

Vlad BOCĂNEȚ

Statistics and probability



Editura UTPRESS
Cluj Napoca, 2023
ISBN 978-606-737-674-6

Vlad BOCĂNEȚ

Statistics and probability

Course textbook



UTPRESS

Cluj-Napoca, 2023

ISBN 978-606-737-674-6



Editura UTPRESS
Str. Observatorului nr. 34
400775 Cluj-Napoca
Tel.: 0264-401.999
e-mail: utpress@biblio.utcluj.ro
www.utcluj.ro/editura

Director: ing. Dan COLȚEA

Recenzia: Prof.dr. Ioan Cristian Chifu
 Prof.dr.ing. Marius Bulgaru

Pregătire format electronic on-line: Gabriela Groza

Copyright © 2023 Editura UTPRESS
Reproducerea integrală sau parțială a textului sau ilustrațiilor din această carte
este posibilă numai cu acordul prealabil scris al editurii UTPRESS.

ISBN 978-606-737-674-6

Coupons

Introduction.....	5
1. Descriptive statistics.....	6
1.1. Frequency.....	9
1.2. Central tendency.....	10
1.3. Variation or spread of data.....	13
2. Data presentation and synthesis methods.....	17
2.1. Text and statistical indicators.....	17
2.2. Tables.....	18
2.3. Viewing data.....	19
2.3.1. Bar chart or columns.....	20
2.3.2. Line diagram.....	22
2.3.3. Scatter plot.....	23
2.3.4. Pie chart.....	24
2.3.5. Histogram.....	25
2.3.6. Boxplot.....	28
2.4. Knowledge check.....	30
Types of data.....	30
Statistical indicators of central tendency.....	32
Statistical indicators of spread.....	34
Viewing data.....	36
3. Statistical events.....	38
3.1. Definition of events.....	38
3.2. Types of events.....	39
3.3. Operations on Events.....	39
3.4. Knowledge check.....	42
4. Probability.....	45
4.1. Conditional probability.....	45
4.2. Multiplication and addition rules.....	46
4.3. Law of total probability.....	46
4.4. Bayes' rule.....	47
4.5. Knowledge check.....	49
5. Random variables.....	53
5.1. Discrete random variables.....	53
5.2. Continuous random variables.....	54
5.3. Cumulative distribution function.....	55
5.4. Discrete and continuous variables.....	56
5.5. Knowledge check.....	57
6. Discrete distributions.....	60
6.1. Uniform distribution.....	60
6.2. Binomial distribution.....	60
Example problem.....	62

6.3.	Hypergeometric distribution.....	63
	Example problem	64
6.4.	Knowledge check.....	65
7.	Continuous distributions	67
7.1.	Uniform distribution.....	67
7.2.	Normal distribution	68
7.3.	Student Distribution	70
7.4.	Chi-square distribution.....	71
7.5.	Knowledge check.....	72
8.	Estimation.....	75
8.1.	Estimating the mean (known population variance).....	78
8.2.	Estimating the mean (population variance unknown)	81
8.3.	Estimating population variance	83
8.4.	Knowledge check.....	87
9.	Statistical process control.....	90
9.1.	The histogram	90
9.2.	Pareto chart.....	91
9.3.	Scatter plot	92
9.4.	Control charts.....	94
	9.4.1. Process capability	94
	9.4.2. Elements of a control chart.....	95
	9.4.3. Types of control charts.....	96
	9.4.4. Interpretation of a control chart.....	104
9.5.	Cause-effect diagram	107
9.6.	Process Diagrams	108
9.7.	Knowledge check.....	109
10.	Correlation and regression	112
10.1.	Correlation.....	112
10.2.	Linear regression	114
10.3.	Knowledge check.....	117
11.	References	120
	Annex 1 - Table for normal distribution (z-values).....	122
	Appendix 2 - Student distribution table (t-values).....	123
	Appendix 3 - Chi-square distribution table (χ^2 values)	124
	Annex 4 - Summary of concepts.....	125
	List of figures	149
	List of tables.....	150

Introduction

This textbook is intended for second-year students of the Industrial Engineering and Industrial Economic Engineering specializations of the Faculty of Industrial Engineering, Robotics and Production Management of the Technical University of Cluj-Napoca. It aims to guide students in assimilating the basic information necessary for their engineering education.

This material is structured according to the Statistics and Probability course taught and is divided into chapters. At the end of each chapter there is a knowledge check section and the correct answers for each question. In the first chapter there is an introduction to the basics of data types, how they can be used in statistical analysis, the central tendency of the data and the characterization of their variability. In the second chapter the topic of data visualization and extracting information from data is addressed. Common visualizations used in statistics and other methods such as text and tables are discussed. Then, in the third chapter, students are introduced to statistical events that form the basis of probabilities. Events and event operations are reviewed here. In the fourth chapter some basic concepts and laws in working with probabilities are presented. The fifth chapter introduces the notion of variables and discusses discrete and continuous random variables. This chapter presents some known types of variables and some examples of their use. Then the notion of probability distribution is introduced. Students will then be introduced to some of the most common discrete distributions: uniform, binomial and hypergeometric in chapter six and to common continuous distributions such as the normal, Student and Chi-square distributions in chapter seven. These notions of distributions are then used in solving a common engineering problem, that of estimating population parameters dealt with in chapter eight. Students will learn to use distribution tables and calculate confidence intervals in order to estimate certain population parameters (mean or variance). The next chapter presents the practical application of statistics in engineering and more specifically in Statistical Process Control (SPC). Some tools already presented in previous chapters are put into a practical framework and others are newly introduced. Finally, Chapter 10 deals with the concepts of correlation and linear regression, which are very important concepts of inferential and predictive statistics.

1. Descriptive statistics

Descriptive statistics is a branch of statistics that involves the collection, organization, presentation, and analysis of data. The purpose of descriptive statistics is to describe the characteristics of a data set through measurements such as measures of central tendency and measures of variance. These measures are used to provide a clearer understanding of the data and to highlight patterns and trends. Descriptive statistics are used in many fields, including market research, social sciences, business, and natural sciences.

The Cambridge Dictionary [1] defines data as:

"Information, especially factual descriptions or numbers, collected for examination, analysis and use to aid decision making, or information in an electronic form that can be stored and used by a computer."

Data, information, and knowledge can be defined in many different ways [2] and are considered to be different. There is also a hierarchy of data, information, and knowledge (Figure 1.1) in which data is the basis of information and knowledge is derived from information. According to [3], [4] data are symbols, information is data processed to be useful and answer questions such as 'who', 'what' and 'when' and knowledge is the application of data and information to answer 'how' questions.

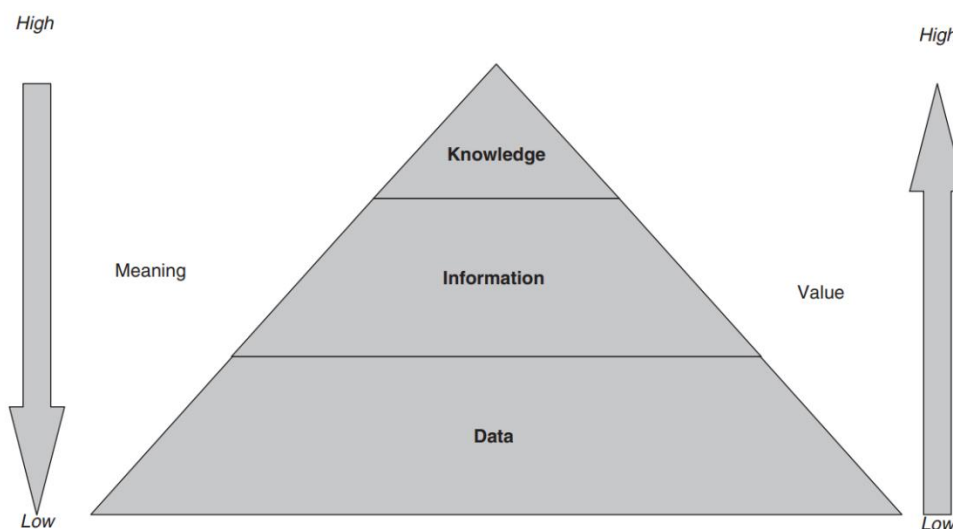


Fig. 1.1. Knowledge-Information-Data Hierarchy [4], [5]

Data sources are all around us and their number is growing as sensors become cheaper and more widely used. Even our smart phones are an aggregation of sensors that continuously collect and analyze acquired data. Companies are becoming more and more data-driven, collecting and analyzing data from both customers and their own production processes, with the aim of gathering insights on which to make better decisions and increase the performance of their businesses.

There are many types of data that can be collected. Some are structured in tables or other data structures, but most are unstructured. Using statistical concepts, we can "dig" into this data and arrive at information that can help us in the decision-making process.

Data can be **numeric** or **non-numeric**. Numeric data are characterized by numbers and usually provide quantitative information. For example, the mass of an object can be expressed in kilograms, which is a numerical quantity. Non-numerical data is usually qualitative and shows certain attributes that cannot be expressed numerically. The color or shape of an object are examples of non-numerical data. Numeric data have a higher information load and can be transformed (if necessary) into non-numeric data. For example, we can describe a wardrobe as weighing 30 kg (numeric information) or simply say it is heavy (qualitative information). Expressing the mass of the cupboard in kilograms tells us exactly how heavy the cupboard is, whereas just saying that it is "heavy" does not give us exact information about its mass, only that we should expect a higher mass.

Depending on how we can characterize a variable (something that can change its value [6]) it can be measured on different levels of measurement: nominal, ordinal, interval, or ratio. Each measurement level adds additional information. The level of measurement is important because it tells us what analysis tools we can use for that variable.

If the measurements of a variable can only be grouped into categories, then we are talking about the **nominal** level of measurement. An item can be in one category or another and this is its characteristic. For example, we have a box of colored pencils and the variable we are looking at is their color. There can be several color categories: black, blue, red, or green, and each pencil has a color belonging to one of these categories. A pencil is considered in one category because it cannot have two colors at the same time (e.g. blue and green). When working with dummy variables, we can count how many observed items we have in each category. The number of items is also called frequency, which we will discuss later in the course.

If the categories can be arranged in an intuitive order, then we say that the data are measured at the **ordinal** level. This gives us one more piece of information about our data, namely order. An example would be the height of Christmas trees by classifying them into one of the categories: *short*, *medium*, or *tall*. As well as knowing that a tree belongs to a certain category, we also know the position of the category in the hierarchy (*short* before *medium* and *medium* before *tall*).

Data measured at nominal and ordinal levels are usually non-numeric. They express **qualitative** or attributive characteristics. Ordinal ones can be given numerical

notations (e.g. instead of small, medium, large we can write 1, 2, 3) but we have to be careful because they are only symbols that help us to observe order and position in the hierarchy. We cannot do calculations with these numbers (addition, subtraction, etc.) In addition to determining the frequency as for nominal variables, for ordinal variables we can assign ranks and do a series of statistical analyzes with ranks.

Numerical data are evaluated using scales and usually have equal intervals. If the scale we use to evaluate a variable does not have a significant "zero", then we are talking about the **interval** level. Significant zero means that the value "0" is just another value on our scale, not symbolizing the lack of a quantity. A common example is temperature. You can measure temperature on a scale with an interval of 1 degree (Celsius for example). Even if we see the value "0" on the scale, this does not mean that if we reach the value 0 we do not have a temperature, just that we have reached another value on the scale. Since we are talking about numeric variables, we can do arithmetic operations with them. Quantities measured on this scale can only be added or subtracted but cannot be divided or multiplied. That's why the expression "*today is twice as cold as yesterday*" is meaningless. Ask yourself, if today is 0°C and tomorrow is twice as hot, what temperature will we have tomorrow? Instead, we can say "*today is 5 degrees colder than yesterday*".

If the scale instead has a "zero" (an origin) signifying the absence of the quantity in question, we are talking about a variable measured at the **ratio** level of measurement. Quantities measured with instruments are usually at this measurement level. Examples of quantities: mass, weight, length, voltage, force, etc. This is the most complete level at which a quantity can be measured. You can do all kinds of operations (addition, subtraction, multiplication, division) and zero means no quantity. For example, if you weigh a bag of potatoes, you can divide it in half or double it, and when you have 0 kg of potatoes, you have no potatoes. Numeric data is also called **quantitative** because it expresses quantities.

Figure 1.2 gives an overview of the different types of data and their measurement levels.

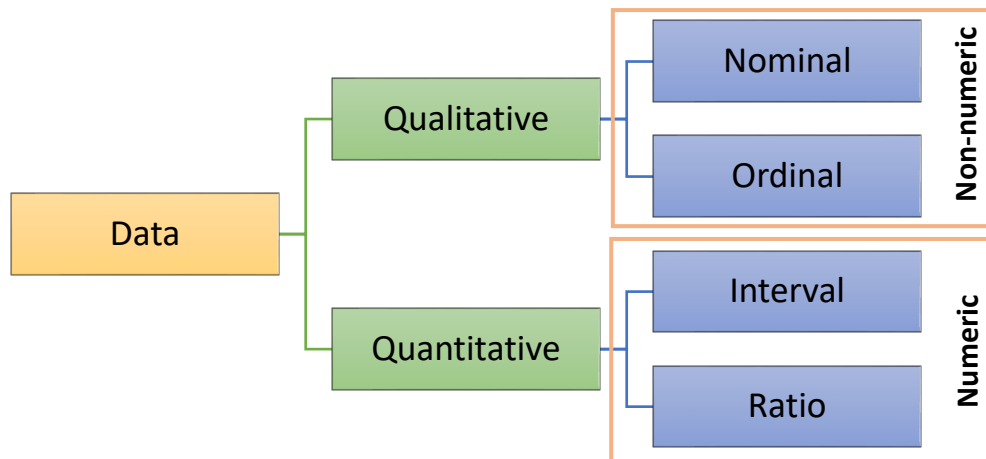


Fig. 1.2. Types of data and their levels of measurement

There are several measures that can be used to describe the data. According to [7] there are four categories:

- Frequency measures (number of occurrences)
- Measures of central tendency (central values around which data accumulate)
- Measures of dissemination (data variation)
- Position measures (position of a value within the data set to which it belongs)

1.1. Frequency

We use the term **frequency** when referring to how often an event occurs or the number of items in a category or interval. We can express frequency by the absolute number of values we observe and say we are determining the **absolute frequency**, or express it as a percentage of the total number of observations, which is the **relative frequency**. This is an important concept that is often used in statistics and data analysis.

