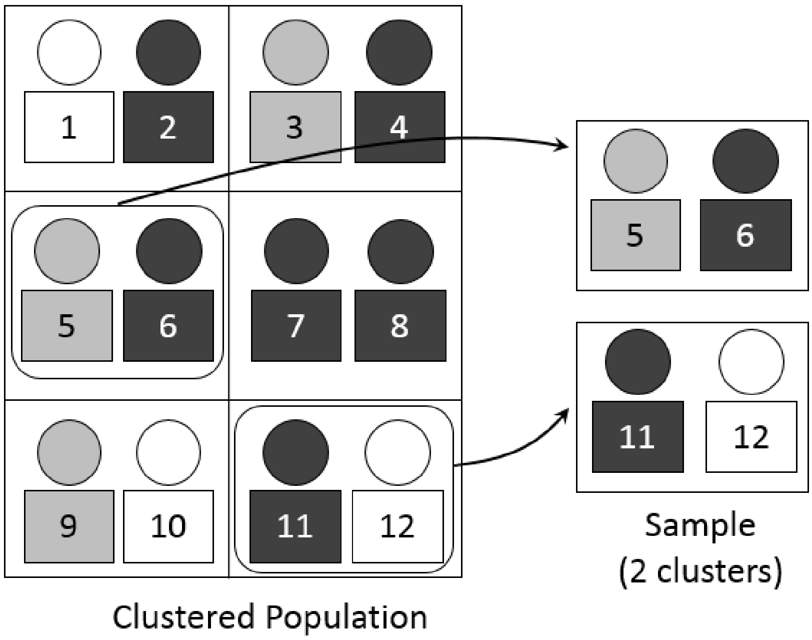


Figure 3.5: The idea of clustered sampling (Dan Kernler).

- **Definition:** Dividing the population into clusters, randomly selecting some clusters, and then including all individuals from the chosen clusters in the sample.

- **Advantages:**

  - Cost-effective, especially for geographically dispersed populations.
  - Reduces logistical challenges compared to other methods.

- **Disadvantages:**

  - Increased variability within clusters compared to other methods.
  - Requires accurate information on cluster characteristics.

Table 3.1: The starwars dataset

```
    name               height           mass            hair_color
 Length:87          Min.   : 66.0   Min.   :  15.00   Length:87
 Class :character   1st Qu.:167.0   1st Qu.:  55.60   Class :character
 Mode  :character   Median :180.0   Median :  79.00   Mode  :character
                    Mean   :174.6   Mean   :  97.31
                    3rd Qu.:191.0   3rd Qu.:  84.50
                    Max.   :264.0   Max.   :1358.00
                    NA's   :6       NA's   :28
  skin_color         eye_color          birth_year          sex
 Length:87          Length:87          Min.   :  8.00   Length:87
 Class :character   Class :character   1st Qu.: 35.00   Class :character
 Mode  :character   Mode  :character   Median : 52.00   Mode  :character
                                       Mean   : 87.57
                                       3rd Qu.: 72.00
                                       Max.   :896.00
                                       NA's   :44
   gender           homeworld           species
 Length:87          Length:87          Length:87
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

3.6 Example - The Star Wars dataset

3.6.1 Get to know the data

3.6.2 Simple Random Sampling

```r
starwars_srswor <- starwars %>%
  sample_n(size = 5)
starwars_srswor
```

```
# A tibble: 5 x 11
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Jek Tono~    180   110 brown      fair       blue              NA <NA>  <NA>
2 Rey           NA    NA brown      light      hazel             NA fema~ femin~
3 Shmi Sky~    163    NA black      fair       brown             72 fema~ femin~
4 C-3PO        167    75 <NA>       gold       yellow           112 none  mascu~
5 Yoda          66    17 white      green      brown            896 male  mascu~
# i 2 more variables: homeworld <chr>, species <chr>
```

3.6.3 Simple Random Sampling with replacment

```r
starwars_srswr <- starwars %>%
  sample_n(size = 5,
           replace = TRUE)
starwars_srswr
```

```
# A tibble: 5 x 11
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Zam Wese~    168    55 blonde     fair, gre~ yellow            NA fema~ femin~
2 Ben Quad~    163    65 none       grey, gre~ orange            NA male  mascu~
3 Ben Quad~    163    65 none       grey, gre~ orange            NA male  mascu~
4 Mas Amed~    196    NA none       blue       blue              NA male  mascu~
5 Cordé        157    NA brown      light      brown             NA <NA>  <NA>
# i 2 more variables: homeworld <chr>, species <chr>
```

3.6.4 Sampling with replacment, sample larger than original data

```
starwars_srswr2 <- starwars %>%
  sample_n(size = 200,
           replace = TRUE)
starwars_srswr2
```

```
# A tibble: 200 x 11
   name      height  mass hair_color skin_color eye_color birth_year sex   gender
   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
 1 Jocasta~     167    NA white      fair       blue              NA fema~ femin~
 2 Ric Olié     183    NA brown      fair       blue              NA male  mascu~
 3 IG-88        200   140 none       metal      red               15 none  mascu~
 4 Jocasta~     167    NA white      fair       blue              NA fema~ femin~
 5 IG-88        200   140 none       metal      red               15 none  mascu~
 6 Cordé        157    NA brown      light      brown             NA <NA>  <NA>
 7 Poe Dam~      NA    NA brown      light      brown             NA male  mascu~
 8 Palpati~     170    75 grey       pale       yellow            82 male  mascu~
 9 Padmé A~     185    45 brown      light      brown             46 fema~ femin~
10 Rey           NA    NA brown      light      hazel             NA fema~ femin~
# i 190 more rows
# i 2 more variables: homeworld <chr>, species <chr>
```

```
mean(starwars$height, na.rm = TRUE)
```

```
[1] 174.6049
```

```
mean(starwars_srswr2$height, na.rm = TRUE)
```

```
[1] 173.172
```

3.6.5 Systematic Sampling

Sample always the $5th$.

```
starwars_syst <- starwars %>%
  slice(seq(sample(1:5, 1),
            nrow(starwars),
            by = 5))
starwars_syst
```

```
# A tibble: 17 x 11
   name      height  mass hair_color skin_color eye_color birth_year sex   gender
   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
 1 Darth V~     202   136 none       white      yellow          41.9 male  mascu~
 2 Biggs D~     183    84 black      light      brown           24   male  mascu~
 3 Han Solo     180    80 brown      fair       brown           29   male  mascu~
 4 Yoda          66    17 white      green      brown          896   male  mascu~
 5 Lando C~     177    79 black      dark       brown           31   male  mascu~
 6 Wicket ~      88    20 brown      brown      brown            8   male  mascu~
 7 Padmé A~     185    45 brown      light      brown           46   fema~ femin~
 8 Watto        137    NA black      blue, grey yellow          NA   male  mascu~
 9 Bib For~     180    NA none       pale       pink            NA   male  mascu~
10 Ben Qua~     163    65 none       grey, gre~ orange          NA   male  mascu~
11 Adi Gal~     184    50 none       dark       blue            NA   fema~ femin~
12 Gregar ~     185    85 black      dark       brown           NA   <NA>  <NA>
13 Barriss~     166    50 black      yellow     blue            40   fema~ femin~
14 Zam Wes~     168    55 blonde     fair, gre~ yellow          NA   fema~ femin~
15 R4-P17        96    NA none       silver, r~ red, blue       NA   none  femin~
16 Tarfful      234   136 brown      brown      blue            NA   male  mascu~
17 Rey           NA    NA brown      light      hazel           NA   fema~ femin~
# i 2 more variables: homeworld <chr>, species <chr>
```

3.6.6 Stratified Sampling

```r
table(starwars$sex)
```

```

        female hermaphroditic           male           none
            16              1             60              6
```

```r
starwars_strat <- starwars %>%
  group_by(sex) %>%
  sample_frac(size = 0.3)
starwars_strat
```

```
# A tibble: 26 x 11
# Groups:   sex [4]
   name      height  mass hair_color skin_color eye_color birth_year sex   gender
   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
 1 Ayla Se~     178  55   none       blue       hazel             48 fema~ femin~
```

```
 2 Luminar~    170  56.2 black      yellow     blue              58 fema~ femin~
 3 Jocasta~    167  NA   white      fair       blue              NA fema~ femin~
 4 Shmi Sk~    163  NA   black      fair       brown             72 fema~ femin~
 5 Taun We     213  NA   none       grey       black             NA fema~ femin~
 6 Finn         NA  NA   black      dark       dark              NA male  mascu~
 7 Rugor N~    206  NA   none       green      orange            NA male  mascu~
 8 Lobot       175  79   none       light      blue              37 male  mascu~
 9 Jar Jar~    196  66   none       orange     orange            52 male  mascu~
10 Qui-Gon~    193  89   brown      fair       blue              92 male  mascu~
# i 16 more rows
# i 2 more variables: homeworld <chr>, species <chr>
```

```
table(starwars_strat$sex)
```

```
female   male   none 
     5     18      2 
```

3.6.7 Clustered Sampling

3.7 Bootstrapping

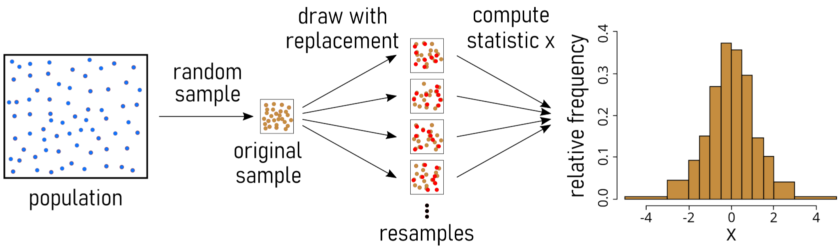


Figure 3.6: The idea of bootstrapping (Biggerj1, Marsupilami)

- **Definition:** Estimating sample statistic distribution by drawing new samples with replacement from observed data, providing insights into variability without strict population distribution assumptions.
- **Advantages:**
  - Non-parametric: Works without assuming a specific data distribution.
  - Confidence Intervals: Facilitates easy estimation of confidence intervals.
  - Robustness: Reliable for small sample sizes or unknown data distributions.

Table 3.2: The starwars dataset with clustered sampling

```
    name               height           mass          hair_color       
 Length:19          Min.   : 97.0   Min.   :  32.0   Length:19         
 Class :character   1st Qu.:169.5   1st Qu.:  75.0   Class :character  
 Mode  :character   Median :178.0   Median :  79.0   Mode  :character  
                    Mean   :173.9   Mean   : 171.2                     
                    3rd Qu.:188.0   3rd Qu.: 116.5                     
                    Max.   :216.0   Max.   :1358.0                     
                                    NA's   :4                          
  skin_color         eye_color           birth_year         sex           
 Length:19          Length:19          Min.   : 19.00   Length:19         
 Class :character   Class :character   1st Qu.: 37.00   Class :character  
 Mode  :character   Mode  :character   Median : 47.00   Mode  :character  
                                       Mean   : 93.29                     
                                       3rd Qu.: 72.00                     
                                       Max.   :600.00                     
                                       NA's   :6                          
    gender           homeworld           species         
 Length:19          Length:19          Length:19         
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  
```

- **Disadvantages:**
  - Computationally Intensive: Resource-intensive for large datasets.
  - Results quality relies on the representativeness of the initial sample (garbage in - garbage out).
  - Cannot compensate for inadequate information in the original sample.
  - Not Always Optimal: Traditional methods may be better in cases meeting distribution assumptions.

4 Inferential Statistics

Inferential statistics involves making predictions, generalizations, or inferences about a population based on a sample of data. These techniques are used when researchers want to draw conclusions beyond the specific data they have collected. Inferential statistics help answer questions about relationships, differences, and associations within a population.

4.1 Hypothesis Testing - Basics

Figure 4.1: We are hypotheses.

Null Hypothesis (H0): This is the default or status quo assumption. It represents the belief that there is no significant change, effect, or difference in the production process. It is often denoted as a statement of equality (e.g., the mean production rate is equal to a certain value).

Alternative Hypothesis (Ha): This is the claim or statement we want to test. It represents the opposite of the null hypothesis, suggesting that there is a significant change, effect, or difference in the production process (e.g., the mean production rate is not equal to a certain value).

4.1.1 The drive shaft exercise - Hypotheses

During the QC of the drive shaft $n = 100$ samples are taken and the diameter is measured with an accuracy of $\pm 0.01mm$. Is the true mean of all produced drive shafts within the specification?

For this we can formulate the hypotheses.

H0: The drive shaft diameter is within the specification.
Ha: The drive shaft diameter is not within the specification.

In the following we will explore, how to test for these hypotheses.

4.2 Confidence Intervals

A Confidence Interval is a statistical concept used to estimate a range of values within which a population parameter, such as a population mean or proportion, is likely to fall. It provides a way to express the uncertainty or variability in our sample data when making inferences about the population. In other words, it quantifies the level of confidence we have in our estimate of a population parameter.

Confidence intervals are typically expressed as a range with an associated confidence level. The confidence level, often denoted as $1 - \alpha$, represents the probability that the calculated interval contains the true population parameter. Common confidence levels include 90%, 95%, and 99%.

There are different ways of calculating CI.

1. For the population mean μ_0 when the population standard deviation σ_0^2 is known ((4.1)).

$$CI = \bar{X} \pm t \frac{\sigma_0}{\sqrt{n}} \tag{4.1}$$

- $\bar{X}$ is the sample mean.
- Z is the critical value from the standard normal distribution corresponding to the desired confidence level (e.g., 1.96 for a 95% confidence interval).
- σ_0 is the populations standard deviation
- n is the sample size

2.For the population mean μ_0 when the population standard deviation σ_0 is Unknown (t-confidence interval), see (4.2).

$$CI = \bar{X} \pm t\frac{sd}{\sqrt{n}} \tag{4.2}$$

- $\bar{X}$ is the sample mean.
- t is the critical value from the t-distribution with $n-1$ degrees of freedom corresponding to the desired confidence level
- sd is the sample standard deviation
- n is the sample size

3. For a population proportion p, see (4.3).

$$CI = \hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.3}$$

- $\hat{p}$ is the sample proportion
- Z is the critical value from the standard normal distribution corresponding to the desired confidence level
- n is the sample size

4. The method for calculating confidence intervals may vary depending on the estimated parameter. Estimating a population median or the differences between two population means, other statistical techniques may be used.

4.2.1 The drive shaft exercise - Confidence Intervals

The 95% CI for the drive shaft data is shown in Figure 4.2. For comparison the histogram with an overlayed density curve is plotted. The highlighted area shows the minimum and maximum CI, the calculated mean is shown as a dashed line.

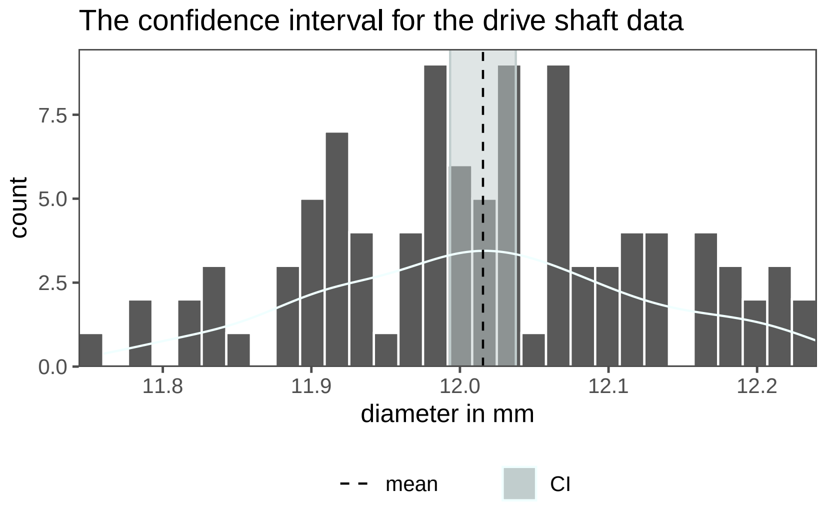


Figure 4.2: The 95% CI for the drive shaft data.

4.3 Significance Level

The significance level α is a critical component of hypothesis testing in statistics. It represents the maximum acceptable probability of making a Type I error, which is the error of rejecting a null hypothesis when it is actually true. In other words, α is the probability of concluding that there is an effect or relationship when there isn't one. Commonly used significance levels include 0.05(5%), 0.01(1%), and 0.10(10%). The choice of α depends on the context of the study and the desired balance between making correct decisions and minimizing the risk of Type I errors.

4.4 False negative - risk

The risk for a false negative outcome is called β - risk. Is is calculated using statistical power analysis. Statistical power is the probability of correctly rejecting a null hypothesis when it is false, which is essentially the complement of beta (β).

$$\beta = 1 - \text{Power} \tag{4.4}$$

4.5 Power Analysis

Statistical power is calculated using software, statistical tables, or calculators specifically designed for this purpose. Generally speaking: The greater the statistical power, the greater is the evidence to accept or reject the H_0 based on the study. Power analysis is also very useful in determining the sample size before the actualy experiments are conducted. Below is an example for a power calculation for a two-sample t-test.

$$\text{Power} = 1 - \beta = P\left(\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > Z_{\frac{\alpha}{2}} - \frac{\delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right)$$

1. Effect Size: This represents the magnitude of the effect you want to detect. Larger effects are easier to detect than smaller ones.

2. Significance Level (α): This is the predetermined level of significance that defines how confident you want to be in rejecting the null hypothesis (e.g., typically set at 0.05).

3. Sample Size (n): The number of observations or participants in your study. Increasing the sample size generally increases the power of the test.

4. Power ($1 - \beta$): This is the probability of correctly rejecting the null hypothesis when it is false. Higher power is desirable, as it minimizes the chances of a Type II error (failing to detect a true effect).

5. Type I Error (α): The probability of incorrectly rejecting the null hypothesis when it is true. This is typically set at 0.05 or 5% in most studies.

6. Type II Error (β): The probability of failing to reject the null hypothesis when it is false. Power is the complement of β ($Power = 1 - \beta$).

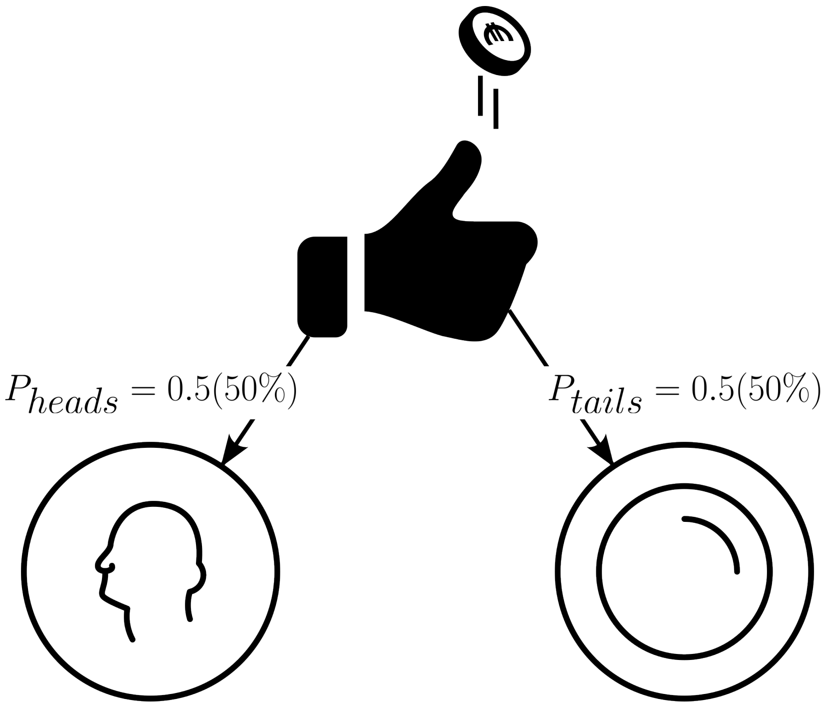


Figure 4.3: The coin toss with the respective probabilites (Champely 2020).

H0: The coin is fair and lands heads 50% of the time.
Ha: The coin is loaded and lands heads more than 50% of the time.

```
pwr.p.test(h = ES.h(p1 = 0.75, p2 = 0.50),
           sig.level = 0.05,
           power = 0.80,
           alternative = "greater")
```

```

     proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.5235988
              n = 22.55126
      sig.level = 0.05
          power = 0.8
    alternative = greater
```

The sample size $n = 23$, meaning 23 coin flips means that the statistical power is 80% at a $\alpha = 0.05$ significance level ($\beta = 1 - power = 0.2 \approx 20\%$). But what if the sample size varies? This is the subject of Figure 4.4. On the `x-axis` the power is shown (or the β-risk on the upper `x-axis`), whereas the sample size `n` is depicted on the `y-axis`. To increase the power by 10% to be 90% the sample sized must be increased by 11. A further power increase of 5% would in turn mean an increase in sample size to be $n = 40$. This highlights the non-linear nature of power calculations and why they are important for experimental planning.

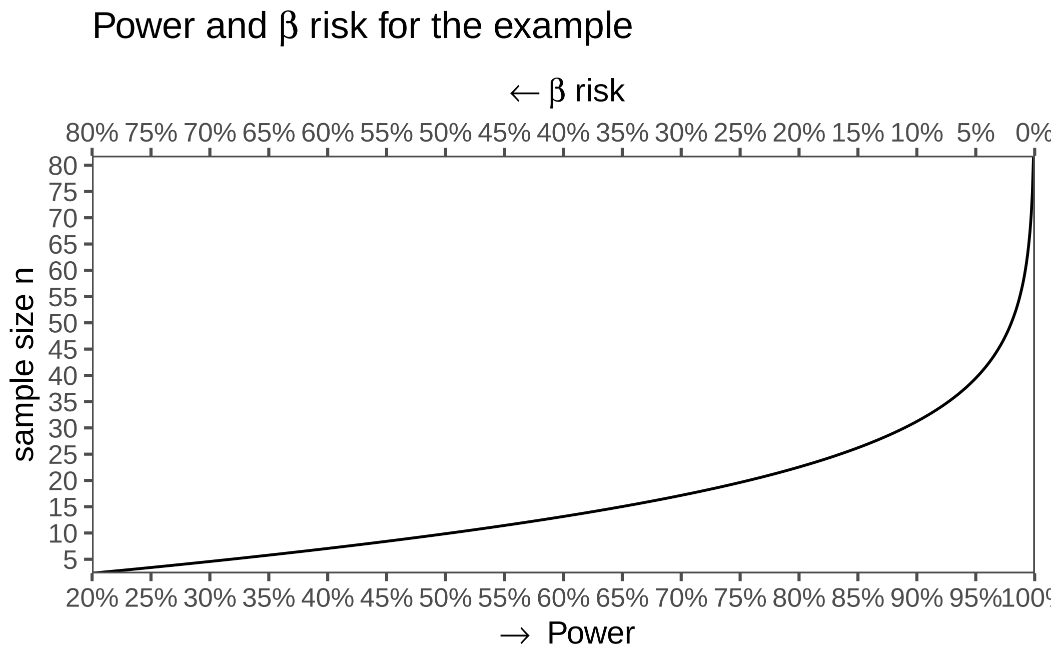


Figure 4.4: The power vs. the sample size

4.5.1 A word on Effect Size

Cohen (Cohen 2013) describes effect size as "the degree to which the null hypothesis is false." In the coin flipping example, this is the difference between 75% and 50%. We could say the effect was 25% but recall we had to transform the absolute difference in proportions to another quantity using the ES.h function. This is a crucial part of doing power analysis correctly: An effect size must be provided on the expected scale. Doing otherwise will produce wrong sample size and power calculations.

When in doubt, Conventional Effect Sizes can be used. These are pre-determined effect sizes for "small", "medium", and "large" effects, see Cohen (2013).

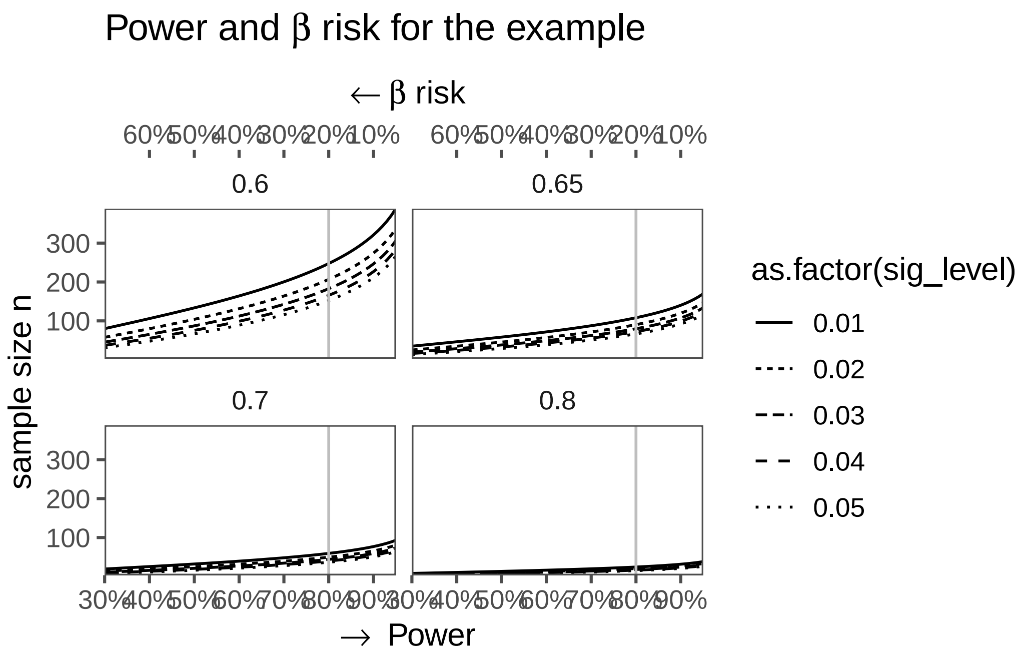


Figure 4.5: The power vs. the sample size for different effect sizes

4.6 p-value

The p-value is a statistical measure that quantifies the evidence against a null hypothesis. It represents the probability of obtaining test results as extreme or more extreme than the ones observed, assuming the null hypothesis is true. In hypothesis testing, a smaller p-value indicates stronger evidence against the null hypothesis. If the p-value is less than or equal to α ($p \leq \alpha$), you reject the null hypothesis. If the p-value is greater than α ($p > \alpha$), you fail to reject the null hypothesis. A common threshold for determining statistical significance is to reject the null hypothesis when $p \leq \alpha$.

The p-value however does not give an assumption about the effect size, which can be quite insignificant (Nuzzo 2014). While the p-value therefore is the probability of accepting H_a as true, it is not a measure of magnitude or relative importance of an effect. Therefore the CI and the effect size should always be reported with a p-value. Some Researchers even claim that most of the research today is false (Ioannidis 2005). In practice, especially in the manufacturing industry, the p-value and its use is still popular. Before implementing any measures in a series production, those questions will be asked. The confident and reliable engineer asks them beforehand and is always his own greatest critique.

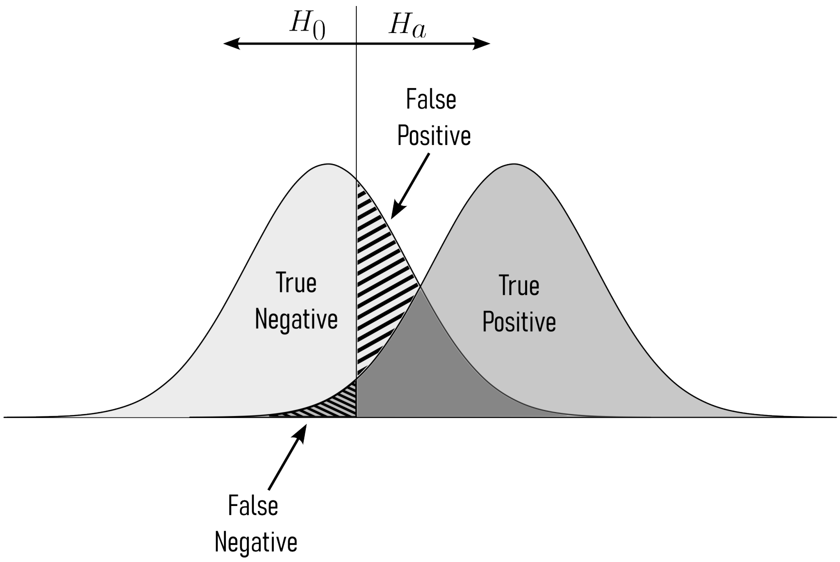


Figure 4.6: Type I and Type II error in the context of inferential statistics.

4.7 Statistical errors

- Type I Error (False Positive, see Figure 4.7):

A Type I error occurs when a null hypothesis that is actually true is rejected. In other words, it's a false alarm. It is concluded that there is a significant effect or difference when there is none. The probability of committing a Type I error is denoted by the significance level α. *Example:* Imagine a drug trial where the null hypothesis is that the drug has no effect (it's ineffective), but due to random chance, the data appears to show a significant effect, and you incorrectly conclude that the drug is effective (Type I error).

- Type II Error (False Negative, see Figure 4.7):

A Type II error occurs when a null hypothesis that is actually false is not rejected. It means failing to detect a significant effect or difference when one actually exists. The probability of committing a Type II error is denoted by the symbol β. *Example:* In a criminal trial, the null hypothesis might be that the defendant is innocent, but they are actually guilty. If the jury fails to find enough evidence to convict the guilty person, it is a Type II error.

Type I Error is falsely concluding, that there is an effect or difference when there is none (false positive). Type II Error failing to conclude that there is an effect or difference when there actually is one (false negative).

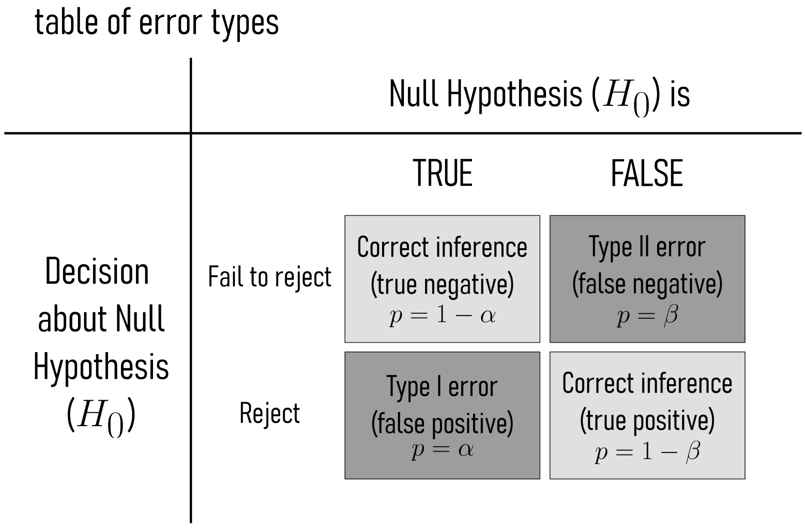


Figure 4.7: The statistical Errors (Type I and Type II).

The relationship between *Type I* and *Type II* errors is often described as a trade-off. As the risk of Type I errors is reduced by lowering the significance level (α), the risk of Type II errors (β) is typically increased (Figure 4.6). This trade-off is inherent in hypothesis testing, and the choice of significance level depends on the specific goals and context of the study. Researchers often aim to strike a balance between these two types of errors based on the consequences and costs associated with each. This balance is a critical aspect of the design and interpretation of statistical tests.

4.8 Parametric and Non-parametric Tests

Parametric and non-parametric tests in statistics are methods used for analyzing data. The primary difference between them lies in the assumptions they make about the underlying data distribution:

1. Parametric Tests:
   - These tests assume that the data follows a specific probability distribution, often the normal distribution.
   - Parametric tests make assumptions about population parameters like means and variances.
   - They are more powerful when the data truly follows the assumed distribution.
   - Examples of parametric tests include t-tests, ANOVA, regression analysis, and parametric correlation tests.
2. Non-Parametric Tests:
   - Non-parametric tests make minimal or no assumptions about the shape of the population distribution.
   - They are more robust and can be used when data deviates from a normal distribution or when dealing with ordinal or nominal data.
   - Non-parametric tests are generally less powerful compared to parametric tests but can be more reliable in certain situations.
   - Examples of non-parametric tests include the Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal-Wallis test, and Spearman's rank correlation.

The choice between parametric and non-parametric tests depends on the nature of the data and the assumptions. Parametric tests are appropriate when data follows the assumed distribution, while non-parametric tests are suitable when dealing with non-normally distributed data or ordinal data. Some examples for parametric and non-parametric tests are given in Table 4.1.

Table 4.1: Some parametric and non-parametric statistical tests.

Parametric Tests	**Non-Parametric Tests**
One-sample t-test	Wilcoxon signed rank test
Paired t-test	Mann-Whitney U test
Two-sample t-test	Kruskal Wallis test
One-Way ANOVA	Welch Test

4.9 Paired and Independent Tests

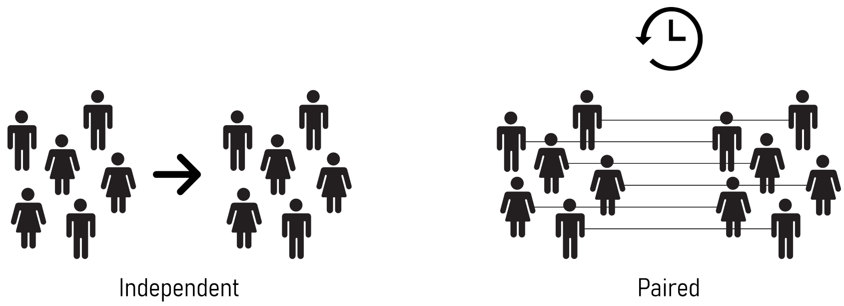


Figure 4.8: The difference between paired and independent Tests.

1. Paired Statistical Test:

- Paired tests are used when there is a natural pairing or connection between two sets of data points. This pairing is often due to repeated measurements on the same subjects or entities.
- They are designed to assess the difference between two related samples, such as before and after measurements on the same group of individuals.
- The key idea is to reduce variability by considering the differences within each pair, which can increase the test sensitivity.

2. Independent Statistical Test:

- Independent tests, also known as unpaired or two-sample tests, are used when there is no inherent pairing between the two sets of data.
- These tests are typically applied to compare two separate and unrelated groups or samples.
- They assume that the data in each group is independent of the other, meaning that the value in one group doesn't affect the value in the other group.

An example for a paired test is, if two groups of data are to be compared in two different points in time (see Figure 4.8).

4.10 Distribution Tests

The importance of testing for normality (or other distributions) lies in the fact that various statistical techniques, such as parametric tests (e.g., t-tests, ANOVA), are based on the assumption of for example normality. When data deviates significantly from a normal distribution, using these parametric methods can lead to incorrect conclusions and biased results. Therefore, it is essential to determine how a dataset is approximately distributed before applying such techniques.

Several tests for normality are available, with the most common ones being the Kolmogorov-Smirnov test, the Shapiro-Wilk test, and the Anderson-Darling test. These tests provide a quantitative measure of how well the data conforms to a normal distribution.

In practice, it is important to interpret the results of these tests cautiously. Sometimes, a minor departure from normality may not affect the validity of parametric tests, especially when the sample size is large. In such cases, using non-parametric methods may be an alternative. However, in cases where normality assumptions are crucial, transformations of the data or choosing appropriate non-parametric tests may be necessary to ensure the reliability of statistical analyses.

Tests for normality do not free you from the burden of thinking for yourself.

4.10.1 Quantile-Quantile plots

Quantile-Quantile plots are a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. They provide a visual comparison between the observed quantiles[1] of the data and the quantiles expected from the chosen theoretical distribution.

A neutral explanation of how QQ plots work:

4.10.1.1 Sample data

In Table 4.2 $n = 10$ datapoints are shown as a sample dataset.

Table 4.2: 10 randomly sampled datapoints for the creation of the QQ-plot

x	smpl_no
-0.56047565	1
-0.23017749	2
1.55870831	3
0.07050839	4

[1]A quantile is a statistical concept used to divide a dataset into equal-sized subsets or intervals.

0.12928774	5
1.71506499	6
0.46091621	7
-1.26506123	8
-0.68685285	9
-0.44566197	10

4.10.1.2 Data Sorting

To create a QQ plot, the data must be sorted in ascending order.

Table 4.3: The sorted data points.

x	smpl_no
-1.26506123	8
-0.68685285	9
-0.56047565	1
-0.44566197	10
-0.23017749	2
0.07050839	4
0.12928774	5
0.46091621	7
1.55870831	3
1.71506499	6

4.10.1.3 Theoretical Quantiles

Theoretical quantiles are calculated based on the chosen distribution (e.g., the normal distribution). These quantiles represent the expected values if the data perfectly follows that distribution.

Table 4.4: The calculated theoretical quantiles

x	smpl_no	x_norm	x_thrtcl
-1.26506123	8	-1.404601888	0.08006985
-0.68685285	9	-0.798376211	0.21232610
-0.56047565	1	-0.665875352	0.25274539
-0.44566197	10	-0.545498338	0.29270541
-0.23017749	2	-0.319572479	0.37464622
0.07050839	4	-0.004316756	0.49827787

0.12928774	5	0.057310762	0.52285118
0.46091621	7	0.405008410	0.65726434
1.55870831	3	1.555994430	0.94014529
1.71506499	6	1.719927421	0.95727718

4.10.1.4 Plotting Points

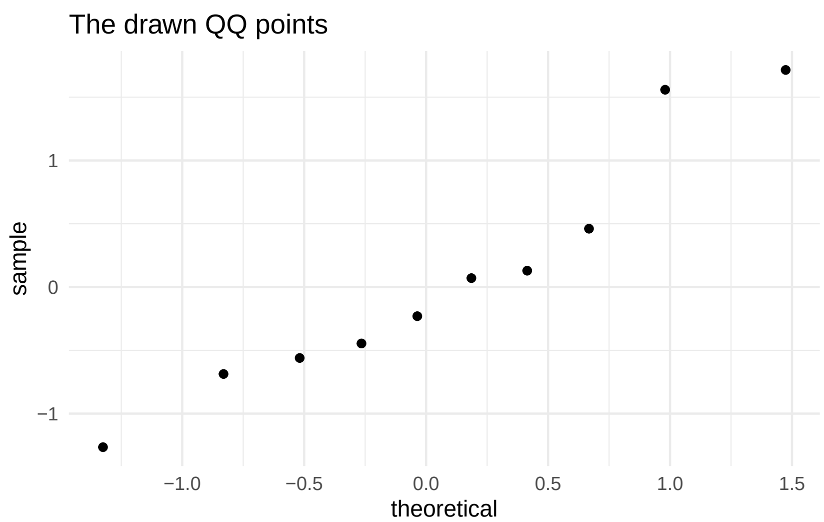


Figure 4.9: The QQ points as calculated before.

For each data point, a point is plotted in the QQ plot. The x-coordinate of the point corresponds to the theoretical quantile, and the y-coordinate corresponds to the observed quantile from the data, see Figure 4.9.

4.10.1.5 Perfect Normal Distribution

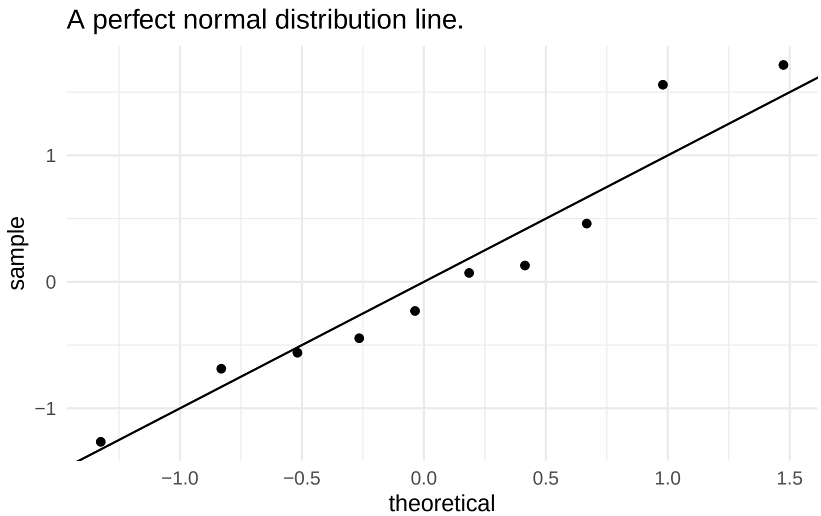


Figure 4.10: A perfect normal distribution would be indicated if all points would fall on this straight line.

In the case of a perfect normal distribution, all the points would fall along a straight line at a 45-degree angle. If the data deviates from normality, the points may deviate from this line in specific ways, see Figure 4.10.

4.10.1.6 Interpretation

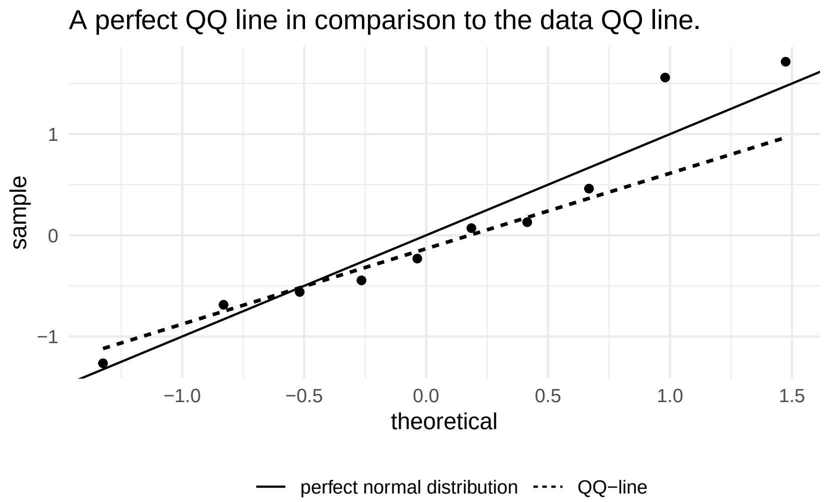


Figure 4.11: The QQ line as plotted using the theoretical and sample quantiles.

Deviations from the straight line suggest departures from the assumed distribution. For example, if points curve upward, it indicates that the data has heavier tails than a normal distribution. If points curve downward, it suggests lighter tails. S-shaped curves or other patterns can reveal additional information about the data's distribution. In Figure 4.11 the QQ-points are shown together with the respective QQ-line and a line of perfectly normal distributed points. Some deviations can be seen, but it is hard to judge, if the data is normally distributed or not.

4.10.1.7 Confidence Interval

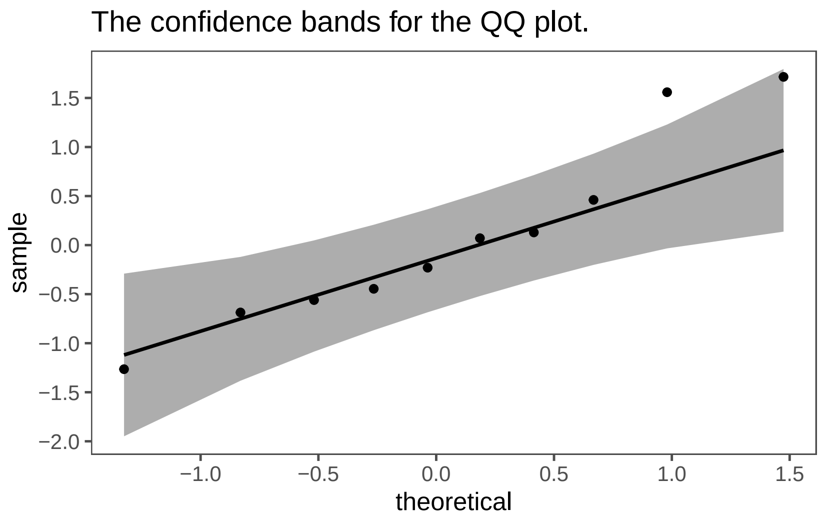


Figure 4.12: The QQ plot with confidence bands.

Because it is hard to judge from Figure 4.11 if the points are normally distributed, it makes sense to get limits for normally disitrbuted points. This is shown in Figure 4.12. The gray area depicts the (95%) confidence bands for a normal distribution. All the points fall into the area, as well as the line. This shows, that the points are likely to be normally distributed.

4.10.1.8 The drive shaft exercise

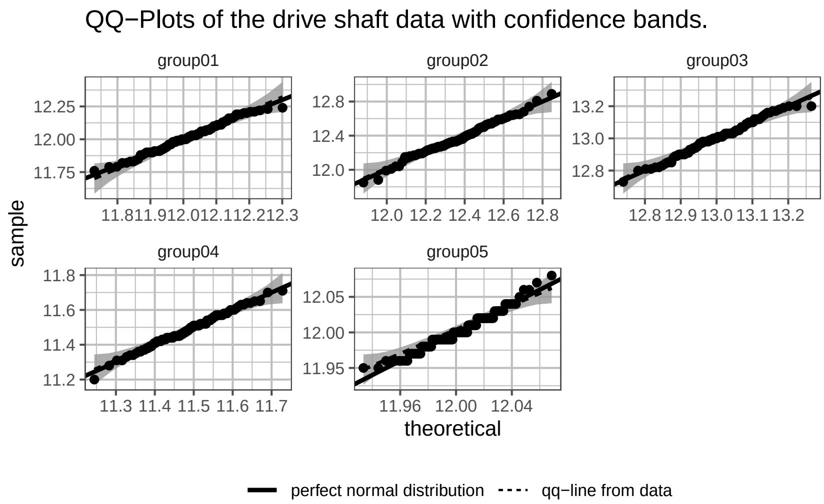


Figure 4.13: The QQ plots for each drive shaft group shown in subplots.

The QQ plot method is extended to the drive shaft exercise in Figure 4.13. In each subplot the plot for the respective group is shown together with the QQ-points, the QQ-line and the respective confidence bands. The scaling for each plot is different to enhance visibility of every subplot. A line for the perfect normal distribution is also shown in solid linestyle. From group 1 ... 4 all points fall into the QQ confidence bands. Group05 differs however. The points from visible categories, which is a strong indicator, that the measurement system may be to inaccurate.

4.10.2 Quantitative Methods

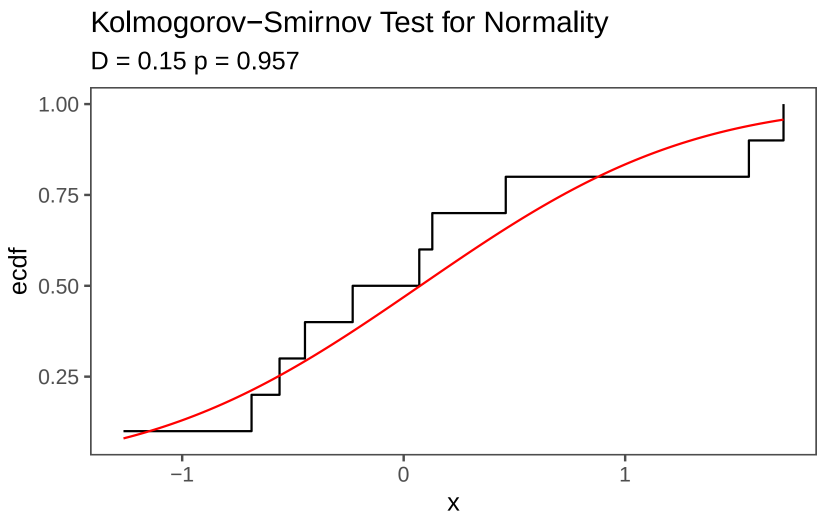


Figure 4.14: A visualisation of the KS test using the 10 datapoints from before

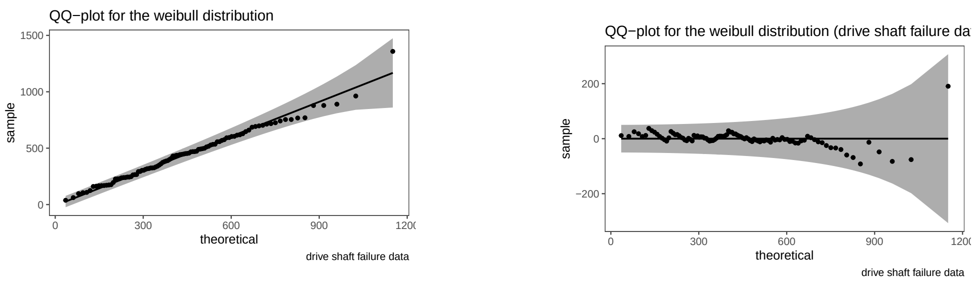
The Kolmogorov-Smirnov test for normality, often referred to as the KS test, is a statistical test used to assess whether a dataset follows a normal distribution. It evaluates how closely the cumulative distribution function of the dataset matches the expected CDF of a normal distribution.

1. **Null Hypothesis (H0):** The null hypothesis in the KS test states that the sample data follows a normal distribution.
2. **Alternative Hypothesis (Ha):** The alternative hypothesis suggests that the sample data significantly deviates from a normal distribution.
3. **Test Statistic (D):** The KS test calculates a test statistic, denoted as D which measures the maximum vertical difference between the empirical CDF of the data and the theoretical CDF of a normal distribution. It quantifies how far the observed data diverges from the expected normal distribution. A visualization of the KS-test is shown in Figure 4.14. The red line denotes a perfect normal distribution, whereas the step function shows the empirical CDF of the data itself.
4. **Critical Value:** To assess the significance of D, a critical value is determined based on the sample size and the chosen significance level (α). If D exceeds the critical value, it indicates that the dataset deviates significantly from a normal distribution.
5. **Decision:** If D is greater than the critical value, the null hypothesis is rejected, and it is concluded that the data is not normally distributed. If D is less than or equal

to the critical value, there is not enough evidence to reject the null hypothesis, suggesting that the data may follow a normal distribution.

It is important to note that the KS test is sensitive to departures from normality in both tails of the distribution. There are other normality tests, like the *Shapiro-Wilk test* and *Anderson-Darling test*, which may be more suitable in certain situations. Researchers typically choose the most appropriate test based on the characteristics of their data and the assumptions they want to test.

4.10.3 Expanding to non-normal disitributions



(a) the QQ-plot for the weibull distribution using the drive shaft failure time data

(b) a detrended QQ-plot

Figure 4.15: The QQ-plot can easily be extended to non-normal distributions.

The QQ-plot can easily be extended to non-normal disitributions as well. This is shown in Figure 4.15. In Figure 4.15a a classic QQ-plot for Figure 2.25 is shown. The same rules as before still apply, they are *only* extended to the weibull distribution. In Figure 4.15b a *detrended* QQ-plot is shown in order to account for visual bias. It is of course known, that the data follows a *weibull* disitribution with a shape parameter $\beta = 2$ and a scale parameter $\lambda = 500$, but such distributional parameters can also be estimated (Delignette-Muller and Dutang 2015).

4.11 Test 1 Variable

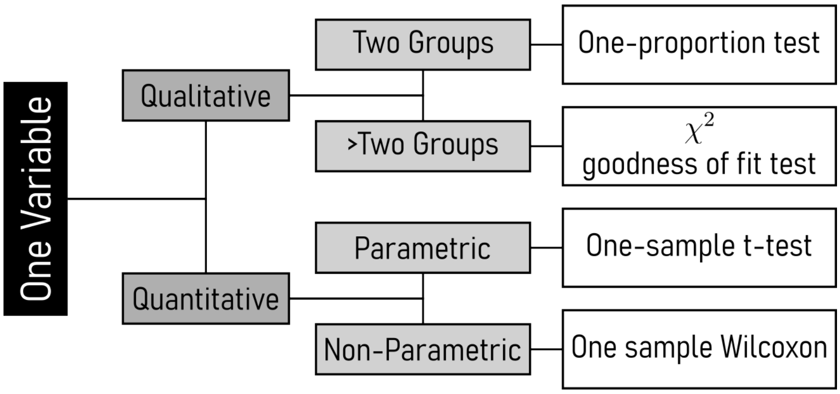


Figure 4.16: Statistical tests for one variable.

4.11.1 One Proportion Test

Table 4.5: The raw data for the proportion test.

Category	Count	Total	plt_lbl
A	35	100	35 counts 100 trials
B	20	100	20 counts 100 trials

The one proportion test is used on categorical data with a binary outcome, such as success or failure. Its prerequisite is having a known or hypothesized population proportion that the sample proportion shall be compared to. This test helps determine if the sample proportion significantly differs from the population proportion, making it valuable for studies involving proportions and percentages.

Table 4.6: The test results for the proportion test.

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	alternative
0.350	0.200	4.915	0.027	1.000	0.018	0.282	two.sided

4.11.2 Chi^2 goodness of fit test

Table 4.7: The raw data for the gof χ^2 test.

group	count_n_observed
group01	100.000
group02	100.000
group03	100.000
group04	100.000
group05	100.000

Table 4.8: The test results for the gof χ^2 test.

statistic	p.value	parameter
0.000	1.000	4.000

The χ^2 goodness of Fit Test (gof) is applied on categorical data with expected frequencies. It is suitable for analyzing nominal or ordinal data. This test assesses whether there is a significant difference between the observed and expected frequencies in your dataset, making it useful for determining if the data fits an expected distribution.

4.11.3 One-sample t-test

The one-sample t-test is designed for continuous data when you have a known or hypothesized population mean that you want to compare your sample mean to. It relies on the assumption of normal distribution, making it applicable when assessing whether a sample's mean differs significantly from a specified population mean.

The test can be applied in various settings. One is, to test if measured data comes from a population with a certain mean (for exampe a test against a specification). To show the application, the *drive shaft data* is employed. In Table 4.9 the *per group* summarised data of the dirve shaft data is shown.

Table 4.9: The raw data for the one sample t-test.

group	mean_diameter	sd_diameter
group01	12.015	0.111
group02	12.364	0.189
group03	13.002	0.102
group04	11.486	0.094
group05	12.001	0.026

One important prerequisite for the One sample t-test normally distributed data. For this, graphical and numerical methods have been introduced in previous chapters. First, a classic QQ-plot is created for every group (see Figure 4.17). From a first glance, the data appears to be normally distributed.

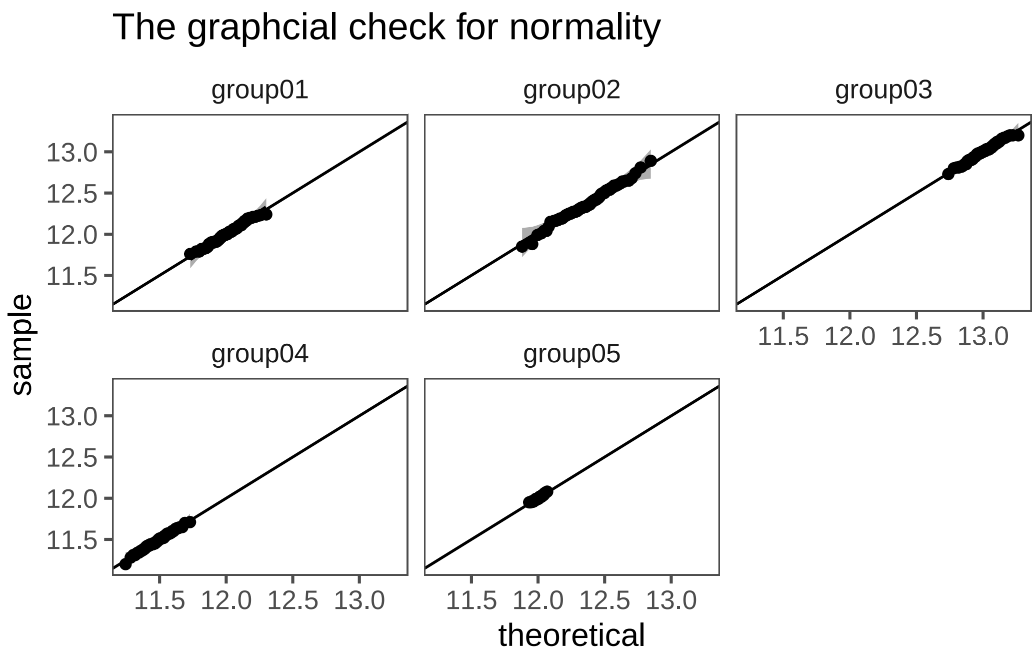


Figure 4.17: The qq-plot for the drive shaft data

A more quantitative approach to tests for normality is shown in Table 4.10. Here, each group is tested with the KS-test for normality. H0 is accepted (the data is normal distributed) because the computed p-value is larger than the significance level ($\alpha = 0.05$).

Table 4.10: The results for the one KS normality test for each group.

group	statistic	p.value	method	alternative
group01	0.048	0.975	Asymptoticone-sampleKolmogorov-Smirnovtest	two-sided
group02	0.067	0.754	Asymptoticone-sampleKolmogorov-Smirnovtest	two-sided
group03	0.075	0.633	Asymptoticone-sampleKolmogorov-Smirnovtest	two-sided

Table 4.10: The results for the one KS normality test for each group.

group	statistic	p.value	method	alternative
group04	0.060	0.862	Asymptoticone-sampleKolmogorov-Smirnovtest	two-sided
group05	0.127	0.081	Asymptoticone-sampleKolmogorov-Smirnovtest	two-sided

There is sufficient evidence to assume normal distributed data within each group. The next step is, to test if the data comes from a certain population mean (μ_0). In this case, the population mean is the specification of the drive shaft at a diameter $= 12mm$.

Table 4.11: The results for the one sample t-test (against mean = 12mm).

group	estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
group01	12.015	1.391	0.167	99.000	11.993	12.038	OneSamplet-test	two.sided
group02	12.364	19.274	0.000	99.000	12.326	12.401	OneSamplet-test	two.sided
group03	13.002	97.769	0.000	99.000	12.982	13.022	OneSamplet-test	two.sided
group04	11.486	−54.441	0.000	99.000	11.468	11.505	OneSamplet-test	two.sided
group05	12.001	0.418	0.677	99.000	11.996	12.006	OneSamplet-test	two.sided

4.11.4 One sample Wilcoxon test

For situations where your data may not follow a normal distribution or when dealing with ordinal data, the one-sample Wilcoxon test is a non-parametric alternative to the t-test. It is used to evaluate whether a sample's median significantly differs from a specified population median.

The wear and tear of drive shafts can occur due to various factors related to the vehicle's operation and maintenance. Some common causes include:

1. **Normal Usage:** Over time, the drive shaft undergoes stress and strain during regular driving. This can lead to gradual wear on components, especially if the vehicle is frequently used.
2. **Misalignment:** Improper alignment of the drive shaft can result in uneven distribution of forces, causing accelerated wear. This misalignment may stem from issues with the suspension system or other related components.

3. **Lack of Lubrication:** Inadequate lubrication of the drive shaft joints and bearings can lead to increased friction, accelerating wear. Regular maintenance, including proper lubrication, is essential to mitigate this factor.

4. **Contamination:** Exposure to dirt, debris, and water can contribute to the degradation of drive shaft components. Contaminants can infiltrate joints and bearings, causing abrasive damage over time.

5. **Vibration and Imbalance:** Excessive vibration or imbalance in the drive shaft can lead to increased stress on its components. This may result from issues with the balance of the rotating parts or damage to the shaft itself.

6. **Extreme Operating Conditions:** Harsh driving conditions, such as off-road terrain or constant heavy loads, can accelerate wear on the drive shaft. The components may be subjected to higher levels of stress than they were designed for, leading to premature wear and tear.

The wear and tear because o the reasons above can be rated on a scale with discrete values from $1 \ldots 5$ with 2 being the reference value. It is therefore interesting, if the wear and tear rating of $n = 100$ drive shafts per group differs *significantly* from the reference value 2. Because we are dealing with discrete data, the one sample t-test can not be used.

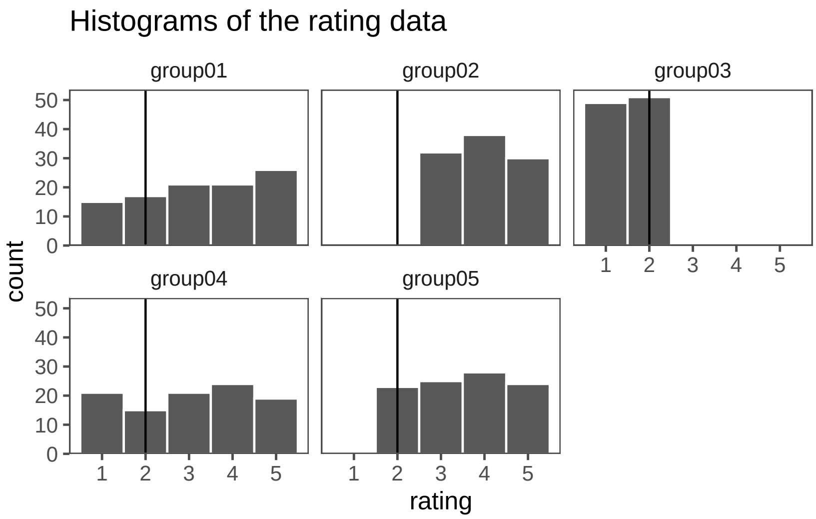


Figure 4.18: The wear and tear rating data histograms.

Table 4.12: The results for the one sample Wilcoxon test for every group against the reference value.

group	statistic	p.value	alternative
group01	3,208.500	0.000	greater
group02	5,050.000	0.000	greater
group03	0.000	1.000	greater
group04	3,203.500	0.000	greater
group05	3,003.000	0.000	greater

Table 4.13: The results for the one sample t-test compared to the results of a one sample Wilcoxon test.

group	t_tidy_p.value	wilcox_tidy_p.value
group01	0.167	0.182
group02	0.000	0.000
group03	0.000	0.000
group04	0.000	0.000
group05	0.677	0.803

4.12 Test 2 Variable (Qualitative or Quantitative)

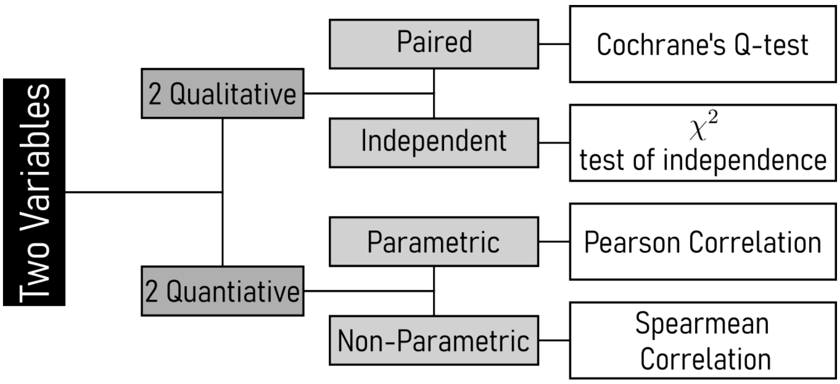


Figure 4.19: Statistical tests for two variables.

4.12.1 Cochrane's Q-test

Cochran's Q test is employed when you have categorical data with three or more related groups, often collected over time or with repeated measurements. It assesses if there is a significant difference in proportions between the related groups.

4.12.2 Chi^2 test of independence

This test is appropriate when you have two categorical variables, and you want to determine if there is an association between them. It is useful for assessing whether the two variables are dependent or independent of each other.

In the context of the drive shaft production the example assumes a dataset with categorical variables like "Defects" (Yes/No) and "Operator" (Operator A/B).

4.12.2.1 Contingency tables

A contingency table, also known as a cross-tabulation or crosstab, is a statistical table that displays the frequency distribution of variables. It organizes data into rows and columns to show the frequency or relationship between two or more categorical variables. Each cell in the table represents the count or frequency of occurrences that fall into a specific combination of categories for the variables being analyzed. It is commonly used in statistics to examine the association between categorical variables and to understand patterns within data sets.

Table 4.14: The contingency table for this example.

Defects	Operator A	Operator B
No	2	3
Yes	3	2

4.12.2.2 test results

With $p \approx 1 > 0.05$ the p-value is greater than the significance level of $\alpha = 0.05$. The H_0 is therefore proven, there is no difference between the operators. The test results are depicted below-

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 0, df = 1, p-value = 1
```