Absolute frequency is the absolute number of items in a category or interval or, if we are talking about events, it is the number of occurrences of an event. For example, if we have 8 apples of which 3 are red and 5 are yellow, then 3 and 5 would be the absolute frequencies for the two categories (red and yellow).

Relative frequency is considered in relation to the total number of items (or occurrences):

$$f_i = \frac{a_i}{n}$$

where:

f_i - the relative frequency of the interval or category i

a_i - the absolute frequency of the interval or category i

n - total number of items considered (from all categories combined)

Relative frequency is expressed in fractions or in decimal form, but most commonly we find it expressed in percentages. Continuing the example above, the 3 red apples would represent $\frac{3}{8}$ or 37.5% of the total number of apples. Yellow apples make up the remaining $\frac{5}{8}$ (or 62.5%) of the apples.

Another way to use frequencies (absolute or relative) is to cumulate them. This is called **cumulative frequency**. It can be calculated starting with the first category (increasing) or the last (decreasing). When determining the cumulative frequency in ascending order, we add the frequencies in each category from the first to the category of interest. The increasing cumulative frequency answers the question: "*How many values do we have in the first i intervals/categories?*". In the case of the decreasing cumulative frequency, we proceed in a similar way, but start from the last category/range instead of the first.

Frequencies are usually given in frequency tables. You can see an example in Table 1.1, expanded from the apple example above.

Table 1.1 - Frequency table (absolute, relative, cumulative ascending and descending)

Categories	Absolute frequency	Relative frequency	Cumulative increasing absolute frequency	Decreasing cumulative absolute frequency
Red apples	11	22%	11	50
Yellow apples	19	38%	30	39
Green apples	8	16%	38	20
Red-yellow apples	12	24%	50	12
Total no. of apples	50	100%		

1.2. Central tendency

Central tendency denotes "the tendency of quantitative data to cluster around a central value" [8]. This tendency gave rise to *the central tendency theory*. The central value around which the data cluster is called a measure of central tendency. The most common are:

- **arithmetic mean** (average)
- **median**
- **mode**

Other useful measures are:

- **root mean square,**
- **geometric mean**
- **harmonic mean**

We will look at each in more detail below.

The arithmetic mean is one of the most commonly used measures of central tendency. To calculate it, add all the values and divide by the number of values. It has several notations such as μ for the population mean and \bar{x} for the sample mean, or more generally with M_x .

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

where x_i are the values for which we calculate the arithmetic mean and n is the number of values.

The median is the value that occupies the central place in an ordered set of values and divides the string into two strings of equal length:

$$M_e = x_{(n+1)/2}$$

where n is the number of values in the string.

Given the following set of values:

$$1, 1, 2, 4, 5, 7, 9$$

then 4 is the median value (we have three values before it and three values after it). If the string has an even number of values, then the median is the arithmetic mean of the values occupying the middle positions of the ordered string:

$$M_e = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Compared to the arithmetic mean, the median has the advantage that it is not as sensitive to outliers (values very different from the rest of the string). For example, for the string:

$$1, 2, 3, 4, 500$$

The arithmetic mean equals 102 while the median is 3, a much more representative value as a central value than the arithmetic mean.

Modal is the value (or category) with the highest frequency. The easiest way to determine the modality is graphically. If we plot the apple frequencies from the previous example, we get the graph in Figure 1.3. The category with the highest frequency contains the modality. In this case, mode is in the yellow apple category.

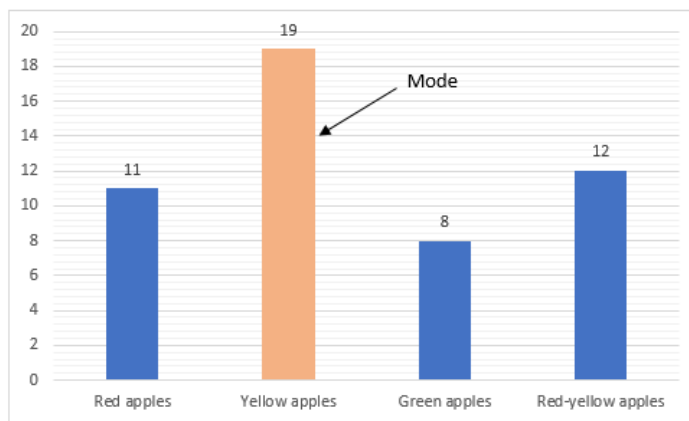


Fig. 1.3. Modal determination by graphical method

In some cases, other measures would be more appropriate to convey the central tendency. The most common are quadratic, geometric, and harmonic averages.

The root **mean square** is the square root of the mean of the squared values. This means that we must first square each value, determine the mean, and extract the square root. The formula looks like this:

$$M_p = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$$

where x_i are the values for which the mean is calculated, and n is the number of values.

The geometric mean can be calculated by multiplying all the values and extracting the n th degree radicals:

$$M_g = \sqrt[n]{\prod_{i=1}^n x_i}$$

where x_i are the values for which the mean is calculated, and n is the number of values.

The geometric mean is used, for example, in finance to determine the average growth rate (or compound annual growth rate). If you want to calculate the average growth rate of an investment over several years, you can use the geometric mean of the annual growth rate to obtain a central value of the growth rate. This is used instead of the arithmetic average which in this case would not be used correctly. It can also be used to calculate the rate of return on an investment over a longer period. For example, if you want to assess the performance of an investment fund over several years, you can use the geometric average of the annual rates of return to get a more accurate picture of its long-term performance. Another example is calculating the average population growth rate. With the geometric mean you can assess long-term demographic trends or forecast future population growth. In general, the geometric mean is useful when a

central measure of a data set containing positive numbers is desired and when it matters whether larger values have a greater impact on the mean than smaller ones.

The harmonic mean is calculated by dividing the number of values by the sum of the inverses of the values:

$$M_a = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

where x_i are the values for which the mean is calculated, and n is the number of values.

The harmonic mean is useful when a central measure of a data set containing positive numbers is desired and when it matters whether smaller values have a larger impact on the mean. The harmonic mean can be used to calculate the average velocity of a moving object. For example, if an object moves at a speed of 40 km/h for one hour, then stops and moves at a speed of 20 km/h for one hour, the average speed of the object over the two hours is the harmonic mean of the two speeds. It can also be used to calculate the average rate of return of an investment portfolio over one year. For example, if an investment portfolio achieved a return of 10% in the first quarter, 5% in the second quarter and 15% in the third quarter, the portfolio's average rate of return for the whole year is the harmonic average of the quarterly rates of return. Another example is for the calculation of average test or exam grades. One can evaluate a student's overall performance by considering that lower grades may have a greater impact on the average than higher grades.

1.3. Variation or spread of data

The data we collect have natural variation, which means they differ from each other. Measures of central tendency cannot capture this variability in the data. For example, we may have a group of students who have the same average for two different subjects (Table 1.2). However, the values for the two subjects are not spread in the same way which can be observed. It is therefore necessary to use measures of spread or variation to get a more accurate understanding of the data.

Table 1.2. A group of students with the same average in 2 different subjects

	Subject 1	Subject 2
Student 1	5	6
Student 2	9	7
Student 3	4	7
Student 4	8	8
Student 5	9	7
Average	7	7

The most used measures of spread are:

- Range
- Interquartile range
- Variance
- Standard deviation

Range is the difference between the maximum and minimum value:

$$R = x_{max} - x_{min}$$

A major advantage of the range is its ease of calculation. It can be used, for example, when you want to determine the difference between the smallest and the largest value. For example, when analyzing the variability of maximum and minimum temperatures in an area. Another example would be to compare the variability of two sets of data, such as the variability of wages in two different companies. The range can also be used when you want to identify possible extreme points in a data set. For example, it can be used when examining the age distribution of employees in a company to identify if there are exceptionally young or old employees.

The disadvantage is that it only considers two values and this makes it sensitive to extreme values (also known as *outliers*).

The interquartile range (IQR) is a measure of variance that measures the difference between the third quartile (75%) and first quartile (25%) of a data set, i.e. the difference between the upper median value and the lower median value. The three quartiles are the values that divide a data set into four equal parts (quartiles). These are:

- First quartile (Q1): this is the value that divides the data set into two parts, so that 25% of the data is less than this value and 75% is greater than it.
- Median or second quartile (Q2): this is the value that divides the data set into two equal parts, so that 50% of the data are less than this value and 50% are greater than it.
- The third quartile (Q3): is the value that divides the data set into two parts, so that 75% of the data is less than this value and 25% is greater than this value.

The formula for calculating the interquartile range is:

$$IQR = Q_3 - Q_1$$

The interquartile range can be used when we want to assess the variability of a data set and eliminate the influence of extreme points. In some applications it is preferable to use IQR instead of the range because it eliminates the variation in extreme points. In addition to applications where one wants to compare the variability of two data sets, it can also be used to identify and eliminate extreme points in a data set. For

example, if you are analyzing the age distribution of employees in a company and want to identify employees who are much older than average or much younger than average.

Generally, the interquartile range is a more robust measure of spread (variation) than the range and is useful in situations with extreme points or significant differences between the means of two data sets.

Variance is another measure of data variance. Unlike the range and interquartile range, it uses all the values in the data set under analysis in the calculation formula.

To calculate the variance, we must first determine the difference between each point and the mean. We then square these differences and average them. The formula for population variance, denoted by σ^2 looks like this:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

where x_i are the values for which the average is calculated, μ is the population mean and n is the number of values.

In the case of sample variance, denoted by s^2 , the formula is corrected by a correction factor that takes into account that the variance is calculated on only a subset of the population:

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

where x_i are the values for which the average is calculated, \bar{x} is the sample mean and n is the number of values.

Variance can be used in various applications such as economics, where it can be used to assess price volatility in financial markets. If a stock has a high variance, this suggests that the price changes often, which can be risky for investors. It can also be used in medicine to assess changes in physiological or laboratory parameters such as blood pressure or blood glucose levels. This information can be useful to assess the effectiveness of a treatment or to identify risk factors for certain conditions. In engineering variance can be used to assess the quality of output in a manufacturing process. A high variance may indicate problems with the production process, such as faulty materials or an inadequate production line. In psychology it can be used to assess differences between individuals in their traits or behaviors, such as to assess the degree of diversity in responses to a personality questionnaire. Another example might be to assess student performance in a class or school. A large variance may indicate significant differences in student performance or problems in the teaching process.

A disadvantage of variance is that the unit of measure in which it is expressed is the square of the unit of measure of the data for which the variance is calculated. We

can easily solve this problem by taking out the square root of the variance. This new indicator is called the standard deviation.

The standard deviation is simply the square root of the variance. This means that we have the following two formulas:

For population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$$

And for the standard deviation of the sample:

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

where x_i are the values for which the average is calculated, μ is the population mean, \bar{x} is the sample mean and n is the number of values.

2. Data presentation and synthesis methods

There are several ways of presenting data, such as using text, tables, and visual methods. Text is the simplest method and can be used to present data in sentences or phrases. Using statistical indicators we can capture and present synthetically certain characteristics of the data, such as the central tendency or spread of the data. Tables involve organizing data into a tabular structure or matrix and can be used to highlight differences and similarities between different data sets. Visual methods include graphs and charts, which can be used to provide a visual representation of data. These can be useful for highlighting trends, patterns and distributions in the data and can provide a clearer picture of the data. Each method has its own advantages and disadvantages, and the choice of method depends on the type of data and the purpose of its presentation.

2.1. Text and statistical indicators

Text and statistical indicators are two important ways to communicate information effectively. Text can be used to explain the context of the data and provide an overview of the problem being addressed. Statistical indicators can provide a clear picture of the distribution of the data and its variation. By using these two approaches together, we can provide the reader with a complete picture of the data and the important conclusions that can be drawn from it.

Statistical indicators are values that provide valuable information about the data we are analyzing. We can use measures of central tendency and variance to get overall information about our data. Table 2.1 summarizes the statistical indicators that are representative in determining the central tendency, spread of the data and characterizing the shape of the data distributions.

Table 2.1. Representative statistical indicators

Measure	Indicator	Formula
Central tendency	Arithmetic mean	$\mu = \frac{\sum_{i=1}^n x_i}{n}$
	Median	Value in the center of the string
	Mode	Maximum frequency
	Root mean square	$M_p = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$
	Geometric mean	$M_g = \sqrt[n]{\prod_{i=1}^n x_i}$

	Harmonic mean	$M_a = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$
	Central value	$x_c = \frac{Max - Min}{2}$
Spread	Min	Lowest value
	Max	Highest value
	Range	$R = x_{max} - x_{min}$
	Variance	$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$
	Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$
Form of distribution	Asymmetry	$g_1 = \frac{\sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$
	Kurtosis	$k = \frac{\sum_{i=1}^n (x_i - \mu)^4}{(\sum_{i=1}^n (x_i - \mu)^2)^2}$

There are, however, limitations to the amount we can effectively convey through text. When we go beyond a certain level of complexity, presenting data as text makes it difficult to understand and convey the intended message. That's why we use tables and visualizations to convey more information effectively.

2.2. Tables

Tables are an important and frequently used means of presenting data in an organized and easy-to-read way. They can be used to highlight differences and similarities between different data sets, as well as to provide an overview of the data. They can be useful in identifying trends and patterns present in the data and can be a valuable resource for analysis and interpretation.