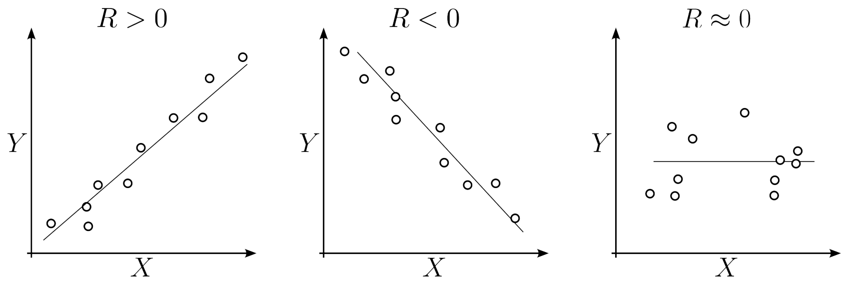


Figure 4.20: Correlation between two variables and the quantification thereof.

4.12.3 Correlation

Correlation refers to a statistical measure that describes the relationship between two variables. It indicates the extent to which changes in one variable are associated with changes in another.

Correlation is measured on a scale from -1 to 1:

- A correlation of 1 implies a perfect positive relationship, where an increase in one variable corresponds to a proportional increase in the other.
- A correlation of -1 implies a perfect negative relationship, where an increase in one variable corresponds to a proportional decrease in the other.
- A correlation close to 0 suggests a weak or no relationship between the variables.

Correlation doesn't imply causation; it only indicates that two variables change together but doesn't determine if one causes the change in the other.

4.12.3.1 Pearson Corrrelation

The pearson correlation coefficient is a normalized version of the covariance.

$$R = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \sigma_y} \tag{4.5}$$

- Covariance is sensitive to **scale** (mm vs. cm)
- Pearson correlation **removes** units, allowing for meaningful comparisons across datasets

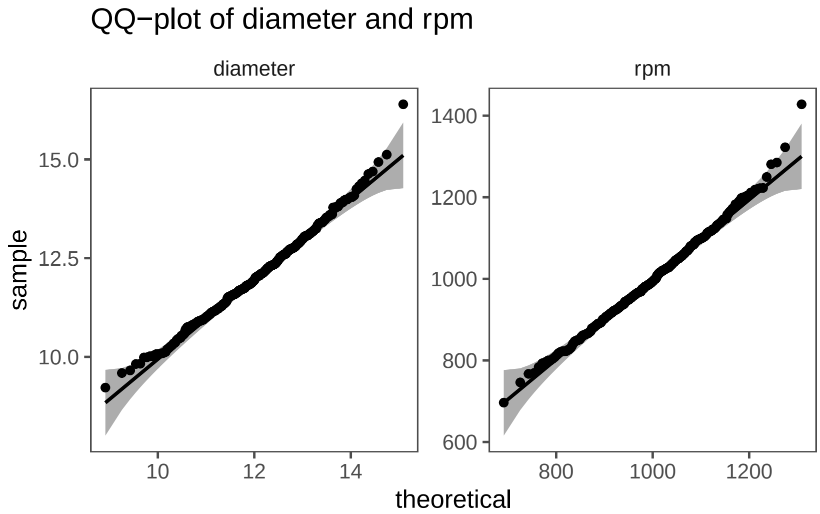


Figure 4.21: The QQ-plot of both variables. There is strong evidence that they are normally distributed.

```
    Pearson's product-moment correlation

data:  drive_shaft_rpm_dia$rpm and drive_shaft_rpm_dia$diameter
t = 67.895, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9406732 0.9578924
sample estimates:
 cor
0.95
```

When you have two continuous variables and want to measure the strength and direction of their linear relationship, Pearson correlation is the go-to choice (Pearson 1895). It assumes normally distributed data and is particularly valuable for exploring linear associations between variables and is calculated via (4.6).

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.6}$$

The Pearson Correlation Coeffcient works best with normal disitributed data. The normal distribution of the data is verified in Figure 4.21.

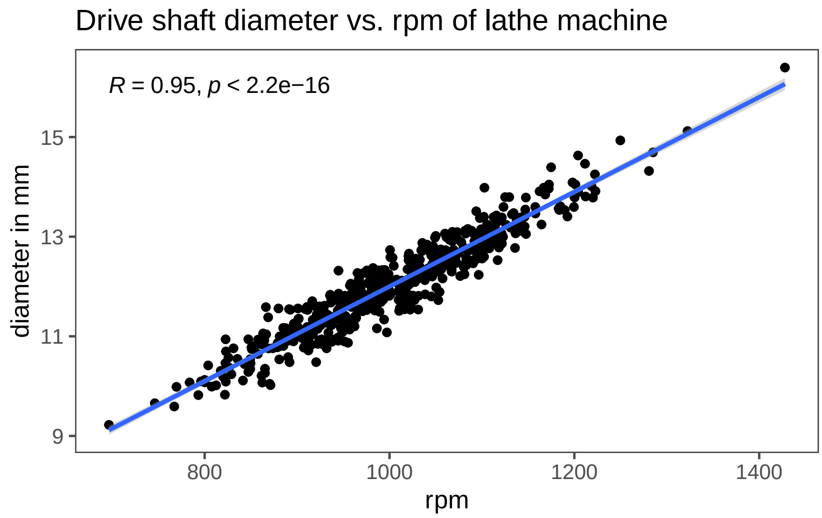


Figure 4.22: Correlation between rpm of lathe machine and the diameter of the drive shaft.

4.12.3.2 Spearman Correlation

Spearman (Spearman 1904) correlation is a non-parametric alternative to Pearson correlation. It is used when the data is not normally distributed or when the relationship between variables is monotonic but not necessarily linear.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{4.7}$$

In Figure 4.23 the example data for a drive shaft production is shown. The `Production_Time` and the `Defects` seem to form a relationship, but the data does not appear to be normally distributed. This can also be seen in the QQ-plots of both variables in Figure 4.24.

The **spearman** correlation coefficient (ρ) is based on the pearson correlation, but applied to **ranked** data

4.12.3.3 Correlation - methodogical limits

While correlation analysis and summary statistics are certainly useful, one must always consider the raw data. The data taken from Davies, Locke, and D'Agostino McGowan (2022) showcases this. The summary statistics in Table 4.15 are practically the same, one would not suspect different underlying data. When the raw data is plotted though

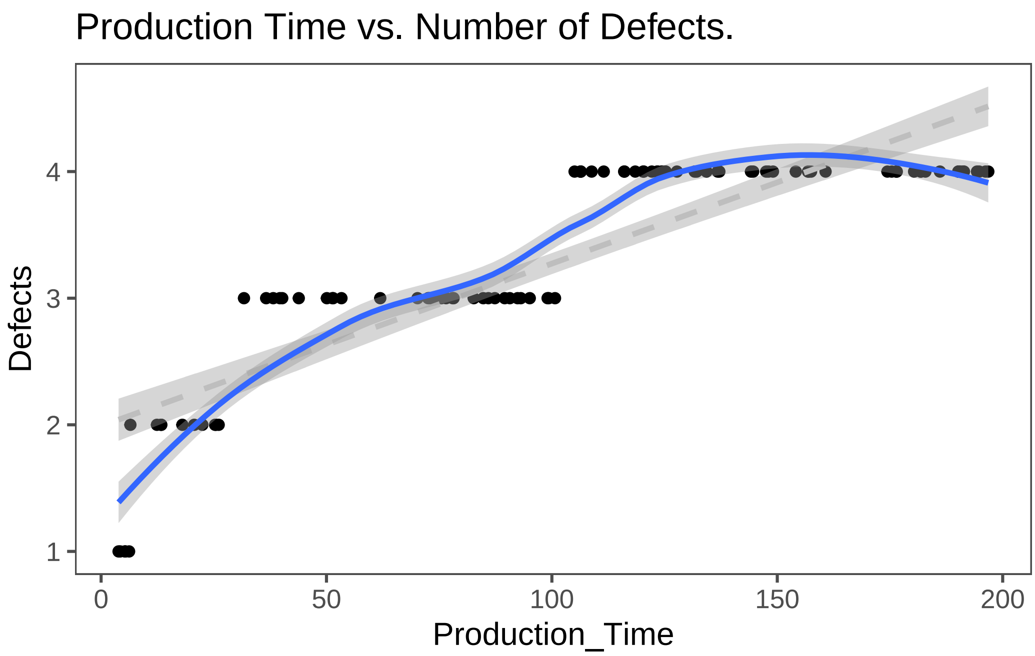


Figure 4.23: The relationship between the production time and the number of defects.

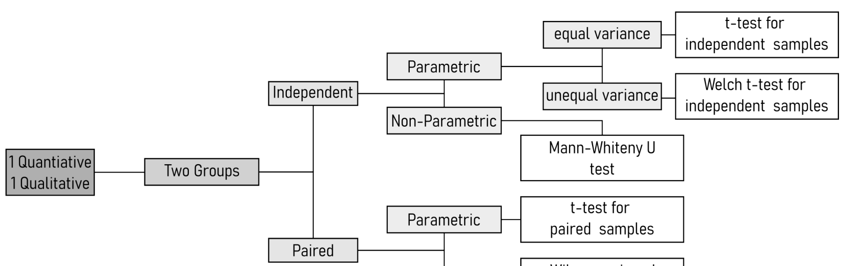
(Figure 4.25), it can be seen that the data appears to be highly non linear, forming different shapes as well as different categories etc.

Always check the raw data.

Table 4.15: The datasauRus data and the respective summary statistics.

dataset	mean_x	mean_y	std_dev_x	std_dev_y	corr_x_y
away	54.266	47.835	16.770	26.940	−0.064
bullseye	54.269	47.831	16.769	26.936	−0.069
circle	54.267	47.838	16.760	26.930	−0.068
dino	54.263	47.832	16.765	26.935	−0.064
dots	54.260	47.840	16.768	26.930	−0.060
h_lines	54.261	47.830	16.766	26.940	−0.062
high_lines	54.269	47.835	16.767	26.940	−0.069
slant_down	54.268	47.836	16.767	26.936	−0.069
slant_up	54.266	47.831	16.769	26.939	−0.069
star	54.267	47.840	16.769	26.930	−0.063
v_lines	54.270	47.837	16.770	26.938	−0.069
wide_lines	54.267	47.832	16.770	26.938	−0.067
x_shape	54.260	47.840	16.770	26.930	−0.066

4.13 Test 2 Variables (2 Groups)



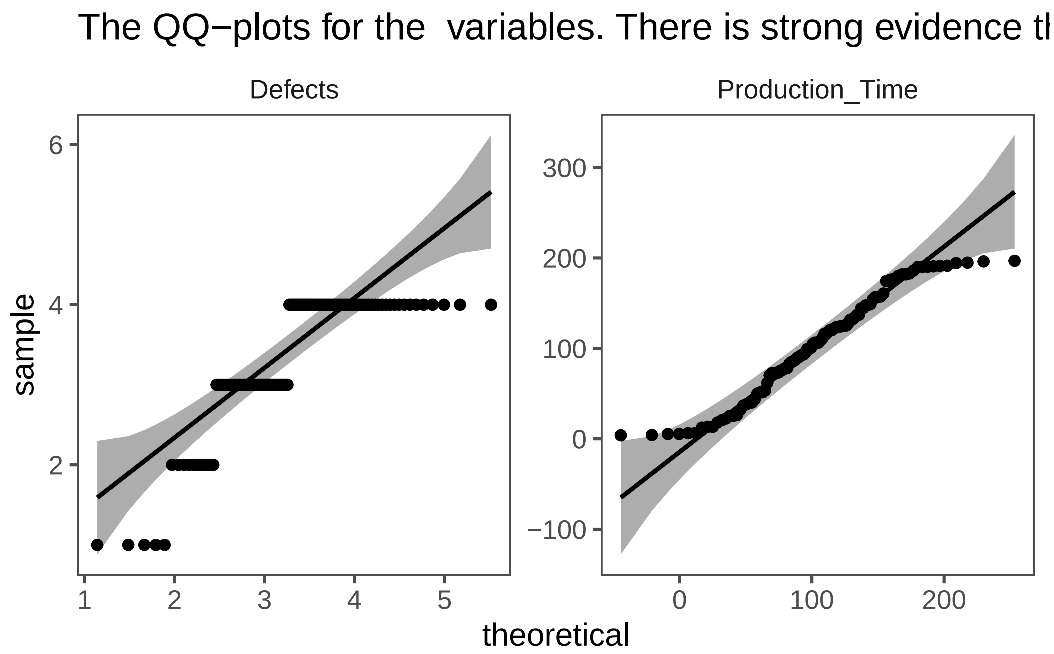


Figure 4.24: The QQ-plots of both variables.

(ANOVA). When the variances are not equal, it can affect the validity of these tests. Two common tests for equal variances are:

Certainly, here are bullet points outlining the null hypothesis, prerequisites, and decisions for each of the three tests:

4.13.1.1 F-Test (Hahs-Vaughn and Lomax 2013)

- **Null Hypothesis:** The variances of the different groups or samples are equal.
- **Prerequisites:**
  - Independence
  - Normality
  - Number of groups = 2
- **Decisions:**
  - $p > \alpha \rightarrow$ fail to reject H0
  - $p < \alpha \rightarrow$ reject H0

```
F test to compare two variances
```

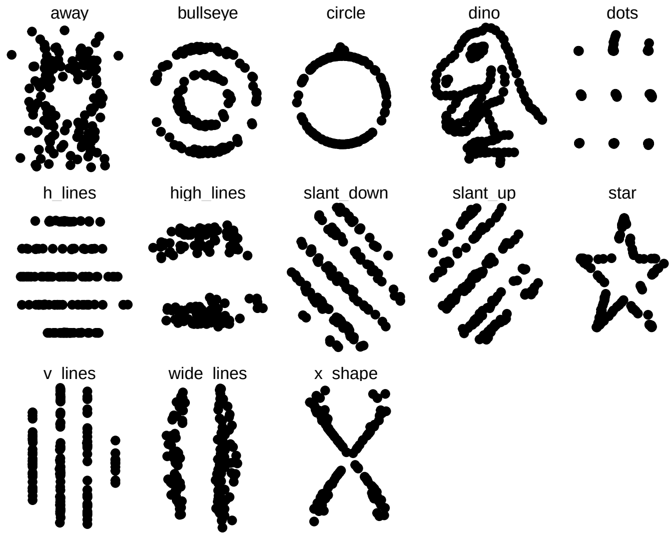


Figure 4.25: The raw data from the datasauRus packages shows, that summary statistics may be misleading.

```
data:  ds_wide$group01 and ds_wide$group03
F = 1.1817, num df = 99, denom df = 99, p-value = 0.4076
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7951211 1.7563357
sample estimates:
ratio of variances
          1.181736
```

4.13.1.2 Bartlett Test (Bartlett 1937)

- **Null Hypothesis:** The variances of the different groups or samples are equal.
- **Prerequisites:**
  - Independence
  - Normality
  - Number of groups > 2
- **Decisions:**
  - $p > \alpha \rightarrow$ fail to reject H0
  - $p < \alpha \rightarrow$ reject H0

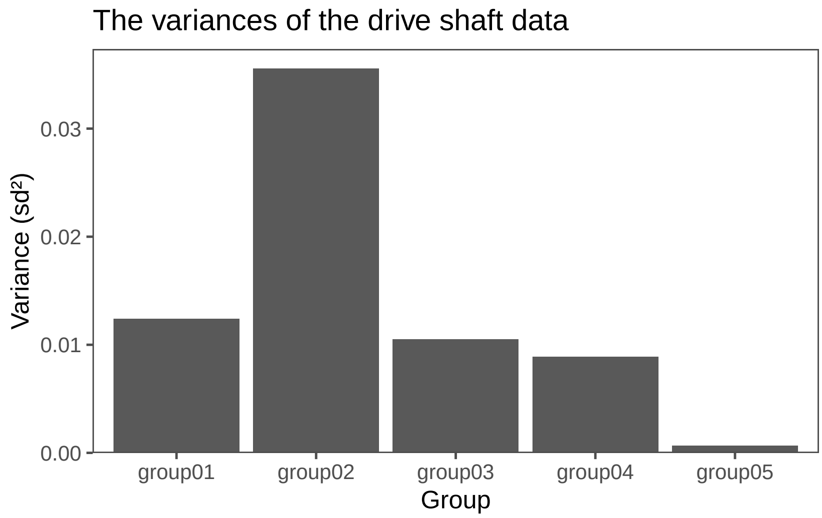


Figure 4.27: The variances (sd^2) for the drive shaft data.

```
    Bartlett test of homogeneity of variances

data:  diameter by group
Bartlett's K-squared = 275.61, df = 4, p-value < 2.2e-16
```

4.13.1.3 Levene Test (Olkin June)

- **Null Hypothesis:** The variances of the different groups or samples are equal.
- **Prerequisites:**
  - Independence
  - Number of groups > 2
- **Decisions:**
  - $p > \alpha \to$ fail to reject H0
  - $p < \alpha \to$ reject H0

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   4  38.893 < 2.2e-16 ***
      495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.13.2 t-test for independent samples

The independent samples t-test is applied when you have continuous data from two independent groups. It evaluates whether there is a significant difference in means between these groups, assuming a normal distribution of the data.

- **Null Hypothesis:** The means of the two samples are equal.
- **Prerequisites:**
  - Independence
  - Normal Distribution
  - Number of groups = 2
  - equal Variances of the groups

First, the variances are compared in order to check if they are equal using the F-Test (as described in Section 4.13.1.1).

```
    F test to compare two variances

data:  group01 %>% pull("diameter") and group03 %>% pull("diameter")
F = 1.1817, num df = 99, denom df = 99, p-value = 0.4076
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7951211 1.7563357
sample estimates:
ratio of variances
          1.181736
```

With $p > \alpha = 0.05$ the H_0 is accepted, the variances are equal.

The next step is to check the data for normality using the KS-test (as described in Section 4.10.2).

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  group01 %>% pull("diameter")
D = 0.048142, p-value = 0.9746
alternative hypothesis: two-sided


    Asymptotic one-sample Kolmogorov-Smirnov test
```

```
data:  group03 %>% pull("diameter")
D = 0.074644, p-value = 0.6332
alternative hypothesis: two-sided
```

With $p > \alpha = 0.05$ the H_0 is accepted, the data seems to be normally distributed.

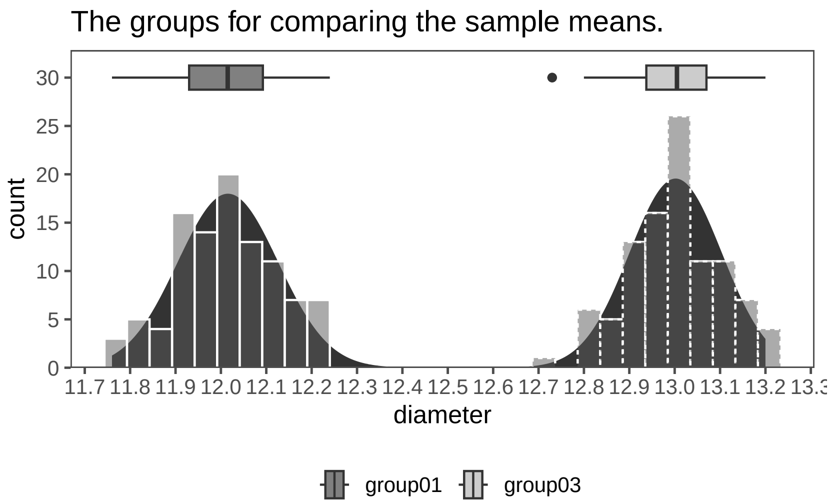


Figure 4.28: The data within the two groups for comparing the sample means using the t-test for independent samples.

The formal test is then carried out. With $p < \alpha = 0.05$ H_0 is rejected, the data comes from populations with different means.

```
    Two Sample t-test

data:  group01 %>% pull(diameter) and group03 %>% pull(diameter)
t = -65.167, df = 198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.0164554 -0.9567446
sample estimates:
mean of x mean of y
  12.0155   13.0021
```

4.13.3 Welch t-test for independent samples

Similar to the independent samples t-test, the Welch t-test is used for continuous data with two independent groups (WELCH 1947). However, it is employed when there are unequal variances between the groups, relaxing the assumption of equal variances in the standard t-test.

- **Null Hypothesis:** The means of the two samples are equal.
- **Prerequisites:**
  - Independence
  - Normal Distribution
  - Number of groups = 2

First, the variances are compared in order to check if they are equal using the F-Test (as described in Section 4.13.1.1).

```
    F test to compare two variances

data:  group01 %>% pull("diameter") and group02 %>% pull("diameter")
F = 0.34904, num df = 99, denom df = 99, p-value = 3.223e-07
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2348504 0.5187589
sample estimates:
ratio of variances
         0.3490426
```

With $p < \alpha = 0.05$ H_0 is rejected and H_a is accepted. The variances are different.

Using the KS-test (see Section 4.10.2) the data is checked for normality.

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  group01 %>% pull("diameter")
D = 0.048142, p-value = 0.9746
alternative hypothesis: two-sided


    Asymptotic one-sample Kolmogorov-Smirnov test
```

```
data:  group02 %>% pull("diameter")
D = 0.067403, p-value = 0.7539
alternative hypothesis: two-sided
```

With $p > \alpha = 0.05$ H_0 is accepted, the data seems to be normally distributed.

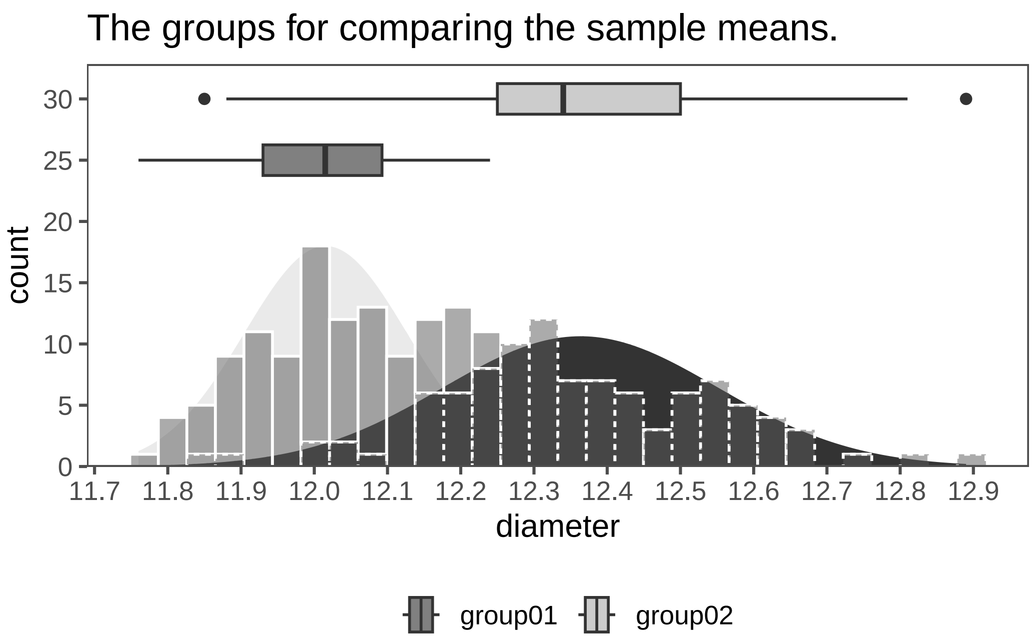


Figure 4.29: The data within the two groups for comparing the sample means using the Welch-test for independent samples.

Then, the formal test is carried out.

```
    Welch Two Sample t-test

data:  group01 %>% pull(diameter) and group02 %>% pull(diameter)
t = -15.887, df = 160.61, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3912592 -0.3047408
sample estimates:
mean of x mean of y
  12.0155   12.3635
```

With $p < \alpha = 0.05$ we reject H_0, the data seems to be coming from different population means, even though the variances are overlapping (and different).

4.13.4 Mann-Whitney U test

For non-normally distributed data or small sample sizes, the Mann-Whitney U test serves as a non-parametric alternative to the independent samples t-test (Mann and Whitney 1947). It assesses whether there is a significant difference in medians between two independent groups.

- **Null Hypothesis:** The medians of the two samples are equal.
- **Prerequisites:**
    - Independence
    - no specific distribution (non-parametric)
    - Number of groups = 2

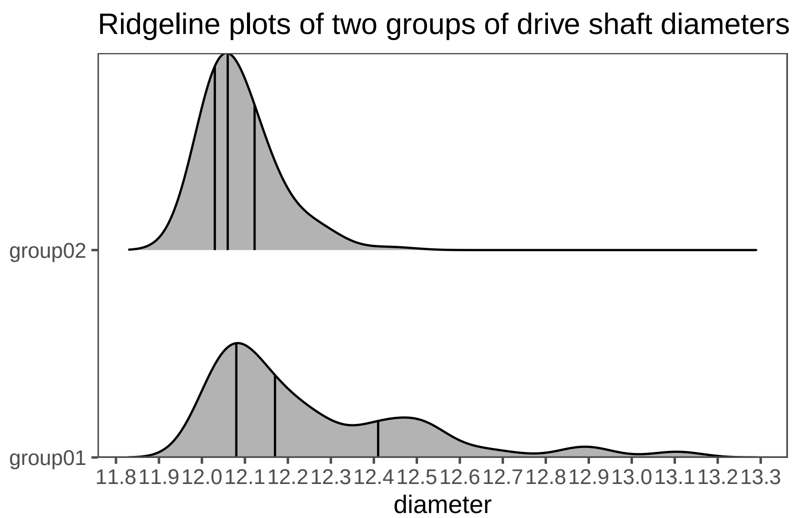


Figure 4.30: The data within the two groups for comparing the sample medians using the Mann-Whitney-U Test.

This time a graphical method to check for normality is employed (QQ-plot, see Section 4.10.1). From the Figure 4.31 it is pretty clear, that the data is not normally distributed. Furthermore, the variances seem to be unequal as well.

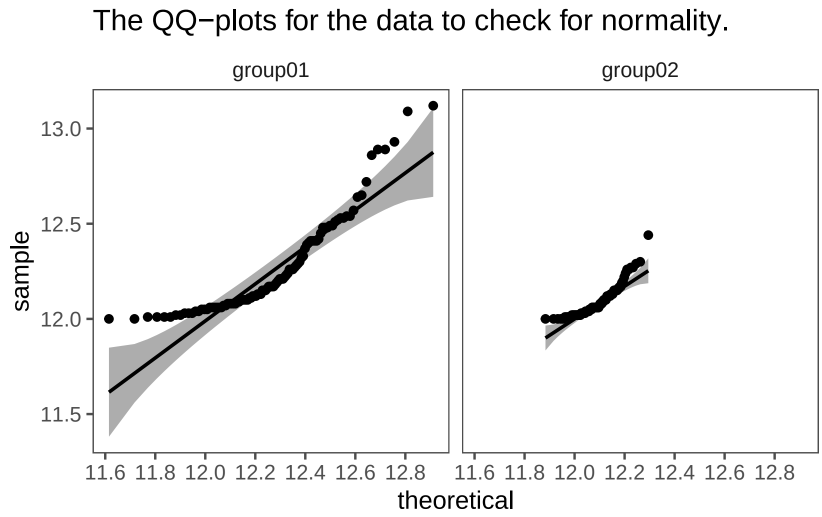


Figure 4.31: The data within the two groups for comparing the sample medians using the Mann-Whitney-U Test.

Then, the formal test is carried out. With $p < \alpha = 0.05$ H_0 is rejected, the true location shift is not equal to 0.

```
    Wilcoxon rank sum test with continuity correction

data:  diameter by group
W = 7396, p-value = 4.642e-09
alternative hypothesis: true location shift is not equal to 0
```

4.13.5 t-test for paired samples

The paired samples t-test is suitable when you have continuous data from two related groups or repeated measures. It helps determine if there is a significant difference in means between the related groups, assuming normally distributed data.

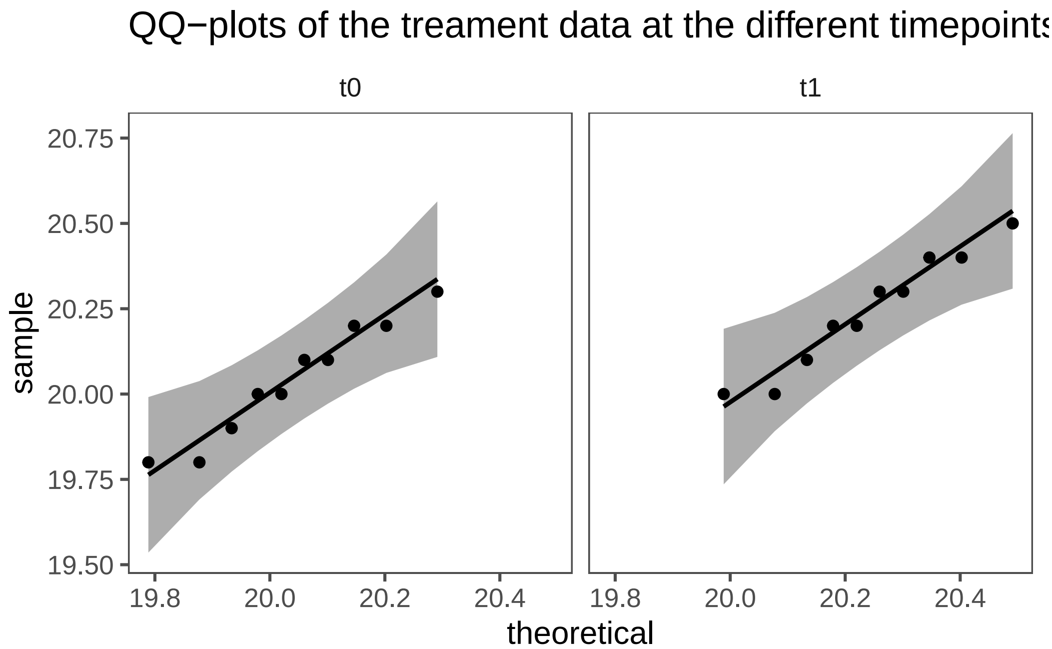
- **Null Hypothesis:** True mean difference is not equal to 0.
- **Prerequisites:**
  - Paired Data
  - Normal Distribution
  - equal variances
  - Number of groups = 2

Using the F-Test, the variances are compared.

```
    F test to compare two variances

data:  diameter by timepoint
F = 1, num df = 9, denom df = 9, p-value = 1
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2483859 4.0259942
sample estimates:
ratio of variances
                 1
```

With $p > \alpha = 0.05$ H_0 is accepted, the variances are equal.

Using a QQ-plot the data is checked for normality.



Without a formal test, the data is assumed to be normally distributed.

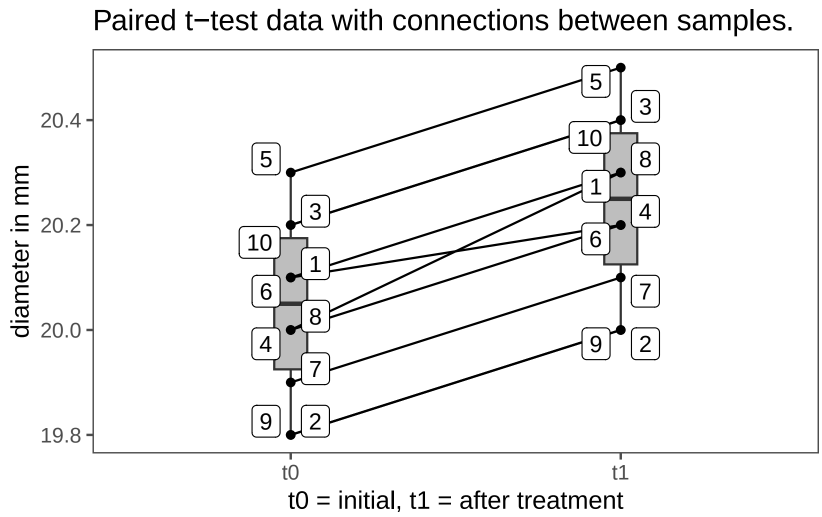


Figure 4.32: A boxplot of the data, showing the connections between the datapoints.

The formal test is then carried out.

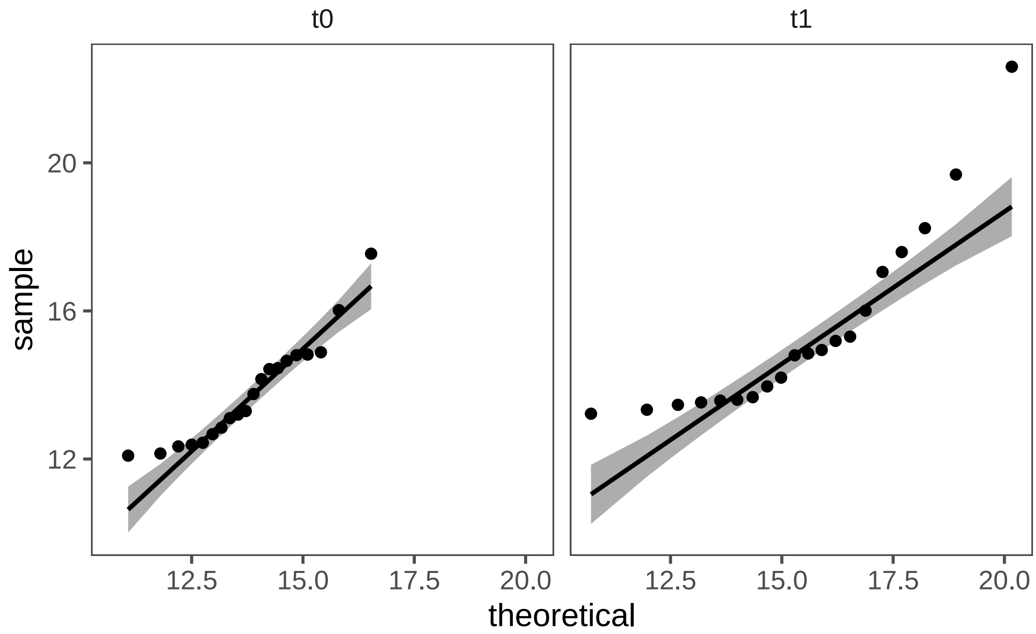
```
# A tibble: 1 x 8
  .y.      group1 group2    n1    n2 statistic    df           p
* <chr>    <chr>  <chr>  <int> <int>     <dbl> <dbl>       <dbl>
1 diameter t0     t1        10    10     -13.4     9 0.000000296
```

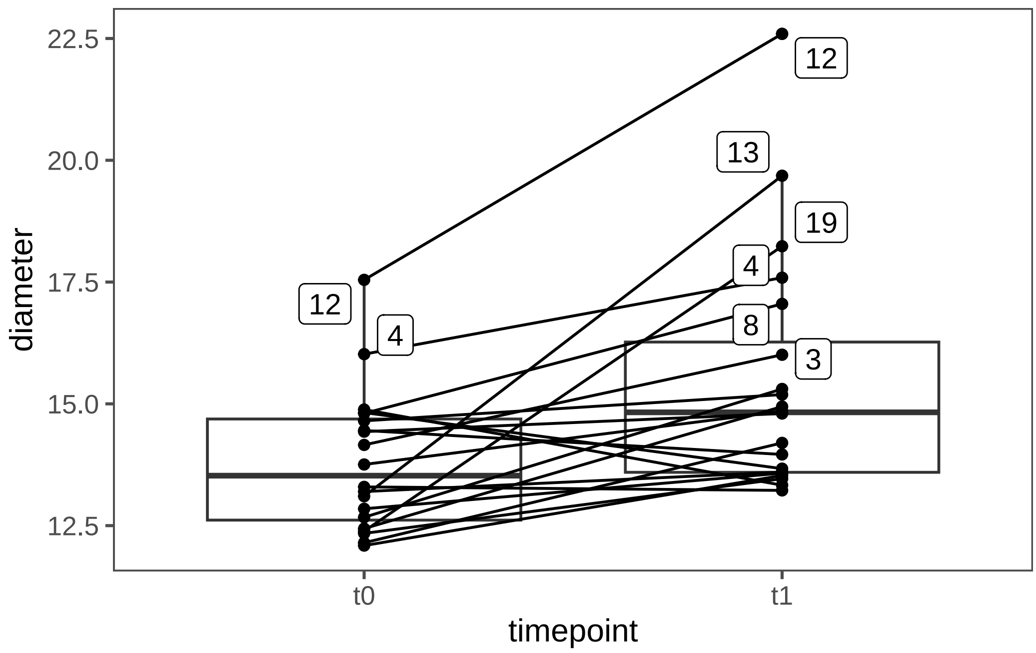
With $p < \alpha = 0.05$ H_0 is rejected, the treatment changed the properties of the product.

4.13.6 Wilcoxon signed rank test

For non-normally distributed data or situations involving paired samples, the Wilcoxon signed rank test is a non-parametric alternative to the paired samples t-test. It evaluates whether there is a significant difference in medians between the related groups.

- **Null Hypothesis:** True mean difference is not equal to 0.
- **Prerequisites:**
    - Paired Data
    - Number of groups = 2






```
# A tibble: 1 x 7
  .y.      group1 group2    n1    n2 statistic       p
* <chr>    <chr>  <chr>  <int> <int>     <dbl>   <dbl>
1 diameter t0     t1        20    20        25 0.00169
```

4.14 Test 2 Variables ($>$ 2 Groups)

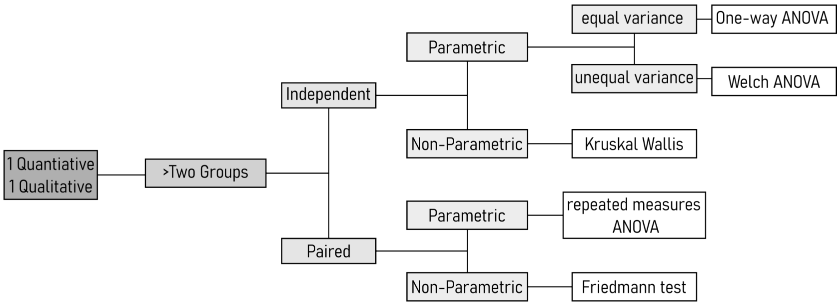


Figure 4.33: Statistical tests for one variable.

4.14.1 Analysis of Variance (ANOVA) - Basic Idea

ANOVA's ability to compare multiple groups or factors makes it widely applicable across diverse fields for analyzing variance and understanding relationships within data. In the context of engineering sciences the application of ANOVA include:

1. **Experimental Design and Analysis:** Engineers often conduct experiments to optimize processes, test materials, or evaluate designs. ANOVA aids in analyzing these experiments by assessing the effects of various factors (like temperature, pressure, or material composition) on the performance of systems or products. It helps identify significant factors and their interactions to improve engineering processes.

2. **Product Testing and Reliability:** Engineers use ANOVA to compare the performance of products manufactured under different conditions or using different materials. This analysis helps ensure product reliability by identifying which factors significantly impact product quality, durability, or functionality.

3. **Process Control and Improvement:** ANOVA plays a crucial role in quality control and process improvement within engineering. It helps identify variations in manufacturing processes, such as assessing the impact of machine settings or production methods on product quality. By understanding these variations, engineers can make informed decisions to optimize processes and minimize defects.

4. **Supply Chain and Logistics:** In engineering logistics and supply chain management, ANOVA aids in analyzing the performance of different suppliers or transportation methods. It helps assess variations in delivery times, costs, or product quality across various suppliers or logistical approaches.

5. **Simulation and Modeling:** In computational engineering, ANOVA is used to analyze the outputs of simulations or models. It helps understand the significance

of different input variables on the output, enabling engineers to refine models and simulations for more accurate predictions.

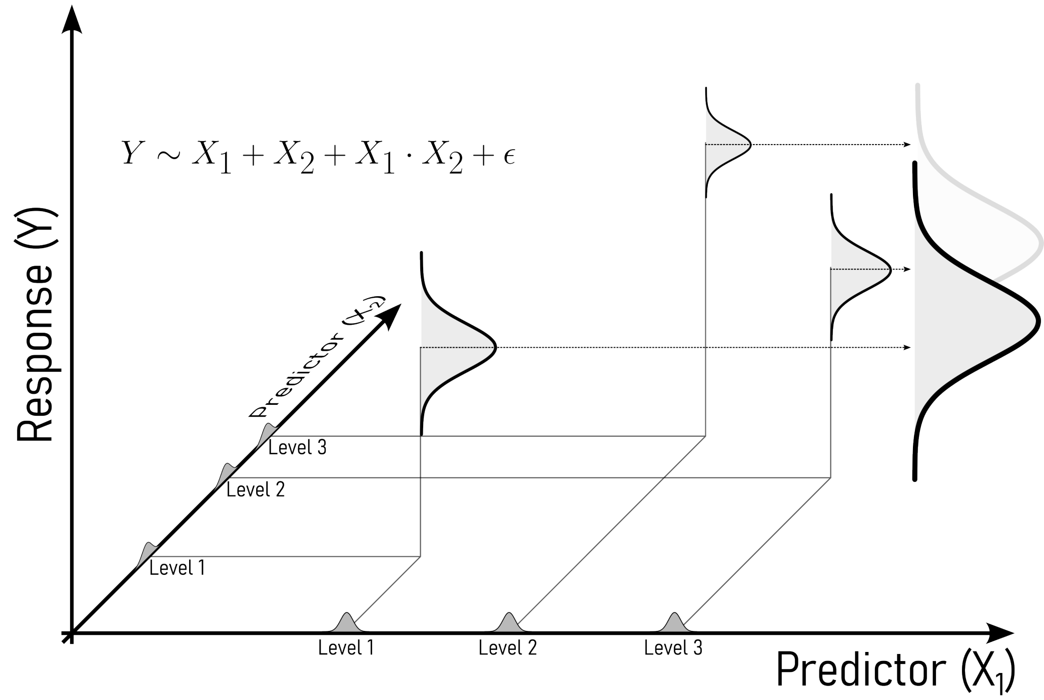


Figure 4.34: The basic idea of an ANOVA.

Across such fields ANOVA is often used to:

Comparing Means: ANOVA is employed when comparing means between three or more groups. It assesses whether there are statistically significant differences among the means of these groups. For instance, in an experiment testing the effect of different fertilizers on plant growth, ANOVA can determine if there's a significant difference in growth rates among the groups treated with various fertilizers.

Modeling Dependencies: ANOVA can be extended to model dependencies among variables in more complex designs. For instance, in factorial ANOVA, it's used to study the interaction effects among multiple independent variables on a dependent variable. This allows researchers to understand how different factors might interact to influence an outcome.

Measurement System Analysis (MSA): ANOVA is integral in MSA to evaluate the variation contributed by different components of a measurement system. In assessing the reliability and consistency of measurement instruments or processes, ANOVA helps in dissecting the total variance into components attributed to equipment variation, operator variability, and measurement error.

As with statistical tests before, the applicability of the ANOVA depends on various factors.

4.14.1.1 Sum of squared error (SSE)

The sum of squared errors is a statistical measure used to assess the goodness of fit of a model to its data. It is calculated by squaring the differences between the observed values and the values predicted by the model for each data point, then summing up these squared differences. The SSE indicates the total variability or dispersion of the observed data points around the fitted regression line or model. Lower SSE values generally indicate a better fit of the model to the data.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.8}$$

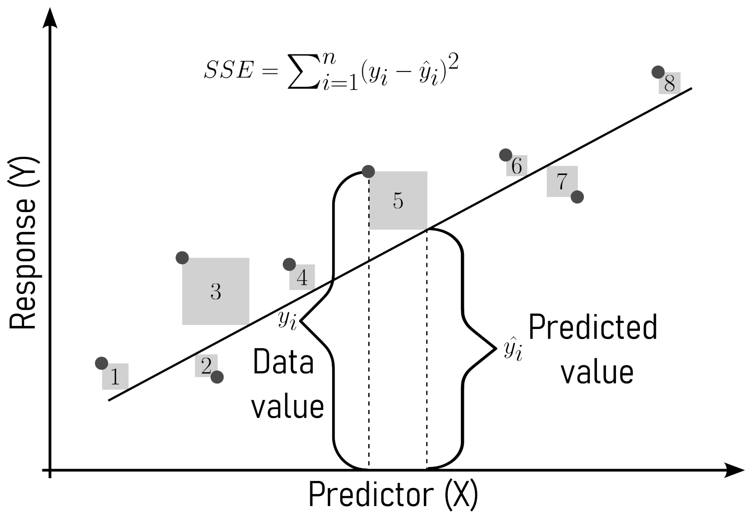


Figure 4.35: A graphical depiction of the SSE.

4.14.1.2 Mean squared error (MSE)

The mean squared error is a measure used to assess the average squared difference between the predicted and actual values in a dataset. It is frequently employed in regression analysis to evaluate the accuracy of a predictive model. The MSE is calculated by taking the average of the squared differences between predicted values and observed values. A lower MSE indicates that the model's predictions are closer to the actual values, reflecting better accuracy.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.9}$$

4.14.2 One-way ANOVA

The one-way analysis of variance (ANOVA) is used for continuous data with three or more independent groups. It assesses whether there are significant differences in means among these groups, assuming a normal distribution.

- **Null Hypothesis:** True mean difference is equal to 0.
- **Prerequisites:**
  - equal variances
  - Number of groups > 2
  - One response, one predictor variable

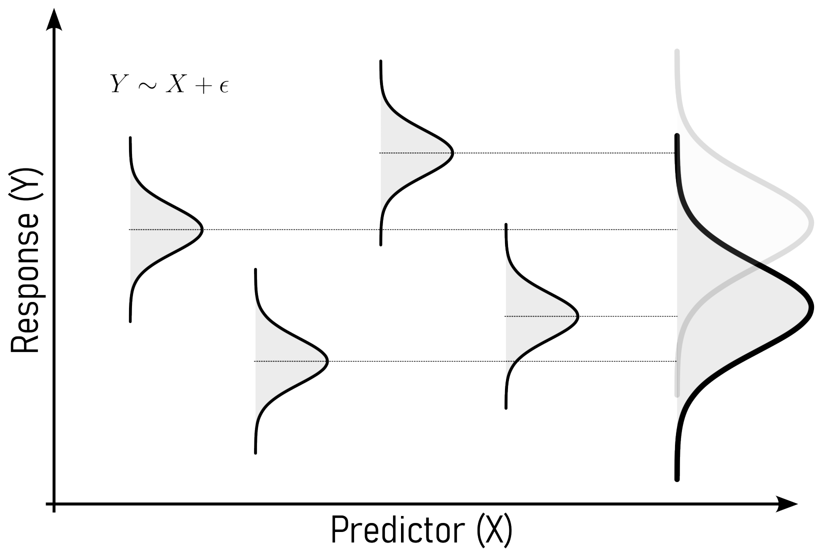


Figure 4.36: The basic idea of a One-way ANOVA.

The most important prerequisite for a One-way ANOVA are equal variances. Because there are more than two groups, the Bartlett test (as introduced in Section 4.13.1.2) is chosen (data is normally distributed).

```
Bartlett test of homogeneity of variances
```

```
data:  diameter by group
Bartlett's K-squared = 275.61, df = 4, p-value < 2.2e-16
```

Because $p < \alpha = 0.05$ the variances are different.

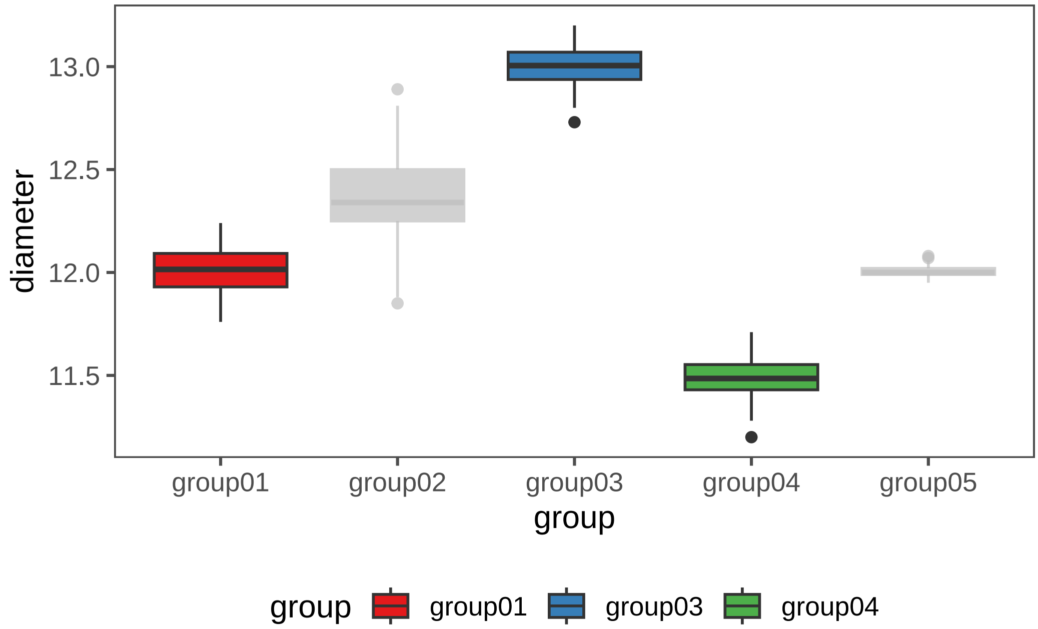


Figure 4.37: The groups with equal variance are highlighted.

```
    Bartlett test of homogeneity of variances

data:  diameter by group
Bartlett's K-squared = 2.7239, df = 2, p-value = 0.2562
```

With $p > \alpha = 0.05$ H_0 is accepted, the variances of `group01`, `group02` and `group03` are equal.

Of course, many software package provide an automated way of performing a One-way ANOVA, but the first will be explained in detail. The general model for a One-way ANOVA is shown in (4.10).

$$Y \sim X + \epsilon \tag{4.10}$$

- H_0: All population means are equal.

- H_a: Not all population means are equal.

For a One-way ANOVA the predictor variable X is the mean ($\bar{x}$) of all datapoints x_i.

First the SSE and the MSE is calculated for the complete model (H_a is true), see Table 4.16. The complete model means, that every mean, for every group is calculated and the SSE according to (4.8) is calculated.

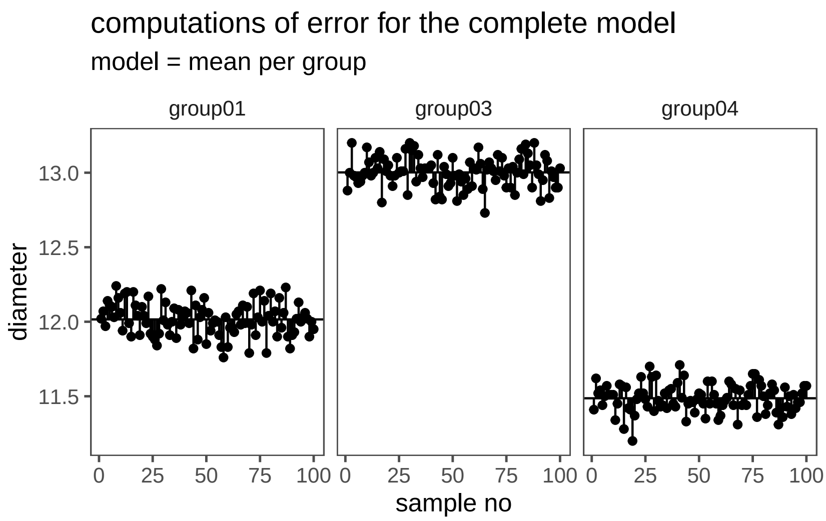


Figure 4.38: Computation of error for the complete model (mean per group as model)

Table 4.16: The SSE and MSE for the complete model.

sse	df	n	p	mse
3.150	297.000	300.000	3.000	0.011

Then, the SSE and the MSE is calculated for the reduced model (H_0 is true). In the reduced model, the mean is not calculated per group, the overall mean is calculated (results in Table 4.17).

Table 4.17: The SSE and MSE from the reduced model.

sse	df	n	p	mse
121.506	299.000	300.000	1.000	0.406

The SSE, df and MSE explained by the complete model are calculated:

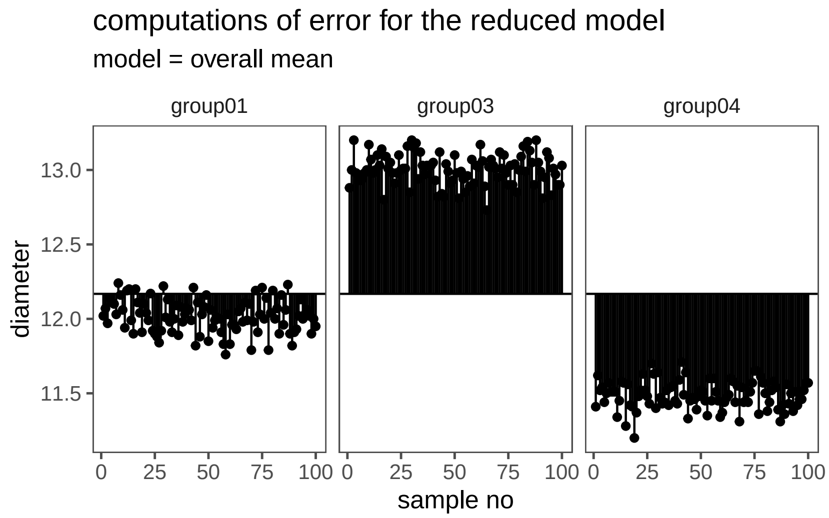


Figure 4.39: Computation of error for the reduced model (overall mean as model)

$$SSE_{explained} = SSE_{reduced} - SSE_{complete} = 118.36 \tag{4.11}$$
$$df_{explained} = df_{reduced} - df_{complete} = 2 \tag{4.12}$$
$$MSE_{explained} = \frac{SSE_{explained}}{df_{explained}} = 59.18 \tag{4.13}$$

The ratio of the variance (MSE) as explained by the complete model to the reduced model is then calculated. The probability of this statistic is afterwards calculated (if H_0 is true).

```
[1] 2.762026e-236
```

The probability of a F-statistic with $pf = 5579.207$ is 0.

A crosscheck with a automated solution (`aov`-function) yields the results shown in Table 4.18.

Table 4.18: The ANOVA results from the aov function.

term	df	sumsq	meansq	statistic	p.value
group	2.000	118.356	59.178	5,579.207	0.000
Residuals	297.000	3.150	0.011	NA	NA

Some sanity checks are of course required to ensure the validity of the results. First, the variance of the residuals must be equal along the groups (see Figure 4.40).

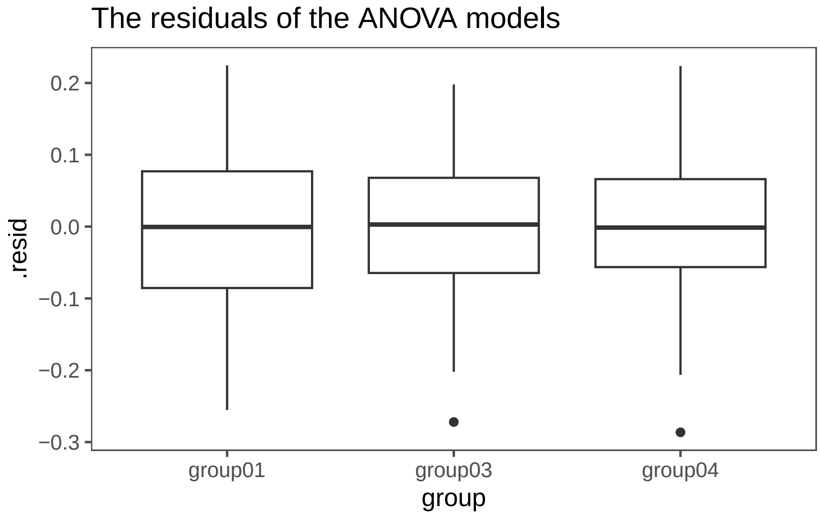


Figure 4.40: The variances of the residuals.

Also, the residuals from the model must be normally distributed (see Figure 4.41).

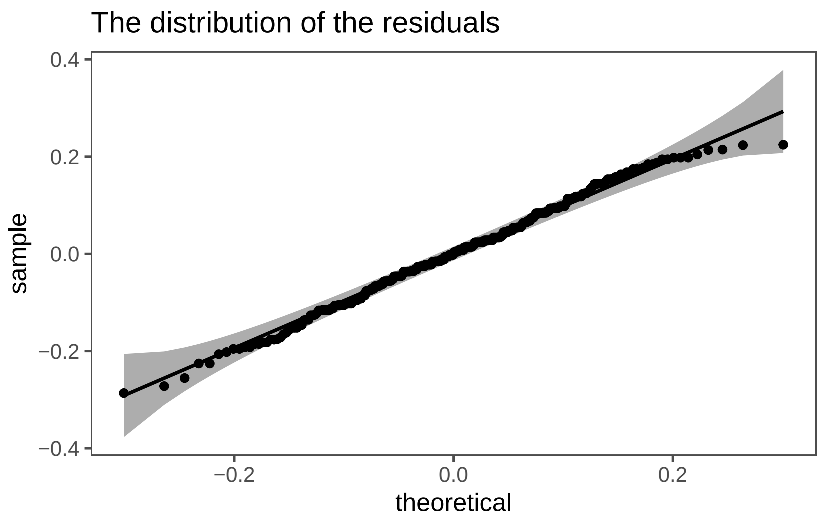


Figure 4.41: The distribution of the residuals.

The model seems to be valid (equal variances of residuals, normal distributed residuals).

With $p < \alpha = 0.05$ H_0 can be rejected, the means come from different populations.

4.14.3 Welch ANOVA

Welch ANOVA: Similar to one-way ANOVA, the Welch ANOVA is employed when there are unequal variances between the groups being compared. It relaxes the assumption of equal variances, making it suitable for situations where variance heterogeneity exists.

- **Null Hypothesis:** True mean difference is not equal to 0.
- **Prerequisites:**
  - Number of groups > 2
  - One response, one predictor variable

The Welch ANOVA drops the prerequisite of equal variances in groups. Because there are more than two groups, the Bartlett test (as introduced in Section 4.13.1.2) is chosen (data is normally distributed).

```
	Bartlett test of homogeneity of variances

data:  diameter by group
Bartlett's K-squared = 275.61, df = 4, p-value < 2.2e-16
```

With $p < \alpha = 0.05$ H_0 can be rejected, the variances are not equal.

The ANOVA table for the Welch ANOVA is shown in Table 4.19.

Table 4.19: The ANOVA results from the ANOVA Welch Test (not assuming equal variances).

num.df	den.df	statistic	p.value	method
4.000	215.085	3,158.109	0.000	One-way analysis ofmeans (not assuming equalvariances)

4.14.4 Kruskal Wallis

Kruskal-Wallis Test: When dealing with non-normally distributed data, the Kruskal-Wallis test is a non-parametric alternative to one-way ANOVA. It is used to evaluate whether there are significant differences in medians among three or more independent groups.

In this example the drive strength is measured using three-point bending. Three different methods are employed to increase the strength of the drive shaft.

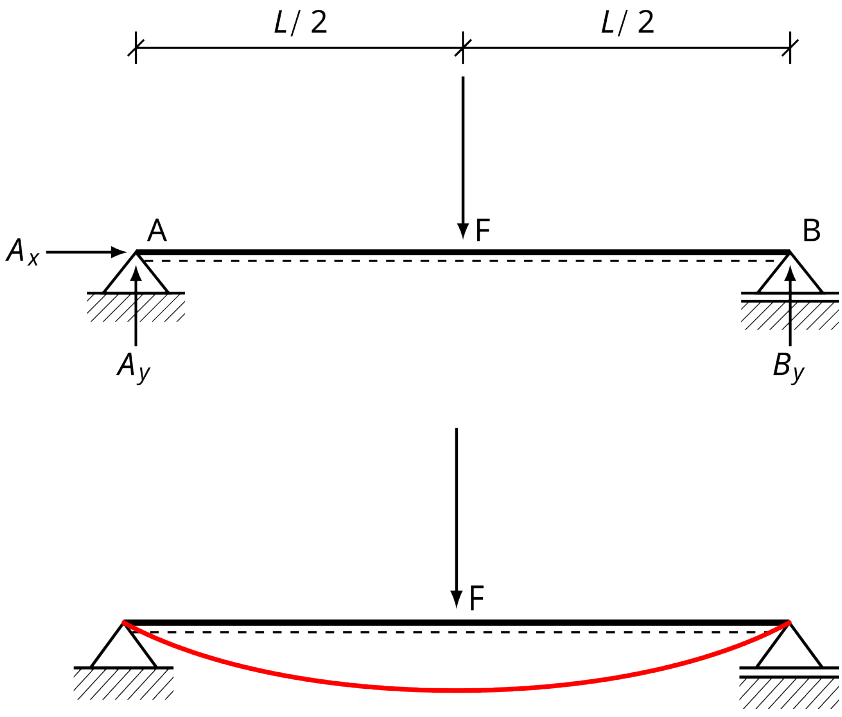


Figure 4.42: The mechanical Background for a three-point bending test

- Method A: baseline material
- Method B: different geometry
- Method C: different material

In Figure 4.43 the raw drive shaft strength data for Method A, B and C is shown. At first glance, the data does not appear to be normally distributed.

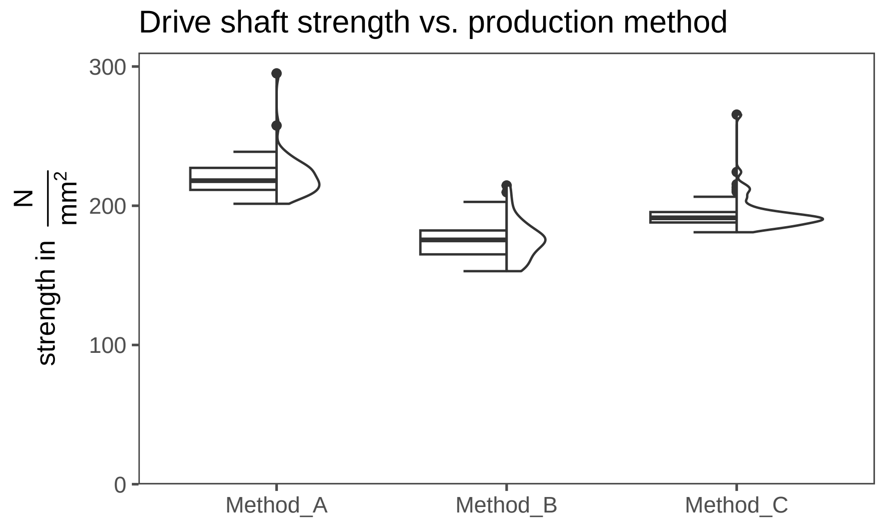


Figure 4.43: The raw data from the drive shaft strength testing.

In Figure 4.44 the visual test for normal distribution is performed. The data does not appear to be normally distributed.

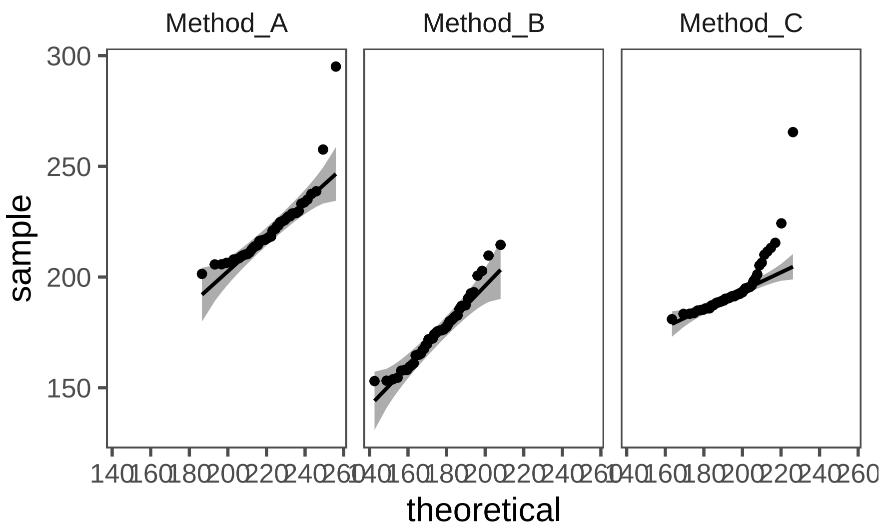


Figure 4.44: The qq-plot for the drive shaft strength testing data.

The Kruskal-Wallis test is then carried out. With $p < \alpha = 0.05$ it is shown, that the groups come from populations with different means. The next step is to find which of the groups are different using a post-hoc analysis.

```
    Kruskal-Wallis rank sum test

data:  strength by group
Kruskal-Wallis chi-squared = 107.65, df = 2, p-value < 2.2e-16
```

The Kruskal-Wallis Test (as the ANOVA) can only tell you, if there is a signifcant difference between the groups, not what groups are different. Post-hoc tests are able to determine such, but must be used with a correction for multiple testing (see (Tamhane 1977))

```
    Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  kw_shaft_data$strength and kw_shaft_data$group

         Method_A Method_B
Method_B < 2e-16  -
Method_C 6.8e-14  2.0e-10

P value adjustment method: bonferroni
```

Because $p < \alpha = 0.05$ it can be concluded, that all means are different from each other.