Creating a table starts with selecting the relevant dataset for analysis. The data should be organized in a coherent way so that the information can be easily read and interpreted. The information is organized into columns and rows, and the column names should be chosen to be meaningful so that the data in each column can be easily identified. Usually, the variables (characteristics) we measure are in columns and the observations (measurements) are in rows. In the example in Table 2.2, toy 1 is round, 23

mm long and red. These are all characteristics of a single object. On the other hand, we see that all toys have only two possible shapes: round and square.

We can either look at one feature of all the items in the set (in a column) or at all the features of a single item (in a row) depending on the information we want to use.

Table 2.2 Example of a table

	Form	Length (mm)	Color
Toy 1	Round	23	Red
Toy 2	Round	27	Green
Toy 3	Square	19	Blue

The format of the tables is important because it can influence how the data is perceived and interpreted. The use of colors and formatting is recommended to draw attention to important information. Overcrowding of tables with too much information should also be avoided. Information should be presented concisely so that users can quickly identify relevant information.

Interpreting a table involves examining the data and identifying trends and patterns. For example, a person may examine a table to identify variation in data over time or differences between two groups of data. Tables can also be used to identify frequencies or proportions of different values in a data set.

When we have little information (such as the example above) it can easily be presented in a table. But when we have large tables with dozens or hundreds of variables and thousands of rows, it can be almost impossible to get information just by looking at the table. It is therefore important to consider the advantages and disadvantages of using tables according to the specific needs of the user and the purpose of the analysis. Sometimes other forms of data presentation, such as graphs or charts, can be more effective than tables in presenting data.

2.3. Viewing data

Data visualization is a powerful method of communicating information that can be used to highlight important trends and patterns. Graphs and charts can be created to illustrate relationships and distributions in the data collected, as well as to highlight variation and differences between groups of data. They can be presented in a variety of formats, such as bar charts, line charts, dot plots and pie charts, depending on the type of data and the information we want to highlight. By using data visualization in a creative and intelligent way, we can provide complex information in an accessible and easy-to-

understand way, which can be extremely valuable in communicating important findings and conclusions. The most used types of graphs are:

1. **Bar chart:** used to compare quantities or frequencies. It consists of a set of vertical or horizontal bars representing the values of variables.
2. **Line chart:** used to illustrate trends and changes in a series of data. It uses a set of points connected by lines, representing the values of variables in a chronological or logical order.
3. **Dot plot:** used to visualize the distribution of data and to identify outliers for two variables.
4. **Pie chart:** used to illustrate the proportions of a whole. It consists of a circle divided into sections, where each section represents a proportion of the whole.
5. **Histogram:** used to illustrate the distribution of continuous data. It consists of a set of bars representing the range of values of the variables.
6. **Boxplot:** used to easily visualize and compare distributions of continuous data.

Next, we will go into more detail for each type of chart and see how they are constructed and when we use them.

2.3.1. Bar chart or columns

The bar chart is an effective way to present data visually. It consists of a set of vertical bars (also called columns) or horizontal bars representing the values of variables. The bar chart is usually used to compare frequencies of variables or to highlight trends in the data.

The bar chart is useful when working with discrete qualitative or quantitative data, where variables are represented by categories or integers. It is recommended to use the bar chart in situations where we have at least four or five categories.

To create a bar chart, follow these steps:

1. Identify the variable you want to represent and its categories.
2. Decide on the width of the bars. The bars must be the same width.
3. Draw a vertical or horizontal axis and mark the values of the variables on this axis.
4. Draw the bars with the appropriate height for each category.

An example of using the bar chart would be to represent the frequencies of occurrence of certain colors in a set of objects. In this case, the variable would be 'color' and the categories would be the possible colors. The bar chart shows the frequency with which each color appears and can highlight preferences or trends in color choice.

There are several types of bar charts, such as single bar chart, multiple bar chart or paired bar chart. Single bar charts can be with columns (Figure 2.1) or horizontal bars (Figure 2.2). These types can be used depending on the purpose and characteristics of the data.

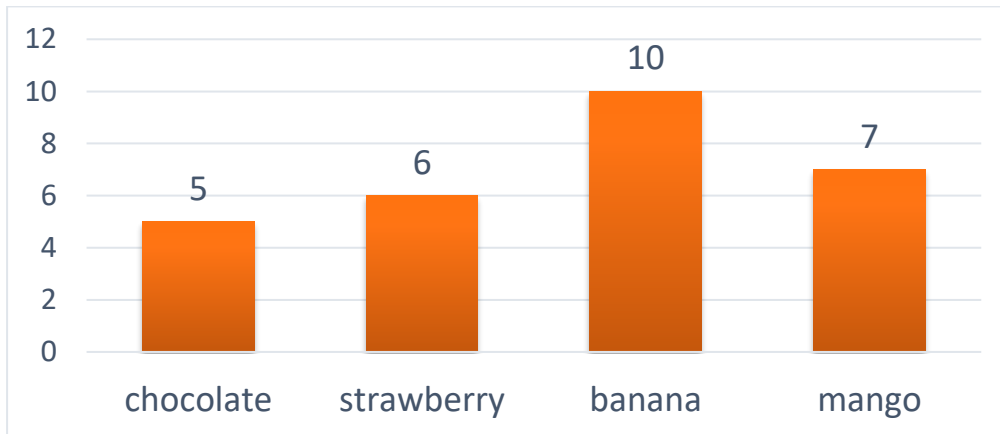


Fig. 2.1. Simple column diagram

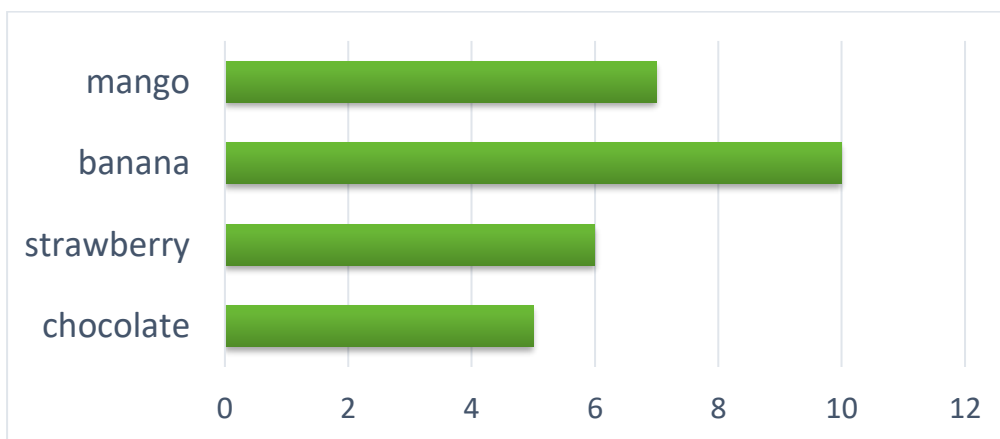


Fig. 2.2. Simple diagram with horizontal bars

In a diagram we can show several instances of the same variable. For example, let's say we sell ice cream in two stores, and we want to compare the sales for the two stores based on the flavors on sale. We can create a chart like the one in Figure 2.3 in which we compare sales for each store based on the flavor sold.

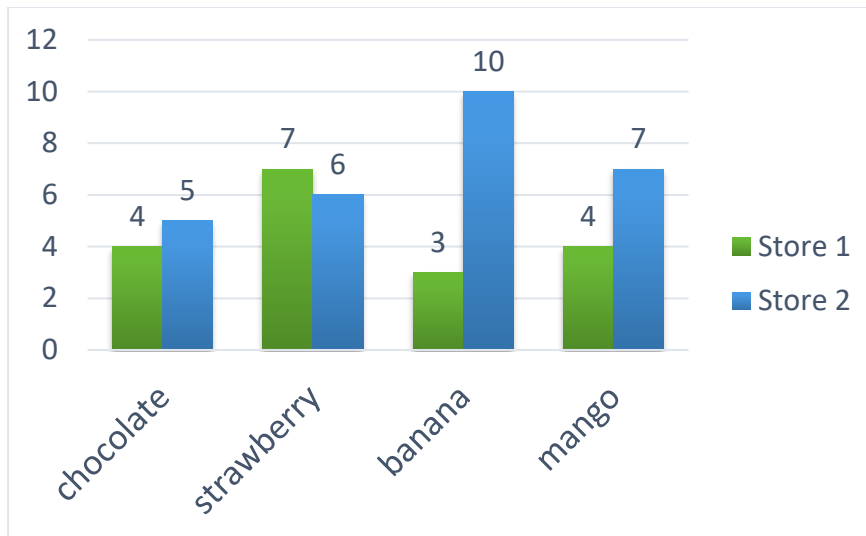


Fig. 2.3. Multi-instance column chart

The same information can be presented as a proportion of a total in the form of stacked columns (Figure 2.4). For each flavor sold we can see what proportion is sold in shop 1 and shop 2 respectively.

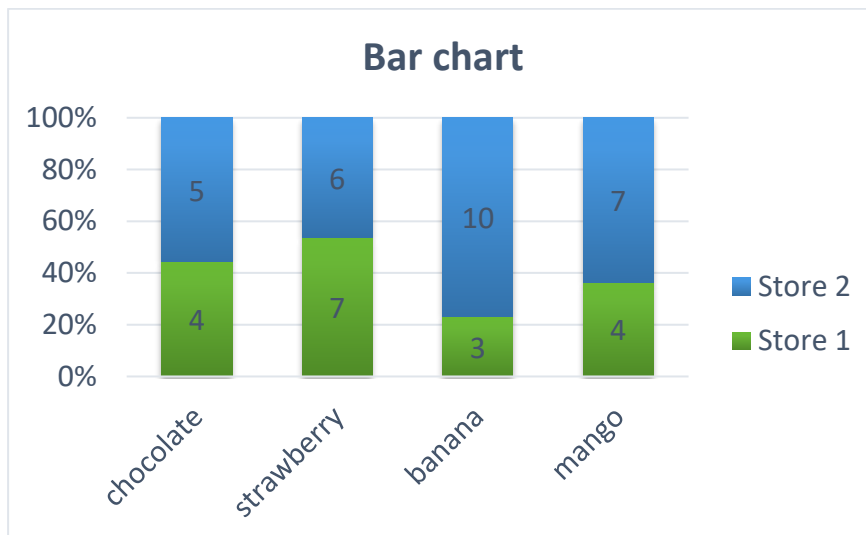


Fig. 2.4. Stacked columns diagram

2.3.2. Line diagram

Line charts are an effective way of showing the evolution of a series of data over a period of time.

To create a line chart, we need to have the data we want to present and organize it in a table with two columns: one for the time values (or an index corresponding to a timeline) and one for the values of the variable of interest. We then draw a coordinate system, where the horizontal axis represents time, and the vertical axis represents the values of the variable of interest. Based on these coordinates, we can mark the points for each value and join them with a line to create the diagram.

There are many examples where line charts are used, such as to show the evolution of average temperatures in a particular city over the year or to track the growth or decline of a company's sales over a given time frame. This graph can also be used to show trends or fluctuations in data collected in a scientific study. For example, Figure 2.5 shows the evolution of the daily maximum price in USDT of a cryptocurrency (BTC) over the period January-February 2022.

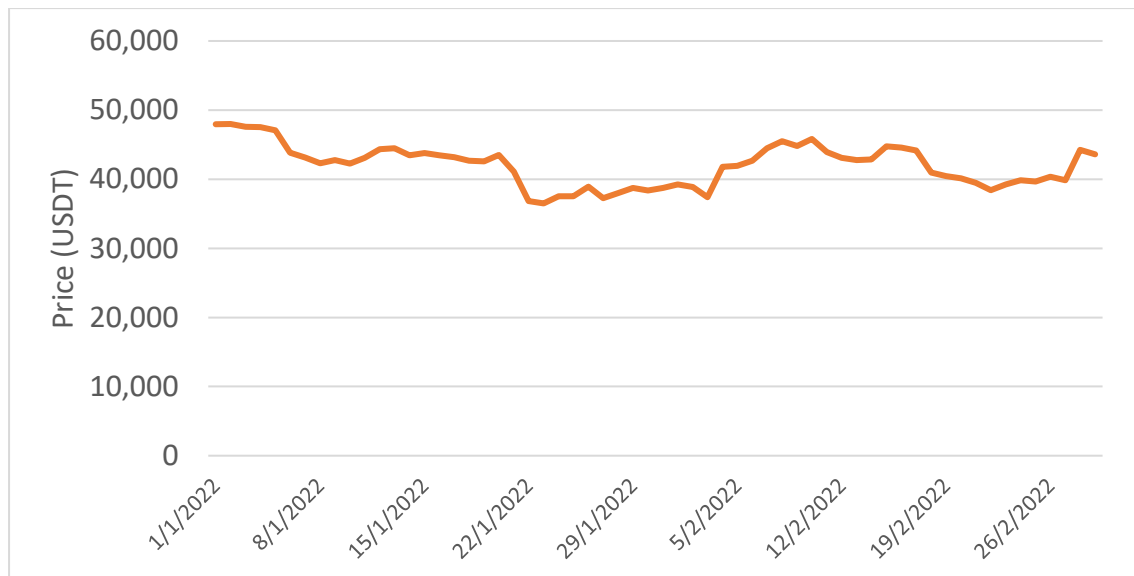


Fig. 2.5. Line chart

In general, line charts are an effective way to communicate complex data in an accessible and easy-to-understand way.

2.3.3. Scatter plot

The dot plot is a graphical representation of data that allows the relationship between two continuous variables to be visualized. This type of graph is mainly used in regression analysis but can also be used to observe the distribution and variation of data.

To create a dot plot, it is necessary to collect the data of the two variables, determine which variable is on the x-axis and which on the y-axis, and place each pair of values on the plot. Usually, a point or symbol is used for each pair of values.

The dot plot can be useful in many situations, such as:

- To analyze the relationship between two variables. If the points are scattered randomly on the graph, then the two variables are not correlated with each other. If the points form a line shape, this indicates a linear relationship between the two variables.
- To identify outliers. Points that are farther apart from each other can easily be seen on the graph.