4.14.5 repeated measures ANOVA

Repeated Measures ANOVA: The repeated measures ANOVA is applicable when you have continuous data with multiple measurements within the same subjects or units over time. It is used to assess whether there are significant differences in means over the repeated measurements, under the assumptions of sphericity and normal distribution.

In this example, the diameter of $n = 20$ drive shafts is measured after three different steps.

- Before Machining
- After Machining
- After Inspection

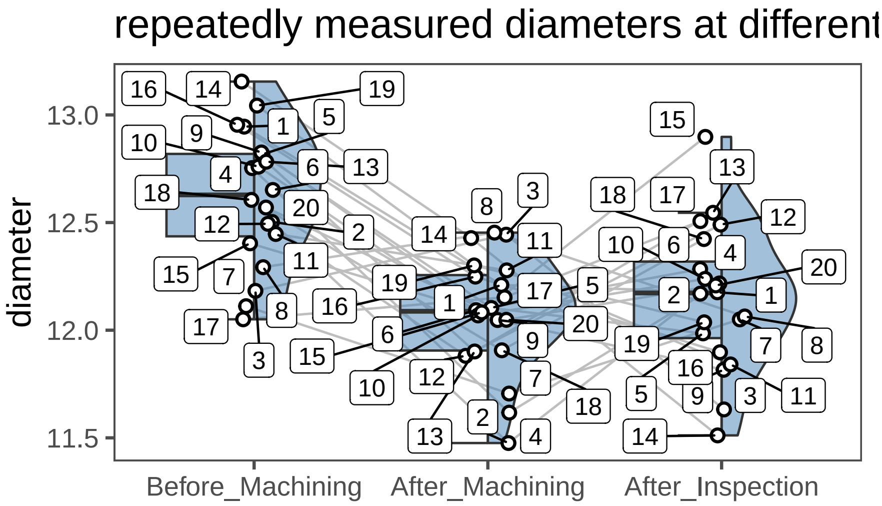


Figure 4.45: The raw data for the repeated measures ANOVA.

First, outliers are identified. There is no strict rule to identify outliers, in this case a classical measure is applied according to (4.14)

$$\text{outlier} = \begin{cases} x_i & > Q3 + 1.5 \cdot IQR \\ x_i & < Q1 - 1.5 \cdot IQR \end{cases} \tag{4.14}$$

```
# A tibble: 1 x 5
  timepoint        Subject_ID diameter is.outlier is.extreme
  <chr>            <fct>         <dbl> <lgl>      <lgl>
1 After_Inspection 15             12.9 TRUE       FALSE
```

A check for normality is done employing the Shapiro-Wilk test (Shapiro and Wilk 1965).

timepoint	variable	statistic	p
After_Inspection	diameter	0.968	0.727
After_Machining	diameter	0.954	0.456
Before_Machining	diameter	0.968	0.741

The next step is to check the dataset for sphericity, meaning to compare the variance of the groups among each other in order to determine the equality thereof. For this the Mauchly Test for sphericity is employed (Mauchly 1940).

```
     Effect     W     p p<.05
1 timepoint 0.927 0.524
```

With $p > \alpha = 0.05$ H_0 is accepted, the variances are equal. Otherwise sphericity corrections must be applied (Greenhouse and Geisser 1959).

The next step is to perform the repeated measures ANOVA, which yields the following results.

Effect	DFn	DFd	F	p	p<.05	ges
timepoint	2.000	36.000	18.081	0.000	*	0.444

With $p < \alpha = 0.05$ H_0 is rejected, the different timepoints yield different diameters. Which groups are different is then determined using a post-hoc test, including a correction for the significance level (Bonferroni 1936).

In this case, the assumptions for a t-test are met, the pairwise t-test can be used.

group1	group2	n1	n2	statistic	df	p	p.adj	signif
After_Inspection	After_Machining	19	19	0.342	18	0.736	1.000	ns
After_Inspection	Before_Machining	19	19	−4.803	18	0.000	0.000	***
After_Machining	Before_Machining	19	19	−6.283	18	0.000	0.000	****

with $p < \alpha = 0.05$ H_0 is rejected for the comparison `Before_Machining - After_Machining` and `After_Inspection - Before_Machining`. It can therefore be concluded that the machining has a significant influence on the diameter, whereas the inspection has none.

4.14.6 Friedman test

The Friedman test is a non-parametric alternative to repeated measures ANOVA (Friedman 1937). It is utilized when dealing with non-normally distributed data and multiple measurements within the same subjects. This test helps determine if there are significant differences in medians over the repeated measurements.

The same data as for the repeated measures ANOVA will be used.

.y.	n	statistic	df	p	method
diameter	20.000	16.900	2.000	0.000	Friedman test

With $p < \alpha = 0.05$ H_0 is rejected, the timepoints play a vital role for the drive shaft parameter.

5 Regression Analysis

Regression analysis is a statistical method used to examine the relationship between one dependent variable and one or more independent variables. It aims to understand how the dependent variable changes when one or more independent variables change.

The core idea is to create a mathematical model that represents this relationship. The model is typically in the form of an equation that predicts the value of the dependent variable based on the values of the independent variables.

There are different types of regression analysis, such as linear regression (when the relationship between variables is linear) and nonlinear regression (when the relationship is not linear). The process involves finding the best-fitting line or curve that minimizes the differences between the predicted values from the model and the actual observed values.

5.1 Linear Regression

$$y = \beta_0 + \beta_1 \cdot X \tag{5.1}$$

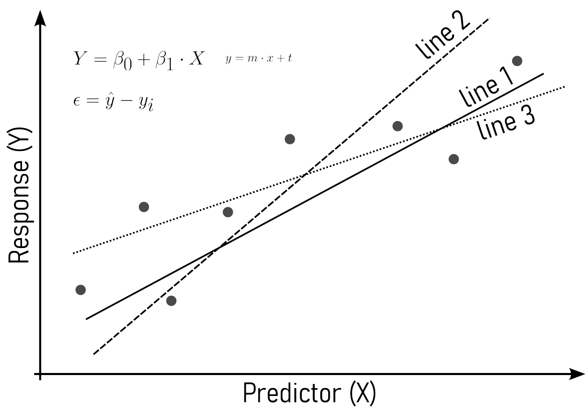


Figure 5.1: The basic idea behind linear regression.

The basic idea behind linear regression is, to find the line of the form $Y = \beta_0 + \beta_1 \cdot X$ that best fits the datapoints. In order to determine the best fit, a criterion to optimize for is needed. This is where residuals come into play.

5.1.1 Residuals

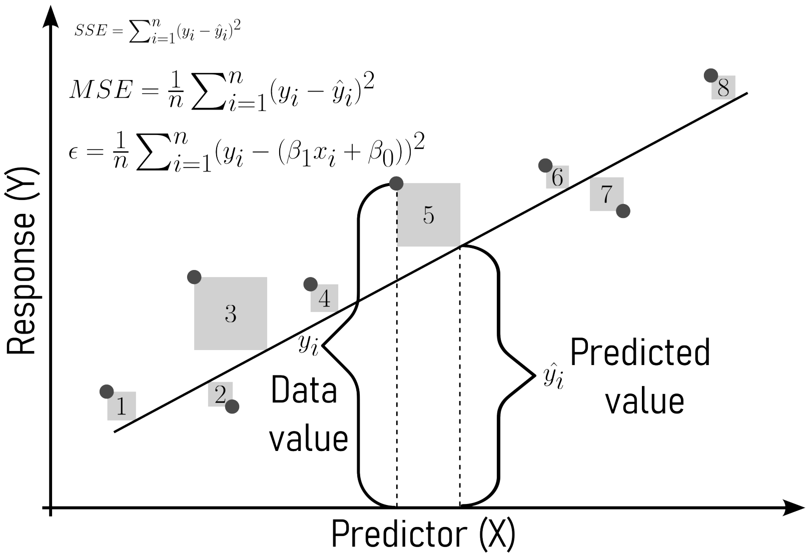


Figure 5.2: The calculation of residuals.

The computation of the residuals is based on (5.2) to the residual sum of squares.

$$RSS = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0))^2 \tag{5.2}$$

5.1.2 Gradient Descent (Ruder 2016)

In linear regression, gradient descent is an iterative optimization process used to minimize the difference between predicted and actual values. It starts with initial coefficients and calculates the gradient of the cost function, representing the error. The coefficients are then updated in the opposite direction of the gradient, with the magnitude of the update controlled by a learning rate. This process is repeated until convergence, gradually refining the coefficients to improve the accuracy of the linear regression model.

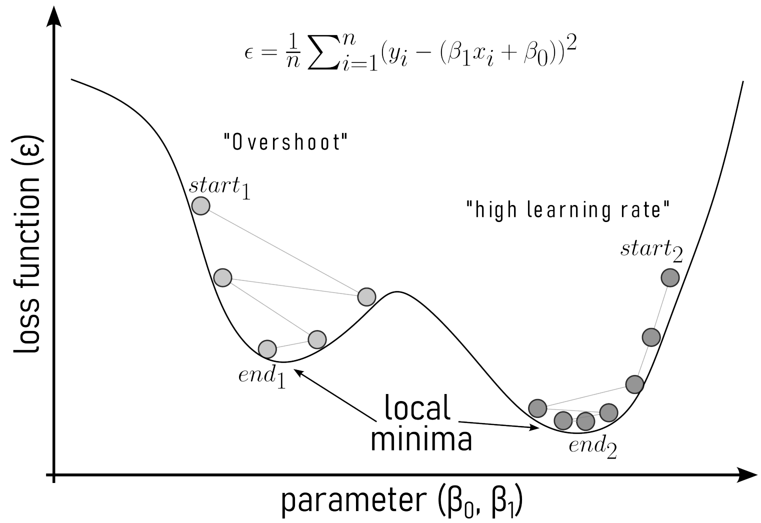


Figure 5.3: An example for the gradient descent algorithm

5.1.3 Model Evaluation and Interpretation

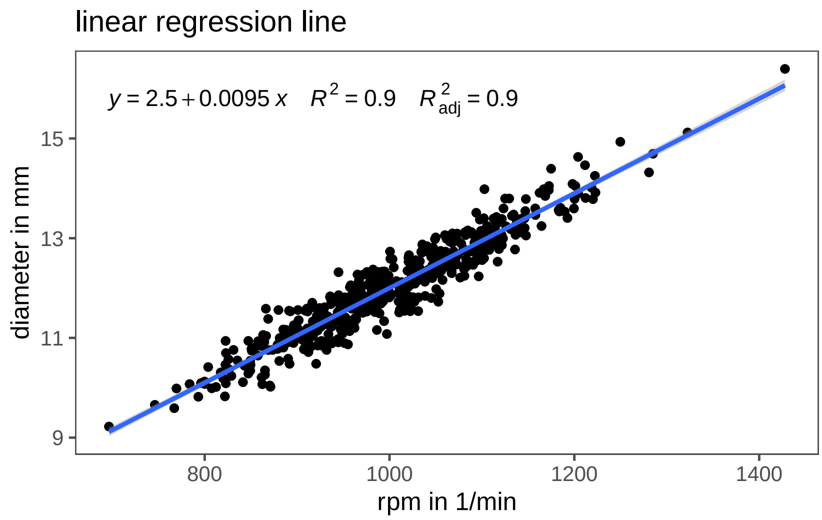


Figure 5.4: The linear regression between rounds per minute (rpm) of the lathing machine and the diameter of the drive shaft.

The coefficient of determination (r^2), is a statistical measure that assesses the proportion of the variance in the dependent variable that is explained by the independent variable(s) in a regression model. It ranges from 0 to 1, where 0 indicates that the model does not explain any variability, and 1 indicates that the model explains all the variability. In other words, r^2 provides insight into the goodness of fit of a regression model, indicating how well the model's predictions match the observed data.

$$r^2 = 1 - \frac{RSS}{SSE} \tag{5.3}$$

The adjusted coefficient of determination, is a modification of the regular r^2 in regression analysis. While r^2 assesses the proportion of variance explained by the independent variables, the $r^2_{adjusted}$ takes into account the number of predictors (k) in the model, addressing potential issues with overfitting according to (5.4).

The $r^2_{adjusted}$ incorporates a penalty for adding unnecessary predictors that do not significantly contribute to explaining the variance in the dependent variable. This adjustment helps prevent an inflated r^2 when including more predictors, even if they don't improve the model significantly.

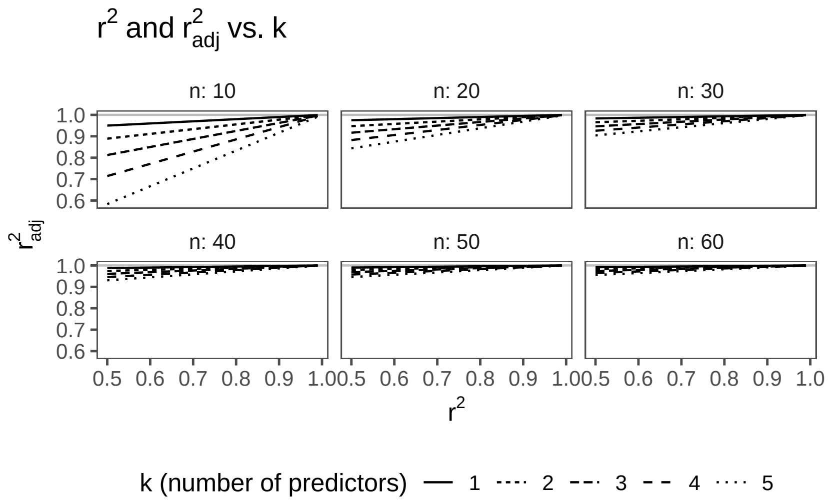


Figure 5.5: The influence of k (number of predictors) on r^2 and $r^2_{adjusted}$.

$$r^2_{adjusted} = 1 - (1 - r^2)\frac{n-1}{n-k-1} \tag{5.4}$$

```
Call:
lm(formula = diameter ~ rpm, data = drive_shaft_rpm_dia)

Residuals:
```

```
     Min       1Q   Median       3Q      Max 
-0.89501 -0.19690 -0.01096  0.21917  1.00742 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.5000000  0.1406190   17.78   <2e-16 ***
rpm         0.0095000  0.0001399   67.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3126 on 498 degrees of freedom
Multiple R-squared:  0.9025,    Adjusted R-squared:  0.9023 
F-statistic:  4610 on 1 and 498 DF,  p-value: < 2.2e-16
```

In linear regression modeling, the absence of a visible pattern in the residuals is desirable because it indicates that the model adequately captures the underlying relationship between the independent and dependent variables. Residuals are the differences between the observed and predicted values, and their randomness or lack of discernible pattern suggests that the model is effectively explaining the variance in the data. A visible pattern in residuals could indicate that the model fails to account for certain patterns or trends, suggesting potential shortcomings or misspecifications in the regression model. Detecting and addressing such patterns in residuals is crucial for ensuring the validity and reliability of the linear regression analysis.

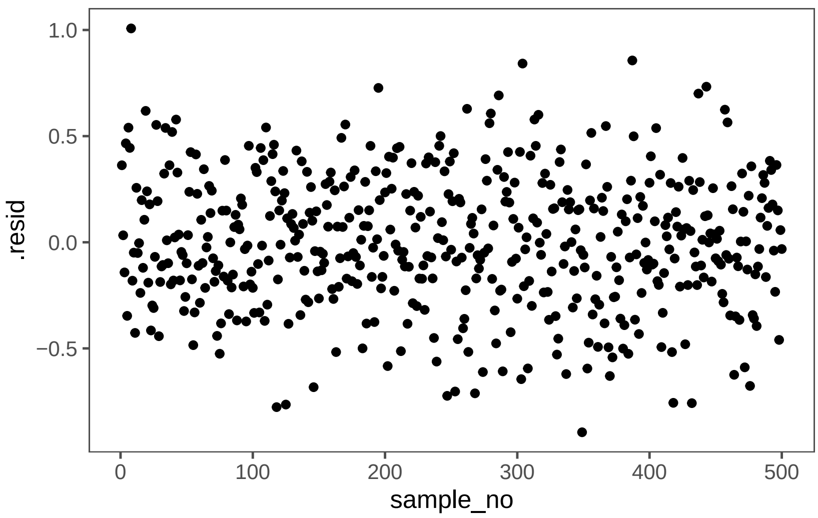


Figure 5.6: There should not be a visible pattern in the residuals.

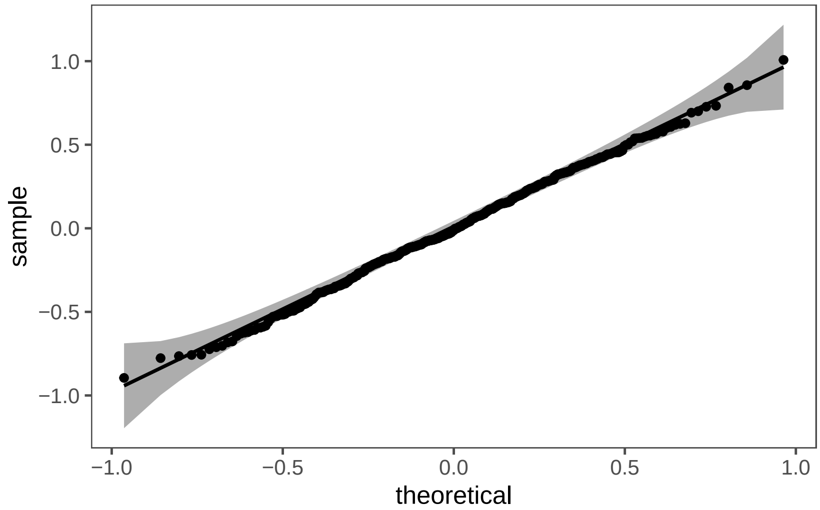


Figure 5.7: The resiuals should be normally distributed.

In linear regression, the assumption of normally distributed residuals is essential for accurate statistical inference, parameter estimation using ordinary least squares, and constructing reliable confidence intervals. Normal residuals indicate that the model appropriately captures data variability and helps identify issues like heteroscedasticity. While departures from normality may not always invalidate results, adherence to this assumption enhances the model's robustness and reliability. If consistently violated, alternative modeling approaches or transformations may be considered.

5.1.4 Hypostesis testing in linear regression

Null Hypothesis (H0): $\beta_1 = 0$

Alternative Hypothesis (Ha): $\beta_1 \neq 0$

Table 5.1: The significance of model parameters.

term	estimate	std.error	statistic	p.value
(Intercept)	2.500	0.141	17.779	0.000
rpm	0.010	0.000	67.895	0.000

In linear regression, t testing of coefficients assesses whether individual regression coefficients significantly differ from zero, providing insights into the significance of each predictor's contribution to the model.

Table 5.2: The significance of the model.

r.squared	adj.r.squared	statistic	p.value	df	df.residual	nobs
0.902	0.902	4,609.692	0.000	1.000	498.000	500.000

In linear regression, the F-test assesses the overall significance of the regression model by comparing the fit of the model with predictors to a model without predictors, helping determine if the regression equation explains a significant proportion of the variance in the dependent variable.

5.2 Multiple linear regression

Table 5.3: The data in a tabular overview including test for normal distribution.

Characteristic	**Overall**			
N = 500[1]	**A**			
N = 165[1]	**B**			
N = 181[1]	**C**			
N = 154[1]	**p-value**			
rpm	999 (932, 1,068)	993 (923, 1,061)	995 (927, 1,074)	1,012 (946, 1,06
diameter	11.95 (11.30, 12.66)	11.90 (11.24, 12.51)	11.98 (11.30, 12.67)	12.01 (11.41, 12.7
feed	40.01 (39.34, 40.67)	39.98 (39.34, 40.63)	39.91 (39.34, 40.65)	40.05 (39.37, 40.7

[1]Median (Q1, Q3)

A short exploratory data analysis of the data for the multiple linear regression is given in Table 5.3.

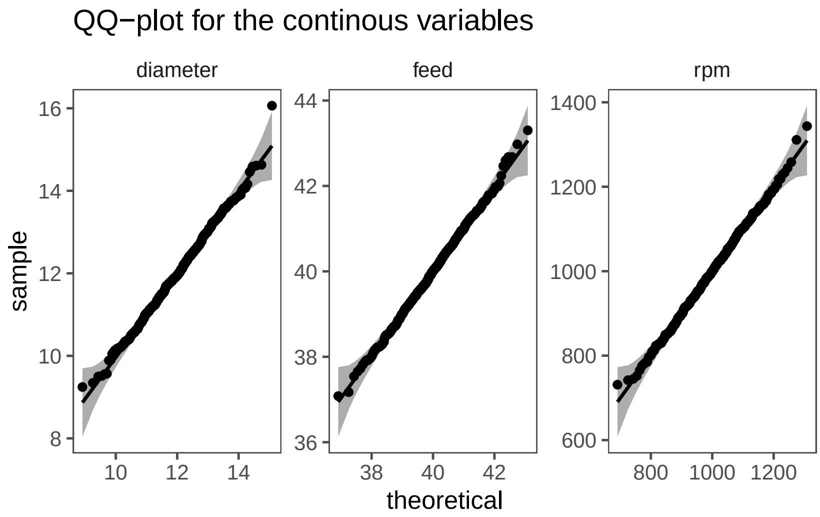


Figure 5.8: The graphical test for normal distribution (QQ-plot)

Figure 5.8 shows the graphical test for normal distribution for the multiple linear regression.

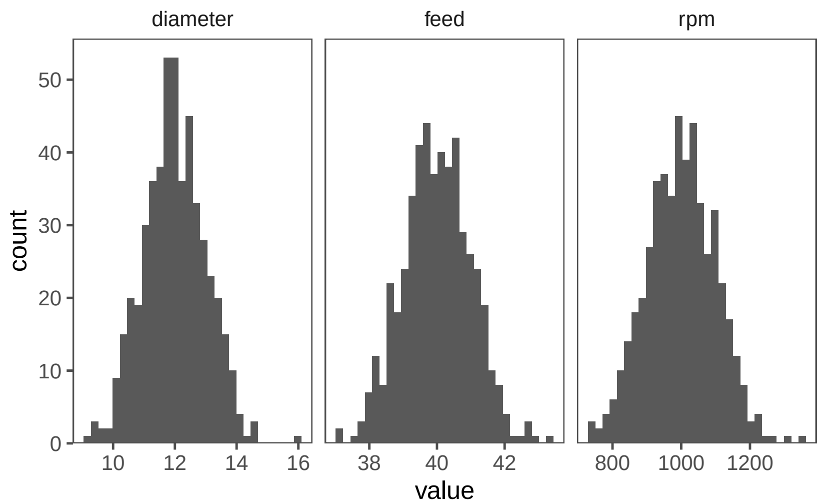


Figure 5.9: The distribution of the output and input parameters.

In Figure 5.9 the distribution of the input data is shown in a histogram.

$$Y \sim rpm + feed + site \tag{5.5}$$

Table 5.4: The output of the multiple linear regression modelling

Characteristic	**Beta**	**95% CI**[1]	**p-value**
rpm	0.00	0.00, 0.01	<0.001
feed	0.44	0.29, 0.58	<0.001
site			
A	0.00	—	
B	0.09	-0.02, 0.20	0.11
C	0.08	-0.03, 0.20	0.15

[1]CI = Confidence Interval

(5.5) shows the general model for the multiple linear regression model. In this example, also the production site (`site A`, `site B` and `site C`) is included to test, if different production sites lead to differently produced drive shafts. The results of the multiple regression are shown in Table 5.4. Whilst the continuous variables appear to be significant ($p < \alpha = 0.05$), the production site does not play a significant rolefor the drive shaft diameter.

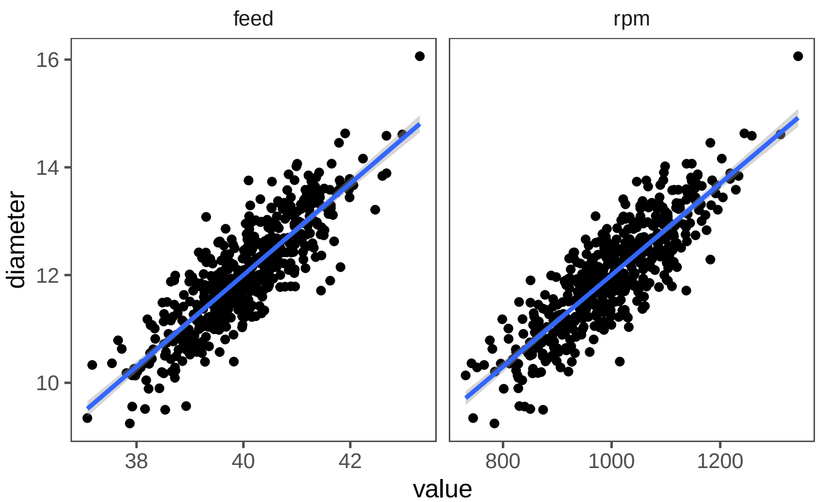


Figure 5.10: The model of the mulitple linear regression

In Figure 5.10 the model is shown to ease the interpretation. With increasing `rpm` or `feed` also the drive shaft diameter increases.

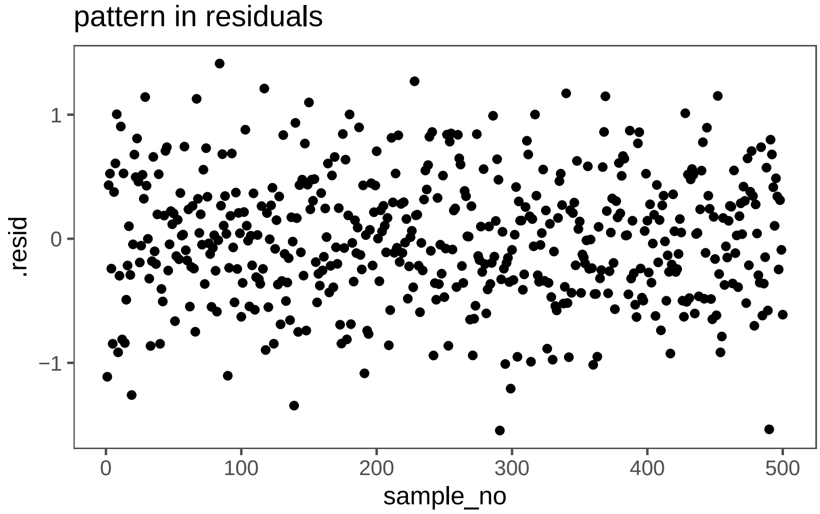


Figure 5.11: The check for pattern in the residuals

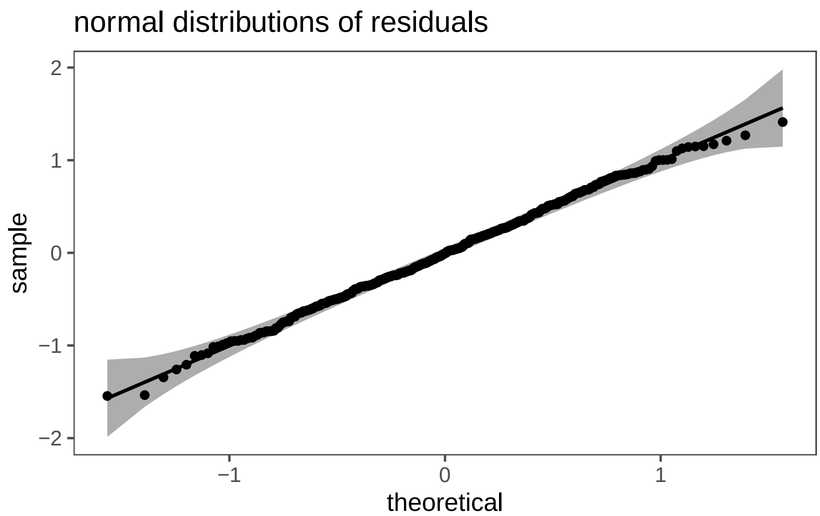


Figure 5.12: The check for normal distribution in the residuals.

In Figure 5.12 the normal distribution of the residuals is confirmed, the model appears to be valid.

5.3 Logistic Regression

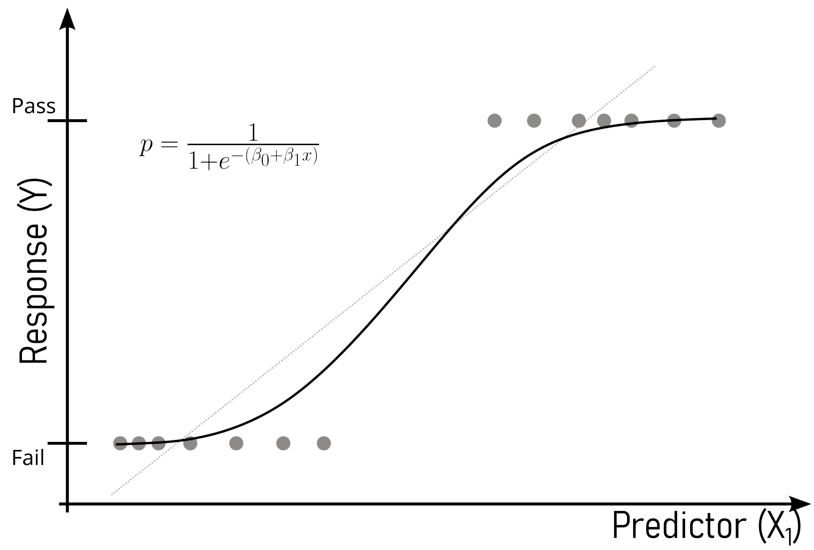


Figure 5.13: The basic idea of logisitic regression.

Logistic regression is a statistical method designed for binary classification problems (Figure 5.13). It models the probability that an observation belongs to a particular class using the sigmoid (logistic) function (5.6). The key steps include:

1. **Probability Modeling:**
   - Model predicts the probability of an instance belonging to a specific class.
2. **Linear Combination:**
   - Combines linearly weighted input features, representing the log-odds of the positive class.
3. **Sigmoid Function:**
   - Transforms the linear combination to ensure output is between 0 and 1.
4. **Decision Boundary:**
   - Threshold probability (usually 0.5) determines class assignment.
5. **Maximum Likelihood Estimation:**
   - Parameters are estimated using maximum likelihood to maximize the likelihood of observed outcomes.
6. **Odds Ratio:**

- Quantifies the impact of each predictor on the odds of the positive class.

Logistic regression is widely used for binary classification tasks in different domains, providing an interpretable way to model the relationship between predictors and a binary outcome.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{5.6}$$

The ordinary linear regression equation is shown in (5.1).

If for y the probabilities P are used they may be > 1 or < 0 which is not possible for P.

To overcome this issue, the odds of $P = \frac{P}{1-P}$ are taken.

$$\begin{aligned} \frac{P}{1-P} &= \beta_0 + \beta_1 x \\ \frac{P}{1-P} &\in 0 \ldots + \infty \end{aligned} \tag{5.7}$$

Restricted variables are not easy to model why (5.7) is expanded to (5.8).

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x \tag{5.8}$$

Which then in turn gives (5.6).

5.3.1 $\beta_0 = 1$ and $\beta_1 = 1$

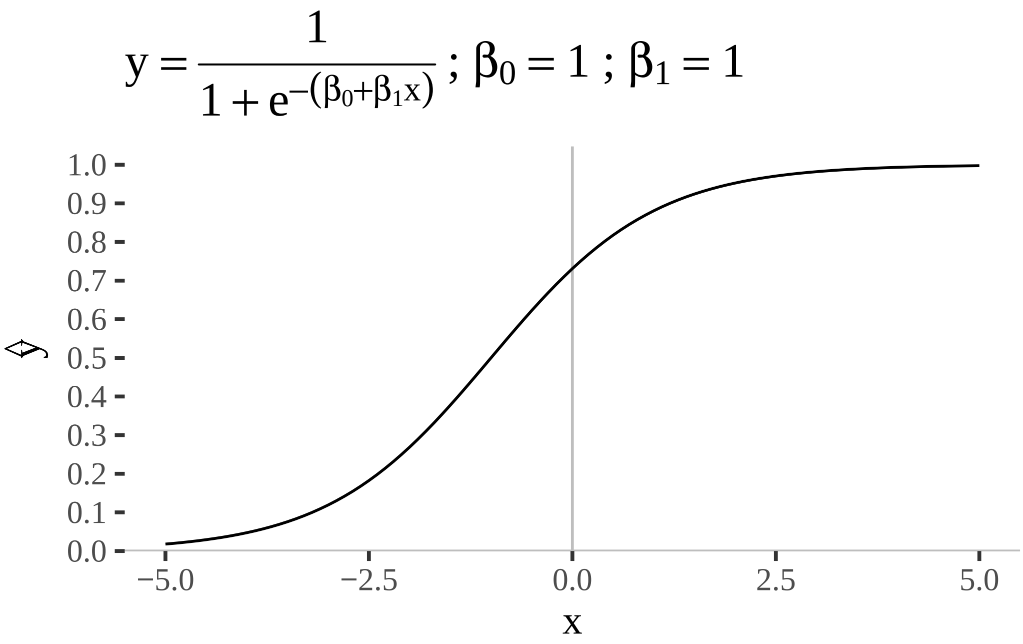


Figure 5.14: The influence of different paramters for the sigmoid function

In order to better understand the influencing factors a small parametric study on β_0 and β_1 is given. Figure 5.14 the sigmoid function $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$ with $\beta_0 = 1$ and $\beta_1 = 1$ is shown as a reference. Please note that the *linear regression* $(\beta_0 + \beta_1 x)$ expands the usual *sigmoid* function which is given by

$$f(x) = \frac{1}{1 + e^{-x}}$$

to model it in the *intercept* and *gradient* kind of logic.

5.3.2 $\beta_0 = 1$ and $\beta_1 = 0 \ldots 5$

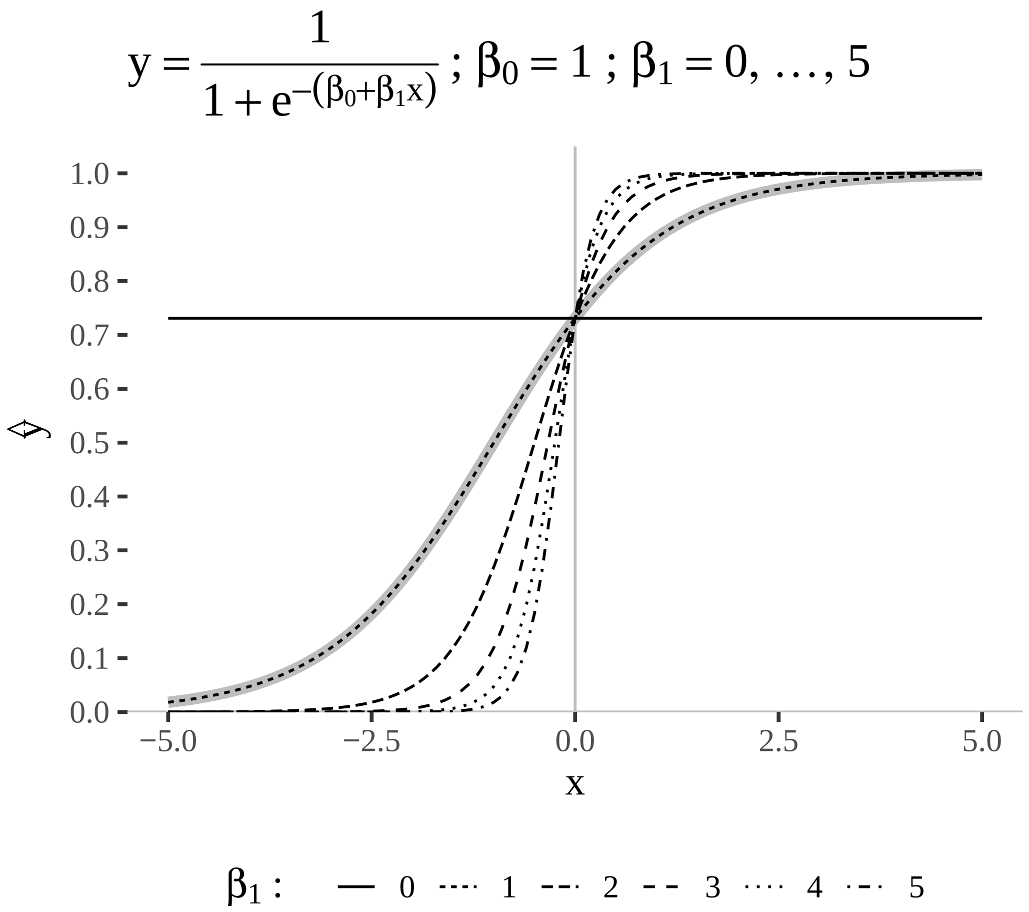


Figure 5.15: The influence of different paramters for the sigmoid function

In the first case of the parametric study the *gradient* parameter is studied by varying it between $0 \ldots 5$ with $step_{size} = 1$. From Figure 5.15 it can be seen, that the *linear regression gradient* parameters varies the characteristic *S*-like shape of the sigmoid function. The higher β_1 is, the more pronounced the *S*-shape becomes. The reference shape for $\beta_0 = 1$ and $\beta_1 = 1$ is shown in light gray in the figure. An interesting effect is visible for a gradient of $\beta_1 = 0$: The function becomes a constant which only depends on the *intercept* (in this case $\beta_0 = 1$).

5.3.3 $\beta_0 = 1$ and $\beta_1 = -5 \dots 0$

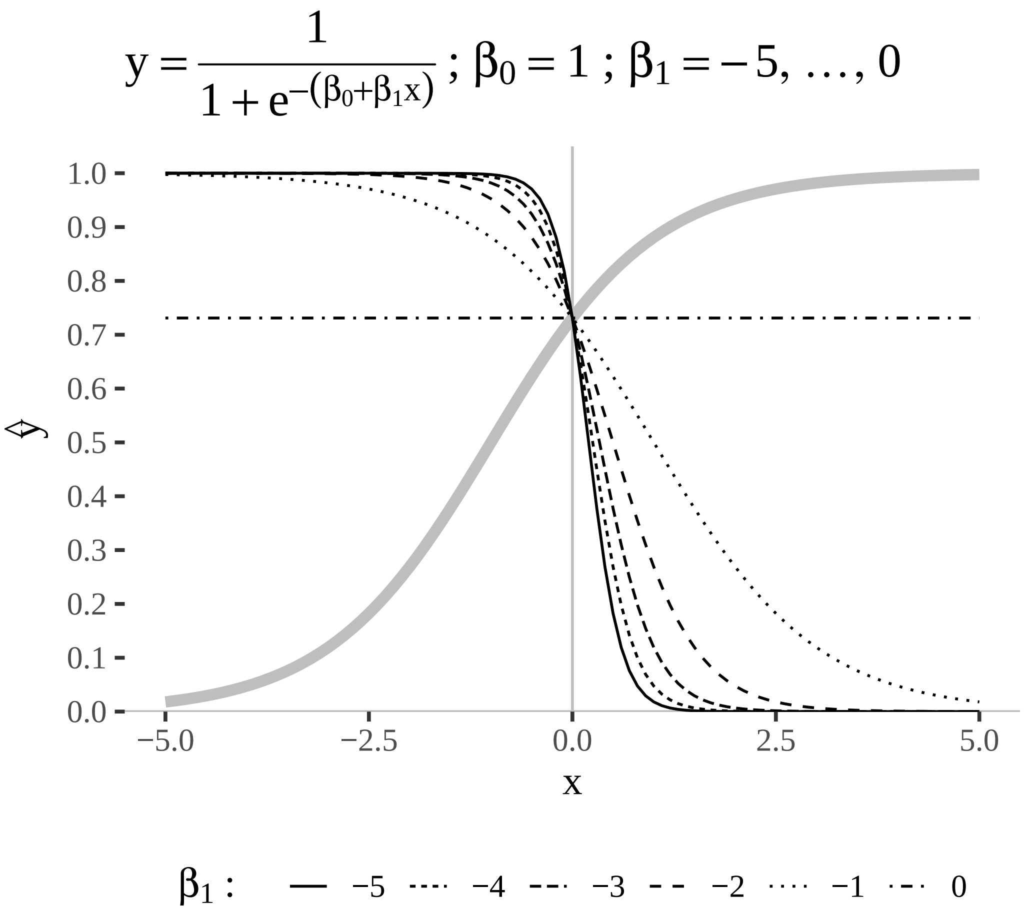


Figure 5.16: The influence of different paramters for the sigmoid function

When the parameter study is expanded to negative values of β_1 ($\beta_1 = -5 \dots 0$) the curve is mirrored and reverses its direction (see Figure 5.16), which is also highlighted by the reference shape for $\beta_0 = 1$ and $\beta_1 = 1$ in light gray. The general interpretation for the influence of this parameter is reversed by stays the same: the larger the deviation from 0 is for β_1, the more pronounced the S-like shape becomes.

5.3.4 $\beta_0 = 0 \ldots 5$ and $\beta_1 = 1$

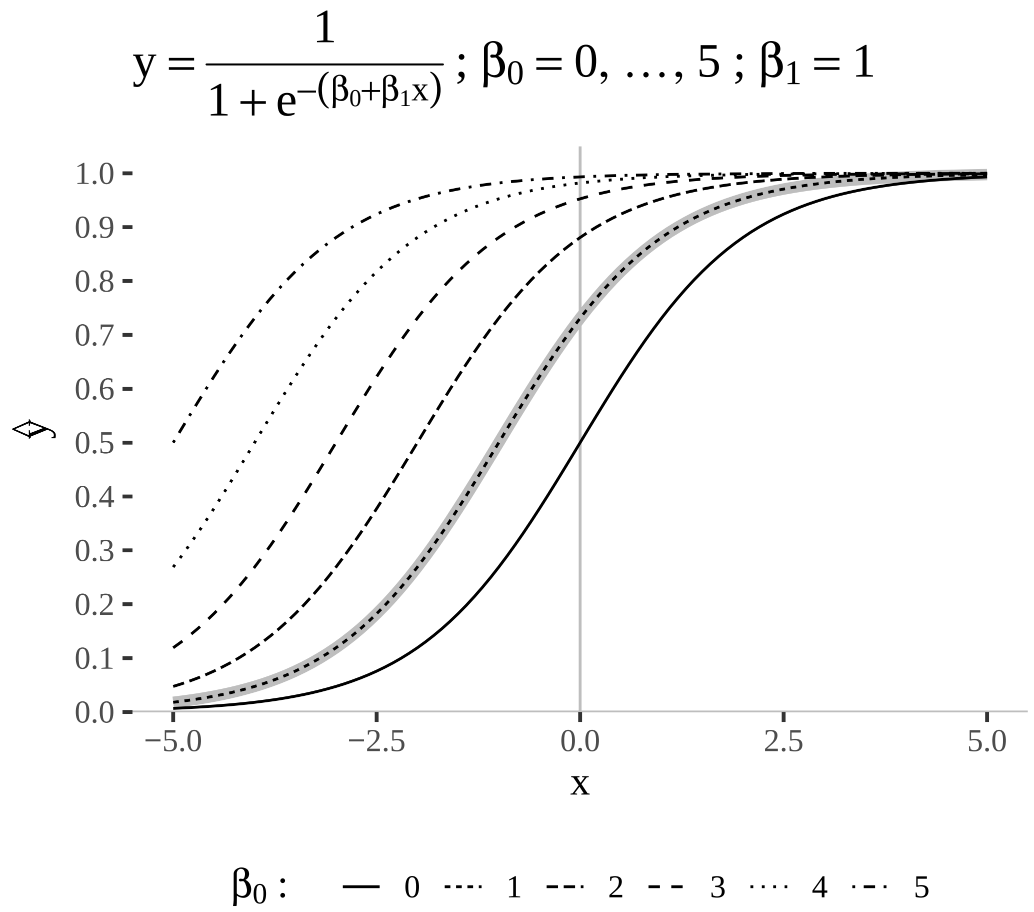


Figure 5.17: The influence of different paramters for the sigmoid function

The second step is to vary the *intercept* (β_1) of the linear regression function that is "hidden" within the sigmoid function. The reference function for $\beta_0 = 1$ and $\beta_1 = 1$ is again shown in light gray in the background in Figure 5.17. It can clearly be seen, that the *intercept* in a sigmoid-function setting can be used as a kind of offset. Whilst the curve is exactly 0.5 at $\beta_0 = 0$, this intersection can be adapted by modeling the intercept. For $\beta_0 > 0$ the intersection point becomes > 0.5.

5.3.5 $\beta_0 = -5 \ldots 0$ and $\beta_1 = 1$

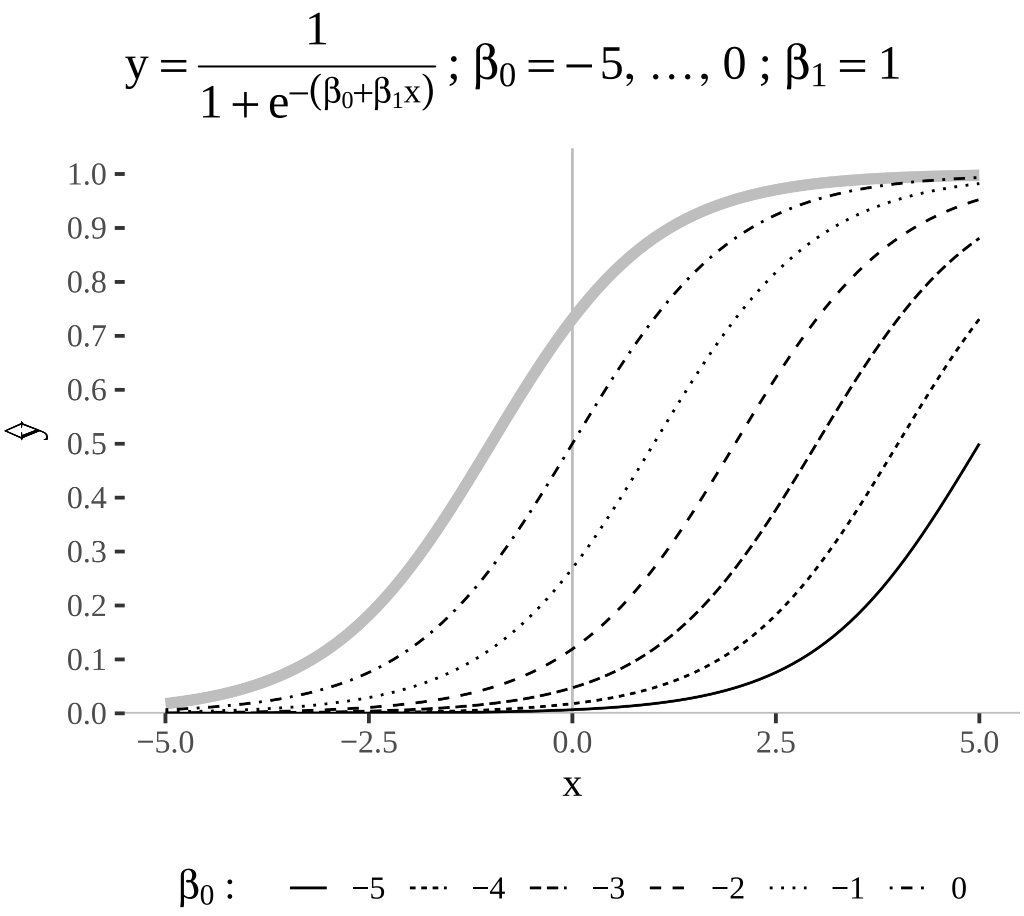


Figure 5.18: The influence of different paramters for the sigmoid function

The reference function for $\beta_0 = 1$ and $\beta_1 = 1$ is again shown in light gray in the background in Figure 5.18. For an *intercept* < 0 the intersection point with the `xaxis` then offsets the curve in the other direction compared with Figure 5.17. For $\beta_0 < 0$ the intersection point becomes < 0.5. In both cases the S-shape like characteristic of the sigmoid function is retained.

5.3.6 Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE) is a statistical method used for estimating the parameters of a model (Starmer 2022). In this approach, the parameter values are chosen to maximize the likelihood function, which represents the probability of observing the given data under the assumed statistical model. The idea is to find the parameter values that make the observed data most probable.

In contrast to the cost function for linear regression (4.9), $\hat{y}_i$ in logistic regression is a non-linear function (5.9).

$$\hat{y} = \frac{1}{1 + e^{-z}} \tag{5.9}$$

Which is why the Maximum Likelihood Estimator is used.

Using the MLE basically means, to try different models (with different model parameters) that maximize the likelihood of the parameters being true. Because it is easier to look for minima (gradient descent), a loss function is formulated that can be used as a loss function.

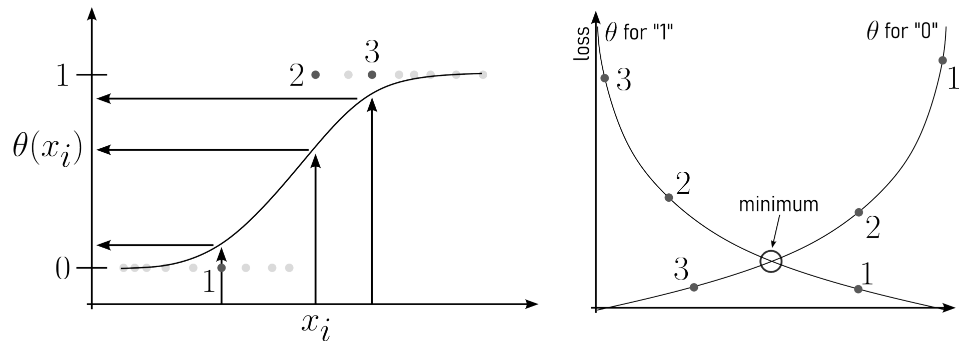


Figure 5.19: The principle of MLE.

$$-\log L(\theta) = -\sum_{i=1}^{n} y \log(\sigma(\theta^T x^i)) + (1 - y) \log(1 - \sigma(\theta^T x^i)) \tag{5.10}$$

5.3.7 Modeling Production Data

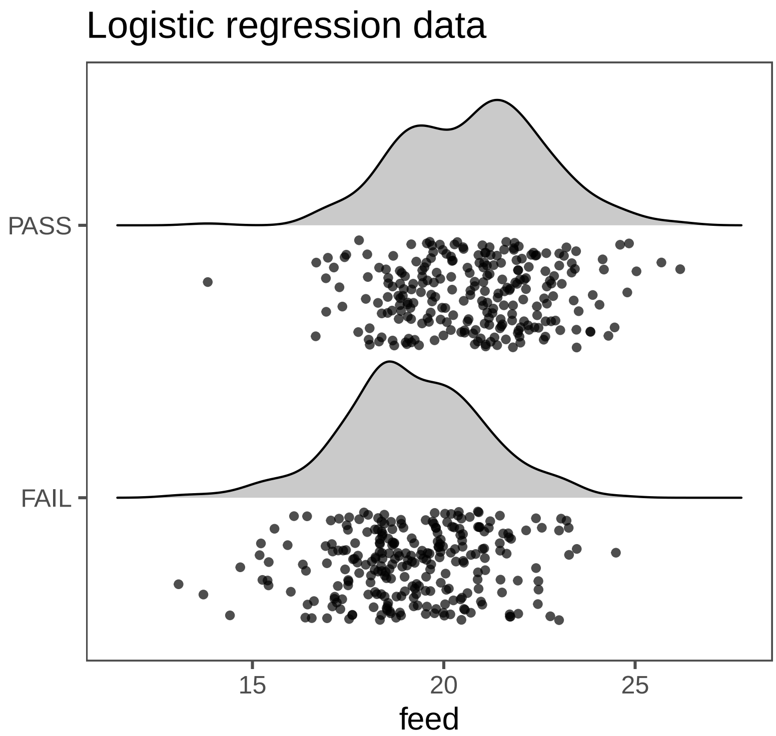


Figure 5.20: The data for the logistic regression data.

In Figure 5.20 the data for the production data. The drive shafts have been rated between `PASS`and `FAIL` and the lathing machine feed has been recorded. The question is now, at which feed the drive shafts start to FAIL.

Table 5.5: The overview of the logistic regression data.

Characteristic	**N = 500**[1]
feed	19.89 (18.55, 21.40)
pass_1_fail_0	
0	256 (51%)
1	244 (49%)

[1]Median (Q1, Q3); n (%)

Table 5.5 shows an overview of the logistic regression data. `PASS` and `FAIL` are fairly similar distributed.

Table 5.6: The modeling of the logisitic regression data.

Characteristic	log(OR)[1]	95% CI[1]	p-value
feed	0.46	0.35, 0.57	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

The model coefficients are shown in Table 5.6. Translated in equation (5.11) and (5.12) we can see, what has been computed.

$$\log(\frac{P}{1-P}) = -9.17 + 0.46x \tag{5.11}$$
$$\frac{P}{1-P} = e^{-9.17+0.46x} \tag{5.12}$$

Therefore the models explains what the odds $\frac{P}{1-P}$ are for a drive shaft to be `FAIL` or `PASS` for a given `feed`.

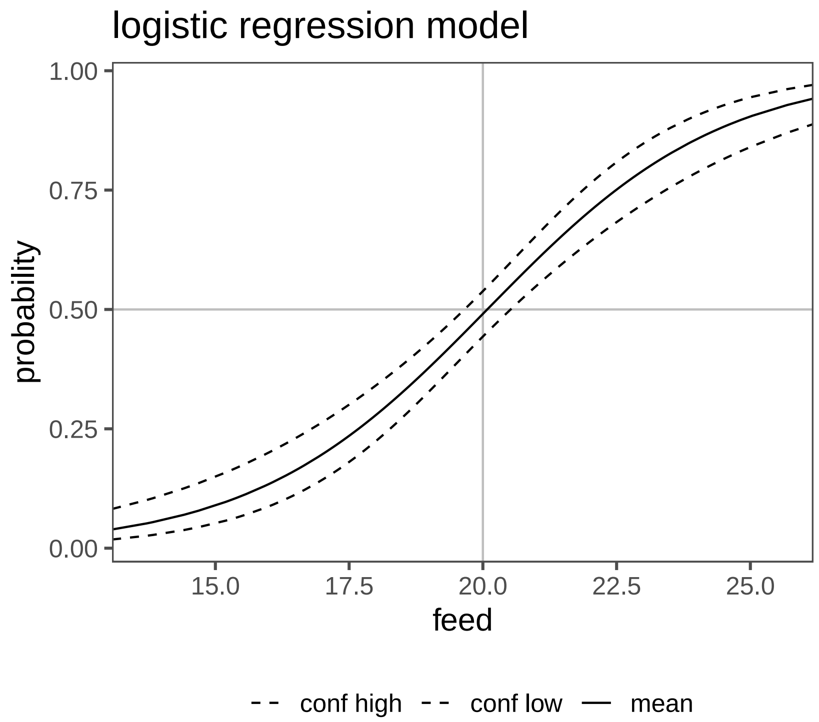


Figure 5.21: The probability (odds) for a drive shaft being PASS or FAIL for a given feed

Figure 5.21 shows the probability for a drive shaft `PASS` or `FAIL` for a given feed as well as the confidence interval of the odds ratio for any given feed. For example the probability for `PASS` at a `feed` of 20 is 49% with a confidence interval of 44% to 54%.

5.3.7.0.1 residuals

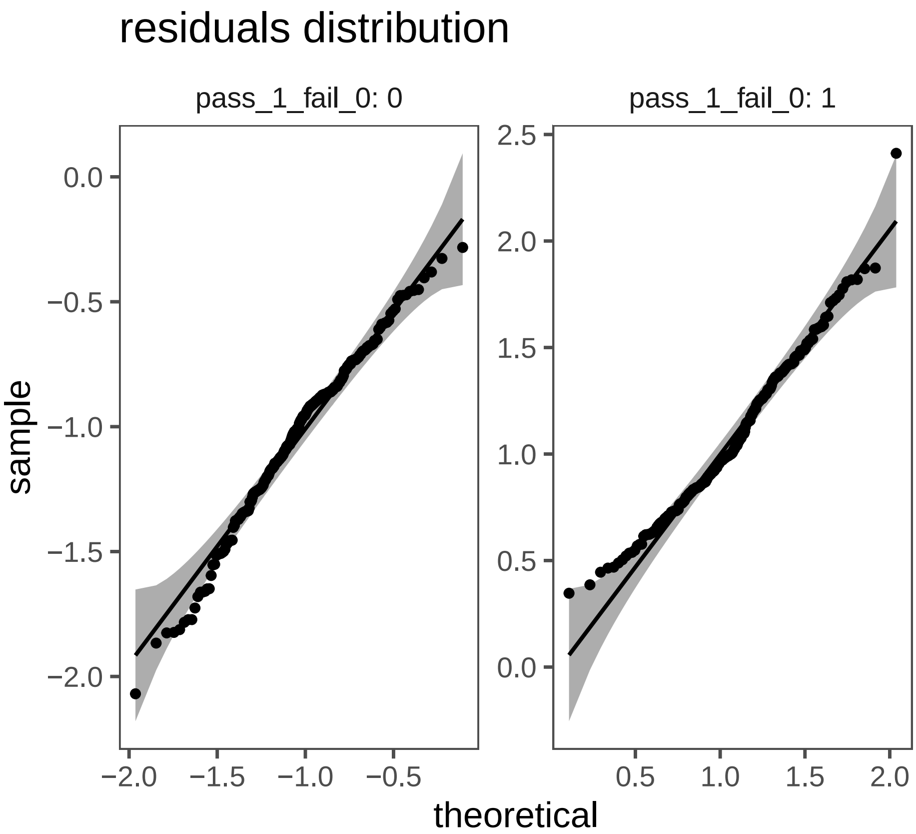


Figure 5.22: Are the residuals of the model normally distributed?

5.3.7.1 Mc Fadden R^2

McFadden's R^2 is a measure used to evaluate the goodness of fit for logistic regression models and is calculated using (5.13).

$$R^2 = 1 - \frac{\log(L_{model})}{\log(L_{null})} = 0.1198876 \tag{5.13}$$

It compares the `model` to the `null`-model. It is much smaller then the coefficient of determination with values ranging between 0.2 … 0.4 already indicating a good model fit in practice.

5.3.7.2 Confusion Matrix

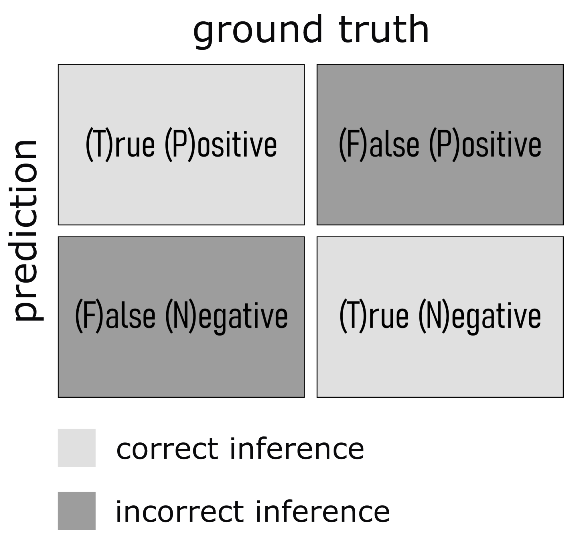


Figure 5.23: A confusion matrix

A confusion matrix is a table used to evaluate the performance of a classification algorithm. It provides a detailed breakdown of the actual versus predicted classifications, enabling the calculation of various performance metrics. The matrix is particularly useful for binary and multiclass classification problems.

On the `x-axis` usually the *ground truth* is depicted whereas on the `y-axis` the predictions of the algorithm are shown. From this several performance metrics can be calculated.

- True Positive (**TP**): The number of positive instances correctly classified as positive.
- False Positive (**FP**): The number of negative instances incorrectly classified as positive (also known as Type I error).
- True Negative (**TN**): The number of negative instances correctly classified as negative.
- False Negative (**FN**): The number of positive instances incorrectly classified as negative (also known as Type II error).

5.3.7.2.1 Accuracy

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Definition The ratio of correctly predicted instances (both true positives and true negatives) to the total instances.

Interpretation Accuracy measures the overall correctness of the model. It indicates the proportion of total predictions that were correct. While accuracy is useful, it can be misleading in cases of imbalanced datasets where one class is more frequent than the other.

5.3.7.2.2 Precision

$$\frac{TP}{TP + FP}$$

Definition The ratio of true positive instances to the total instances predicted as positive.

Interpretation Precision, also known as positive predictive value, measures the accuracy of positive predictions. It is the proportion of correctly identified positive instances out of all instances predicted as positive. High precision indicates a low false positive rate.

5.3.7.2.3 Recall

$$\frac{TP}{TP + FN}$$

Definition The ratio of true positive instances to the total actual positive instances.

Interpretation Recall measures the model's ability to correctly identify all positive instances. It is the proportion of correctly identified positive instances out of all actual positive instances. High recall indicates a low false negative rate.

5.3.7.2.4 Specificity

$$\frac{TN}{TN + FP}$$

Definition The ratio of true negative instances to the total actual negative instances.

Interpretation Specificity measures the model's ability to correctly identify negative instances. It is the proportion of correctly identified negative instances out of all actual negative instances. High specificity indicates a low false positive rate.

5.3.7.2.5 F1 Score

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Definition The harmonic mean of precision and recall.

Interpretation The F1 Score combines precision and recall into a single metric. It provides a balance between the two, particularly useful when you need to take both false positives and false negatives into account. The F1 score is especially helpful when the class distribution is uneven or when you seek a balance between precision and recall.

5.3.7.2.6 Summary on metrics

- **Accuracy** is best for overall performance but can be misleading for imbalanced datasets.
- **Precision** is crucial when the cost of false positives is high.
- **Recall** is important when the cost of false negatives is high.
- **Specificity** complements recall, providing insight into the true negative rate.
- **F1 Score** offers a balanced measure, useful when both precision and recall are important.

5.3.7.3 Confusion Matrix in practice

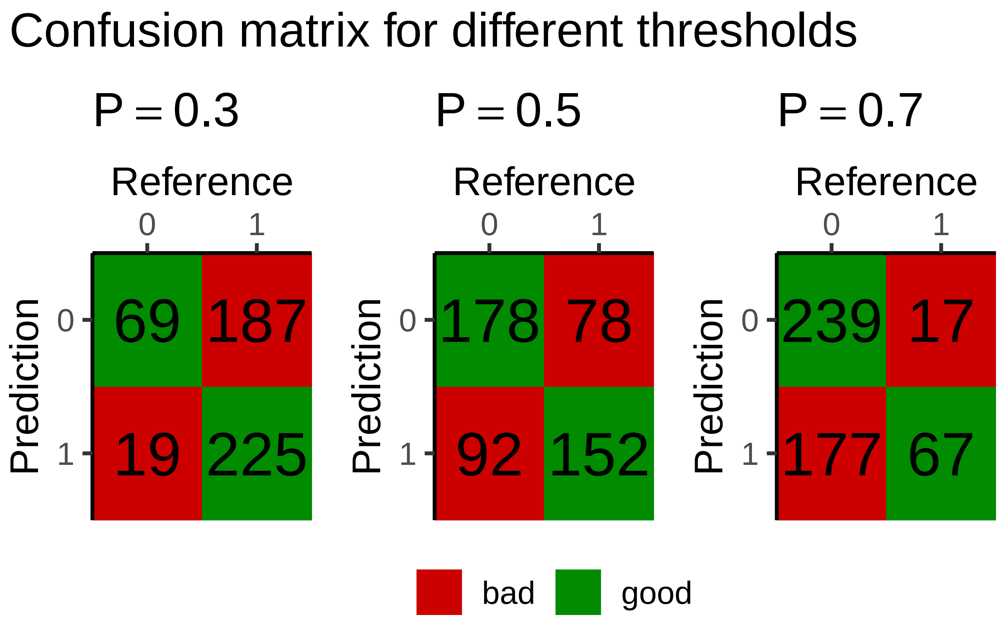
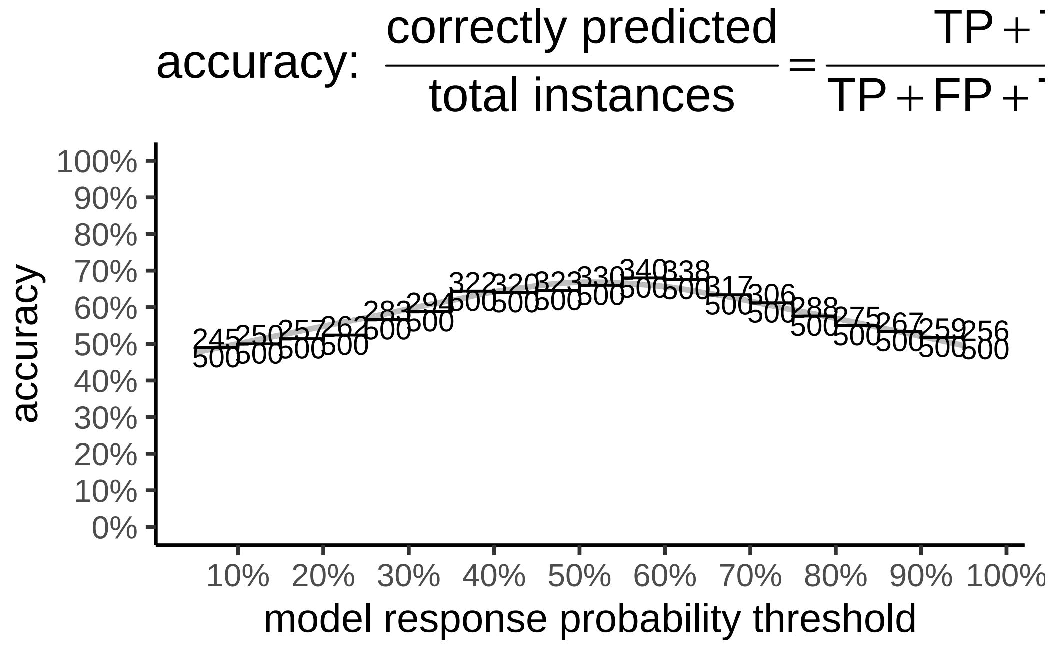


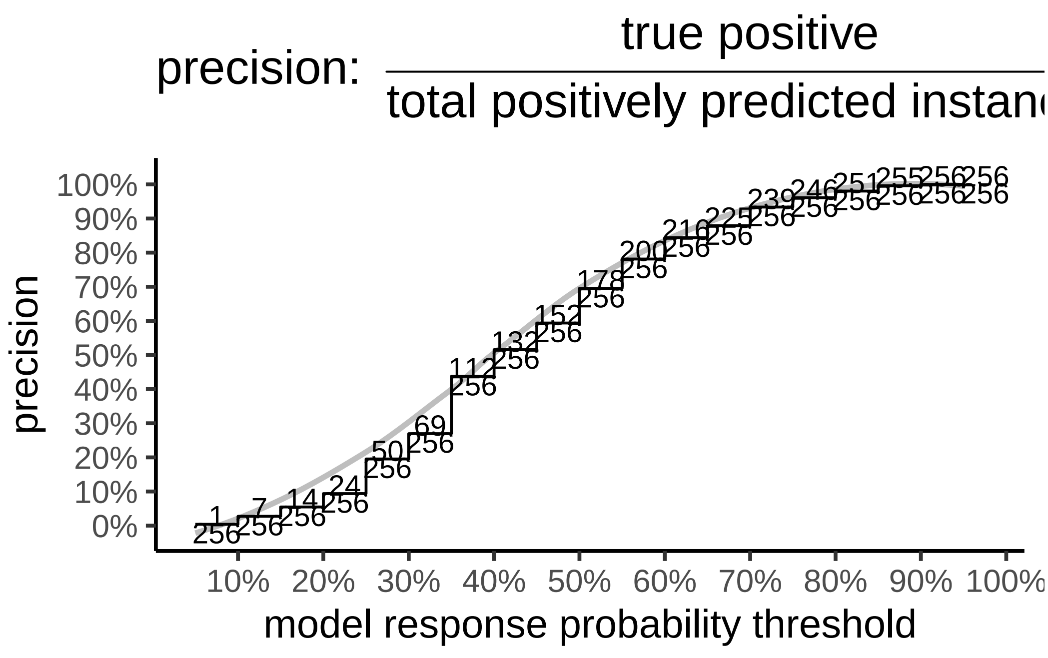
Figure 5.24: Confusion matrices at different probability thresholds

Figure 5.24 shows three different confusion matrices at different probability threshold for the logistic regression model and the respective True Positive, False Positive, True Negative and False Negative rates. On the `x-axis` the reference is depicted and the *true* classes, being 0 for `FAIL` and 1 for `PASS` parts. The `y-axis` shows the prediction of the respective model with the classes again being 0 for `FAIL` and 1 for `PASS`. The *probability threshold* $P = 0.3 \dots 0.7$ is the classification threshold of the model. The logistic regression model computes a *Probability* based on the *Predictor* variable (`feed`). This threshold then classifies the product as `pass` or `fail`

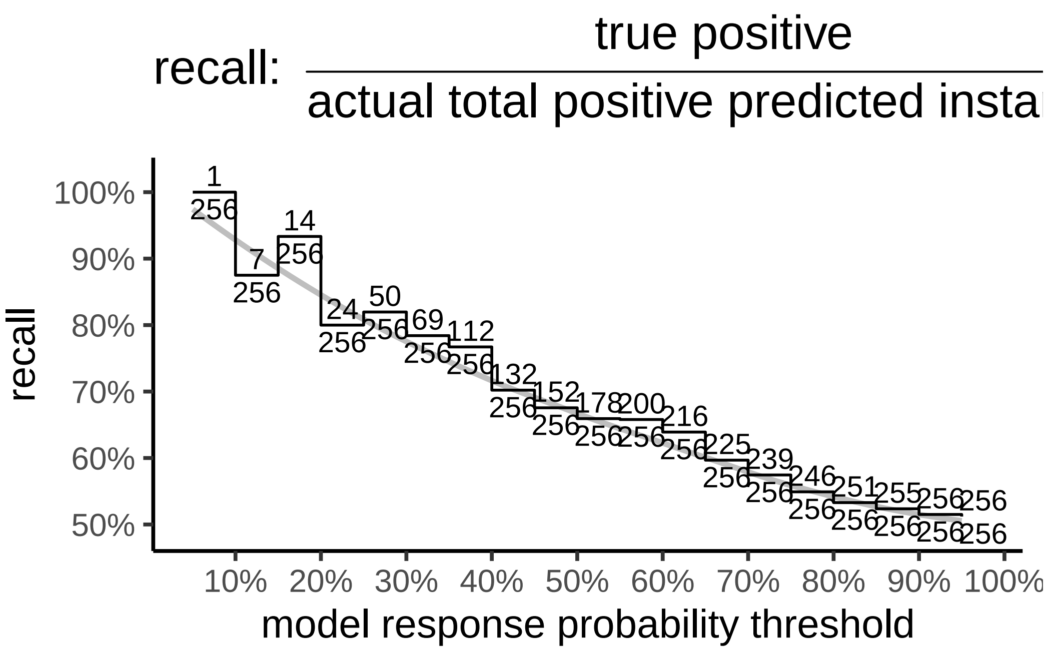
5.3.7.4 Accuracy, correct classification rate, proportion of correct predictions



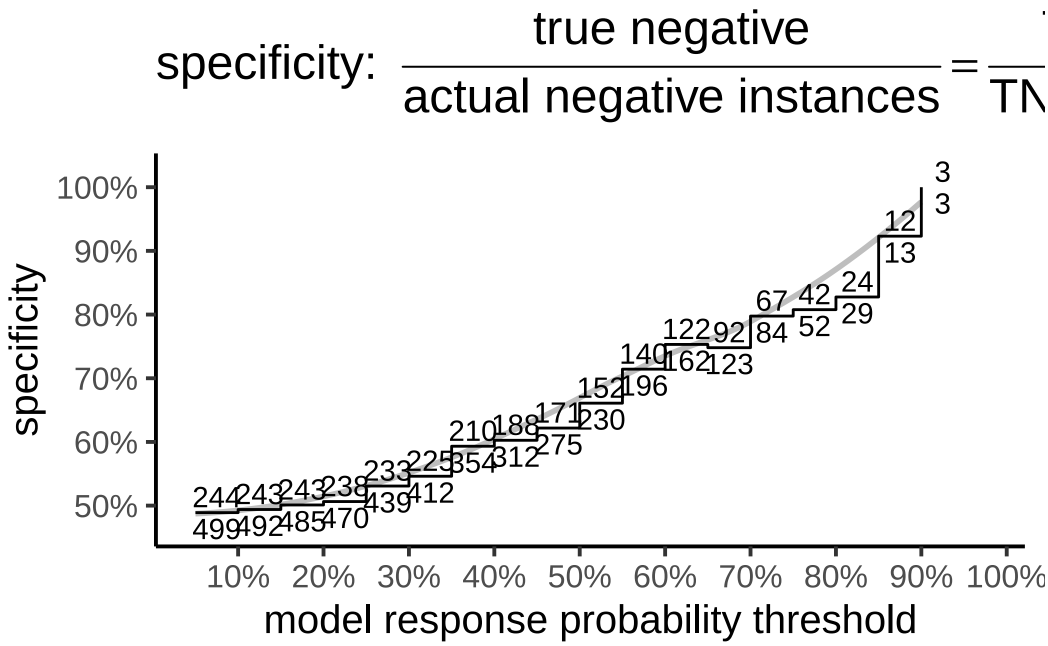
5.3.7.5 Precision



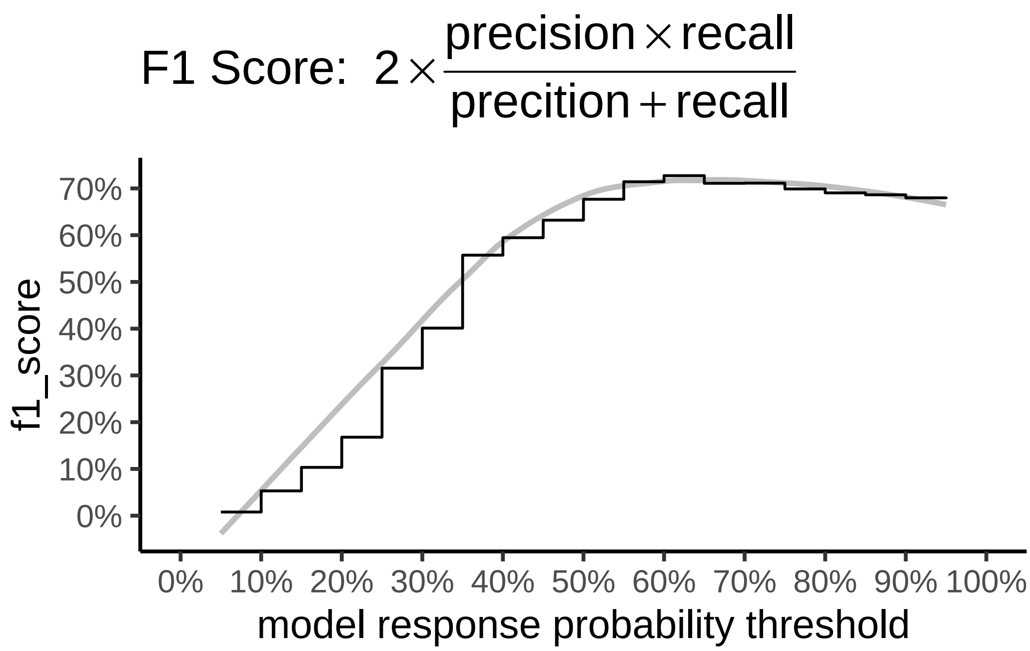
5.3.7.6 Recall, True positive rate, sensitivity, hit rate, detection rate



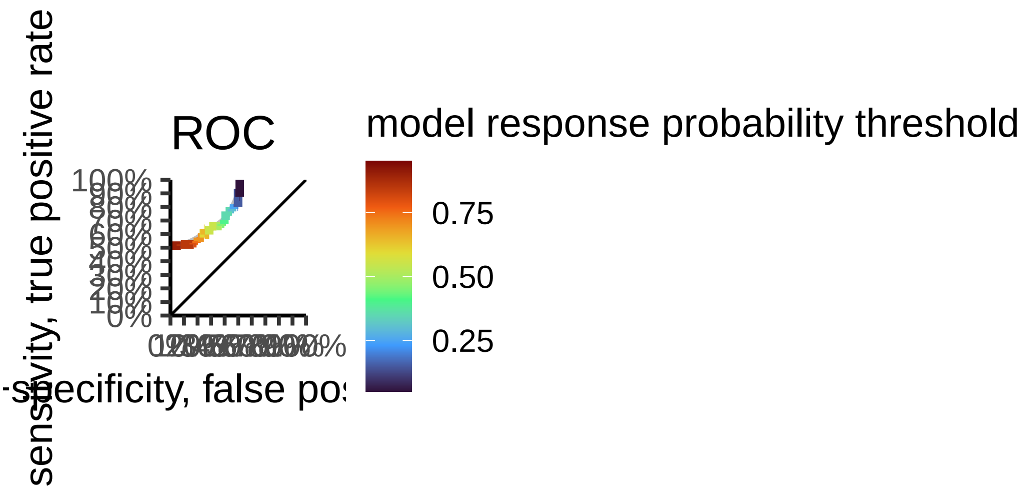
5.3.7.7 Specificity, true negative rate, selectivity, true negative fraction, 1 - false positive rate



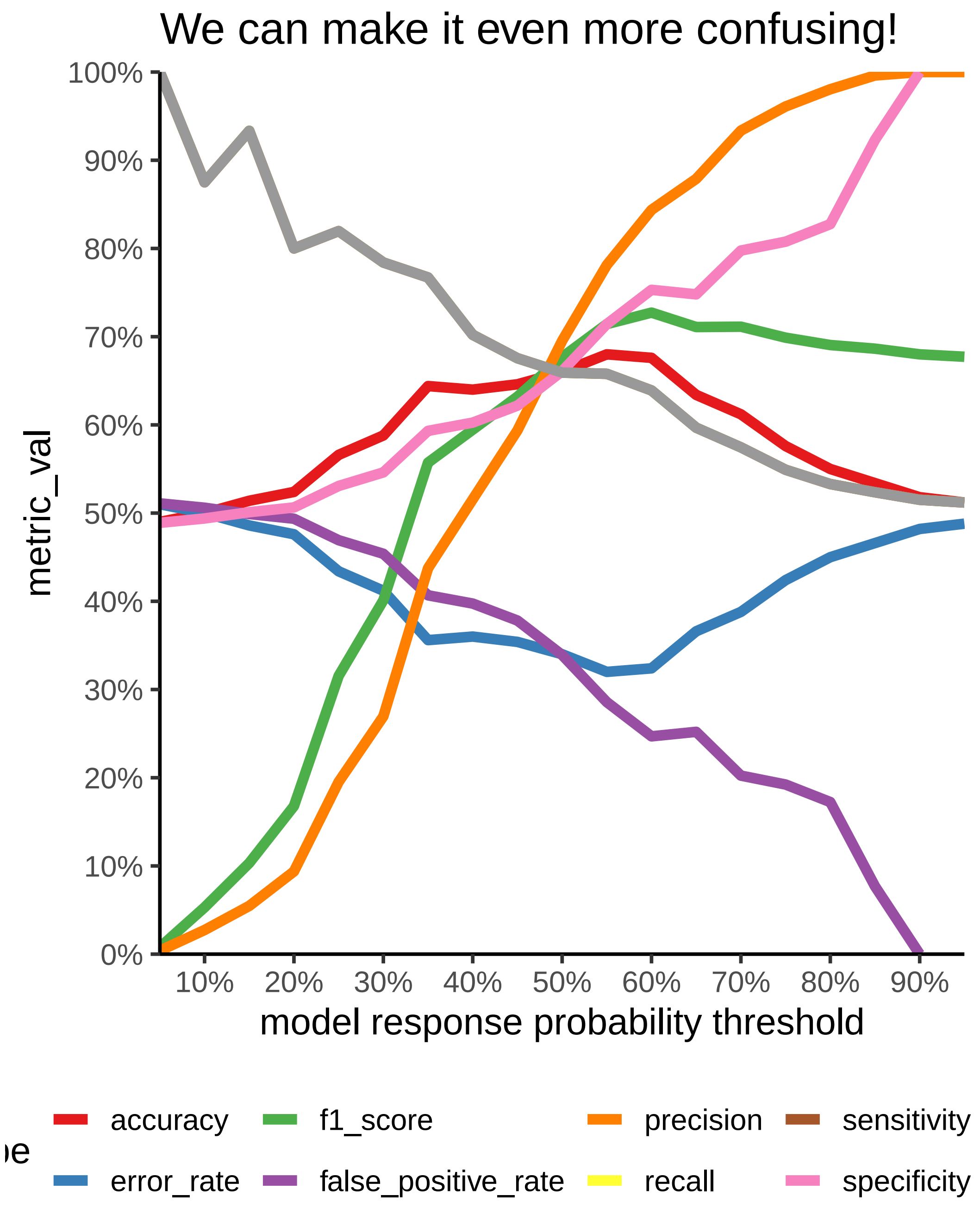
5.3.7.8 F1 Score, harmonic mean of precision and recall



5.3.7.9 Receiver Operator Curve (ROC)



5.3.7.10 METRICSSS!!!!!



6 Chose a statistical Test

One Proportion Test: Used for binary categorical data to compare a sample proportion to a known population proportion.

Chi-Square Goodness of Fit Test: Assesses whether observed categorical data frequencies match expected frequencies.

One Sample T-Test: Compares a sample mean to a known or hypothesized population mean for continuous data, assuming a normal distribution.

One Sample Wilcoxon Test: Non-parametric test for continuous data or ordinal data to compare a sample's median to a known population median.

Cochran's Q Test: Evaluates proportions in three or more related categorical groups, often with repeated measures.

Chi-Square Test of Independence: Determines if two categorical variables are associated.

Pearson Correlation: Measures linear relationships between two continuous variables, assuming normal distribution.

Spearman Correlation: Non-parametric alternative for non-linear or non-normally distributed data.

T-Test for Independent Samples: Compares means of two independent groups for continuous data, assuming normal distribution.

Welch T-Test for Independent Samples: Used when variances between two independent groups are unequal.

Mann-Whitney U Test: Non-parametric alternative for comparing two independent groups with non-normally distributed data.

T-Test for Paired Samples: Compares means of two related groups or repeated measures, assuming normal distribution.

Wilcoxon Signed Rank Test: Non-parametric alternative for paired data or non-normally distributed data.

One-Way ANOVA: Compares means of three or more independent groups for continuous data, assuming normal distribution.

Welch ANOVA: Utilized when variances between groups being compared are unequal.

Kruskal-Wallis Test: Non-parametric alternative for comparing three or more independent groups with non-normally distributed data.

Repeated Measures ANOVA: For continuous data with multiple measurements within the same subjects over time.

Friedman Test: Non-parametric alternative for analyzing non-normally distributed data with repeated measures.

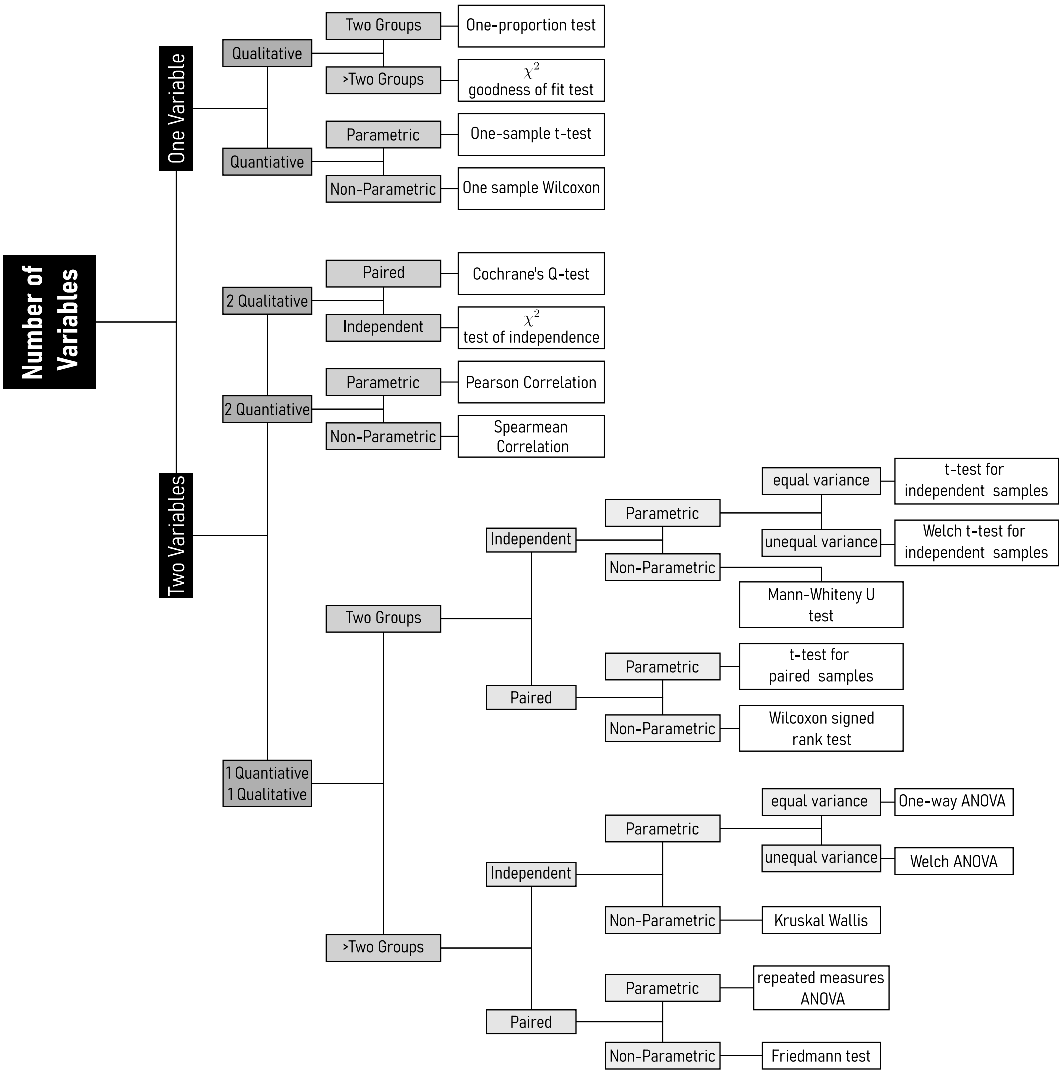


Figure 6.1: Roadmap to choosse the right test

7 Production Statistics

7.1 Introduction to Production Statistics

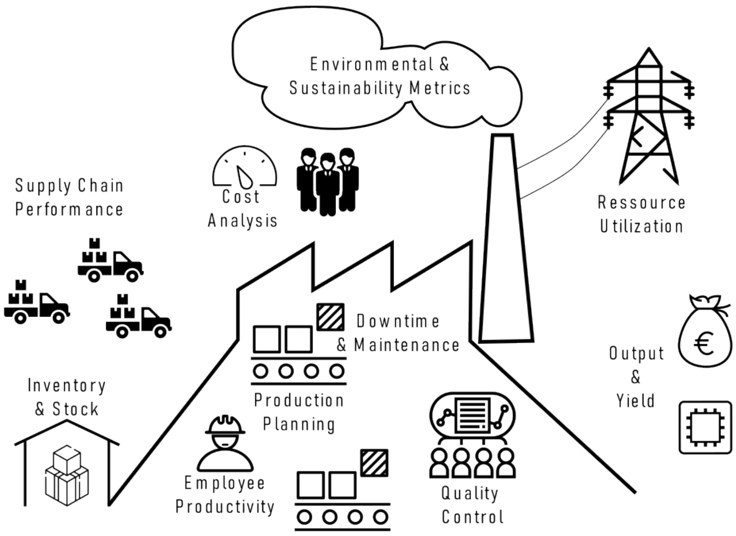


Figure 7.1: What Production Statistics tries to quanitfy.