- To observe data variation. If the points are spread evenly over the graph, then the distribution of the data is homogeneous. If the dots are denser in a particular area, this indicates less variation in that region.

For example, we are interested in looking at the association between height (in centimeters) and mass (in kilograms) of a group of people. After collecting the data, we place height on the horizontal axis and mass on the vertical axis. For each person we have a pair of height-mass values (e.g., 157 cm and 44 kg) plotted as a dot (Figure 2.6). By plotting these dots, we can observe associations or deviations. In this case we can see that as height increases, mass increases: usually taller people have more mass. There are also exceptions, such as the last person (marked on the graph with a purple square) who, although taller, has a lower mass.

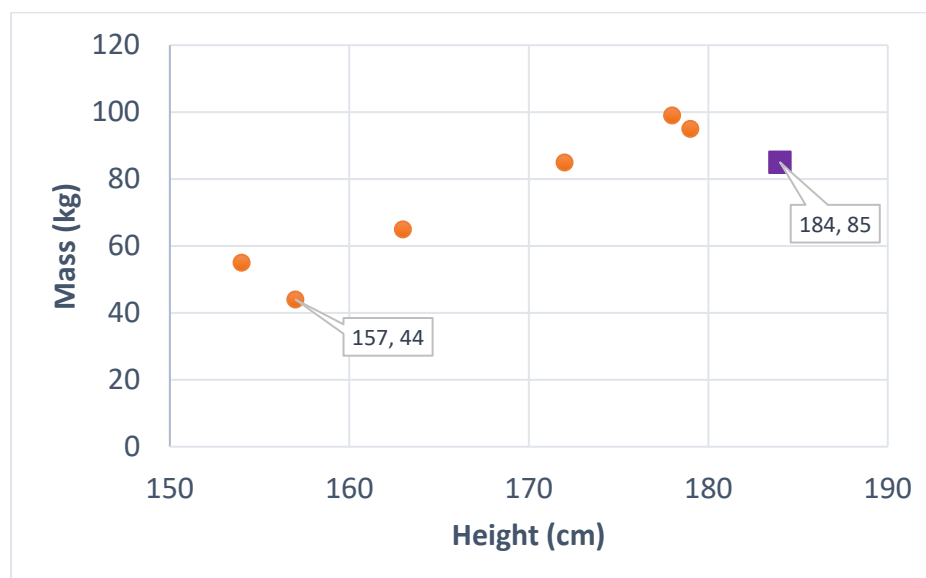


Fig. 2.6. Scatter plot

The dot plot is a useful method of visualizing data and can be used in many fields, including statistics, economics, biology, and others.

2.3.4. Pie chart

Pie chart, also known as a proportion chart, is a type of chart commonly used to show the proportion of each category of data in a data set by representing each category as a pie chart.

The pie chart is useful when we want to visualize the proportions (relative frequency) of different categories in a data set, rather than the absolute number (absolute frequency). This is useful because it allows a quick and easy visualization of the distribution of the data and can be used to make comparisons between different categories.

To create a pie chart, you need to:

1. Calculate the proportion of each category in the dataset.
2. Convert the proportions into degrees by multiplying by 360.
3. Draw a circle and divide it into sectors so that the angle of each sector corresponds to the proportion for that category.
4. Label each sector with the name or label corresponding to the category.

The pie chart can be used in many different areas, including:

- In business, to illustrate the proportion of sales per product or to show how money is spent in a company.
- In research, to illustrate the distribution of different answers to a question or to show the demographic distribution of a population.
- In education, to illustrate the proportion of students learning different subjects or to show the distribution of grades in a class.

Let's say we're doing a survey, and we want to find out the pizza preferences of a group of students in the dorm. Once we collect the data, we can present it as a pie chart, as in figure 2.7. From the figure we can see that the most preferred pizza by the students is Marguerita with 30% of the responses. Although we do not visualize the absolute frequencies, we can see how the proportions vary between the different categories of a whole.

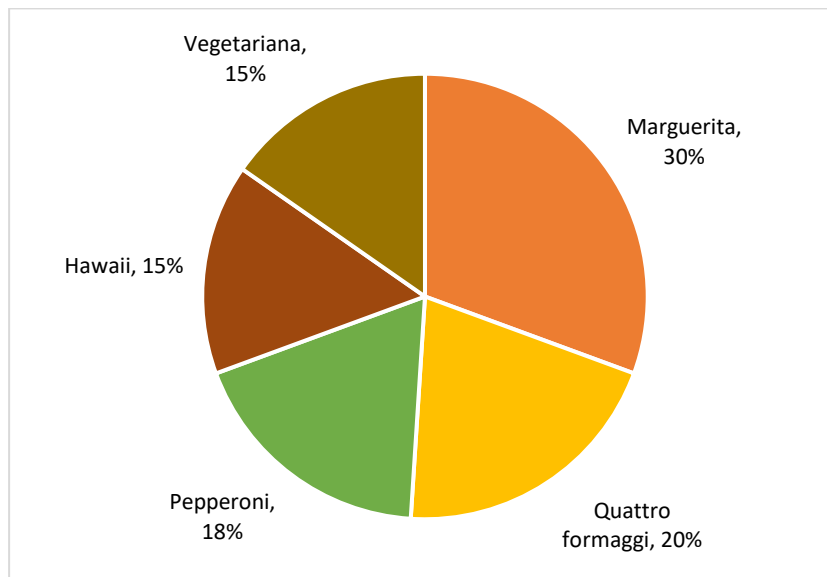


Fig. 2.7. Pie chart

In general, the pie chart can be used to illustrate proportions and distributions in a visual and understandable way.

2.3.5. Histogram

A histogram is a bar chart showing the frequency with which certain values occur in a continuous data set. With this type of graph, we can visualize the frequency

distribution of a continuous variable. For example, we can use a histogram to see how the ages of a population are distributed in age ranges.

To create a histogram, you need to:

1. We determine the number of intervals best suited to the data we want to view
2. We divide the range by the number of intervals
3. We calculate the ends of each interval
4. We determine the frequency with which the values occur in each interval
5. We draw the bars according to the determined frequency.

The number of intervals can affect how the distribution looks. There are several methods for choosing the optimal number of intervals, the most used of which are:

- **Square root rule:** The optimal number of intervals (K) is approximately radical of the number of observations (n): the formula is:

$$K = \sqrt{n}$$

- **Sturges' rule:** The optimal number of intervals is given by the formula:

$$K = 1 + 3.322 * \log_2 (n)$$

- **Rice's rule:** The optimal number of intervals is given by the formula:

$$K = 2\sqrt[3]{n}$$

Whichever method we use, the number of intervals must be the nearest integer to the result obtained from the calculation. We can't have 4.3 intervals; we have either 4 or 5 intervals.

An example of using a histogram is to show the weight distribution of individuals in a group of adults. The histogram can show whether the distribution is normal or not, as well as whether there are outliers that could be anomalies or measurement errors. Normal distribution will be discussed later in the course when we look at types of distributions.

Assume we have the following set of data for the weights of the adults in the study:

70, 68, 73, 64, 72, 69, 76, 77, 75, 71, 68, 70, 72, 73, 74, 70, 71, 75, 73, 72

To create the histogram, we will use Sturges' rule for the number of intervals, we will go through the following steps:

- a. We determine the number of observations in the data set: $n = 20$
- b. We calculate the optimal number of intervals using Sturges' rule:

$$K = 1 + 3.322 * \log_2 (n) \approx 5.32 \approx 6$$

c. We divide the range of the values into 6 equal intervals:

$$A = 77 - 64 = 13$$

$$d = \frac{13}{6} = 2.166$$

d. Using d we determine the ends of the intervals

Table 2.3. The 6 ranges and the limits for each range

Interval	Limit
Interval 1	64 - 66.17
Interval 2	66.17 - 68.33
Interval 3	68.33 - 70.5
Interval 4	70.5 - 72.67
Interval 5	72.67 - 74.83
Interval 6	74.83 - 77

e. We determine the frequency with which the values occur in each interval

Table 2.4. Frequency table for the 6 intervals

Interval	Limit	Frequency
Interval 1	64 - 66.17	1
Interval 2	66.17 - 68.33	2
Interval 3	68.33 - 70.5	4
Interval 4	70.5 - 72.67	5
Interval 5	72.67 - 74.83	4
Interval 6	74.83 - 77	4

f. We draw the bars according to the determined frequency (Figure 2.8)

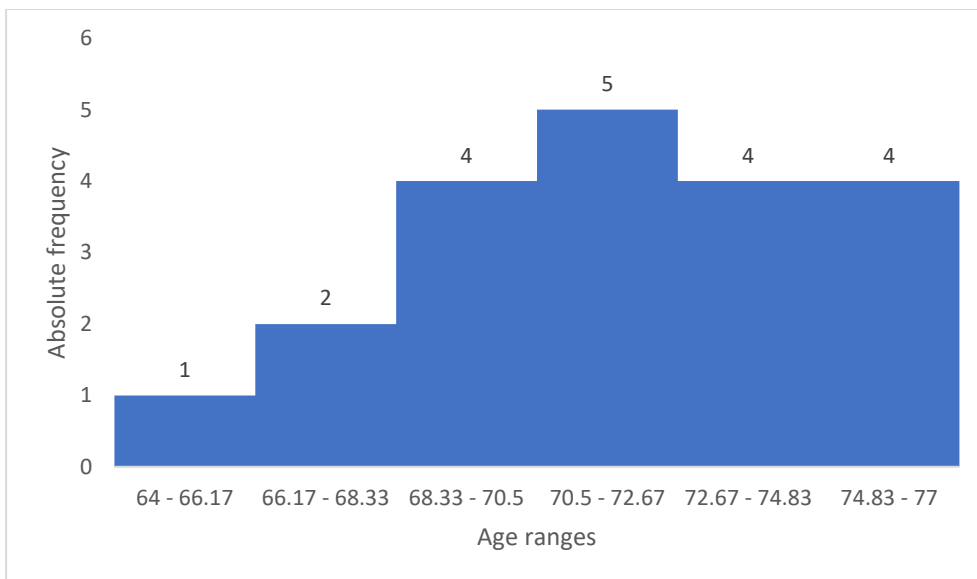


Fig. 2.8. Age distribution in age ranges

We note that most of the values are concentrated around the range 70.50 - 72.67. The histogram can help us identify and eliminate outliers that could be anomalies or measurement errors.

2.3.6. Boxplot

A boxplot is a graph that gives an overview of the distribution of a continuous variable and can be used to identify outliers and compare the distributions of several groups.

To create a boxplot, the following steps are required:

1. Determine the minimum, maximum and median (central value of the data) values for the variable of interest.
2. Calculation of quartiles Q1 and Q3.
3. Calculate the interquartile range (IQR), which is the difference between Q3 and Q1.
4. Identify outliers (outliers), which are data that are more than 1.5 times IQR away from Q1 or Q3.
5. Create the graph, which consists of a rectangle representing the interquartile range (Q1 - Q3), a vertical line representing the median, and segments extending from the rectangle to the smallest and largest points that are not outliers. Outliers are represented by dots or circles.

Boxplot can be used in several cases, including:

1. **Comparison of distributions:** Used to compare the distributions of two or more continuous variables. This can be useful to see if there are significant

differences between the distributions and whether they have the same shape.

2. **Identifying outliers:** Outliers can be significant in data analysis as they may indicate errors or problems in data collection or recording.
3. **Viewing skewed distributions:** Provides an overview of the median and interquartile range. This can be important in identifying differences between symmetric and asymmetric distributions.
4. **Comparison of several groups:** The boxplot can be used to compare the distributions of multiple groups, whether they are experimental, control or other groups. It can help identify significant differences between groups and can be useful in the process of interpreting the data.

For example, using this type of graph we can compare the height of students in two classes (Figure 2.9). From the graph we see that students in class B are generally taller, but we also see that we have a more spread-out distribution of values.

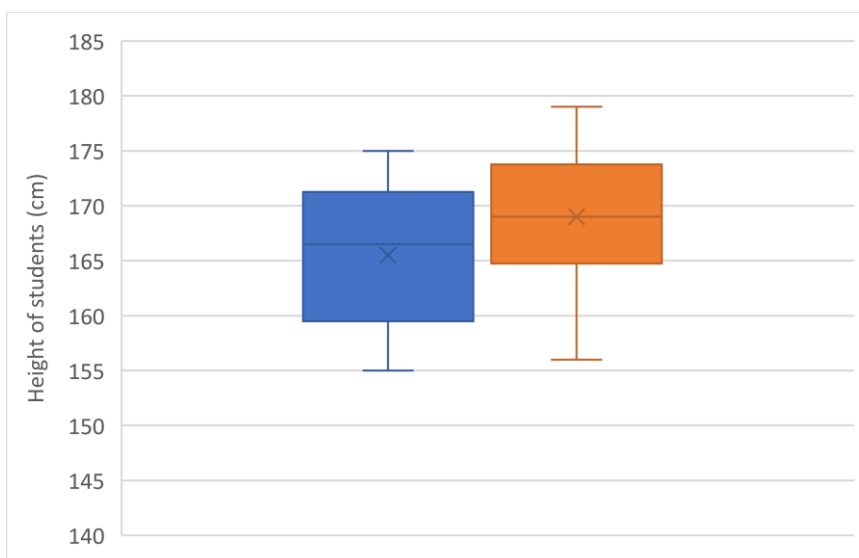


Fig. 2.9. Comparison of height distributions of students from two different classes (Class A – left, class B – right)

The boxplot is a useful tool to visualize the distribution of a continuous variable and can be used in many different situations to compare and visualize distributions of continuous variables.

In addition to the types of graphs presented in this chapter, there are many other types of graphs that can be used to visualize data. Geographical data can be visualized on maps, the evolution over time of some variables can be visualized using spaghetti graphs or we can visualize different outcomes and probabilities related to an experiment using the tree graph. However, the graphs shown are the most common types of graphs and form the basis for other types of visualizations. The important thing is to choose the right type of graph depending on the nature of the data you want to represent.

2.4. Knowledge check

Types of data

1. What are the four levels of data measurement?
 - a) Nominal, ordinal, interval, statistical
 - b) Numeric, nominal, ordinal, cardinal
 - c) Nominal, ordinal, interval, ratio

2. What is the biggest difference between interval and ratio data?
 - a) Interval data have an absolute zero, which is not the case for ratio data.
 - b) Ratio data have an absolute zero, which is not the case for interval data.
 - c) There are no significant differences between interval and rational data.

3. If data were collected on people's opinion of satisfaction with a new product (on a scale of 1 to 5), what level of measurement would this data be?
 - a) Nominal level
 - b) Ordinal level
 - c) Interval level

4. What level of measurement would data on the marital status of some subjects be?
 - a) Nominal level
 - b) Ordinal level
 - c) Interval level

5. If data is collected on the temperature of a place at different times of the day, what level of measurement would this data be?

6. What is the difference between nominal and ordinal data?

7. What kind of date would a person's date of birth be?

8. What level of measurement would be the data on a test score?

9. What kind of data would be the height of a person measured in cm?
 - a) Nominal level
 - b) Ordinal level
 - c) Interval level
 - d) Ratio level

10. What level of measurement would be the ranking data from a sports competition?
 - a) Nominal level
 - b) Ordinal level
 - c) Interval level
 - d) Ratio level

Correct answers

1. c) Nominal, ordinal, interval, ratio
2. b) Ratio data have an absolute zero, which is not the case for interval data.
3. b) Ordinal level
4. a) Nominal level
5. Answer: Data about the temperature of a place is interval level.
Answer: Nominal data is used to identify or classify items into categories, while ordinal data also has a natural order.
7. Answer: Date of birth is interval level because year 0 does not mean no time.
- 8, Answer: The data about grades is ratio level.
9. d) Ratio level
10. b) Ordinal level

Statistical indicators of central tendency

1. What is the central tendency and how is it calculated?
2. What is the average and how is it calculated?
3. What is the median and how is it calculated?
4. What is the mode and how is it calculated?
5. What is the difference between the mean and median of a data set?
6. Which of the following is not a measure of central tendency?
 - a. Media
 - b. Mode
 - c. Variance
 - d. Median
7. Which of the following is the most appropriate measure of the central tendency of a nominal variable?
 - a. Media
 - b. Mode
 - c. Median
 - d. All of the above
8. What happens to the average if a large extreme value is added?
 - a. Average increases
 - b. Average decreases
 - c. Average does not change
 - d. Average becomes negative
9. What is the median of the data set: 3, 7, 9, 11, 12 ?
 - a. 7
 - b. 9
 - c. 11
 - d. 10
10. What is the modal of the data set: 2, 2, 3, 5, 5, 6, 7?
 - a. 7
 - b. 5
 - c. 2
 - d. 6

Correct answers

1. The central tendency is a measure of the position of the values of a distribution and is used to represent their central value. The most used measures of central tendency are the mean, median and mode.
2. The mean is a measure of central tendency that represents the sum of the values in a distribution divided by the number of values. The mean may be affected by extreme values, called outliers.
3. The median is a measure of central tendency that represents the value in the middle of an ordered increasing or decreasing distribution. If the number of values is even, the median is the average of the two central values.
4. The mode is a measure of central tendency that represents the value that occurs most frequently in a distribution. If all values occur with the same frequency, the distribution is bimodal or multimodal.
5. The arithmetic mean of a data set is determined by dividing the sum of the values by the number of values, while the median is the middle value. The mean is affected by extreme values, while the median is not.
6. c. Dispersal
7. b. Mode
8. a. Average increases
9. b. 9
10. b. 5, as this is the value that occurs most frequently in the distribution, three times.

Statistical indicators of spread

1. What is the definition of the spread indicator "range"?
2. How is the interquartile range calculated and what information does it provide?
3. What is variance and how is it calculated?
4. What is the relationship between variance and standard deviation?
5. When should we use variance instead of standard deviation?
6. Which of the following statements is true for the interquartile range?
 - a. Represents the difference between the minimum and maximum value of the data set.
 - b. Represents the difference between the third and first quartiles.
 - c. Represents the difference between the mean and median of the data set.
 - d. Represents the difference between two consecutive values in the data set.
7. Which of the following statements is false about the standard deviation?
 - a. Represents the square root of variance.
 - b. Measures the variability or spread of values relative to the mean.
 - c. It is expressed in the same units as the average.
 - d. It is always less than or equal to the variance.
8. If all values in a data set are equal, then:
 - a. The variance is zero, but the standard deviation is not zero.
 - b. The standard deviation is zero, but the variance is not zero.
 - c. Both variance and standard deviation are zero.
 - d. Variance and standard deviation cannot be calculated in this situation.
9. If we add an extremely large or extremely small value to a data set, how will the standard deviation change?
 - a. It will increase
 - b. It will decrease
 - c. It will not change
 - d. It will become negative
10. Which statement is true about variance and standard deviation?
 - a. Variance and standard deviation are always equal.
 - b. Variance and standard deviation may be equal in some situations.
 - c. Variance and standard deviation are always different.
 - d. Variance and standard deviation are measured in the same unit.

Correct answers

1. The range is the difference between the maximum and minimum value of a data set.
2. The interquartile range is calculated by the difference between the third and first quartiles and provides information about the spread of values in the middle of the data set.
3. Variance is the average of the squares of the differences between each value in the data set and its mean and is calculated by dividing the sum of these differences by the total number of values in the data set.
4. Variance is the square of the standard deviation.
5. Variance is preferable when we want to compare the variation between two or more data sets that have different units of measurement, while standard deviation is preferable when we want to compare the variation between two or more data sets that have the same unit of measurement.
6. b. Represents the difference between the third and first quartiles.
7. d. It is always less than or equal to the variance.
8. c. Both variance and standard deviation are zero.
9. a) It will grow.
10. b) Variance and standard deviation may be equal in certain situations.

Viewing data

1. What is a line chart and when is it best to use it?
2. How can we compare two variables using a dot plot?
3. What is a bar chart and when is it useful?
4. When is a dot plot useful?
5. What type of graph is best for showing trends over time?
 - a. Pie chart
 - b. Line diagram
 - c. Histogram
 - d. Dot diagram
6. What is the best way to visualize the distribution of a continuous variable?
 - a. Pie chart
 - b. Line diagram
 - c. Histogram
 - d. Dot diagram
7. What is a pie chart and when is it useful?
 - a. A graph showing the distribution of proportions of a variable
 - b. A graph showing trends over time
 - c. A graph comparing two variables
 - d. A graph showing the relationship between two variables
8. What is a dot plot and how can it be used?
 - a. A graph showing the distribution of a variable
 - b. A graph comparing two variables
 - c. A graph showing the relationship between two variables
 - d. A graph showing trends over time
9. What is a boxplot and how can it be used to analyze data?
 - a. A graph showing the distribution of a variable
 - b. A graph comparing two discrete variables
 - c. A graph showing the relationship between two variables
 - d. A graph showing trends over time
10. What is the best way to show the distribution of a nominal variable?
 - a. Pie chart
 - b. Line chart
 - c. Histogram
 - d. Bar chart

Correct answers

1. A line chart is a graph that uses lines to show the time evolution from one value to another. It is useful when we want to see the evolution of a variable over time.
2. A dot plot is useful to show the relationship between two continuous variables. We can see if there is an association between the two variables and if so, what kind of association.
3. A bar chart is useful when we want to compare quantities between different categories or groups. For example, we can use a bar chart to compare the sales of two companies in a given quarter.
4. A dot plot is useful to visualize the relationship between continuous variables.
5. b) Line graph
6. c) Histogram
7. a) A graph showing the distribution of proportions of a variable
8. c) A graph showing the relationship between two variables
9. a) A graph showing the distribution of a variable
10. d) Bar chart

3. Statistical events

To understand the notion of probability, we need to define some terms. When we talk about probabilities, we first conduct a statistical experiment. **An experiment** is a procedure that happens under certain predefined conditions. In an experiment, we can have one or more **trials**. Each trial has an **outcome**, which is called an **event** and which we can observe. The set of all possible events in an experiment is called the **sample space**.

A common example of an experiment is a coin toss. We can toss it several times, which means we have several trials. The results can be *heads* or *tails*, which are the possible outcomes. Heads and tails represent the whole field of events when the experiment consists of a single coin toss.

3.1. Definition of events

There are several types of events such as simple events, which are single events, such as rolling a die and getting an even number, and compound events, which are composed of several simple events, such as getting two even numbers in two separate rolls of the die. When we roll two dice, we can get a combination of numbers. The event where the sum of the numbers equals 7 is a compound event.

Some events have special meanings, such as **the impossible event** and **the certain event**. If the outcome cannot occur in an experiment, it is called an impossible event. For example, we cannot get the number 7 when rolling a 6-sided die. A certain event is the opposite of the impossible event; it is certain to happen. For example, getting an odd or even number when rolling a die is a certain event.

Events are denoted with capital letters, and their contents are listed within curly braces:

$$A = \{1, 2, 3\}$$

Event A consists of elements 1, 2, and 3.

A compound event may have several elements or groups of elements:

$$B = \{(1, 2), (3, 4)\}$$

Event B contains two groups of events (1, 2) and (3, 4). Each group has two elements each.

A useful way of representing events is by using Venn diagrams. Venn diagrams are graphical representations used to illustrate sets of events. Let's say we have the following two events:

$$A = \{1, 2, 3\}$$

$$B = \{4, 5, 6\}$$

The sample space (denoted by S) is:

$$S = \{1, 2, 3, 4, 5, 6\}$$

We can draw a Venn diagram like the one in Figure 3.1.



Fig. Representation of events by Venn diagrams

If the sample space consists of only two sets that have no elements in common, they are called disjoint sets. Events A and B in the example above are disjoint crowds.

3.2. Types of events

Events can be either **non-mutually exclusive** or **mutually exclusive**. Non-mutually exclusive events can occur simultaneously in the same experiment, while mutually exclusive events cannot. If we flip a coin once, the event of getting heads and the event of getting tails are mutually exclusive. We can have either one or the other. On the other hand, if we flip a coin twice, the same two events become non-mutually exclusive. In one trial, you might get heads, and in the next, tails.

Events can be **dependent** or **independent**. If one event has an impact on another, then they are dependent. If not, they are independent. If you have an urn with 5 black balls and 5 white balls, drawing one ball from the urn changes the ratio of white to black balls, which impacts the outcome of the second draw. In this case, the second draw depends on the first draw. On the other hand, if we toss a coin twice, the result of the first toss has no effect on the result of the second toss.

3.3. Operations on Events

There are two types of operations we are interested in using events: **union** and **intersection**.

The union (Figure 3.2) is the combination of the elements of two or more sets (or events). It is denoted by U and can be interpreted as addition (+). For events A and B we have:

$$A = \{1, 2\}$$

$$B = \{3, 4\}$$

$$A \cup B = \{1, 2, 3, 4\}$$

And we read "A or B" or "A union B"

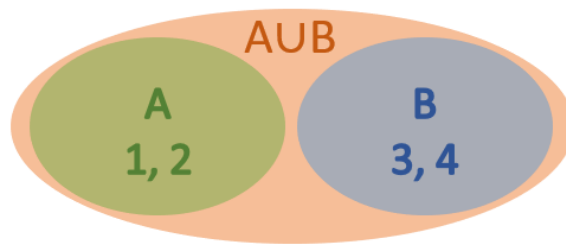


Fig. 3.2. Union of two events

The intersection (Figure 3.3) is the set of elements common to two or more sets. It is denoted by \cap and is interpreted as multiplication (*). For two events A and B, their intersection is:

$$A = \{1, 2, 3\}$$

$$B = \{3, 4\}$$

$$A \cap B = \{3\}$$

And we read "A and B" or "A intersection B"

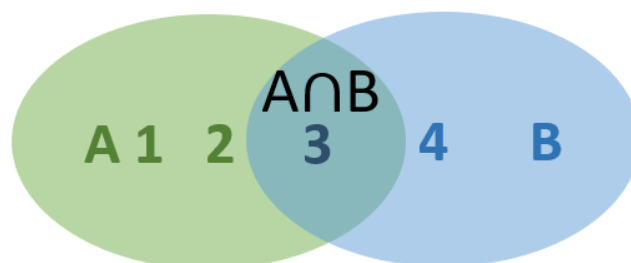


Fig. 3.3. Intersection of two events

The complement (Figure 3.4) of an event includes all outcomes in the sample space that are not in that event. If we have an event A, its complement is marked with \bar{A} and represents all elements that are not part of A.

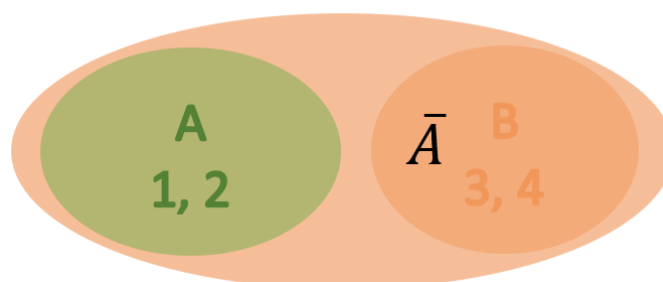


Fig. 3.4. Complement of an event

The complement of the certain event is the impossible event. Conversely, the complement of the impossible event is the certain event.