1. **Output and Yield** statistics refer to the measurement of both the *quantity* and *quality* of products or services produced during a specific period. This includes tracking metrics such as the *number of units produced*, *yield rates*, and *defect rates*, as well as assessing production *cycle times.*

2. **Resource Utilization** statistics involve the monitoring and analysis of how efficiently resources such as *labor*, *machinery*, *materials*, and *energy* are used in production processes. Key metrics in this category include *machine uptime*, *downtime*, and overall resource *efficiency.*

3. **Quality Control** statistics play a vital role in evaluating the quality of products or services by tracking *defects*, *errors*, and *variations* in the production process. These statistics encompass *defect rates*, *reject rates*, and variation analysis to ensure products meet specified quality standards.

4. **Cost Analysis** through production statistics involves assessing the cost-effectiveness of production processes. This includes analyzing production *costs*, *overhead expenses*, and calculating the *cost per unit produced*. Such data aids in making informed decisions related to cost reduction and budgeting.

5. **Inventory and Stock** statistics pertain to the management of inventory levels and *turnover rates*. These statistics also encompass *lead times* and tracking *stock-outs*, which are crucial for efficient inventory management and ensuring product availability.

6. **Production Planning** statistics are essential for optimizing production processes. Metrics include *capacity utilization*, *order fulfillment rates*, and production *lead times*. This data assists in scheduling and ensuring the efficient use of resources.

7. **Downtime and Maintenance** statistics track equipment *breakdowns*, *maintenance schedules*, and production *interruptions*. Monitoring such data is vital for minimizing production downtime and ensuring equipment operates efficiently.

8. **Employee Productivity** statistics evaluate workforce performance and efficiency. Metrics such as *output per worker* and *labor efficiency* are used to assess employee contributions and identify areas for improvement, including *training needs*.

9. **Supply Chain Performance** statistics extend beyond production to evaluate the entire supply chain, including suppliers, logistics, and distribution. Metrics like *lead times*, *order fulfillment rates*, and supplier performance data help ensure the efficiency of the supply chain.

10. **Environmental and Sustainability Metrics** encompass *resource consumption*, *waste generation*, and *environmental impact*. They are used to assess an organization's environmental footprint and implement sustainable practices.

7.2 Control Charts for Variables

7.2.1 The production

In Figure 7.2 the drive shaft production and the behaviour of the mission critical parameter `diameter` is shown over time.

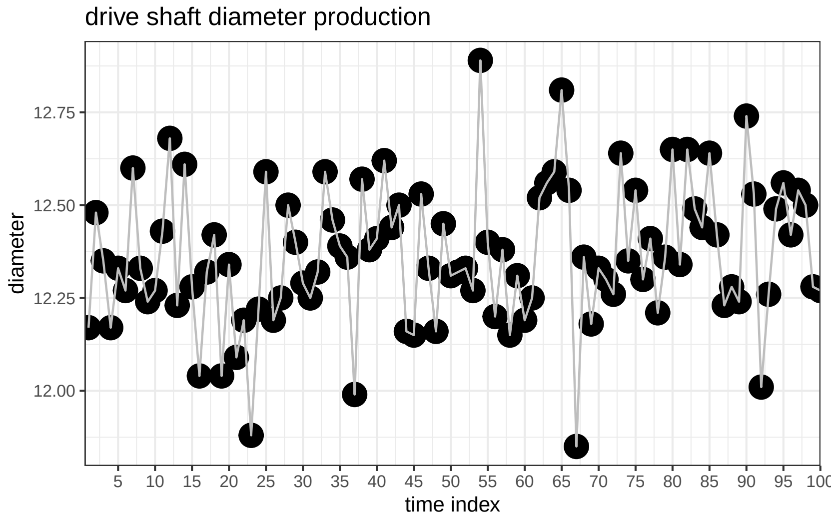


Figure 7.2: The drive shaft production over time

7.2.2 Run Chart

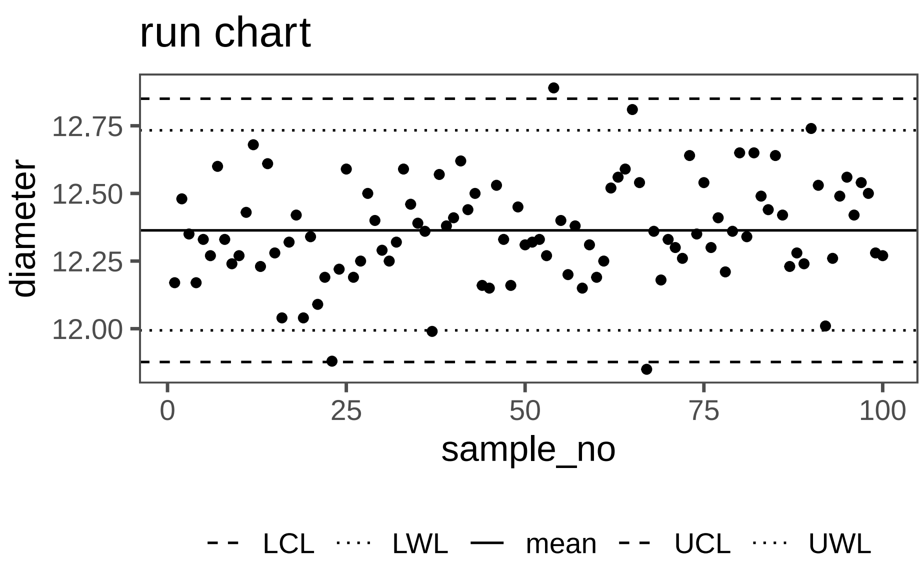


Figure 7.3: A run chart with control and warning limits without subgroups.

$$UCL = \bar{x} + 2.58\frac{sd(x)}{\sqrt{n}} \text{ with } n = 1 \tag{7.1}$$

$$LCL = \bar{x} - 2.58\frac{sd(x)}{\sqrt{n}} \text{ with } n = 1 \tag{7.2}$$

$$UWL = \bar{x} + 1.96\frac{sd(x)}{\sqrt{n}} \text{ with } n = 1 \tag{7.3}$$

$$LWL = \bar{x} - 1.96\frac{sd(x)}{\sqrt{n}} \text{ with } n = 1 \tag{7.4}$$

In Shewhart (Shewhart and Deming 1986) charts for statistical process control, control limits such as the Upper Control Limit (UCL), Lower Control Limit (LCL), Upper Warning Limit (UWL), and Lower Warning Limit (LWL) play a crucial role. These limits establish boundaries for normal process variability. By incorporating confidence intervals, such as 97% or 99%, into these limits, a statistical framework is added, providing a nuanced understanding of process variability. A 97% confidence interval implies that 97% of data points should fall within the calculated range, while a 99% interval accommodates 99%. This approach enhances the sensitivity of Shewhart charts, aiding in the timely detection of significant process shifts. The choice of confidence level depends on the desired balance between false alarms and the risk of missing genuine deviations from the norm.

7.2.3 X-bar chart

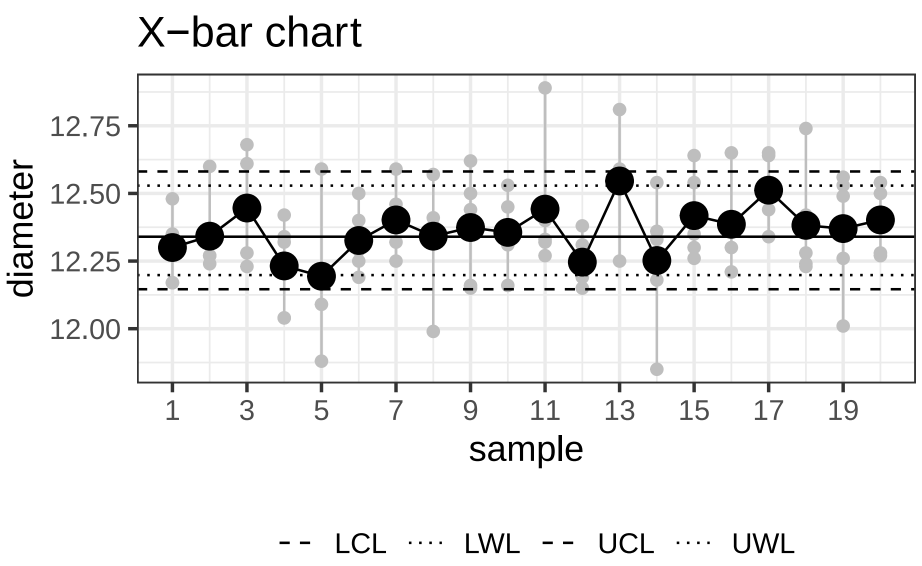


Figure 7.4: A X-bar chart with control and warning limits based on subgroups of $n = 5$

$$UCL = \bar{x} + 2.58\frac{sd(x)}{\sqrt{n}} \text{ with } n = 5 \tag{7.5}$$

$$LCL = \bar{x} - 2.58\frac{sd(x)}{\sqrt{n}} \text{ with } n = 5 \tag{7.6}$$

$$UWL = \bar{x} + 1.96\frac{sd(x)}{\sqrt{n}} \text{ with } n = 5 \tag{7.7}$$

$$LWL = \bar{x} - 1.96\frac{sd(x)}{\sqrt{n}} \text{ with } n = 5 \tag{7.8}$$

An X-bar chart is a statistical tool for quality control, used to monitor process stability over time. It involves collecting data, calculating subgroup means, determining control limits, and plotting the data on a chart. By monitoring points relative to the control limits, it helps identify shifts in the process mean, allowing corrective action for consistent quality.

It is effective in quality control because it focuses on detecting changes in the process mean. By setting statistical control limits, it distinguishes between common and special causes of variation. When data points fall outside these limits, it signals the presence of external factors, prompting corrective action. The chart's visual representation of data points over time facilitates early issue detection, supporting a proactive approach to maintaining process stability and continuous improvement in quality control.

7.2.4 S-Chart

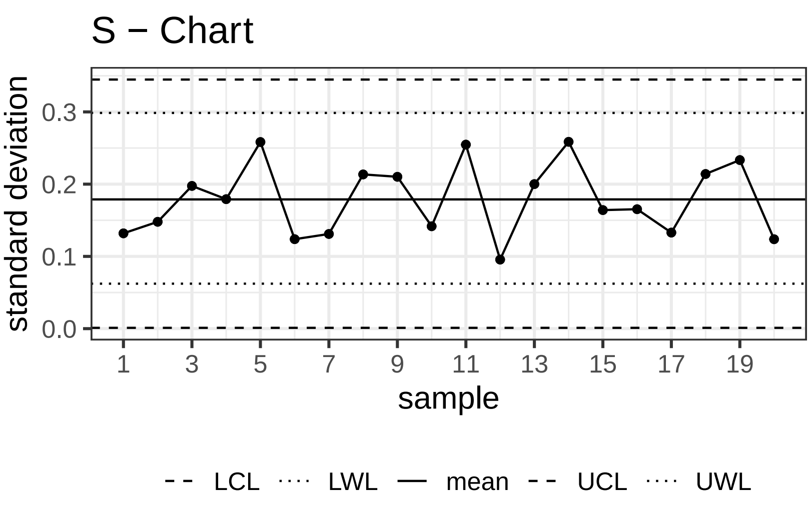


Figure 7.5: The s chart with control and warning limits.

$$UCL = \sigma * \sqrt{\frac{\chi^2_{1-\beta=0.995;n-1}}{n-1}} \text{ with } n = 5 \tag{7.9}$$

$$LCL = \sigma * \sqrt{\frac{\chi^2_{1-\beta=0.005;n-1}}{n-1}} \text{ with } n = 5 \tag{7.10}$$

$$UWL = \sigma * \sqrt{\frac{\chi^2_{1-\beta=0.975;n-1}}{n-1}} \text{ with } n = 5 \tag{7.11}$$

$$LWL = \sigma * \sqrt{\frac{\chi^2_{1-\beta=0.025;n-1}}{n-1}} \text{ with } n = 5 \tag{7.12}$$

An S chart, or standard deviation chart, is a type of control chart used in statistical process control. It is designed to monitor the variability or dispersion of a process over time. The S chart displays the sample standard deviation of a process by plotting it against time or the sequence of samples. Similar to other control charts, it typically includes a central line representing the average standard deviation and upper and lower control limits. The S chart is useful for detecting shifts or trends in the variability of a process, allowing for timely adjustments or interventions if needed.

7.3 Control Charts for Attributes

7.3.1 NP Chart

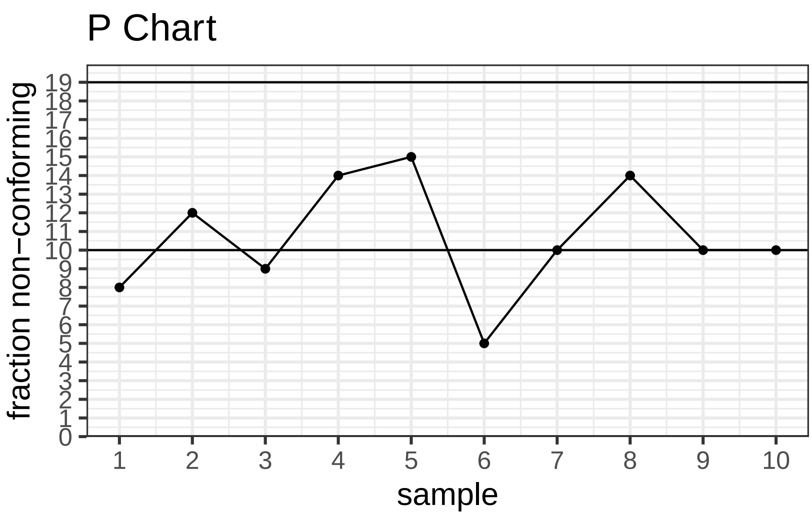


Figure 7.6: A NP-Chart with control limits.

$$CL = n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})} \tag{7.13}$$

An NP chart, also known as a Number of Defects Per Unit chart, is a statistical tool used in quality control to monitor the number of defects or errors in a process over time. It is commonly employed in manufacturing and other industries to assess the stability and performance of a production process. The chart typically displays the number of defects observed in a sample of units or products, allowing for the identification of trends, patterns, or variations in the defect rates. This information aids in quality improvement efforts by enabling organizations to take corrective actions and maintain consistent product or service quality.

7.3.2 P Chart

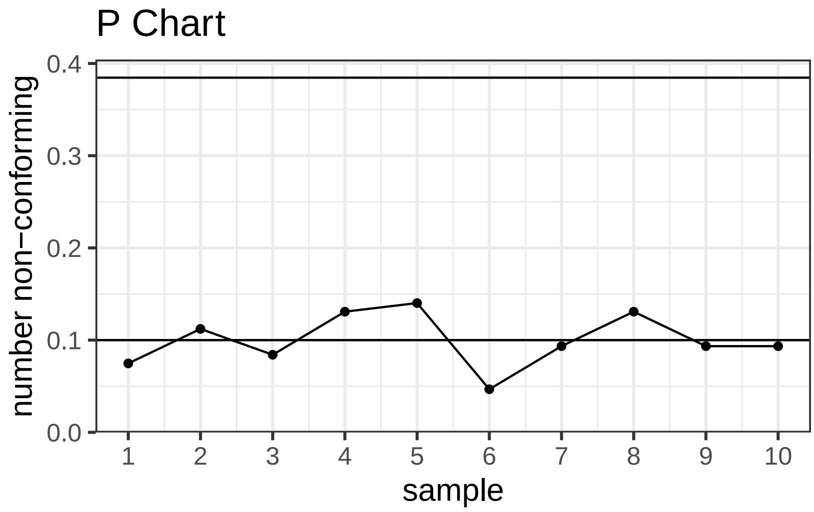


Figure 7.7: A P-Chart with control limits.

$$CL = \bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{7.14}$$

The P chart is designed to track the proportion of nonconforming items or defects within a sample or subgroup over consecutive periods. The chart typically consists of a horizontal axis representing time periods and a vertical axis representing the proportion of nonconforming items. It helps identify variations and trends in the process, allowing for timely corrective actions when necessary.

P charts are commonly used in industries where the output is binary, such as the presence or absence of a specific attribute, and provide a visual representation of the process's performance, aiding in quality improvement efforts.

7.4 Process Capability and Six Sigma

7.4.1 How good is good enough?

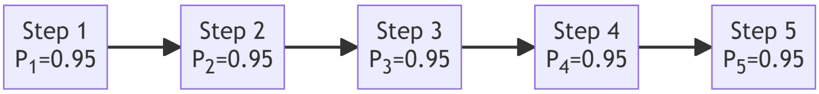


Figure 7.8: What are the joint probabilities?

A success rate of 95% per step (Figure 7.8) sounds at first glance like a successful process. After all, having a 95% chance of winning the lottery would be awesome. Yet, the question is: What are the joint probabilities when we connect five steps sequentially? From previous chapters we know that the joint probability can be calculated in (7.15).

$$P_{ges} = P_1 * P_2 * P_3 * P_4 * P_5 = 0.95^{(n=5)} = 0.774 \approx 77.4\% \tag{7.15}$$

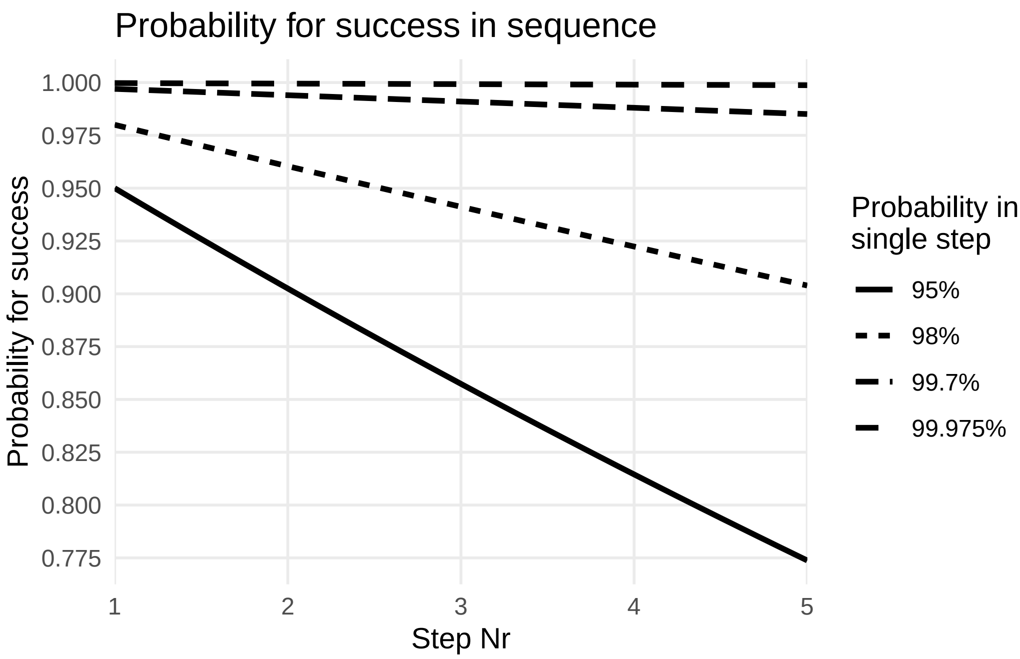


Figure 7.9: Probabilities for success in sequence.

The joint probability for n-steps in sequence can therefore be estimated using (7.15) and visually represented in Figure 7.9. On the x-axis the number of steps is depicted whereas on the y-axis the joint probability is shown for the respective step index. As also calculated in (7.15) after $n = 5$ steps the joint probability for a good part drops to around 77%, which is not acceptable. Figure 7.9 shows that not even 98% probability for a good part for a single step results in an acceptable joint probability ($P = 0.98^{n=5} = 0.904$). A staggering probability of 99.7% for a single step is necessary to still reach a probability for a good part of 98%, and this is only true for $n = 5$steps. For an acceptable parts per milltion (ppm) rate the acceptable single step probability is 99.975% as shown in Figure 7.9.

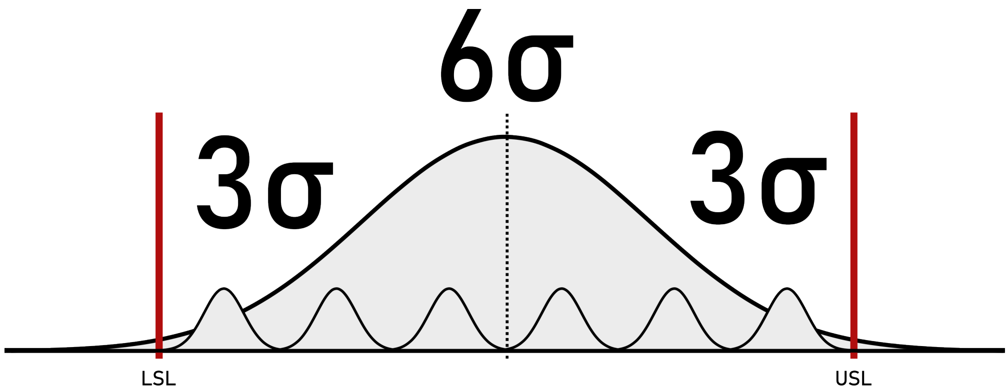


Figure 7.10: The origin of the term Six Sigma (6σ)

What that means in a tolerance-specification setting is shown in Figure 7.10. In order to ensure a 99.975% for a continuous variable, the process variation (here measured as process standard deviation) must fit **at least 6** times into the actual tolerance/specification window of the Critical to Quality (CTQ) measure. Additionally, this is only true if the process is *centered*. The term 6σ carries this inherent property for a $0ppm$ production, which is favoured by many, but achieved by few.

7.4.2 The Six Sigma Project Model (DMAIC)

The Six Sigma Project model consists of five phases in total: (D)efine, (M)easure, (A)nalyse, (I)mprove, (C)ontrol. In essence these project phases are the application of the scientific method, but in a systematic and industry friendly way.

The *Define* Phase involves setting the project's goals and objectives, identifying key stakeholders, developing a high-level process map, and defining customer requirements and critical-to-quality (CTQ) characteristics. Additionally, the project scope is established, and a project charter is developed to guide the overall initiative.

In the *Measure* Phase, key process metrics are identified, and relevant data is collected to assess the current state of the process. This phase includes analyzing process capabil-

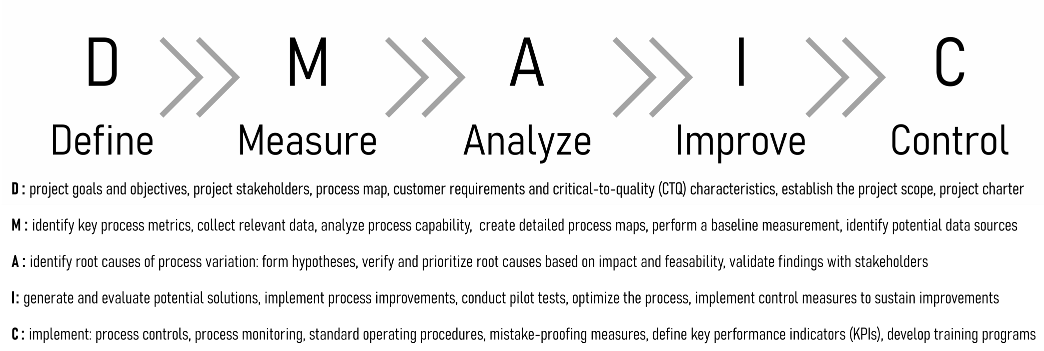


Figure 7.11: DMAIC Process

ity, creating detailed process maps, performing baseline measurements, and identifying potential data sources to ensure comprehensive data collection.

During the *Analyze* Phase, potential root causes of process variation are identified through data analysis using statistical tools. Hypotheses for root causes are developed and verified through further data analysis. Root causes are then prioritized based on their impact and feasibility, and findings are validated with stakeholders to ensure accuracy and relevance.

The *Improve* Phase focuses on generating and evaluating potential solutions for process improvement. Implementing these improvements involves developing an implementation plan, conducting pilot tests if applicable, and optimizing the process based on feedback. Control measures are implemented to sustain the improvements achieved.

Finally, the *Control* Phase involves developing control plans to monitor process performance continuously. This includes establishing process controls and standard operating procedures, implementing mistake-proofing measures, and defining key performance indicators (KPIs). Additionally, training programs for process stakeholders are developed, and a system for ongoing monitoring and feedback is established to ensure the process remains effective over time.

7.4.3 Process Capability - idea

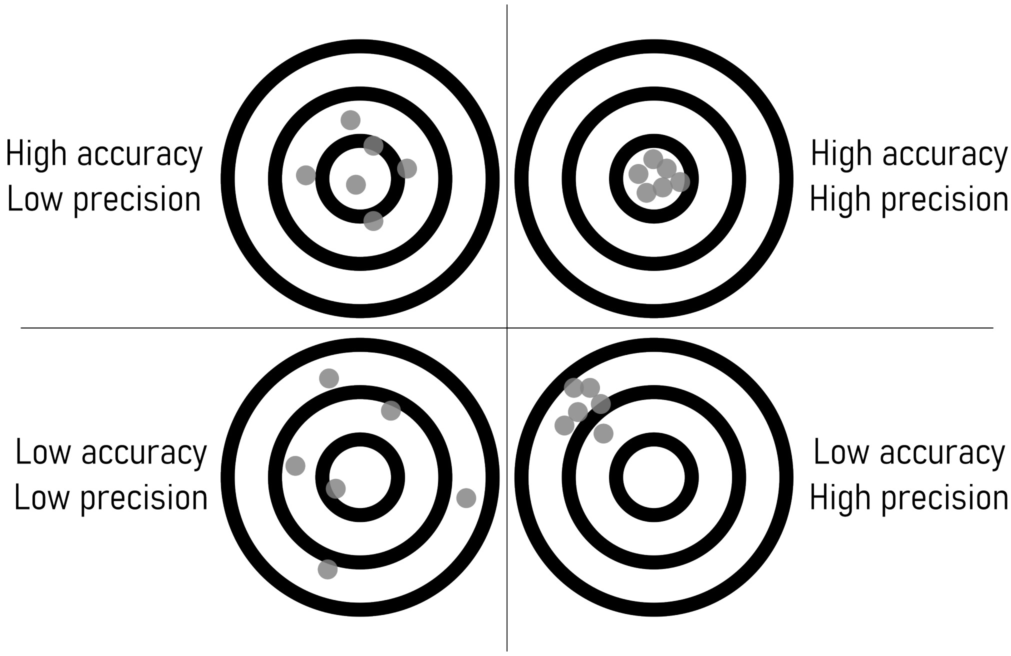


Figure 7.12: The idea of process capabilities

Process capability refers to the ability of a process to consistently produce outputs that meet predetermined specifications or requirements. It is a measure of how well a process performs relative to its specifications. The general idea behind process capability is to assess the inherent variability of a process and determine whether it is capable of producing products or services within the desired quality limits.

1. **Specification Limits**: These are the predetermined limits or requirements for a particular process output, defining the range within which the product or service should fall to meet customer expectations.
2. **Process Variation**: This refers to the natural variability inherent in the process. Sources of variation can include factors such as machine performance, material properties, human factors, and environmental conditions.
3. **Process Capability Indices**: These are statistical measures used to quantify the relationship between process variation and specification limits. Common indices include C_p, C_{pk}, P_p, and P_{pk}, which provide insights into whether a process is capable of meeting specifications and how well it is centered within the specification limits.
4. **Assessment and Improvement**: Once process capability is assessed, steps can be taken to improve it if necessary. This may involve reducing process variation,

adjusting process parameters, implementing quality control measures, or redesigning the process altogether.

Overall, the goal of analyzing process capability is to ensure that processes are capable of consistently delivering products or services that meet customer requirements, minimize defects, and optimize quality and efficiency.

7.4.4 High Accuracy - Low Precision

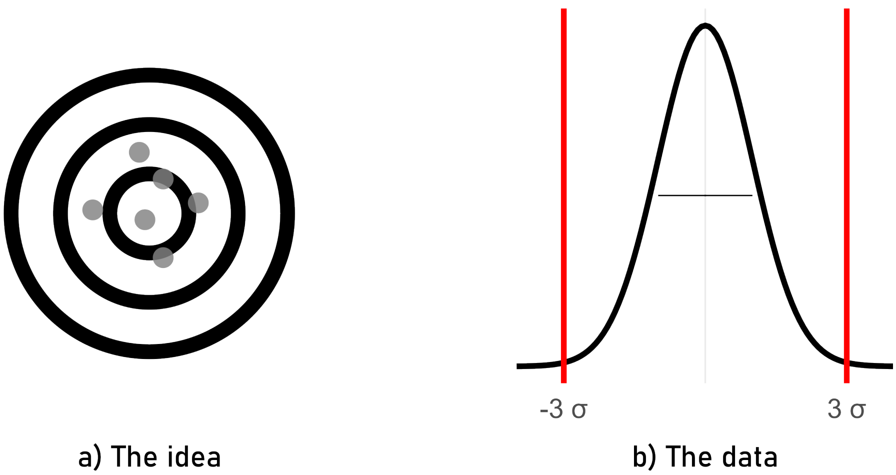


Figure 7.13: The spreaded - High Accuracy, Low Precision

In this scenario, the process consistently produces results that are very close to the target or desired value (high accuracy). However, the variation among individual measurements is large, meaning they are not tightly clustered around the target value (low precision). For example, if a machine consistently produces parts with dimensions close to the desired specifications but with significant variation in each part's dimensions, it exhibits high accuracy but low precision.

7.4.5 Low Accuracy - Low Precision

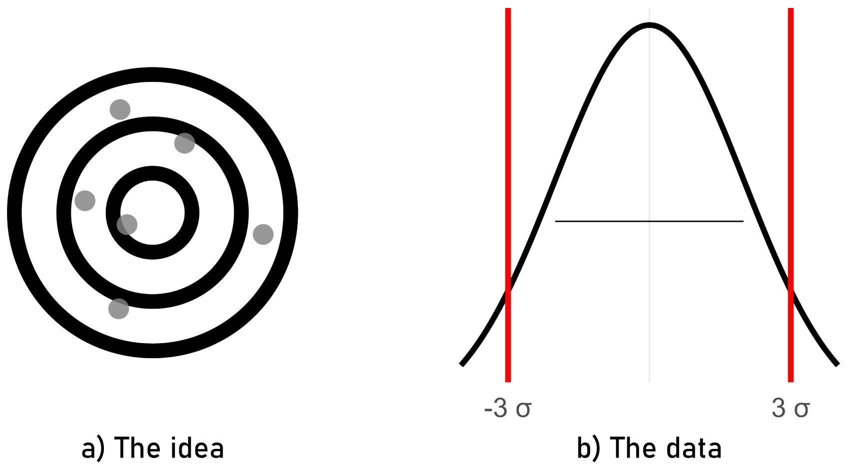


Figure 7.14: The worst - Low Accuracy, Low Precision

Here, the process consistently produces results that are far from the target or desired value (low accuracy). Additionally, the variation among individual measurements is large, indicating low precision. An example could be a manufacturing process that consistently produces parts with dimensions that are both far from the desired specifications and vary significantly from one part to another.

7.4.6 Low Accuracy - High Precision

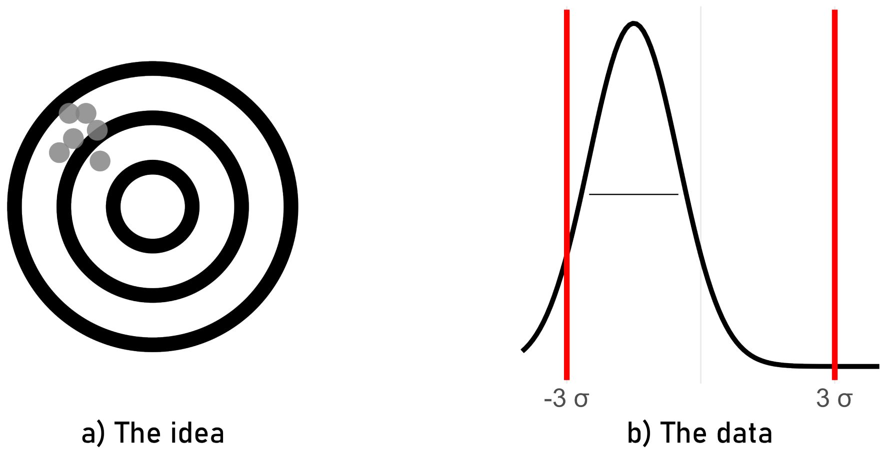


Figure 7.15: The missing the mark - Low Accuracy, High Precision

This scenario involves a process that consistently produces results that are tightly clustered around a single point, but that point is far from the target or desired value (low accuracy). For instance, if a weighing scale consistently displays a weight that is slightly off from the true weight but shows very little variation between repeated measurements, it demonstrates low accuracy but high precision.

7.4.7 High Accuracy - High Precision

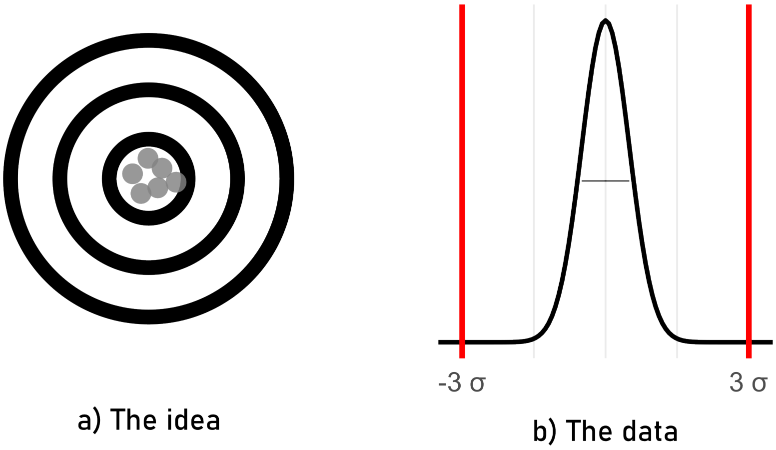


Figure 7.16: The desired - High Accuracy, High Precision

This is the ideal scenario where the process consistently produces results that are both very close to the target or desired value (high accuracy) and tightly clustered around that value (high precision). For example, a manufacturing process that consistently produces parts with dimensions very close to the desired specifications and with minimal variation between individual parts exhibits both high accuracy and high precision.

7.4.8 Computing Process Capabilities

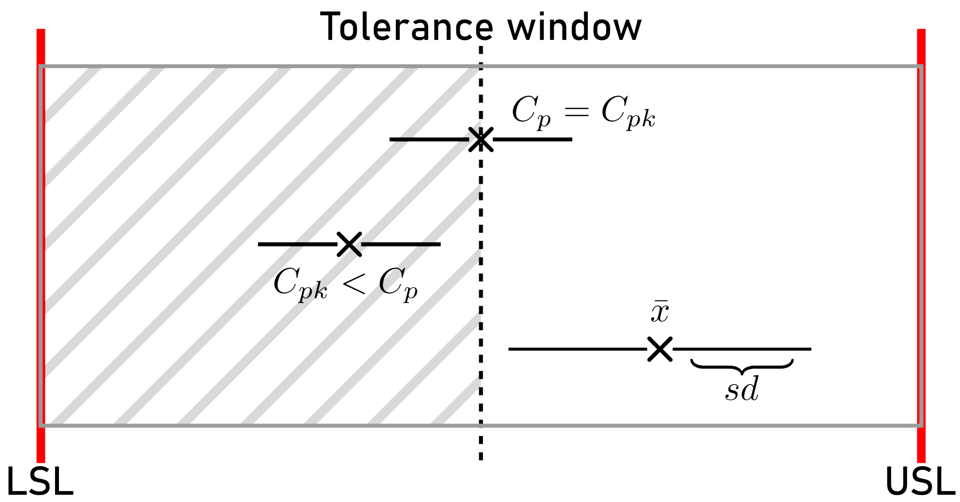


Figure 7.17: The idea to calculate the C_{pk}

$$C_p = \frac{USL - LSL}{6 * sd} \tag{7.16}$$

$$C_{pk} = \frac{\min(USL - \bar{x}, \bar{x} - LSL)}{3 * sd} \tag{7.17}$$

C_p compares the spread of the process variation to the width of the specification limits (7.16). A C_p value greater than 1 indicates that the process spread fits within the specification limits, suggesting that the process has the *potential* to meet specifications. However, C_p does not take into account the process mean, so it does *not* provide information about process centering. For a more comprehensive assessment of process capability, both C_p and C_{pk} are often used together.

The C_{pk} value indicates the capability of the process relative to the specified limits (7.17). A C_{pk} value greater than 1 indicates that the process spread (6 standard deviations) fits within the specification limits. A value less than 1 indicates that the process spread exceeds the specification limits, indicating potential issues with meeting specifications. A higher C_{pk} value indicates better process capability.

7.4.9 Process Capabilities and ppm

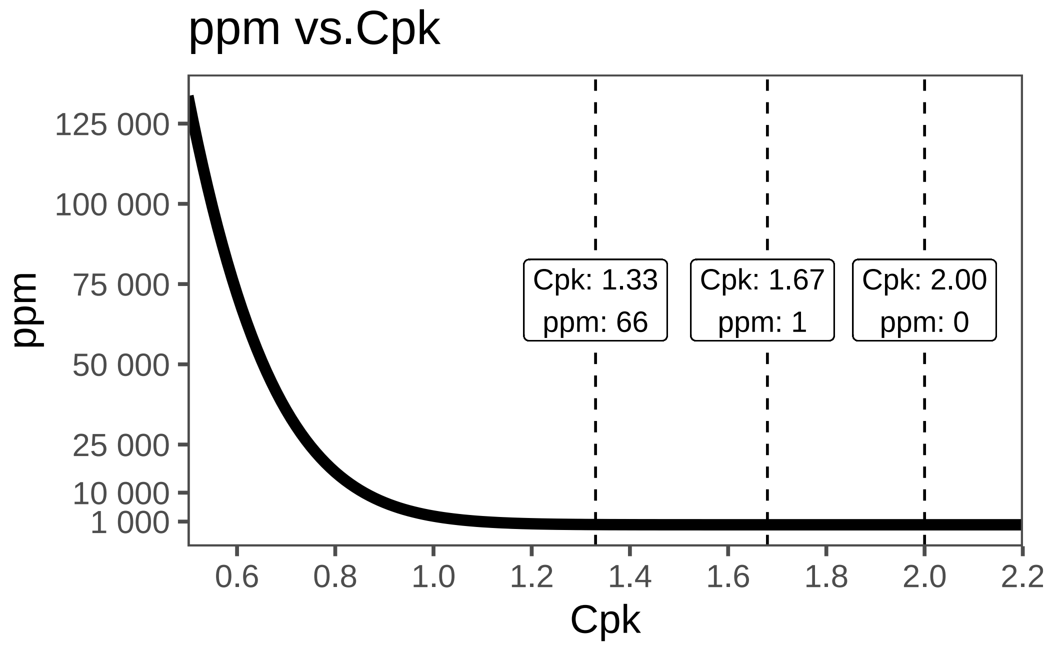


Figure 7.18: The failed parts per million vs. the C_{pk}

Process capability and parts per million (PPM) are closely related metrics used to assess the performance of manufacturing processes. They provide a statistical measure of how well a process can produce output within specified limits. PPM is a measure of the number of defective parts per million produced by the process. The connection between process capability indices and PPM can be understood through statistical distributions, primarily the normal distribution, and the concept of defects or non-conformance.

The connection between process capability indices and PPM can be established through the Z-score (Z-standardization), which translates process capability into the probability of defects.

1. Using C_p: Assuming the process is centered and follows a normal distribution: $Z = 3C_p$. The corresponding PPM can be found from standard normal distribution tables. For example, if $C_p = 1$, then $Z = 3$, and the area under the normal curve beyond 3 standard deviations on either side is approximately 0.0027, or $2700PPM$.

2. Using C_{pk} C_{pk} directly relates to the Z-score: $Z = 3C_{pk}$. The PPM can be calculated using the cumulative distribution function for the normal distribution. For example, if $C_{pk} = 1.33$, then $Z = 3 \times 1.33 = 3.99$. Using standard normal distribution tables, the area beyond $Z = 3.99$ is approximately 0.000066, or $66ppm$.