Intersection and Union have the following properties:

$$A \cup A = A \quad A \cap A = A$$

$$A \cup B = B \cup A \quad A \cap B = B \cap A$$

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cup \bar{A} = E \quad A \cap \bar{A} = \Phi$$

$$A \cup E = E \quad A \cup \Phi = A$$

$$E \cap \Phi = \Phi \quad E \cup \Phi = E$$

$$\overline{(A \cap B)} = \bar{A} \cap \bar{B} \quad \overline{(A \cup B)} = \bar{A} \cup \bar{B}$$

3.4. Knowledge check

1. Which of the following best describes an event in the context of probability?
 - a. A random process
 - b. An impossible situation
 - c. A specific outcome or set of outcomes
 - d. A mathematical statement
2. What is a compound event?
 - a. An event that can never happen
 - b. An event consisting of one possible outcome
 - c. An event combining several simple events
 - d. An event that necessarily occurs
3. Which of the following is an example of a simple event?
 - a. Obtaining two different die faces
 - b. Drawing a red card from a deck of cards
 - c. Meeting a person named "John"
 - d. Reaching a temperature of -100 degrees Celsius
4. What is the sample space?
 - a. A set of impossible events
 - b. A set containing all possible events in a statistical experiment
 - c. A geographical area with unusual properties
 - d. A term used in geography
5. What is the union of two events?
 - a. An event comprising elements present in both events
 - b. An event comprising elements present in either of the events
 - c. The event that does not appear in either of the two events
6. What is the intersection of two events?
 - a. An event comprising elements present in both events
 - b. An event comprising elements present in either of the events
 - c. The event that does not appear in either of the two events
7. What is the complementary event?
 - a. The event that occurs most frequently in an experiment
 - b. The event that never occurs in an experiment
 - c. The set of outcomes not included in the original event
8. What is a sequence of independent events?
 - a. Events that cannot happen at the same time
 - b. Events that do not influence each other

- c. Events that always happen together
9. What is a series of dependent events?
- a. Events that cannot happen at the same time
 - b. Events that do not influence each other
 - c. Events that depend on one or more previous events
10. What is it when two events are mutually exclusive?
- a. They cannot occur at the same time
 - b. They are independent and do not influence each other
 - c. They have elements in common

Correct answers

1. a) A random process
2. c) An event combining several simple events
3. b) Drawing a red card from a deck of cards
4. b) A set containing all possible events in a statistical experiment
5. b) An event comprising elements present in either of the events
6. a) An event comprising elements present in both events
7. c) Event not appearing in the original event
8. b) Events that do not influence each other
9. c) Events dependent on one or more previous events
10. a) They cannot occur at the same time

4. Probability

Probability is a way of quantifying the chance of an event occurring or not occurring. It is usually expressed numerically either as decimal values between 0 and 1 or as percentages between 0% and 100%. The extreme values are theoretical values and correspond to the impossible event (0%) and the certain event (100%). The probability of an event X is denoted by $P(X)$.

To determine the probability of an event occurring, we need to know the number of favorable events and the total number of events in an experiment. The probability of an event occurring is the number of favorable events divided by the total number of equally probable events:

$$P(X) = \frac{\text{number of favorable events}}{\text{number of equally likely events}}$$

For example, when rolling a six-sided die, we have six equally likely events (numbers 1 to 6). If we are interested in getting the number 6, then we say this is our favorable event. The probability of rolling a 6 on a six-sided die is therefore $1/6$.

4.1. Conditional probability

The outcome of some events may depend on the outcome of other events. These are called dependent events. In this case, the probability of the occurrence of one event depends on the probability of the occurrence of another event. This is called **conditional probability**. If we have an event A that depends on the outcome of event B , then we say: "*the probability of event A occurring, **given that** event B has occurred*" and write $P(A|B)$.

Given that event B has occurred, the relevant part of A is where it intersects with B . As a result, the probability of A occurring given that B occurred is the probability of A intersecting B relative to the probability of B occurring:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Suppose we have a standard deck of 52 playing cards and we want to calculate the probability of drawing an ace, assuming we know that the card drawn is a spade card. In this case, event A is "*drawing an ace*" and event B is "*drawing a spade card*".

There are 4 aces in the deck and 13 spade cards. Only one of these aces is also a spade card (Ace of Spades). So, the probability of drawing an ace from the deck is $P(A) = \frac{4}{52}$ and the probability of drawing a spade card is $P(B) = \frac{13}{52}$. The probability of drawing the Ace of Spades, which is both ace and spade, is $P(A \cap B) = \frac{1}{52}$.

Using the conditional probability formula, we have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

So, the probability of drawing an ace, assuming we have already drawn a spade card, is 1/13.

4.2. Multiplication and addition rules

Depending on the type of event (dependent, independent, mutually exclusive, non-mutually exclusive), the calculation of intersection and union differs. Table 4.1 summarizes the calculations for each case.

Table 4.1. Multiplication and addition rules according to event types

Operation	Type of event	Probability calculation
Union	Mutually exclusive	$P(A \cup B) = P(A) + P(B)$
	Non-mutually exclusive and independent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B)$
	Non-mutually exclusive and dependent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B A)$
Intersection	Non-mutually exclusive and independent	$P(A \cap B) = P(A) * P(B)$
	Compatible and dependent	$P(A \cap B) = P(A) * P(B A)$

4.3. Law of total probability

The law of total probability (or total probability formula) is a method used to calculate the probability of an event A, considering all possible ways in which it could occur. If B1, B2, B3, ... Bn are mutually exclusive events and cover the whole field of events S, then for any event A we have:

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

This means that (Figure 4.1), given four events (B1 , B2 , B3 , B4) occupying the entire sampling space and one event dependent on these four, the probability of event A occurring is the sum of the probabilities of A occurring in each of the four events.

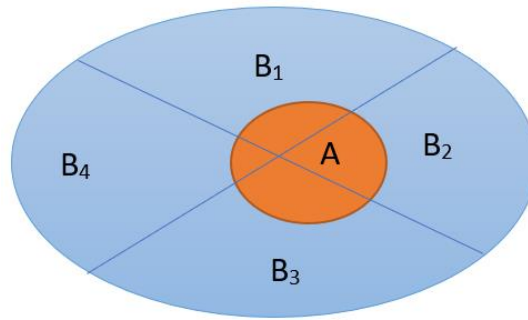


Fig. 4.1. Law of total probability

For example, if we have a country with four distinct regions ($B_1 - B_4$) and we are interested in the mountainous area (A), then the total area with mountains is:

$$A = \sum_{i=1}^4 A \cap B_i$$

4.4. Bayes' rule

Bayes' rule is a mathematical method that allows us to update the probability of an event based on new information that becomes available. The rule is based on the idea that the probability of an event can be calculated differently depending on the information available. Specifically, it allows us to calculate the probability of an event A based on the initial probability of event A and a new piece of information B that becomes available.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

where:

- $P(A|B)$ is the probability of event A given information B ;
- $P(B|A)$ is the probability of information B given event A ;
- $P(A)$ is the initial probability of event A ;
- $P(B)$ is the probability of information B .

Bayes' rule is widely applied across various fields, including medicine, computer science, engineering, and economics. For example, in medicine, Bayes' rule can be used to calculate the probability of a patient having a particular disease based on their symptoms and medical history. In engineering, Bayes' rule can be used to calculate the reliability of a complex system, based on available data about its individual components.

Let's say you work in a factory that produces light bulbs. Each worker has his or her own job and produces light bulbs which are then collected and put together in a

container. The container is then taken to quality control. At the quality check, one bulb is found to be defective. What is the probability that it is coming from your station?

We denote by $P(D)$ the probability that the bulb is defective. Then $P(X)$ is the probability that the bulb came from your station. Suppose, for simplicity, that there is only one other worker and the probability that a bulb came from his station is $P(Y)$.

Because there are only two workers (you and Y) the sum of their probabilities must equal 100%. If $P(X)$ and $P(Y)$ are the probabilities of selecting a bulb from your station and the other worker's station respectively, then

$$P(X) + P(Y) = 1$$

If we draw a Venn diagram it looks like in Figure 4.2.

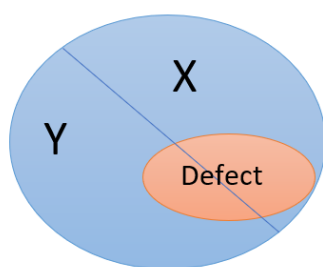


Fig. 4.2. Bayes rule explained in an example

At the top are parts made by X, some of which are defective, and at the bottom are bulbs made by Y, some of which are also defective.

If we consider this from the point of view of the event, we can write:

$$P(X \cap D) = P(D \cap X)$$

Since D depends on X, this can be written as:

$$P(X) * P(D|X) = P(D) * P(X|D)$$

The probability we are interested in is $P(X|D)$, i.e. the probability of a faulty light bulb coming from your station. This we extract from the formula above:

$$P(X|D) = \frac{P(D|X) * P(X)}{P(D)}$$

In the context of Bayes' Rule, the prior probability ($P(A)$) is the probability of the hypothesis before considering the evidence, and the posterior probability ($P(A|B)$) is the probability of the hypothesis given the new evidence. ($P(A|B)$).

4.5. Knowledge check

1. What is the conditional probability of event A given event B?
 - a. Probability of event A and event B
 - b. Probability of event A given that event B happened
 - c. Probability of event B given that event A happened
2. When are two events considered independent?
 - a. Events that cannot happen at the same time
 - b. Events that do not influence each other
 - c. Events that depend on one or more previous events
3. What is the probability of the intersection of two independent non-mutually exclusive events?
 - a. $P(A) + P(B)$
 - b. $P(A) - P(B)$
 - c. $P(A) * P(B)$
4. What is the law of total probability?
 - a. Method to calculate the probability of a union of events
 - b. Method to calculate the probability of an intersection of events
 - c. Method of calculating the probability of an event given other events
5. What is a conditional event?
 - a. The event that depends on other events
 - b. The event that does not depend on other events
 - c. The event that cannot occur in conjunction with other events
6. What is the prior probability?
 - a. Probability of an event with updated information
 - b. Probability of an event given other events
 - c. Probability of an event before we have any new information
7. What does the posterior probability represent?
 - a. Probability of a given event
 - b. Probability of an event after taking into account other new information
 - c. Probability of an event given other events
8. A medical test is positive for a particular disease 95% of the time when the person is ill, and 2% of the time when the person is healthy. If a person tests positive, what is the probability that they are really ill knowing that 3% of people have the disease?
 - a. The probability is about 60%.
 - b. The probability is about 50%.

- c. The probability is about 75%.
 - d. The probability is about 5%.
9. In a city there are two taxi companies, A and B. 80% of A's taxis are yellow and 60% of B's taxis are yellow. If 60% of the taxis in the city are from company A, what is the probability that a randomly selected yellow taxi is from company B?
- a. 0.25
 - b. 0.40
 - c. 0.50
 - d. 0.33
10. If 25% of a college's students have Apple laptops, and 60% of students with Apple laptops also have an iPhone, what is the probability that a randomly selected student will have an Apple laptop and an iPhone?

Correct answers

1. b) Probability of event A given that event B happened
2. b) Events that do not influence each other
3. c) $P(A) * P(B)$
4. c) Method of calculating the probability of a given event being other events
5. a) Event dependent on other events
6. c) Probability of an event before we have any new information
7. b) Probability of an event after taking into account other new information
8. a) The probability is approx. 60%.

Rationale: The probability that a person is truly ill given a positive test result is given by the Bayes formula, so:

$$P(Bolnav | Pozitiv) = \frac{P(Pozitiv | Bolnav) * P(Bolnav)}{P(Pozitiv | Bolnav) * P(Bolnav) + P(Pozitiv | Sănătos) * P(Sănătos)}$$

Using the data in the problem statement, it can be calculated that:

$$P(Bolnav | Pozitiv) = \frac{0.95 * 0.03}{0.95 * 0.03 + 0.02 * 0.97} / () = 0.595 \approx 0.60$$

9. d) 0.33. The law of total probability says that the probability of an event can be calculated as the sum of the probabilities conditional on the events that can influence it. In this case, the events are $Y = \text{"yellow taxi"}$, $A = \text{"company A"}$, $B = \text{"company B"}$. We can use the formula of the law of total probability:

$$P(Y) = P(Y|A) * P(A) + P(Y|B) * P(B)$$

Substituting known values into this formula, we get:

$$P(Y) = (0.8 * 0.6) + (0.6 * 0.4) = 0.48 + 0.24 = 0.72$$

The probability that a randomly selected yellow taxi is from company B is:

$$P(B|Y) = \frac{P(Y|B) * P(B)}{P(Y)} = \frac{0.6 * 0.4}{0.72} = 0.33$$

10. To find the probability that a randomly selected student has an Apple laptop and an iPhone, we need to use the conditional probability theorem. We denote the event "has Apple laptop" by A and the event "has iPhone" by B. We have $P(A) = 0.25$, the probability that a randomly selected student has an Apple laptop, and $P(B|A) = 0.6$, the probability

that a student who has an Apple laptop also has an iPhone. We want to find $P(A \cap B)$, the probability that a randomly selected student has both an Apple laptop and an iPhone. Using the conditional probability formula, we have:

$$P(A \cap B) = P(B|A) * P(A) = 0.6 * 0.25 = 0.15$$

Therefore, the probability of a randomly selected student having both an Apple laptop and an iPhone is 0.15 or 15%.

5. Random variables

A **random variable** is a mathematical function that assigns a numerical value to each possible outcome of an event in a statistical experiment. Depending on the experiment, random variables can be discrete or continuous. We say that a random variable is **discrete** when only certain values can be obtained. For example, when rolling a die, the possible outcomes are the numbers 1 through 6. On the other hand, if the result of the experiment is a real number, we say that the random variable is **continuous**. "For instance, measuring the length of an object yields a continuous value, which can theoretically have an infinite number of decimal places.

5.1. Discrete random variables

Discrete random variables can only take on certain values and have a certain probability associated with them. If our experiment consists of rolling a die 100 times, for example, we can determine the probability of getting a number on the die. To determine this probability, we divide the number of occurrences of a specific face by the total number of rolls. For instance, if we roll the die 100 times and the number six appears 20 times, the probability of rolling a six is 20/100 or 20%.