7.5 The role of measurement accuracy in production

7.5.1 Measurement Errors

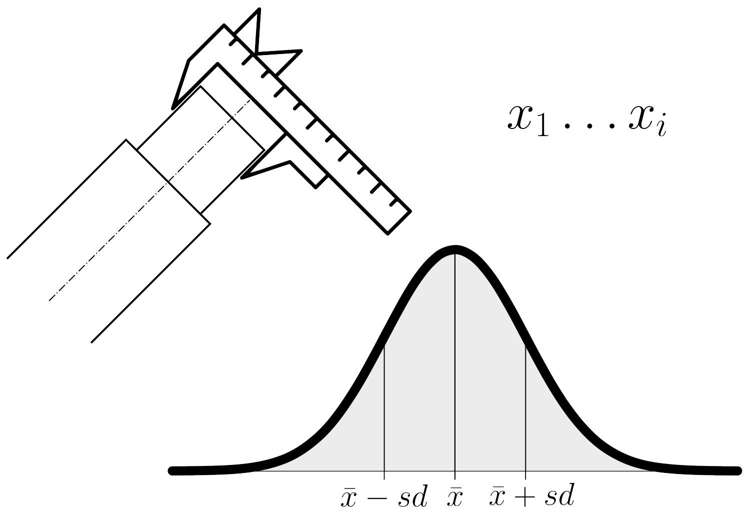


Figure 7.19: Measurement Errors arise during every measurement.

In scientific experiments and real-world measurements, there are often inherent sources of random error (Nuzzo 2014). These errors can introduce variability into measurements, and the accumulation of these errors often conforms to a normal distribution. For instance, when measuring the diameter of an object with a caliper, small measurement errors can cause the observed values to follow a normal distribution. Even during such a simple measurement some random errors may include:

1. Parallax Error: Parallax can introduce random errors if the observer's eye is not consistently aligned with the scale or graduations during measurements.
2. Dirt or Debris: Foreign particles or debris on the measuring surfaces can lead to random measurement errors by causing slight variations in the contact points between the caliper and the object.
3. Jaw Alignment: Small variations in the alignment of the caliper jaws from one measurement to another can introduce random errors in measurements.
4. Material Deformation: When measuring soft or deformable materials, random errors can occur due to variations in the material's response to pressure during different measurements.
5. Human Error: Random errors can arise from misreading the scale or not positioning the caliper precisely on the object, especially if different operators are involved.

6. Slop or Play in the Jaws: Variability in the amount of play or slop in the caliper's jaws from one measurement to another can lead to random errors in measurements.

7.5.2 Significant Digits in Production

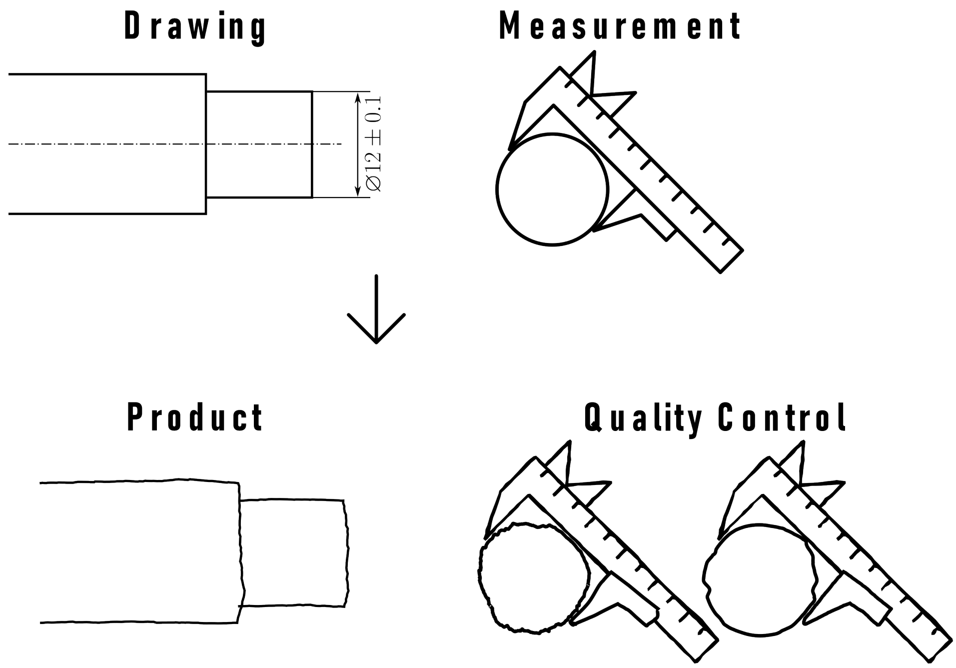


Figure 7.20: Drawings and specifications are just an approximation of reality.

Significant digits, or significant figures, are vital for precision and quality in production. They ensure precision, quality, and consistency in production, leading to better efficiency and customer satisfaction. Significant digits indicate the precision of measurements, ensuring products meet quality standards and specifications.

Applications:

1. Quality Control: Accurate measurements ensure consistent product quality.
2. Tolerances: Precise tolerances (e.g., $\pm 0.05mm$) must be adhered to.
3. Fit and Interchangeability: Parts must fit together correctly, requiring precise measurements.
4. Calibration: Instruments must match the required significant digits for accuracy.
5. Documentation: Accurate recording of measurements is essential for quality reports and compliance.
6. Training: Employees must understand and apply significant digits to maintain standards.

Best Practices:

- Reduce Human Error: Training and audits are essential.
- Use Proper Instruments: Ensure tools can measure accurately.
- Control Environment: Manage factors like temperature and humidity.
- Follow Rounding Rules: Apply proper rounding to maintain precision.

7.5.2.1 General Rule of Thumb

To maintain accuracy and avoid overestimating the precision of results, it's advisable not to report more significant digits than justified by the precision of the input measurements.

7.5.2.2 Rule of Ten

In practical terms, for a number to be considered significant, it should be at least ten times greater than the smallest unit of measure (i.e., the least significant digit). This helps in avoiding overestimating the precision and ensures that the reported figures are meaningful.

7.5.2.3 Addition and Subtraction

When performing addition or subtraction, the result should be reported with the same number of decimal places as the measurement with the fewest decimal places. For instance, if you add 12.11 (two decimal places) to 0.4 (one decimal place), the result should be reported with one decimal place, as 12.5.

7.5.2.4 Multiplication and Division

When performing multiplication or division, the result should be reported with the same number of significant digits as the measurement with the fewest significant digits. For example, if you multiply 2.34 (three significant digits) by \$0.0\$5 (one significant digit), the result should be reported with one significant digit, as 0.1.

7.5.2.5 edge cases

Significant digits can help with edge cases that naturally occur during measurement processes. As depicted in Figure 7.21, the first two measurements are well within specification. The third measurement can actually not be interpreted, as the measurement instrument seems not to be fit for purpose. The fourth measurement shows, that the product is within the specification, it always holds the number with the smallest number

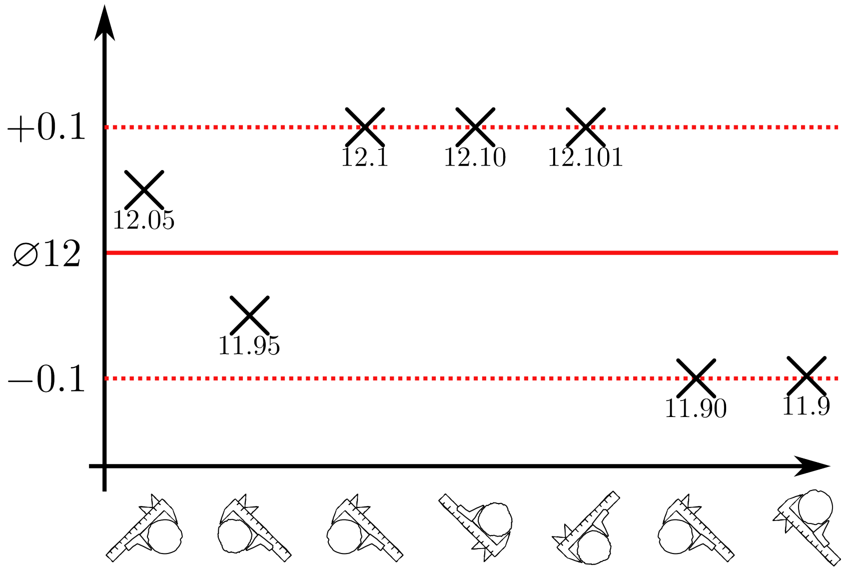


Figure 7.21: Edge cases during measuring a simple part.

of digits. The measurement of the fifth product is just within specification, the gage that shoed the last reading is not accurate enough.

There are many rules involved in these kind of edge cases including the rounding of number. It is referred to (Standards, (U.S.), and SEMATECH. 2002) or the national standards for more elaborate discussions about this manner.

7.5.3 Measurement System Analysis Type I

In conducting a Measurement System Analysis Type I (MSA1), the initial step involves focusing on gage as the sole source of variation. To achieve this, 50 measurements are performed, each repeated on a reference part. This process allows for the isolation and assessment of the gage's impact on the overall measurement system, ensuring that any observed variability is attributed solely to the gage. The process of doing a MSA1 is fairly standardized.

7.5.3.1 Potential Capability index C_g

From a MSA1 the potential Measurement System Capability Index C_g can be computed via (7.18).

$$C_g = \frac{K/100 * Tol}{L * \sigma} \tag{7.18}$$

Tol Tolerance
C_g Capability Gage
K percentage of the tolerance (20%)
σ standard deviations of the tolerance
L number of standard deviations that represent the process (6×)

7.5.3.1.1 Capability index with systematic error C_{gk}

Very similar to the process capability, a C_g gives only the *potential* capability as it does not include if the measures are centered around a mean. This is overcome by computing the Measurement Capability Index with systematic error C_{gk}, which incorporates the mean via (7.19).

$$C_{gk} = \frac{(0.5 * K/100) * Tol - |\bar{x} - x_{true}|}{3 * \sigma} \tag{7.19}$$

Tol Tolerance
$\bar{x}$ mean of the measurements
K percentage of the tolerance (20%)
x_{true} the "true" value of the reference (calibration)
σ standard deviation of the measurements

7.5.3.2 MSA1 example

Table 7.1: The summary of the raw data for the MSA1.

Characteristic	**N = 50**[1]
measured_data	20.303 (0.005)

[1]Mean (SD)

In Table 7.1 the raw data that was collected during the experiments is depicted, whereas in Figure 7.22 the same data is shown in graphical format.

On the `x-axis` the measurement index is shown, the `y-axis` shows the measurment value. One of the main advantages of a MSA1 is, that a reference value is known, because the values are taken agains a standard reference normal. This true value (`x_true` in Figure 7.22, dashed black line) allows the estimation of a systematic error. The 20% tolerance (7.19) is shown as dashed green line. This is the reduced tolerance in which the gage shall be capable to produce good measurement values.

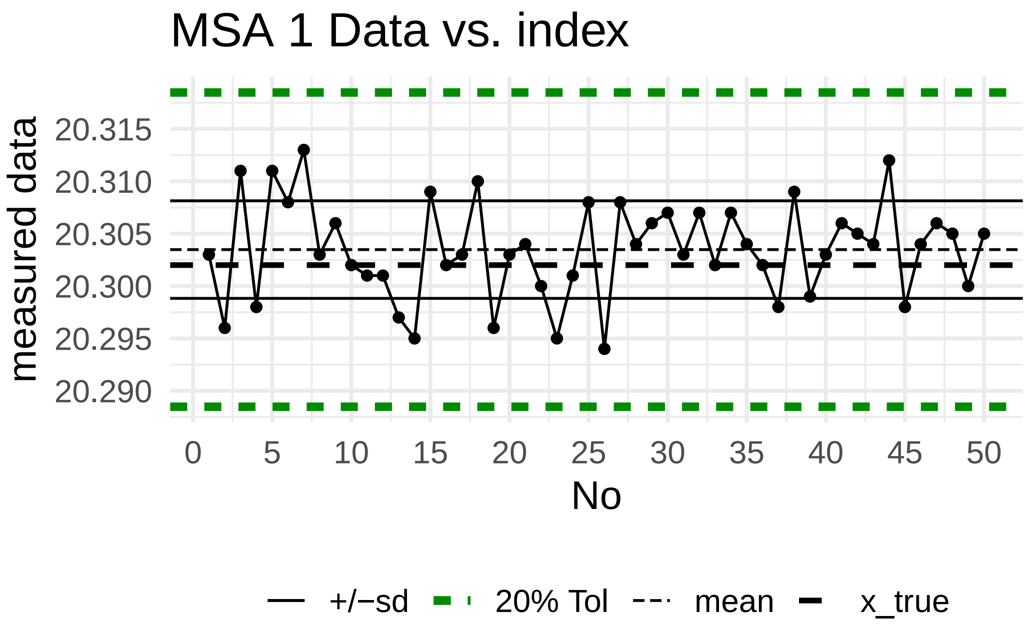


Figure 7.22: The data as measured during the MSA1 with all measures included.

7.5.3.2.1 Data Distribution

Measurement errors are often assumed to be normally distributed due to the CLT and the nature of random processes involved. The CLT states that the sum of many independent, random variables tends to follow a normal distribution, even if the original variables are not normally distributed. Measurement errors typically result from the combination of numerous small, independent errors, such as instrument precision, environmental factors, and human mistakes. This aggregation leads to a normal distribution of the overall errors.

Additionally, many error sources are random and independent, further supporting the normal distribution assumption. The normal distribution is mathematically convenient, being fully described by its mean and variance, which simplifies statistical analysis and hypothesis testing. Empirical evidence across various fields also shows that measurement errors often approximate a normal distribution.

While the normal distribution is a useful assumption, it may not always be valid. In cases with asymmetric errors, heavy tails, or significant outliers, other distributions may be more appropriate. Nonetheless, for many practical purposes, assuming a normal distribution for measurement errors is reasonable and effective.

7.5.3.2.2 computed values

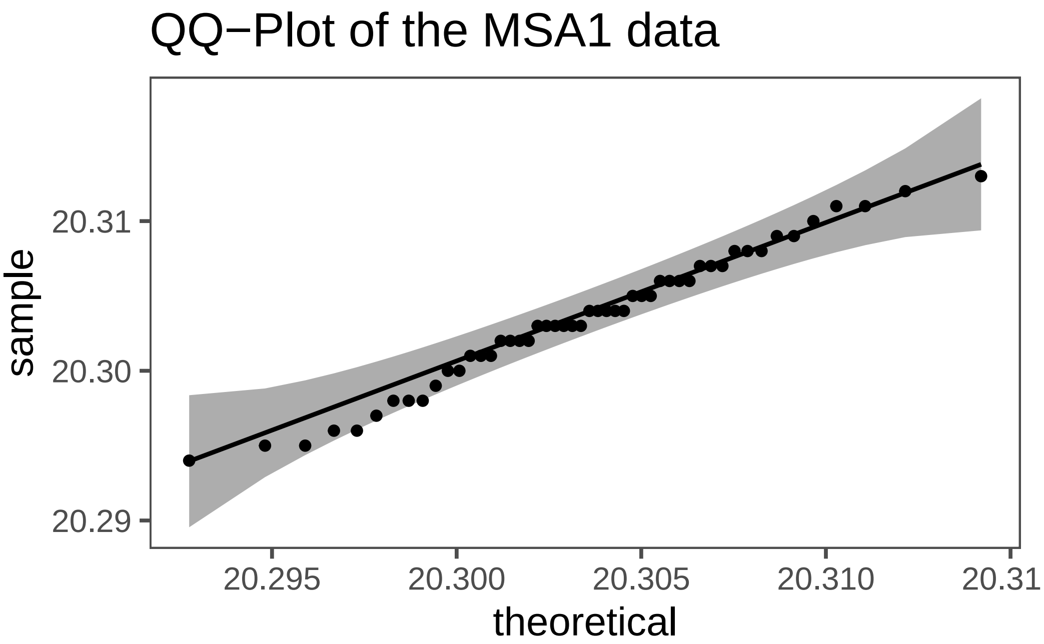


Figure 7.23: By definition, measurement errors shoul be normally disitrbuted.

Table 7.2: C_g, C_{gk} for the measured values

Cg	Cgk
2.13	2.02

In Table 7.2 the numeric values for C_g and C_{gk} are shown. Both values are well above 1.33 which indicates that the gage is fit for the measurement purpose at hand (defined by the tolerance). The *potential* gage capability (C_g) is greater than the *actual* gage capability C_{gk} which implies a systematic error, but the numeric values being > 2 there seems not to be any reason to take serious action. If the systematic error is significant could be tested using the *t-test for one variable*.

7.5.4 Measurement System Analysis Type II (Gage R&R)

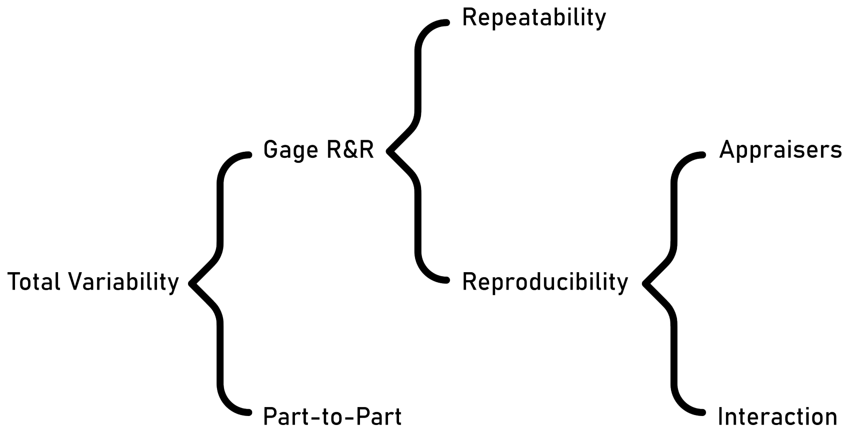


Figure 7.24: The general principle of a gage R & R

A Gage R&R study assesses the variation in measurements from a specific process by measuring the same parts multiple times with the same instrument by different operators. It helps determine the reliability of the measurement system and identifies areas for improvement.

7.5.4.1 Definitions

Accuracy The closeness of agreement between a test result and the accepted reference value(Cano, Moguerza, and Redchuk 2012).

Trueness The closeness of agreement between the average value obtained from a large series of test results and an accepted reference value(Cano, Moguerza, and Redchuk 2012).

Precision The closeness of agreement between independent test results obtained under stipulated conditions(Cano, Moguerza, and Redchuk 2012).

Repeatability Precision under repeatability conditions (where independent test results are obtained using the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time)(Cano, Moguerza, and Redchuk 2012).

Reproducibility Precision under reproducibility conditions (where test results are obtained using the same method on identical test items in different laboratories with different operators using different equipment)(Cano, Moguerza, and Redchuk 2012).

7.5.4.2 Introductory example

- A battery manufacturer makes several types of batteries for domestic use.
- Voltage is **C**ritical **T**o **Q**uality (CTQ)

- the parts are the batteries $a = 3$
- the appraisers are the voltmeters $b = 2$
- measurement is taken three times $n = 3$
- $a \times b \times n = 3 \times 2 \times 3 = 18$ measurements

7.5.4.3 The data

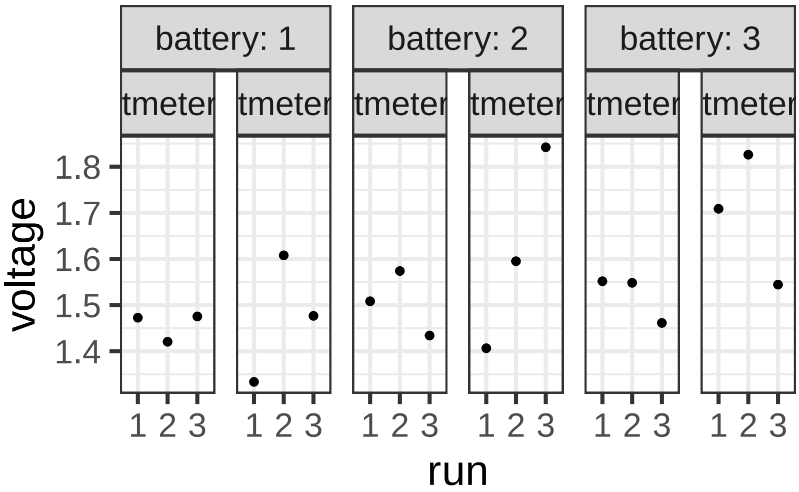


Figure 7.25: The data from the 18 experiments for the GageR&R

7.5.4.4 The analysis

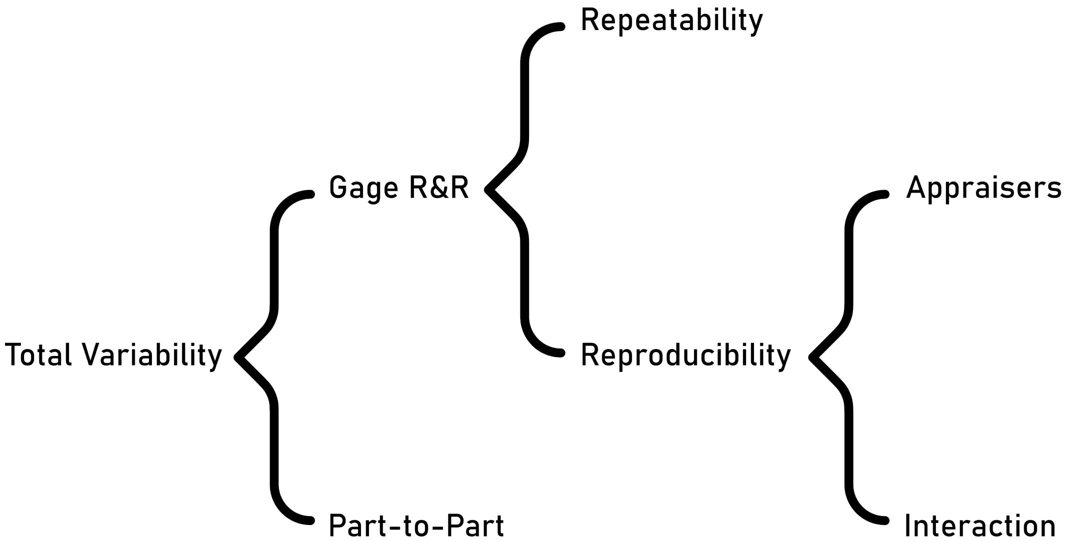
```
anova(lm(voltage ~ battery + voltmeter + battery * voltmeter,
         data = ss.data.batteries))
```

```
Analysis of Variance Table

Response: voltage
                  Df   Sum Sq  Mean Sq F value Pr(>F)
battery            2 0.063082 0.031541  1.9939 0.1788
voltmeter          1 0.044442 0.044442  2.8095 0.1195
```

```
battery:voltmeter  2 0.018472 0.009236  0.5839 0.5728
Residuals         12 0.189821 0.015818
```

WOW!

7.5.4.5 Variance decomposition - the theory



7.5.4.5.1 Repeatability

$$\sigma^2_{Repeatability} = MSE \tag{7.20}$$

- directly obtainable in ANOVA table

7.5.4.5.2 Reproducibility

$$\sigma^2_{Reproducibilty} = \sigma^2_{Appraiser} + \sigma^2_{Interaction} \tag{7.21}$$

$$\sigma^2_{Appraiser} = \frac{MSB - MSAB}{a \times n} \tag{7.22}$$

$\sigma^2_{Appraiser}$ Variance introduced by appraisers
MSB Mean of squares - B
$MSAB$ Mean squares of interaction - AB
a number of levels for factor - number of batteries: 3
n number of replicated measures: 3

$$\sigma^2_{Interaction} = \frac{MSBA - MSE}{n} \tag{7.23}$$

$\sigma^2_{Interaction}$ Variance introduced by interaction
$MSAB$ Mean squares of interaction - AB
MSE Mean squares of error
n number of replicated measures: 3

7.5.4.5.3 Gage R&R

$$\sigma^2_{Gage\ R\&R} = \sigma^2_{Repeatability} + \sigma^2_{Reproducibility} \tag{7.24}$$

All variance is calculated that comes from the Gage!

Are we finished?

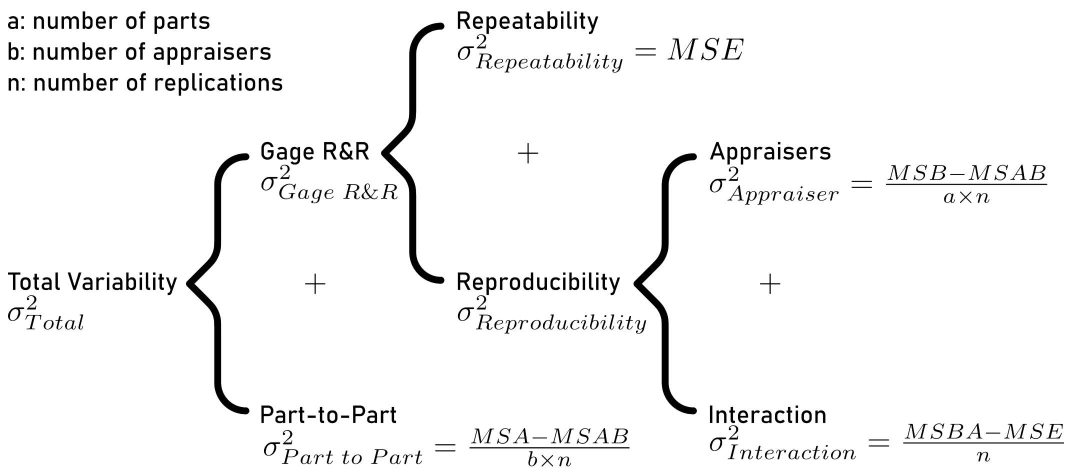
We measure *something*, so what about the part?

7.5.4.5.4 Part to Part

$$\sigma^2_{Part\ to\ Part} = \frac{MSA - MSAB}{b \times n} \tag{7.25}$$

$\sigma^2_{Part\ to\ Part}$ Variance introduced by the parts
MSA Mean of squares - A
$MSAB$ Mean squares of interaction - AB
b number of appraisers - number of voltmeters: 2
n number of replicated measures: 3

7.5.4.5.5 Total Variability



7.5.4.6 Variance decomposition - the values

$$
\begin{aligned}
\sigma^2_{Repeatability} &= 0.0158 \\
\sigma^2_{Appraiser} &= 0.0039 \\
\sigma^2_{Interaction} &= 0 < 0 \rightarrow 0 \\
\sigma^2_{Reproducibility} &= 0.0039 \\
\sigma^2_{Gage\ R\&R} &= 0.0197 \\
\sigma^2_{Part\ to\ Part} &= 0.0037 \\
\sigma^2_{Total} &= 0.0234
\end{aligned}
$$

7.5.4.7 Gage R&R “standardized output”

7.5.4.7.1 AVNOVA table

```
                  Df  Sum Sq Mean Sq F value Pr(>F)
battery            2 0.06308 0.03154   3.415  0.227
voltmeter          1 0.04444 0.04444   4.812  0.160
battery:voltmeter  2 0.01847 0.00924   0.584  0.573
Repeatability     12 0.18982 0.01582
Total             17 0.31582
```

7.5.4.7.2 ANOVA reduced model

```
              Df  Sum Sq Mean Sq F value Pr(>F)
battery        2 0.06308 0.03154   2.120  0.157
voltmeter      1 0.04444 0.04444   2.987  0.106
Repeatability 14 0.20829 0.01488
Total         17 0.31582
```

7.5.4.7.3 Variance decomposition

```
                      VarComp %Contrib
Total Gage R&R    0.018162959    86.74
  Repeatability   0.014878111    71.05
  Reproducibility 0.003284848    15.69
    voltmeter     0.003284848    15.69
Part-To-Part      0.002777127    13.26
Total Variation   0.020940086   100.00
```

7.5.4.7.4 Study Variance

```
                      StdDev  StudyVar %StudyVar %Tolerance
Total Gage R&R    0.13477002 0.8086201     93.13      80.86
  Repeatability   0.12197586 0.7318552     84.29      73.19
  Reproducibility 0.05731359 0.3438816     39.61      34.39
    voltmeter     0.05731359 0.3438816     39.61      34.39
Part-To-Part      0.05269846 0.3161907     36.42      31.62
Total Variation   0.14470690 0.8682414    100.00      86.82
```

7.5.4.7.5 ndc - number of distinct categories

```
[1] 1
```

7.5.4.7.6 standardized graphical output

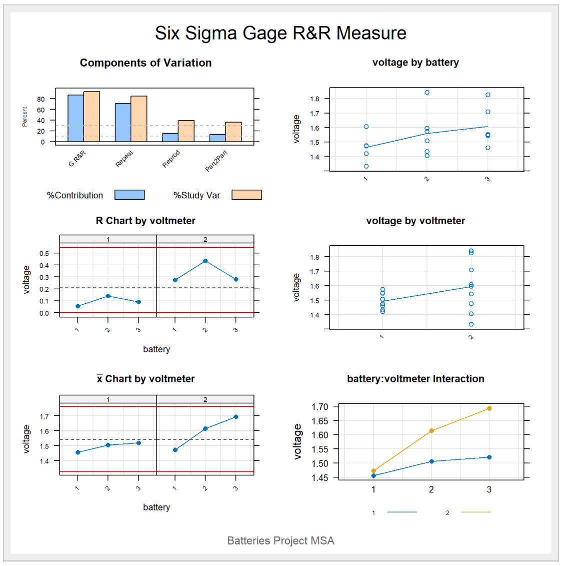


Figure 7.26: A standardized graphical output after a complete GageR&R

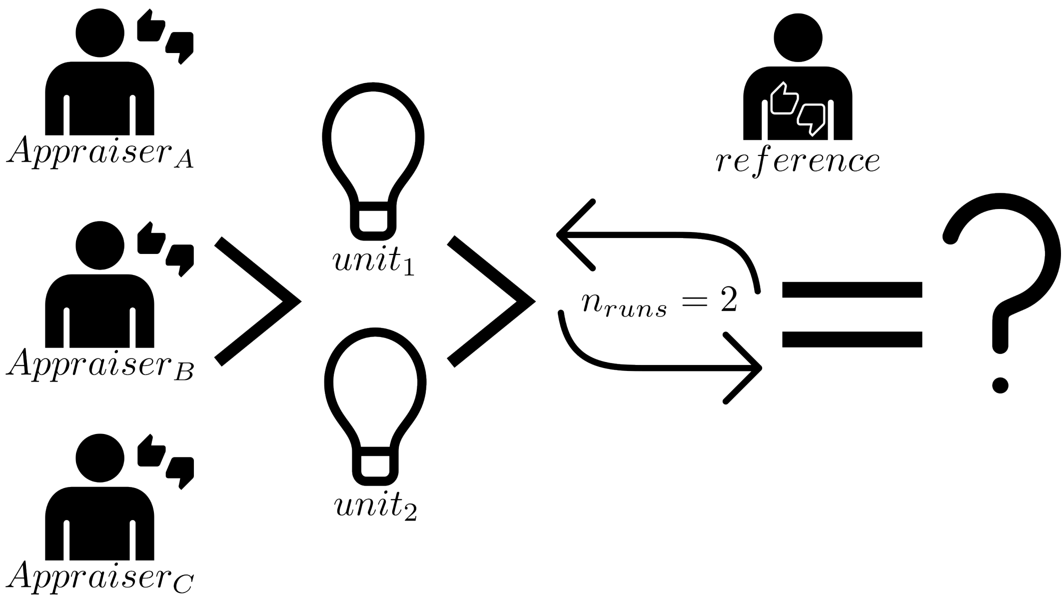
7.5.4.8 Gage R&R in the classroom

- 3 parts
- 3 volunteers
- 1 recorder
- 1 gage
- 10 experiments
- 3 repetitions
- randomize the trials
- now do it

7.5.4.9 Attribute Agreement Analysis

Attribute Agreement Analysis (AAA) is a statistical method used to evaluate the agreement among multiple observers when assigning categorical ratings to items. It involves defining attributes, selecting observers, collecting ratings, and analyzing the data to determine the level of agreement. This helps ensure the reliability of assessments and informs decision-making processes.

7.5.4.9.1 Setup



7.5.4.9.2 Results

Table 7.3

appraiser	runs	units	reference	results
1	1	3	bad	bad
1	1	1	good	good
1	1	2	bad	good
2	1	3	bad	good
2	1	1	good	good
2	1	2	bad	good
1	2	3	good	good
1	2	1	bad	bad
1	2	2	bad	bad
2	2	3	good	bad
2	2	1	bad	bad
2	2	2	bad	good

7.5.4.9.3 Overall agreement

$$Agreement_{overall} = 100 \times \frac{X}{N} \tag{7.26}$$

X number of times appraisers agree with reference
N number of rows with valid data

$$Agreement_{overall} = 58.3\%$$

7.5.4.9.4 Appraiser Agreement

$$Agreement_{appraiser} = 100 \times \frac{X}{N} \tag{7.27}$$

X number of times the single appraisers agrees with reference
N_i number of runs for the i-th appraiser

$$Appraiser_1 = 83.3\%$$
$$Appraiser_2 = 33.3\%$$

7.5.4.9.5 Reference Agreement

$$Agreement_{reference} = 100 \times \frac{X}{N} \tag{7.28}$$

X number of times result agrees with the reference
N_i number of runs for the i-th result

$$Reference_{bad} = 50\%$$
$$Reference_{good} = 75\%$$

7.5.4.9.6 Run agreement

$$Agreement_{run} = 100 \times \frac{X}{N} \tag{7.29}$$

X number of reference agreement in runs
N_i number of runs for the i-th run

$$Reference_1 = 50\%$$
$$Reference_2 = 66.7\%$$

7.5.4.9.7 Appraiser and reference agreement

$$Agreement_{appraiser\,ref} = 100 \times \frac{X}{N} \tag{7.30}$$

X number of reference agreement in for appraisers in reference class
N_i number of agreements for the i-th appraiser and the i-th standard

Table 7.4

appraiser	reference	overall_agreement
1	bad	75.00%
1	good	100.00%
2	bad	25.00%
2	good	50.00%

7.5.4.9.8 graphical representation

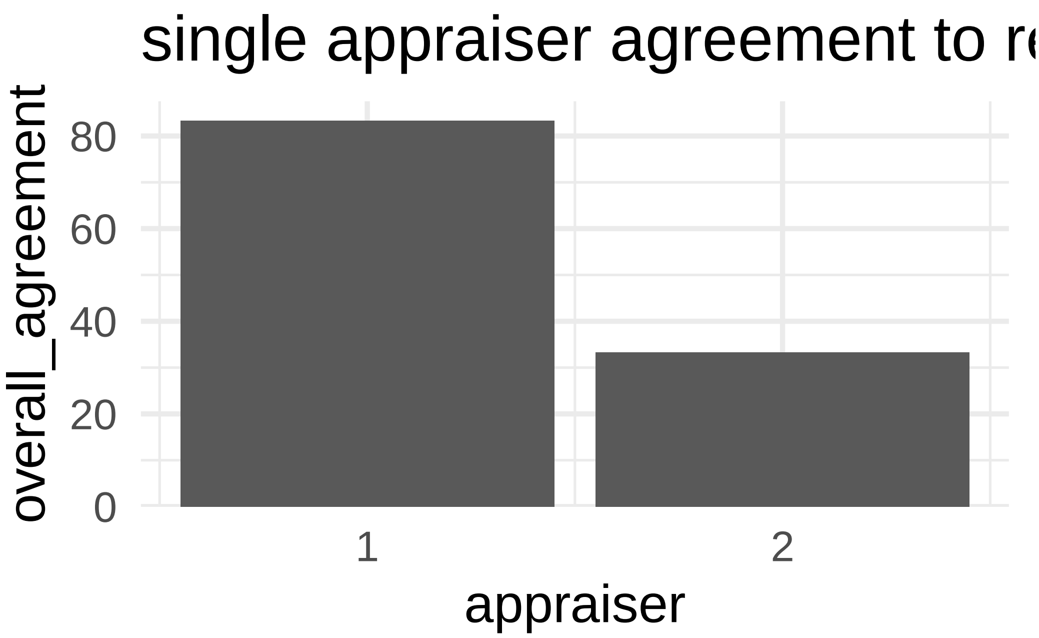


Figure 7.27: Single appraiser agreement to reference.

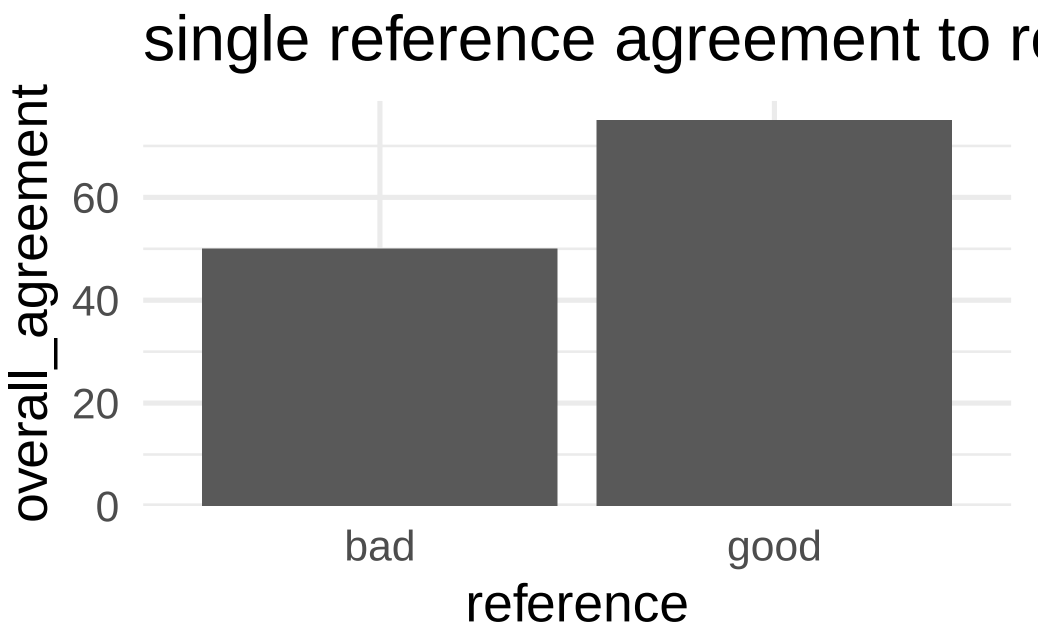


Figure 7.28: How good is the agreement in the reference?

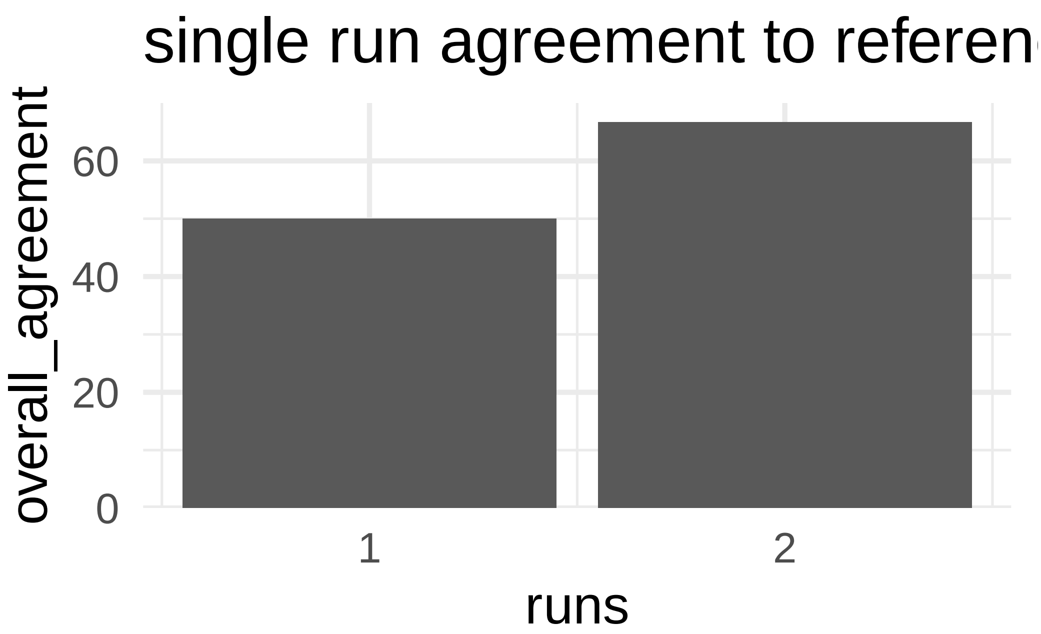


Figure 7.29: Single run agreement to reference.

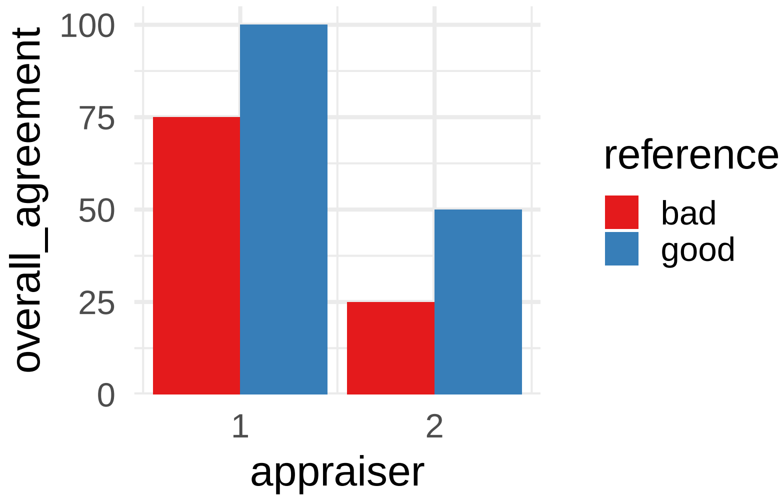
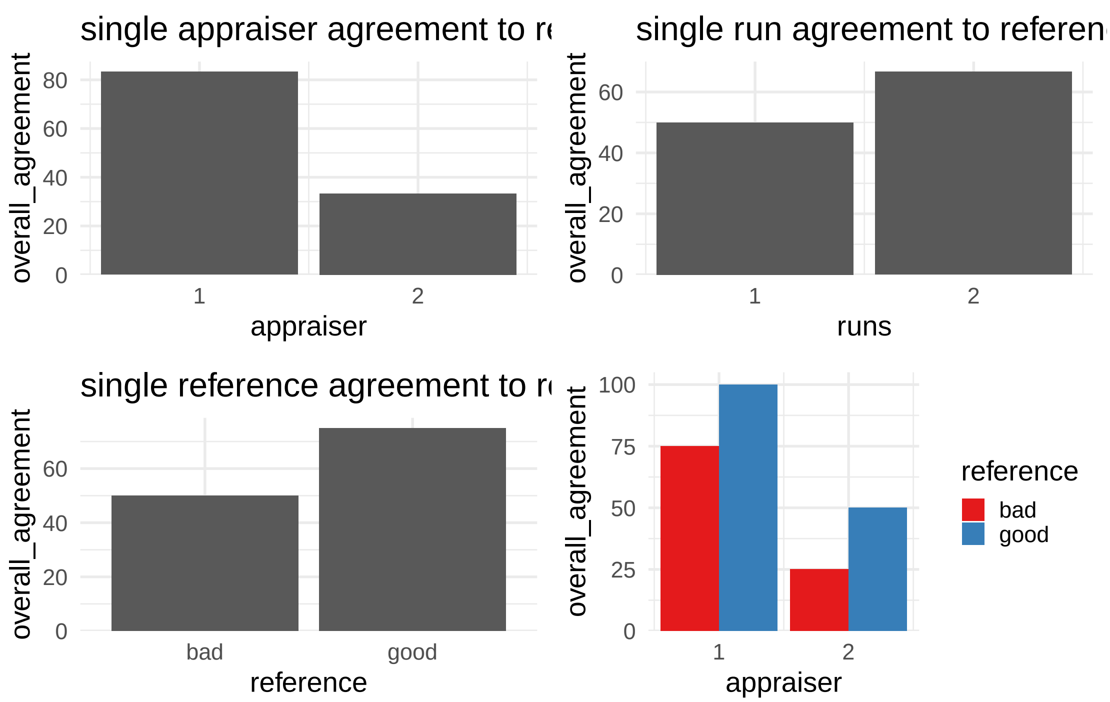


Figure 7.30: Appraiser ref agreement



8 Introduction to Design of Experiments (DoE)

8.1 (O)ne (F)actor (A)t a (T)ime

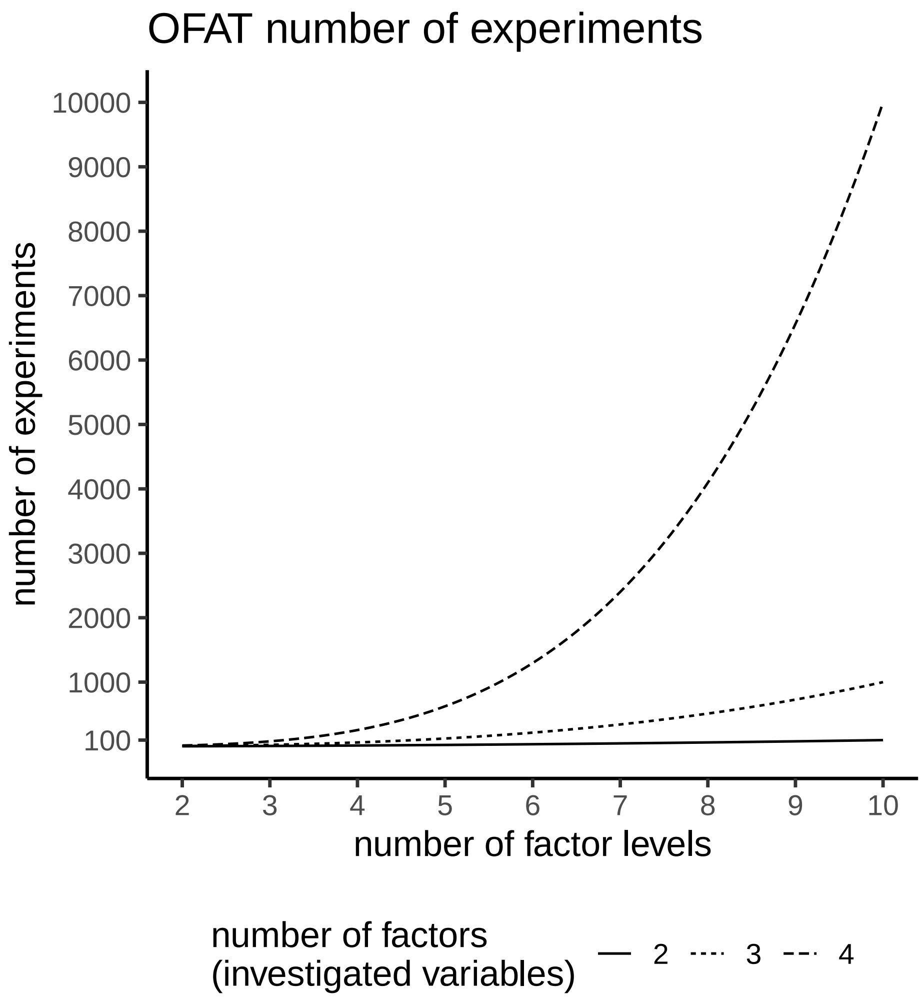


Figure 8.1: OFAT quickly becomes cumbersome

8.2 curse of dimensionality

$$n_{experiments} = n_{levels}^{n_{factors}} \tag{8.1}$$

8.3 Concept of ANOVA

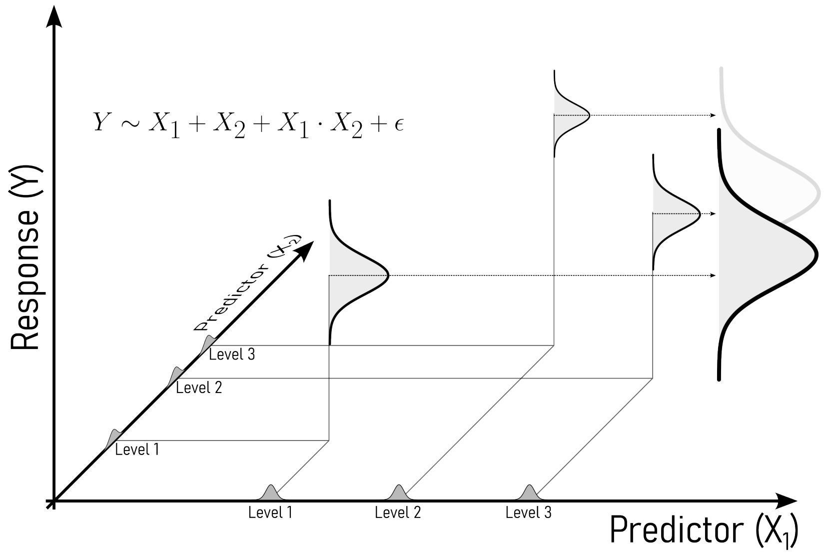


Figure 8.2: classical ANOVA concept

8.4 Basics of Experimental Design

Design of Experiments

8.5 Experimental planning strategies

1.No planning

- bad way of conducting an experiment
- happens often enough (trial-and-error approach)

2.Plan everything at the beginning

- after definition the entire budget is allocated to perform all possible experiments

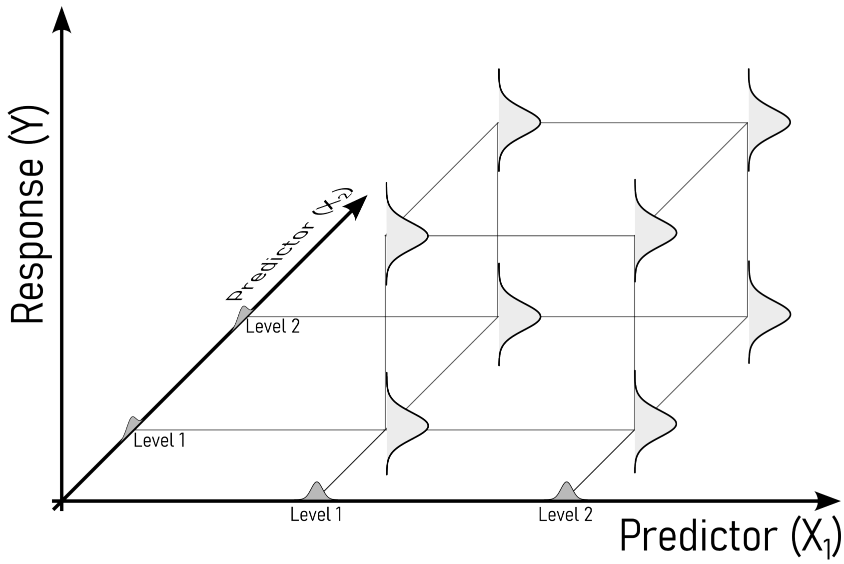


Figure 8.3: The connection between ANOVA and DoE.

- does not take into account intermediate results
- spend money on experiments that contributed nothing to our knowledge of the process

3.Sequential planning

- first stage, a reduced number of trials will be conducted to make decisions about the next stage
- first stage should consume between 25% and 40% of the budget
- most of the budget should be spent in subsequent stages, taking into account previous results.

8.6 pizza dough example

- representation of factors and levels for a designed experiment
- example: pizza dough
  - food manufacturer is looking for the best recipe for its main product: pizza dough sold in retailers
  - three factors shall be determined: `flour`, `salt`, baking powder: `bakPow`
  - response will be determined by experts as `score`
  - factors are to be set `low` (−) and `high` (+)

8.7 design matrix

Table 8.1: The design matrix for the pizza dough experimentation

flour	salt	bakPow	score
-	-	-	NA
+	-	-	NA
-	+	-	NA
+	+	-	NA
-	-	+	NA
+	-	+	NA
-	+	+	NA
+	+	+	NA

Be bold, but not stupid!

8.7.1 progressive experimentation

- OFAT
  - will leave out **interactions** of variables
- 2^k: two-level factor experimentation
- including replications

1. *Screening* experiments: to select the most important factors
2. *Characterizing* experiments: to study the model (residuals) of $Y = f(X)$
3. *Optimization* experiments: operational minimum value for the process

8.8 Model assumptions

- randomization!

Table 8.2: The randomized design matrix for experimental runs

flour	salt	bakPow	score	ord
+	-	+	NA	1

-	-	-	NA	2
-	+	+	NA	3
+	+	-	NA	4
+	+	+	NA	5
-	-	+	NA	6
-	+	-	NA	7
+	-	-	NA	8

8.9 experimental model

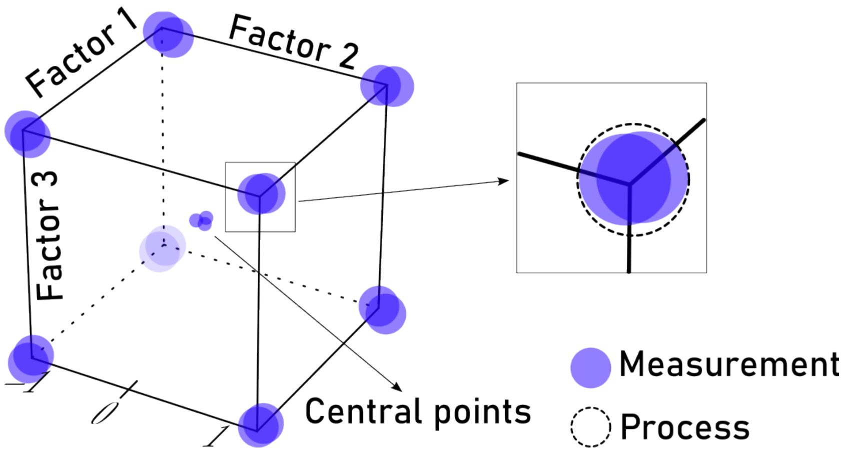


Figure 8.4: The experimental model for a DoE

8.10 analytical model

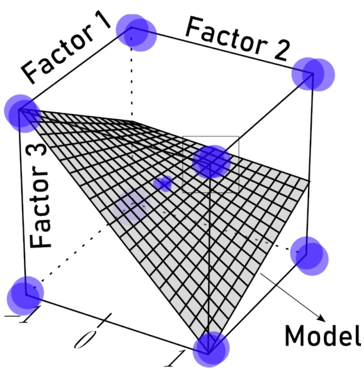


Figure 8.5: The experimental model with the fitted linear model.

8.11 2^k factorial Designs

k number of factors to be studied, all with 2 levels

n number of replications $\rightarrow$ total number of experiments $= n \times 2^k$

$A, B, \ldots$ factors (uppercase latin letters)

$\alpha, \beta, \ldots$ main effects

8.12 complete analytical model

- three factors, n replications

$$\begin{aligned} y_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{kl} \\ &\quad + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl} \end{aligned} \tag{8.2}$$

$$i = 1, 2 \quad j = 1, 2 \quad k = 1, 2 \quad l = 1 \ldots n$$

$$\epsilon_{ijkl} \sim N(0, \sigma) \tag{8.3}$$

μ global mean of the response

α_i effect of factor A at level i

β_j effect of factor B at level j

γ_k effect of factor C at level k

$(\alpha\beta)_{ij}$ effect of the interaction of factors A and B at levels i and j

$(\alpha\gamma)_{ik}$ effect of the interaction of factors A and C at levels i and k

$(\beta\gamma)_{jk}$ effect of the interaction of factors B and C at levels j and k

$(\alpha\beta\gamma)_{ijk}$ effect of the interaction of factors A, B and C at levels i, j and k

ϵ_{ijkl} random error component of the model

8.12.1 pizza dough example raw data

"… bake the pizza for 9min at 180°C …"

repl	flour	salt	bakPow	score	ord
1	-	-	-	5.33	2
1	+	-	-	6.99	4
1	-	+	-	4.23	8
1	+	+	-	6.61	5
1	-	-	+	2.26	1
1	+	-	+	5.75	6
1	-	+	+	3.26	3
1	+	+	+	6.24	7
2	-	-	-	5.70	2
2	+	-	-	7.71	4
2	-	+	-	5.13	8
2	+	+	-	6.76	5
2	-	-	+	2.79	1
2	+	-	+	4.57	6
2	-	+	+	2.48	3
2	+	+	+	6.18	7

(Cano, Moguerza, and Redchuk 2012)

8.12.2 pizza dough example summarised data

flour	salt	bakPow	mean_score
-	-	-	5.515
-	-	+	2.525
-	+	-	4.680
-	+	+	2.870
+	-	-	7.350
+	-	+	5.160
+	+	-	6.685
+	+	+	6.210

8.12.3 pizza dough recipe full model

```
doe.model1 <- lm(score ~ flour + salt + bakPow +
flour * salt + flour * bakPow +
```

```
salt * bakPow + flour * salt * bakPow,
data = ss.data.doe1)

summary(doe.model1)
```

```
Call:
lm(formula = score ~ flour + salt + bakPow + flour * salt + flour *
    bakPow + salt * bakPow + flour * salt * bakPow, data = ss.data.doe1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5900 -0.2888  0.0000  0.2888  0.5900

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.5150     0.3434  16.061 2.27e-07 ***
flour+                 1.8350     0.4856   3.779 0.005398 **
salt+                 -0.8350     0.4856  -1.719 0.123843
bakPow+               -2.9900     0.4856  -6.157 0.000272 ***
flour+:salt+           0.1700     0.6868   0.248 0.810725
flour+:bakPow+         0.8000     0.6868   1.165 0.277620
salt+:bakPow+          1.1800     0.6868   1.718 0.124081
flour+:salt+:bakPow+   0.5350     0.9712   0.551 0.596779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4856 on 8 degrees of freedom
Multiple R-squared:  0.9565,    Adjusted R-squared:  0.9185
F-statistic: 25.15 on 7 and 8 DF,  p-value: 7.666e-05
```

8.12.4 pizza dough recipe elimination model

```
doe.model2 <- lm(score ~ flour + bakPow,data = ss.data.doe1)

summary(doe.model2)
```

```
Call:
lm(formula = score ~ flour + bakPow, data = ss.data.doe1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.84812 -0.54344  0.06063  0.44406  0.86938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8306     0.2787  17.330 2.30e-10 ***
flour+        2.4538     0.3219   7.624 3.78e-06 ***
bakPow+      -1.8662     0.3219  -5.798 6.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6437 on 13 degrees of freedom
Multiple R-squared:  0.8759,    Adjusted R-squared:  0.8568
F-statistic: 45.87 on 2 and 13 DF,  p-value: 1.288e-06
```

8.12.5 pizza dough statistical model

$$\widehat{score} = 4.83 + 2.45 \times flour - 1.87 \times bakPow \tag{8.4}$$

$$\widehat{score} = 5.12 + 1.23 \times flour - 0.93 \times bakPow \tag{8.5}$$

8.12.6 main effect plot

8.12.7 interaction plot

8.12.8 model validity

8.12.8.1 residual patterns

8.12.8.2 residual distribution

```r
shapiro.test(doe.model2_aug$.resid)
```

```
    Shapiro-Wilk normality test

data:  doe.model2_aug$.resid
W = 0.90652, p-value = 0.1023
```

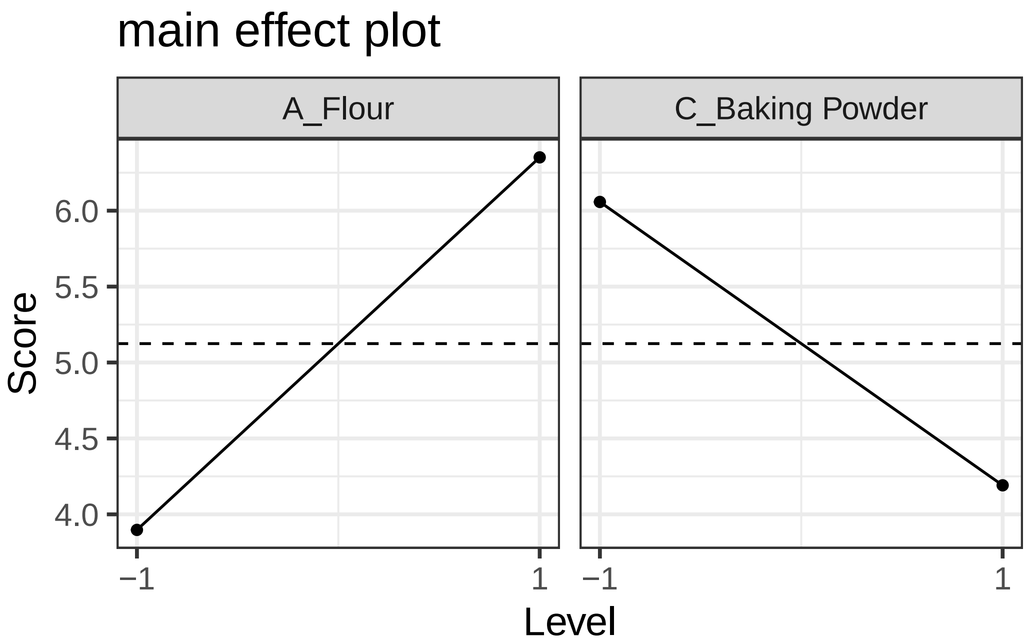


Figure 8.6: The main effect plot for the pizza dough model

8.13 Design of Experiments for process improvement

What if ...

... not all influencing factors (X) on the process have been identified?

... some X depend on external conditions and are not under control?

robust design

... means also including *noise* factors that are not under our control.

8.13.1 pizza dough example

- pizzas came out pretty bad as reported by the customers
- pizza quality heavily relies on baking conditions! ($T = 180°C, t = 9min$)
- almost **nobody** followed the recipe
- noise factors are included with two levels
  - $7min$ and $11min$ as $t+$ and $t-$
  - $160°C$ and $200°C$ as $T+$ and $t-$

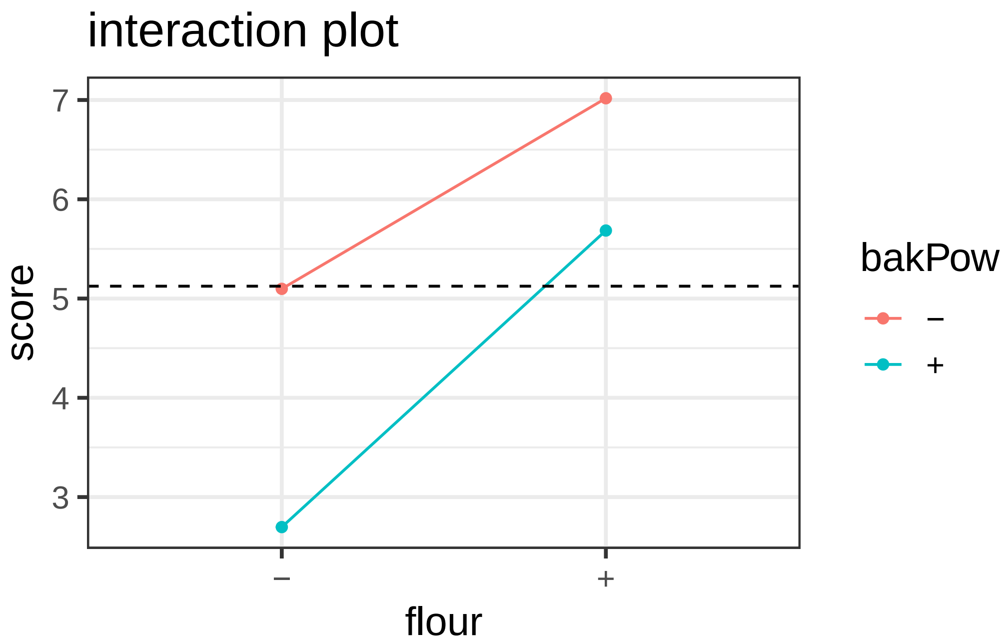


Figure 8.7: The interaction plot for the pizza dough model

- 2^5 factorial design with 2 replications = 64 experimental runs

8.14 linear model - first run

```
Call:
lm(formula = score ~ (. - repl)^3, data = ss.data.doe2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20094 -0.32937  0.02625  0.35656  1.07187

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.16906    0.42203   7.509 5.09e-09 ***
flour+                0.07406    0.54902   0.135  0.89340
salt+                -1.47219    0.54902  -2.681  0.01078 *
bakPow+              -1.43219    0.54902  -2.609  0.01293 *
temp+                 2.56156    0.54902   4.666 3.75e-05 ***
time+                 1.49594    0.54902   2.725  0.00967 **
```

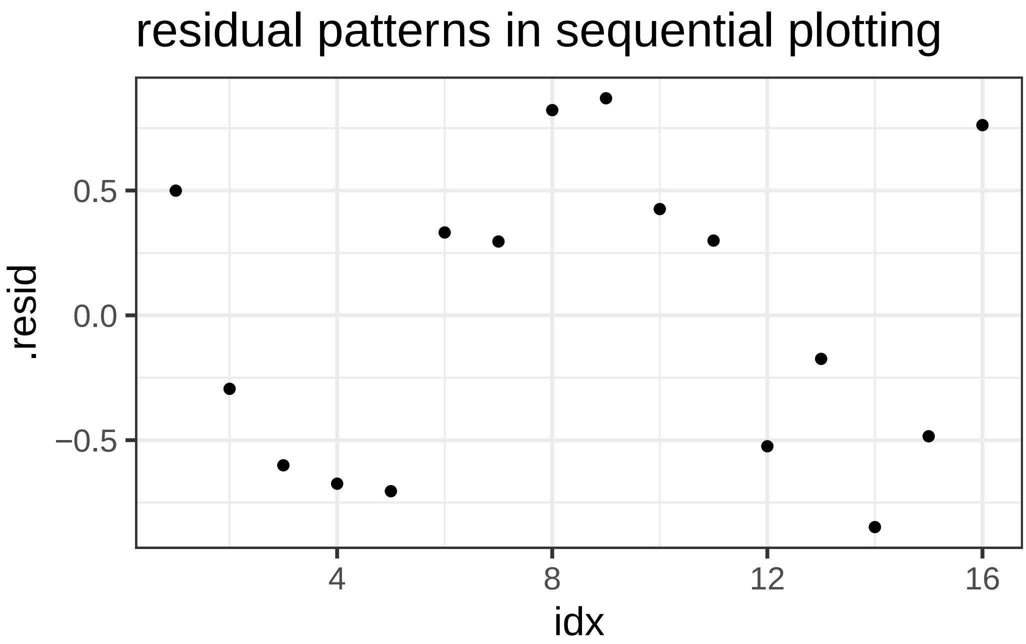


Figure 8.8: Check for any pattern in the model residuals

```
flour+:salt+           1.71000    0.66214   2.583  0.01378 *
flour+:bakPow+         2.14000    0.66214   3.232  0.00254 **
flour+:temp+          -1.26250    0.66214  -1.907  0.06414 .
flour+:time+           0.46375    0.66214   0.700  0.48796
salt+:bakPow+          0.89250    0.66214   1.348  0.18567
salt+:temp+           -0.19500    0.66214  -0.294  0.76998
salt+:time+            1.38625    0.66214   2.094  0.04302 *
bakPow+:temp+         -1.17000    0.66214  -1.767  0.08526 .
bakPow+:time+         -1.30375    0.66214  -1.969  0.05628 .
temp+:time+           -3.91125    0.66214  -5.907 7.64e-07 ***
flour+:salt+:bakPow+   0.14875    0.66214   0.225  0.82346
flour+:salt+:temp+     1.52375    0.66214   2.301  0.02696 *
flour+:salt+:time+    -1.11875    0.66214  -1.690  0.09930 .
flour+:bakPow+:temp+   0.22375    0.66214   0.338  0.73728
flour+:bakPow+:time+   0.09125    0.66214   0.138  0.89112
flour+:temp+:time+     0.30125    0.66214   0.455  0.65172
salt+:bakPow+:temp+   -0.33125    0.66214  -0.500  0.61977
salt+:bakPow+:time+    0.33625    0.66214   0.508  0.61451
salt+:temp+:time+     -1.04375    0.66214  -1.576  0.12324
bakPow+:temp+:time+    2.19125    0.66214   3.309  0.00205 **
---
```

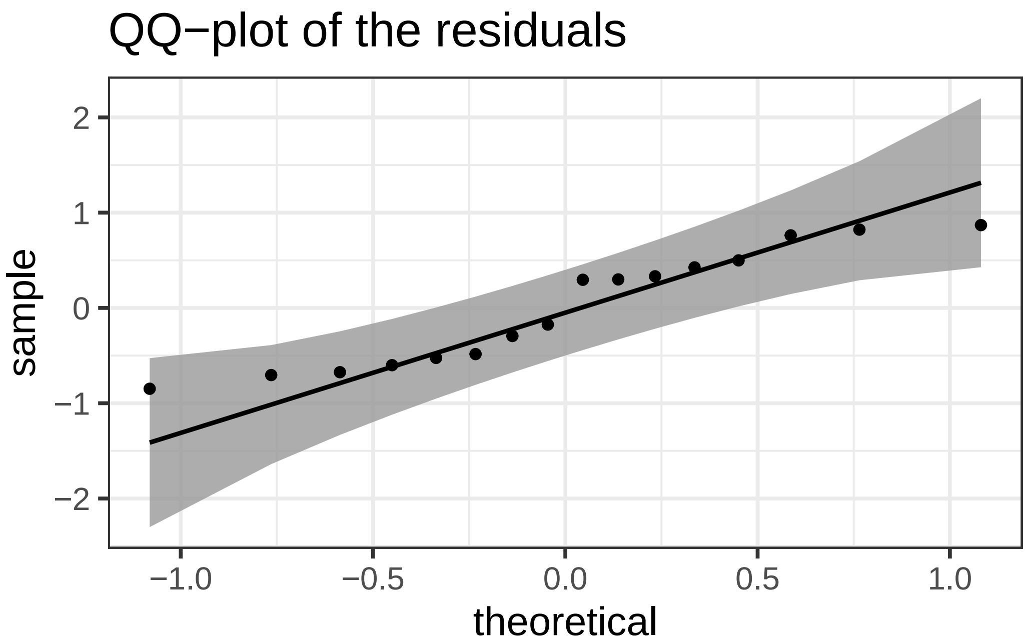


Figure 8.9: Check for the residuals normality (QQ plot)

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6621 on 38 degrees of freedom
Multiple R-squared:  0.9037,    Adjusted R-squared:  0.8404
F-statistic: 14.27 on 25 and 38 DF,  p-value: 1.428e-12
```

8.15 linear model - stepwise elimination

8.15.1 get rid of non-significant

```
selectionvar <- step(model.prob1, method="backwards")
```

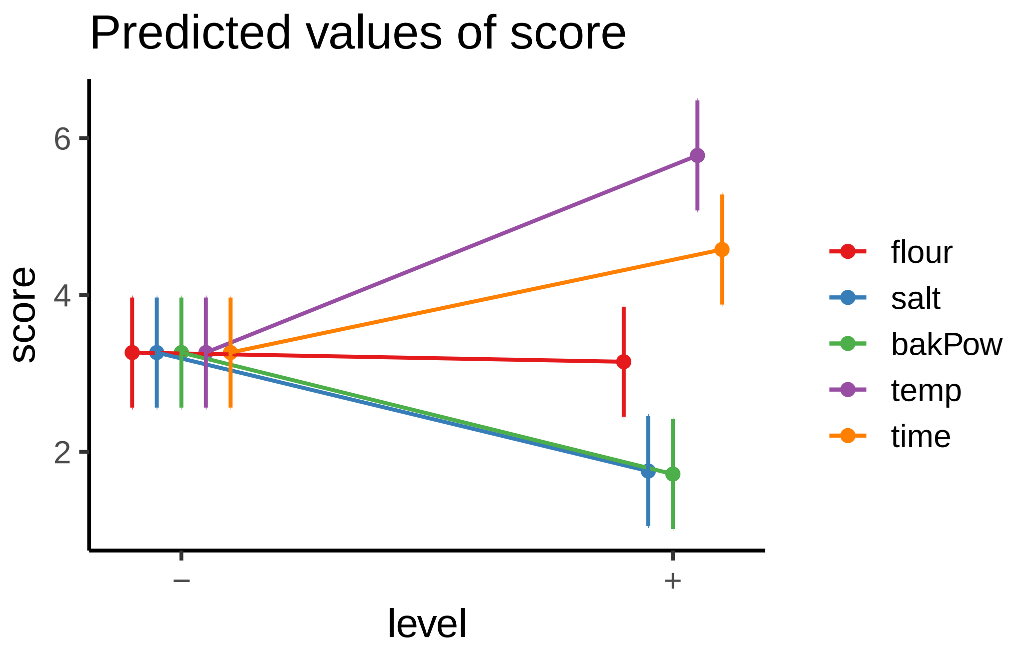
```
Start:  AIC=-34.13
score ~ ((repl + flour + salt + bakPow + temp + time) - repl)^3

                    Df Sum of Sq    RSS     AIC
- flour:bakPow:time  1    0.0083 16.669 -36.102
- flour:salt:bakPow  1    0.0221 16.683 -36.049
```

```
- flour:bakPow:temp  1    0.0501 16.710 -35.942
- flour:temp:time    1    0.0908 16.751 -35.787
- salt:bakPow:temp   1    0.1097 16.770 -35.714
- salt:bakPow:time   1    0.1131 16.773 -35.701
<none>                           16.660 -34.134
- salt:temp:time     1    1.0894 17.750 -32.080
- flour:salt:time    1    1.2516 17.912 -31.498
- flour:salt:temp    1    2.3218 18.982 -27.784
- bakPow:temp:time   1    4.8016 21.462 -19.926

Step:  AIC=-36.1
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:bakPow +
    flour:salt:temp + flour:salt:time + flour:bakPow:temp + flour:temp:time +
    salt:bakPow:temp + salt:bakPow:time + salt:temp:time + bakPow:temp:time

                    Df Sum of Sq    RSS     AIC
- flour:salt:bakPow  1    0.0221 16.691 -38.017
- flour:bakPow:temp  1    0.0501 16.719 -37.910
- flour:temp:time    1    0.0908 16.759 -37.755
- salt:bakPow:temp   1    0.1097 16.779 -37.682
- salt:bakPow:time   1    0.1131 16.782 -37.670
<none>                           16.669 -36.102
- salt:temp:time     1    1.0894 17.758 -34.050
- flour:salt:time    1    1.2516 17.920 -33.469
- flour:salt:temp    1    2.3218 18.991 -29.756
- bakPow:temp:time   1    4.8016 21.470 -21.902

Step:  AIC=-38.02
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:temp +
    flour:salt:time + flour:bakPow:temp + flour:temp:time + salt:bakPow:temp +
    salt:bakPow:time + salt:temp:time + bakPow:temp:time

                    Df Sum of Sq    RSS     AIC
- flour:bakPow:temp  1    0.0501 16.741 -39.826
- flour:temp:time    1    0.0908 16.782 -39.670
- salt:bakPow:temp   1    0.1097 16.801 -39.598
- salt:bakPow:time   1    0.1131 16.804 -39.585
<none>                           16.691 -38.017
- salt:temp:time     1    1.0894 17.780 -35.971
```

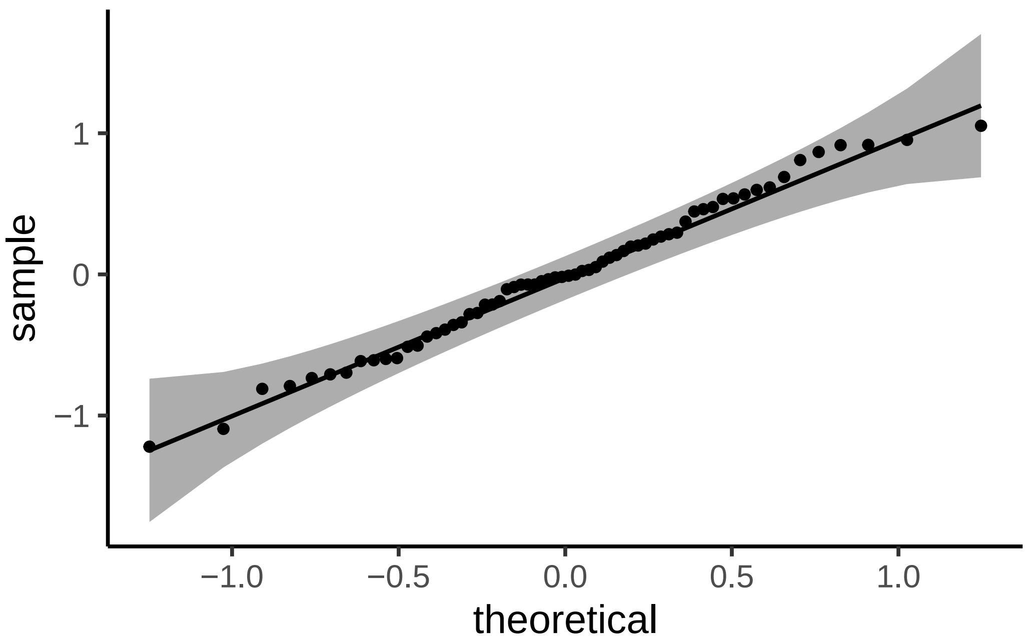
```
- flour:salt:time    1    1.2516 17.942 -35.390
- flour:salt:temp    1    2.3218 19.013 -31.682
- bakPow:temp:time   1    4.8016 21.492 -23.836

Step:  AIC=-39.83
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:temp +
    flour:salt:time + flour:temp:time + salt:bakPow:temp + salt:bakPow:time +
    salt:temp:time + bakPow:temp:time

                   Df Sum of Sq    RSS     AIC
- flour:temp:time   1    0.0908 16.832 -41.480
- salt:bakPow:temp  1    0.1097 16.851 -41.408
- salt:bakPow:time  1    0.1131 16.854 -41.395
<none>                          16.741 -39.826
- salt:temp:time    1    1.0894 17.830 -37.791
- flour:salt:time   1    1.2516 17.993 -37.211
- flour:salt:temp   1    2.3218 19.063 -33.513
- bakPow:temp:time  1    4.8016 21.543 -25.687
- flour:bakPow      1   22.5032 39.244  12.699

Step:  AIC=-41.48
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:temp +
    flour:salt:time + salt:bakPow:temp + salt:bakPow:time + salt:temp:time +
    bakPow:temp:time

                   Df Sum of Sq    RSS     AIC
- salt:bakPow:temp  1    0.1097 16.941 -43.064
- salt:bakPow:time  1    0.1131 16.945 -43.051
<none>                          16.832 -41.480
- salt:temp:time    1    1.0894 17.921 -39.466
- flour:salt:time   1    1.2516 18.083 -38.889
- flour:salt:temp   1    2.3218 19.154 -35.209
- bakPow:temp:time  1    4.8016 21.633 -27.418
- flour:bakPow      1   22.5032 39.335  10.847

Step:  AIC=-43.06
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:temp +
```

```
    flour:salt:time + salt:bakPow:time + salt:temp:time + bakPow:temp:time

                   Df Sum of Sq    RSS     AIC
- salt:bakPow:time  1    0.1131 17.054 -44.638
<none>                          16.941 -43.064
- salt:temp:time    1    1.0894 18.031 -41.075
- flour:salt:time   1    1.2516 18.193 -40.502
- flour:salt:temp   1    2.3218 19.263 -36.844
- bakPow:temp:time  1    4.8016 21.743 -29.094
- flour:bakPow      1   22.5032 39.445   9.025

Step:  AIC=-44.64
score ~ flour + salt + bakPow + temp + time + flour:salt + flour:bakPow +
    flour:temp + flour:time + salt:bakPow + salt:temp + salt:time +
    bakPow:temp + bakPow:time + temp:time + flour:salt:temp +
    flour:salt:time + salt:temp:time + bakPow:temp:time

                   Df Sum of Sq    RSS     AIC
<none>                          17.054 -44.638
- salt:temp:time    1    1.0894 18.144 -42.675
- flour:salt:time   1    1.2516 18.306 -42.106
- flour:salt:temp   1    2.3218 19.376 -38.469
- salt:bakPow       1    3.7588 20.813 -33.891
- bakPow:temp:time  1    4.8016 21.856 -30.762
- flour:bakPow      1   22.5032 39.558   7.208
```

8.15.2 main effect and interaction



8.15.3 check residuals



8.15.4 pragmatic result

Table 8.5: The pragmatic results for the DoE

flour	salt	bakPow	score	T1t1	T2t1	T1t2	T2t2	Mean	SD
-	-	-	5.515	3.675	5.120	4.185	3.900	4.479	0.6352559
+	-	-	7.350	3.370	4.520	5.050	2.940	4.646	0.9814615
-	+	-	4.680	0.955	4.910	5.295	1.170	3.402	2.3394319
+	+	-	6.685	3.590	5.895	5.625	3.870	5.133	1.1827299
-	-	+	2.525	1.915	3.055	1.725	1.700	2.184	0.6446882
+	-	+	5.160	3.140	5.010	5.535	2.900	4.349	1.3216617
-	+	+	2.870	1.215	1.860	3.040	1.310	2.059	0.8388223
+	+	+	6.210	5.805	6.110	5.980	5.965	6.014	0.1249667

9 References

Bartlett, Maurice Stevenson. 1937. "Properties of Sufficiency and Statistical Tests." *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* 160 (901): 268–82. https://doi.org/10.1098/rspa.1937.0109.

Bonferroni, C. E. 1936. *Teoria Statistica Delle Classi e Calcolo Delle Probabilità*. Pubblicazioni Del r. Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze. Seeber. https://books.google.de/books?id=3CY-HQAACAAJ.

Cano, Emilio L., Javier M. Moguerza, and Andrés Redchuk. 2012. "Six Sigma with r" Not available: Not available. https://doi.org/10.1007/978-1-4614-3652-2.

Champely, Stephane. 2020. *Pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr.

Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587.

Davies, Rhian, Steph Locke, and Lucy D'Agostino McGowan. 2022. *datasauRus: Datasets from the Datasaurus Dozen*. https://CRAN.R-project.org/package=datasauRus.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (4): 1–34. https://doi.org/10.18637/jss.v064.i04.

Friedman, Milton. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (December): 675–701. https://doi.org/10.1080/01621459.1937.10503522.

Greenhouse, Samuel W., and Seymour Geisser. 1959. "On Methods in the Analysis of Profile Data." *Psychometrika* 24 (June): 95–112. https://doi.org/10.1007/bf02289823.

Hahs-Vaughn, Debbie L., and Richard G. Lomax. 2013. *An Introduction to Statistical Concepts*. Routledge. https://doi.org/10.4324/9780203137819.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

J. Bibby, E. J. G. Pitman. 1980. "Some Basic Theory for Statistical Inference." *The Mathematical Gazette* 64 (428): 138–38. https://doi.org/10.2307/3615104.

Johnson, Norman Lloyd. 1994. *Continuous Univariate Distributions*. Wiley.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *The Annals of Mathematical Statistics*. https://doi.org/10.1214/aoms/1177730491.

Mauchly, John W. 1940. "Significance Test for Sphericity of a Normal n-Variate Distribution." *The Annals of Mathematical Statistics* 11 (2): 204–9. http://www.jstor.org/stable/2235878.

Meyna, Arno. 2023. *Sicherheit Und Zuverlässigkeit Technischer Systeme*. Carl Hanser Verlag GmbH & Co. KG. https://doi.org/10.3139/9783446468085.fm.

Nuzzo, Regina. 2014. "Scientific Method: Statistical Errors." *Nature* 506 (7487): 150–52. https://doi.org/10.1038/506150a.

Olkin, Ingram. June. *Contributions to Probability and Statistics*. Stanford Univ Pr.

Pearson, Karl. 1895. "Note on Regression and Inheritance in the Case of Two Parents." *Proceedings of the Royal Society of London Series I* 58: 240–42.

Ramalho, Joao. 2021. *industRial: Data, Functions and Support Materials from the Book "industRial Data Science"*. https://CRAN.R-project.org/package=industRial.

Ruder, Sebastian. 2016. "An Overview of Gradient Descent Optimization Algorithms." http://arxiv.org/pdf/1609.04747.pdf.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (December): 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Shewhart, Walter Andrew, and William Edwards Deming. 1986. *Statistical Method from the Viewpoint of Quality Control*. Courier Corporation.

Spearman, C. 1904. "The Proof and Measurement of Association Between Two Things." *The American Journal of Psychology*. https://doi.org/10.2307/1412159.

Standards, National Institute of, Technology (U.S.), and International SEMATECH. 2002. "NIST/SEMATECH Engineering Statistics Handbook," January. https://doi.org/10.18434/M32189.

Starmer, J. 2022. *The StatQuest Illustrated Guide to Machine Learning!!!: Master the Concepts, One Full-Color Picture at a Time, from the Basics All the Way to Neural Networks. BAM!* Packt Publishing, Limited. https://books.google.de/books?id=gWRGzwEACAAJ.

Student. 1908. "The Probable Error of a Mean." *Biometrika* 6 (1): 1. https://doi.org/10.2307/2331554.

Taboga, Marco. 2017. *Lectures on Probability Theory and Mathematical Statistics - 3rd Edition*. Createspace Independent Publishing Platform.

Tamhane, Ajit C. 1977. "Multiple Comparisons in Model i One-Way Anova with Unequal Variances." *Communications in Statistics - Theory and Methods* 6 (January): 15–32. https://doi.org/10.1080/03610927708827466.

"The R Graph Gallery – Help and Inspiration for r Charts." 2022. https://r-graph-gallery.com/.

Tiedemann, Frederik. 2022. *Gghalves: Compose Half-Half Plots Using Your Favourite Geoms*. https://CRAN.R-project.org/package=gghalves.

Weibull, Waloddi. 1951. "A Statistical Distribution Function of Wide Applicability." *Journal of Applied Mechanics* 18 (3): 293–97. https://doi.org/10.1115/1.4010337.

WELCH, B. L. 1947. "The Generalization of "STUDENT's" Problem When Several Different Population Variances Are Involved." *Biometrika*. https://doi.org/10.1093/biomet/34.1-2.28.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science*. "O'Reilly Media, Inc.".

Wilke, Claus O. 2022. *Ggridges: Ridgeline Plots in 'Ggplot2'*. https://CRAN.R-project.org/package=ggridges.