Random variables are denoted with capital letters. We can represent all possible outcomes and their associated probabilities in a probability distribution table. For the dice example we can write:

$$X: \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.15 & 0.2 & 0.1 & 0.25 & 0.2 \end{pmatrix}$$

One thing to note is that the probability of an event occurring is its relative frequency. If we plot it graphically, we get the diagram in Figure 5.1.

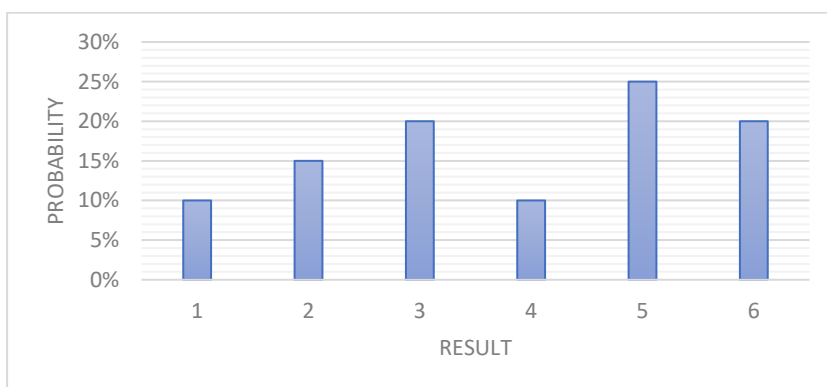


Fig. 5.1. Distribution of results of rolling a die 100 times

This is how the probabilities are arranged or distributed. This is called a distribution. The graph above helps us visualize the probability distribution for this die. If it's a fair die (not intervened on in such a way as to favor a particular outcome), all sides have the same probability of occurrence. If we roll it many times, the probabilities

will all approach 1/6 and the graph will become flat (all columns are approximately the same height). The function that assigns the probability to each value in X is called the **distribution function**.

When we are interested in a particular value of a random variable (for example, $X = 3$), we use the distribution function to determine it, $P(X = 3) = 0.2$.

The probabilities for any value in X must be greater than or equal to zero:

$$P(X = x) \geq 0, \text{ for all } x \text{ in } X$$

The sum of all probabilities must equal 1:

$$\sum P(x) = 1$$

Discrete random variables are an important part of probability theory and statistics and are used in a variety of fields including data science, economics and engineering.

5.2. Continuous random variables

Random variables that can assume any real number within a given interval are called continuous random variables. In this case, it's not feasible to create a table listing specific values and their probabilities, as there are infinitely many possible values. Consequently, the probability of the random variable taking any specific value is effectively zero. We can still plot the probability graphically. Since all the values are close together, the graph is a curve (Figure 5.2).

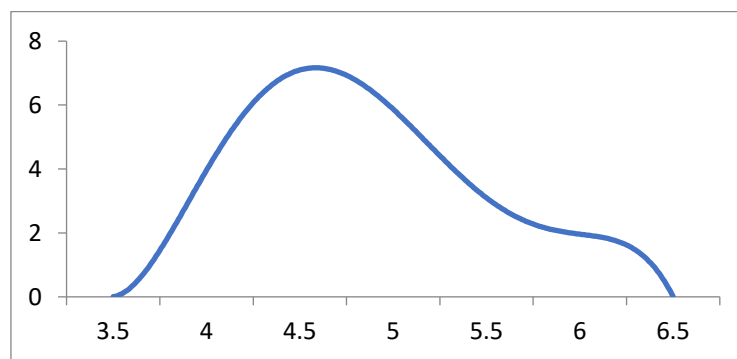


Fig. 5.2. Example curve for a continuous distribution

The area under the curve between two values a and b represents the probability that an event x is between these two values and can be determined using an integral:

$$P(a < x < b) = \int_a^b f(x)dx$$

Suppose you want to determine the probability that a coffee machine will give you between 95 and 100 ml of coffee. Given that the amount of coffee dispensed can be any real number within a certain range, this represents a continuous random variable. By

taking several measurements (say 100), we can determine the function, $f(x)$, that gives us the curve and then write:

$$P(95 < x < 100) = \int_{95}^{100} f(x)dx$$

By solving the integral, we can determine the probability that x takes values between 95 and 100.

5.3. Cumulative distribution function

Sometimes we need to know the probability of getting all the values up to a particular one. In a dice game, we might end up winning if we roll any number up to 4. To determine what our probability of winning is, we need to add up the probabilities of all the values up to 4 (1, 2, 3 and 4). This sum is the cumulative probability. The function that gives us the distribution of these cumulative probabilities is called the **cumulative distribution function**. If we take the random variable X from the previous example:

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.15 & 0.2 & 0.1 & 0.25 & 0.2 \end{pmatrix}$$

and write the cumulative probability for each value, we get the following table:

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0.1 & 0.25 & 0.45 & 0.55 & 0.75 & 1 \end{pmatrix}$$

If we are interested in the cumulative probability up to a certain value, (say 3) we look for this value in the first row (3) and read the associated probability (0.45). This means that the probability of getting all values up to 3 is 0.45, or 45%.

In the case of a continuous random variable we use the integral. We start from $-\infty$ to the desired value.

$$P(x < a) = \int_{-\infty}^a f(x)dx$$

We can plot the probability distribution as in Figure 5.3 for both discrete and continuous variables.



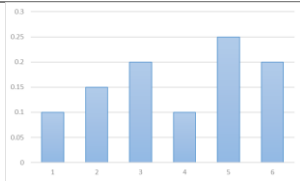
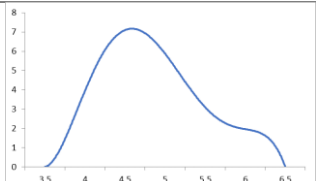
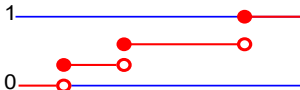
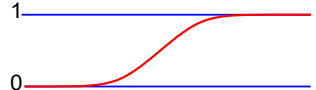
Fig. 5.3. Cumulative distribution function for a discrete (left) and continuous (right) function [9].

In the case of the discrete random variable (left), each probability is associated with a certain value and we have gaps between the resulting values. The sum of the probabilities is 1. In the case of the continuous random variable (right) we have a continuous curve starting at 0 and ending at 1.

5.4. Discrete and continuous variables

If we put the two types of variables side by side, we can see the differences between them (Table 5.1).

Table 5.1. Comparison between discrete and continuous variables

	Discrete random variables	Continuous random variables
Probabilities can be tabulated	yes	no
Probability of a specific value can be determined	yes	no
Probability distribution		
Cumulative distribution function		

Discrete and continuous random variables can have an infinite number of distribution modes. However, there are certain distributions that are more commonly encountered. Learning about these distributions and knowing when to use them can be very helpful in understanding processes and making decisions.

Distributions can be discrete or continuous depending on the type of variable we are analyzing.

5.5. Knowledge check

1. A random variable is:
 - a. A variable that changes over time.
 - b. A numerical result of a random process.
 - c. A deterministic numerical value.

2. Which of the following is a discrete random variable?
 - a. The weight of a random apple.
 - b. Number of pupils in a class.
 - c. The time it takes to run a marathon.

3. What is the main difference between discrete and continuous random variables?
 - a. Discrete variables can only take integer values, while continuous variables can take any value.
 - b. Discrete variables can take any value, while continuous variables can only take specific values.
 - d. There is no difference; both are types of random variables.

4. Which of the following is an example of a continuous random variable?
 - a. Rolling a dice.
 - b. Number of cars in a car park.
 - c. The temperature in a room at one time.

5. If you flip a fair coin three times, the random variable representing the count of heads is:
 - a. A continuous random variable.
 - b. A discrete random variable.
 - d. Neither discrete nor continuous.

6. Which of the following can be represented by a continuous random variable?
 - a. Number of emails received in an hour.
 - b. Height of pupils in a school.
 - c. Number of rainy days in a year.

7. A probability distribution function for a discrete random variable:
- Shows the probability that the variable is continuous over an interval.
 - List each of the possible outcomes and the probability associated with each.
 - It is the integral of the probability density function.
8. The sum of all probabilities for all possible values of a discrete random variable must be equal:
- 1
 - 0
 - 100
9. A real-world example of a discrete random variable is:
- The amount of milk in a glass.
 - Number of textbooks on a shelf.
 - Speed of a moving car.
10. The probability of a continuous random variable taking any single specific value is:
- Exactly 1.
 - Exactly 0.
 - Any positive number greater than 0

Correct answers

1. b. A numerical result of a random process.
2. b. Number of pupils in a class.
3. a. Discrete variables can only take integer values, while continuous variables can take any value.
4. c. Temperature in a room at a given time.
5. b. Discrete random variable.
6. b. Height of pupils in a school.
7. b. List each of the possible outcomes and the probability associated with each.
8. a. 1
9. b. Number of textbooks on a shelf.
10. b. Exactly 0.

6. Discrete distributions

The function given by the probability distribution can help us to obtain valuable information about our data and guide us in our decision-making process. There are a few distributions that are more commonly encountered in everyday life. Therefore, understanding a few fundamental aspects of these distributions is beneficial.

Some of the most common distributions of discrete variables that we will discuss next are:

- Uniform distribution
- Binomial distribution
- Hypergeometric distribution

6.1. Uniform distribution

The uniform distribution has the same probability for all outcomes (Figure 6.1). Each outcome has probability $1/n$ where n is the number of outcomes.

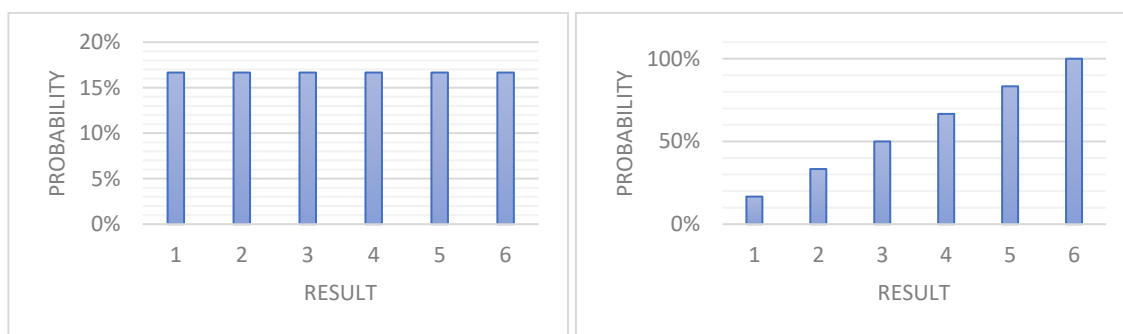


Fig. Probability function and cumulative distribution function for uniform distribution.

An example of a uniform distribution is rolling a die. Possible outcomes are: 1, 2, 3, 4, 5, 6. Each outcome has a probability of $1/6$ if the die is fair (not biased).

The cumulative distribution function for the uniform distribution increases steadily across categories.

6.2. Binomial distribution

The binomial distribution is used whenever we have only two possible outcomes with repeated trials and the observations are independent. We focus on only one outcome, called *success*. For example, consider flipping a coin, which has two sides: heads and tails. By flipping the coin several times we have a binomial experiment. We look at just one outcome, say we observe the number of heads that appear. This is the "success" result.

Here are the notations we will use:

- p - the probability of success on a single trial
- q – the probability of failure on a single trial ($q = 1 - p$)
- n - number of trials
- k - number of successes at a given time

It is called binomial distribution because the probability follows Newton's binomial formula:

$$(a + b)^n = C_n^0 a^n b^0 + C_n^1 a^{n-1} b^1 + \dots + C_n^n a^0 b^n$$

For example, if from an urn containing two types of balls (red and blue) we draw a ball twice, **putting the ball back each time**, the random variable for the number of red balls drawn is:

$$X: \begin{pmatrix} 0 & 1 & 2 \\ q^2 & 2pq & p^2 \end{pmatrix}$$

The probability for each outcome is each term of its decomposition $(q + p)^2$.

In general, the probability of a given outcome k (distribution function) can be calculated with the formula:

$$P(X = k) = C_n^k p^k q^{n-k}$$

where C_n^k is combinations of n taken as k :

$$C_n^k = \frac{n!}{k!(n - k)!}$$

The cumulative probability function has the following formula:

$$P(X \leq k) = \sum C_n^k p^k q^{n-k}$$

The plots of the distribution and probability functions are shown in the figure below.

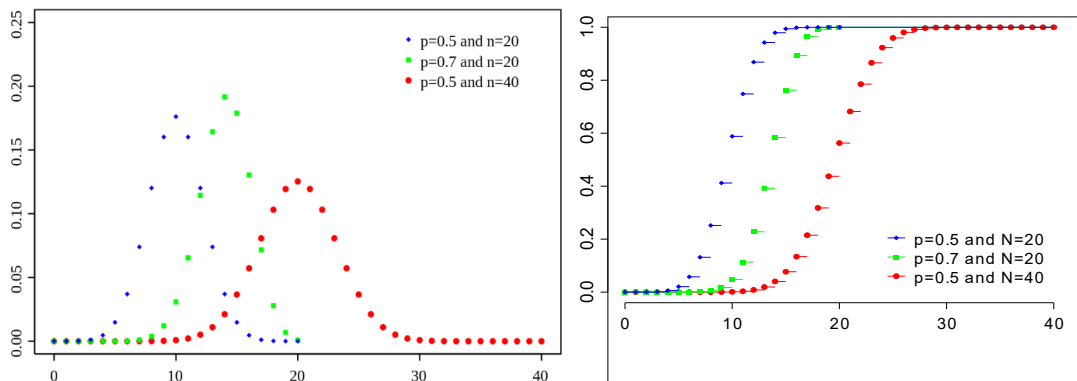


Fig. Distribution (left) and probability (right) functions of the binomial distribution [10].

A common example in manufacturing is when working with parts. If you divide the parts into two categories, bad and good, then we can have:

p - probability of removing a bad part on a single trial

q - probability of getting a good part on a single trial

The binomial distribution is often used to model situations where we have a fixed number of independent trials, each with only two possible outcomes and a constant probability of success for each trial. An important use case for the binomial distribution is in the context of acceptance sampling, which is a quality control procedure used to determine whether a batch of parts meets certain standards.

In acceptance sampling, a sample of parts is randomly selected from a larger lot, and each part in the sample is inspected for defects or other quality problems. The acceptance condition is then based on the number of defective parts in the sample, which is modelled using the binomial distribution.

The acceptance condition is typically expressed in terms of two parameters: sample size (n) and maximum allowed number of defects (c). The sample is accepted if the number of defective parts (X) in the sample is less than or equal to the maximum allowed number of defects (c). Mathematically, this can be expressed as:

$$P(X \leq c) \geq A$$

where $P(X \leq c)$ is the cumulative probability of having at most c defective parts in the sample, and A is the specified acceptance level, which is usually a small probability value (e.g., 0.05 or 0.01).

In other words, the batch of parts is accepted if the probability of observing c or fewer defective parts in the sample is greater than or equal to the specified acceptance level. If the probability is less than the acceptance level, then the lot is rejected and further inspection or corrective action may be required.

Example problem

We want to buy a batch of parts from a supplier. He declares his scrap coefficient (percentage of bad parts) is **p = 5%**. We extract $n = 3$ parts, **each time putting the part back**. What is the probability of getting at most one bad part?

The 4 possible outcomes are 0, 1, 2, or 3 bad parts.

Using the distribution function formula given above, we obtain the following random variable:

$$X: \left(\begin{array}{cccc} 0 & 1 & 2 & 3 \\ C_3^0 q^3 p^0 & C_3^1 q^2 p^1 & C_3^2 q^1 p^2 & C_3^3 q^0 p^3 \end{array} \right)$$

After doing the calculations we get the probabilities for each possible outcome:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ 85.74\% & 13.54\% & 0.71\% & 0.01\% \end{pmatrix}$$

The cumulative distribution function is:

$$X: \begin{pmatrix} 0 & 1 & 2 & 3 \\ 85.74\% & \mathbf{99.28\%} & 99.99\% & 100\% \end{pmatrix}$$

This means that the probability of getting at most 0 bad parts is 85.74%, the probability of getting at most 1 bad part is 99.28% and so on. The correct answer in this case is 99.28%.

6.3. Hypergeometric distribution

The hypergeometric distribution is similar to the binomial distribution in that it deals with two possible outcomes but differs in the dependency of trials. The difference is that the events depend on each other.

Notations:

- p - probability of success
- q - probability of failure
- n - number of objects
- m - number of trials
- a - number of successes
- b - number of failures
- k - number of successes at a given point

The distribution function has the following formula:

$$P(X = k) = \frac{C_a^k C_b^{m-k}}{C_n^m}$$

The cumulative distribution function has the following formula:

$$P(X \leq k) = \frac{1}{C_n^m} \sum C_a^k C_b^{m-k}$$

$$P(X \leq k) = F(k) = \frac{1}{C_n^m} \sum_{k=0}^k C_a^k C_b^{m-k}$$

The plots of the two functions, distribution and cumulative, are shown in Figure 6.3.

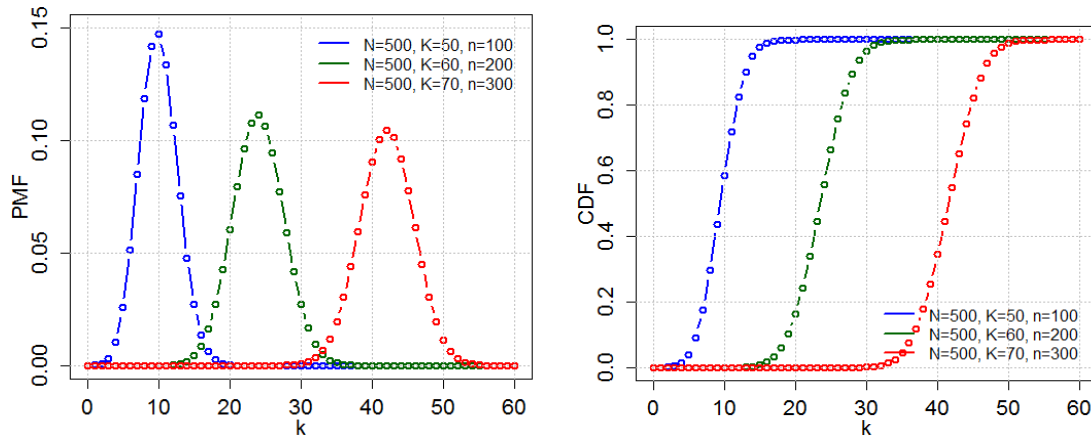


Fig. Distribution function and cumulative distribution function for hypergeometric distribution [11].

Example problem

Suppose you want to buy a batch of $n = 100$ parts from a supplier. He says that his scrap coefficient (percentage of bad parts) is $p = 5\%$. We extract $m = 3$ parts, **without putting the part back**. Determine the probability of getting at most one bad part.

The 4 possible outcomes are 0, 1, 2, or 3 bad parts.

The expected number of bad parts (a) is calculated as:

$$a = n * p = 100 * 5\% = 5$$

and the number of good parts (b):

$$b = n * q = n - a = 100 * 95\% = 95$$

Using the distribution function formula given above, we obtain the following random variable:

$$X: \left(\begin{array}{cccc} 0 & 1 & 2 & 3 \\ \frac{C_5^0 * C_{95}^3}{C_{100}^3} & \frac{C_5^1 * C_{95}^2}{C_{100}^3} & \frac{C_5^2 * C_{95}^1}{C_{100}^3} & \frac{C_5^3 * C_{95}^0}{C_{100}^3} \end{array} \right)$$

After doing the calculations we get the probabilities for each possible outcome:

$$X: \left(\begin{array}{cccc} 0 & 1 & 2 & 3 \\ 85.60\% & 13.81\% & 0.59\% & 0.01\% \end{array} \right)$$

The probability function is:

$$X: \left(\begin{array}{cccc} 0 & \mathbf{1} & 2 & 3 \\ 85.60\% & \mathbf{99.41\%} & 99.99\% & 100\% \end{array} \right)$$

This means that the probability of getting at most 0 bad parts is 85.60%, the probability of getting at most 1 bad part is 99.41% and so on. So the correct answer is 99.41%.

6.4. Knowledge check

1. What is discrete uniform distribution?
 - a. A probability distribution where each outcome is equally likely
 - b. A probability distribution in which events are independent
 - c. A probability distribution that models the probability of success in a finite number of independent trials
2. What is the binomial distribution?
 - a. A probability distribution that models the number of successes in a continuous process
 - b. A probability distribution that models the probability of success in a finite number of independent trials
 - c. A probability distribution that models the probability of having a given number of successes in a fixed number of independent trials
3. What is the hypergeometric distribution?
 - a. A probability distribution that models the probability of having a given number of successes in a fixed number of independent trials
 - b. A probability distribution that models the probability of having a given number of successes in a sample drawn from a finite and specified population.
 - c. A probability distribution that models the number of successes in a continuous process
4. What is the probability function of the uniform distribution?
 - a. $f(x) = 1/n$, where n is the number of possible events
 - b. $f(x) = x/n$, where n is the number of possible events
 - c. $f(x) = n$, where n is the number of possible events
5. What is the difference between binomial and hypergeometric distributions?
 - a. The binomial distribution models the probability of having a given number of successes in a fixed number of independent trials, while the hypergeometric distribution models the probability of having a given number of successes in a sample drawn from a finite and specified population.
 - b. The binomial distribution models the probability of having a given number of successes in a sample drawn from a finite and specified population, while the hypergeometric distribution models the probability of having a given number of successes in a fixed number of independent trials.
 - c. The binomial and hypergeometric distributions are the same and model the same scenarios.

Correct answers:

1. a) A probability distribution where each outcome is equally likely
2. c) A probability distribution that models the probability of having a given number of successes in a fixed number of independent trials
3. b) A probability distribution that models the probability of having a given number of successes in a sample drawn from a finite and specified population.
4. a) $f(x) = 1/n$, where n is the number of possible events
5. a) The binomial distribution models the probability of having a given number of successes in a fixed number of independent trials, while the hypergeometric distribution models the probability of having a given number of successes in a sample drawn from a finite and specified population.

7. Continuous distributions

Continuous distributions describe the probability distribution of a continuous random variable, which can assume any value within a continuous range of real numbers. These distributions are different from discrete distributions, which apply to discrete random variables. Some of the best known continuous distributions include:

- Uniform distribution (continued)
- Normal distribution
- Student distribution
- Chi-square distribution

The uniform distribution applies to random variables that have an equal probability of occurring between two limits. The normal distribution, also known as the Gaussian distribution, is one of the most important continuous distributions and is used in a wide range of fields, from the social sciences to finance and engineering. The Student's distribution is mainly used in statistical analysis when we have small sample sizes, while the Chi-square distribution is used in many fields, such as statistical hypothesis testing or analysis of experimental data.

7.1. Uniform distribution

The continuous uniform distribution is one of the simplest, but still very important, continuous distributions and is used in many fields, from social sciences to natural sciences and technology. This distribution describes random variables that have a uniform probability distribution between two limits, so every value in the interval has the same probability of being chosen. This means that the probability of getting a particular value from the interval is proportional to the length of the interval and does not depend on any other characteristic of the distribution.

"In a uniform distribution, every interval of the same length within the range [a, b] has the same probability. Outside the interval [a, b], the probability is 0. The distribution and cumulative functions are shown in Figure 7.1. and have the following formulae:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b \\ 1 & \text{for } x > b \end{cases}$$

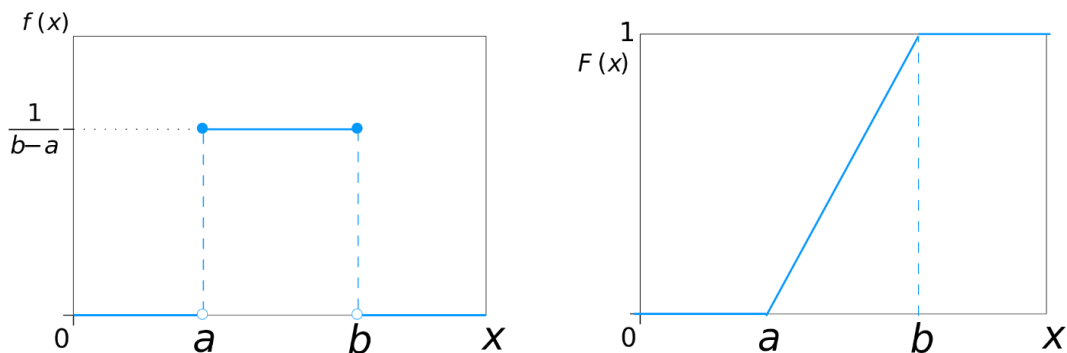


Fig. 7.1 Distribution (left) and cumulative (right) function for continuous uniform distribution [12]

7.2. Normal distribution

The normal distribution, also known as the Gaussian distribution, is one of the most important and commonly used continuous distributions. It is characterized by the symmetric curve around its mean, which represents the center of the distribution, and the standard deviation, which measures how much variability there is in the data of the distribution. This distribution is used in a wide range of fields, from social sciences and economics to natural sciences and technology. It is used to model data that is approximately symmetric and to make probability estimates or statistical inference. Because many natural and social phenomena follow this distribution, the normal distribution is of fundamental importance in data analysis and data-driven decision making in a variety of fields.

The Probability Density Function has the formula:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where σ^2 is the population variance and μ is the population mean.

The Cumulative Distribution Function is described by Eq:

$$F(x) = \int_{-\infty}^x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

The distribution function has a bell shape and therefore the normal distribution is also called the bell distribution. The distribution function and the cumulative distribution function look as in Figure 7.2.

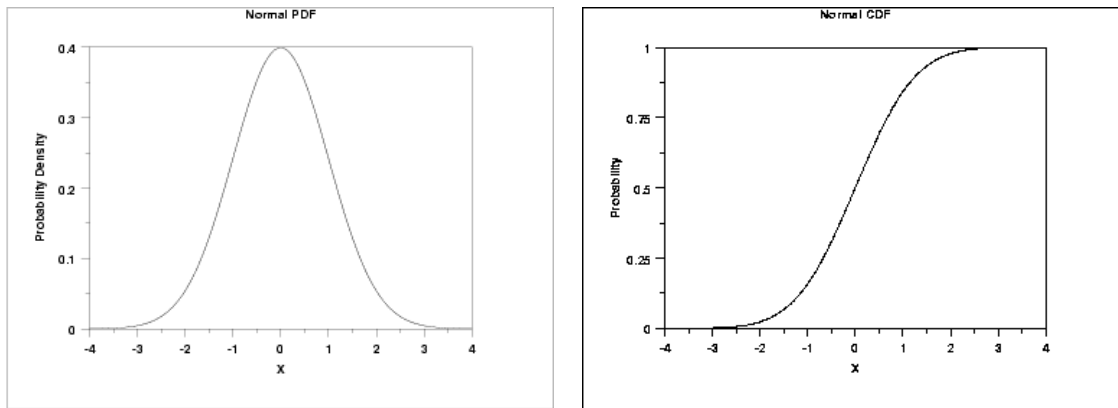


Fig. 7.2. Probability Density Function (left) and Cumulative Distribution Function (right) for the normal distribution [13]

The normal distribution is characterized by two important parameters: the mean (μ) and the variance (σ^2). Changing the mean shifts the curve left or right on the horizontal axis. By changing the variance the curve becomes either thinner and taller or wider and flatter (Figure 7.3.). You can see what effect these two parameters have on the shape of the curve on several websites that have an interactive normal distribution curve, such as this one: http://onlinestatbook.com/2/Calculators/normal_dist.html

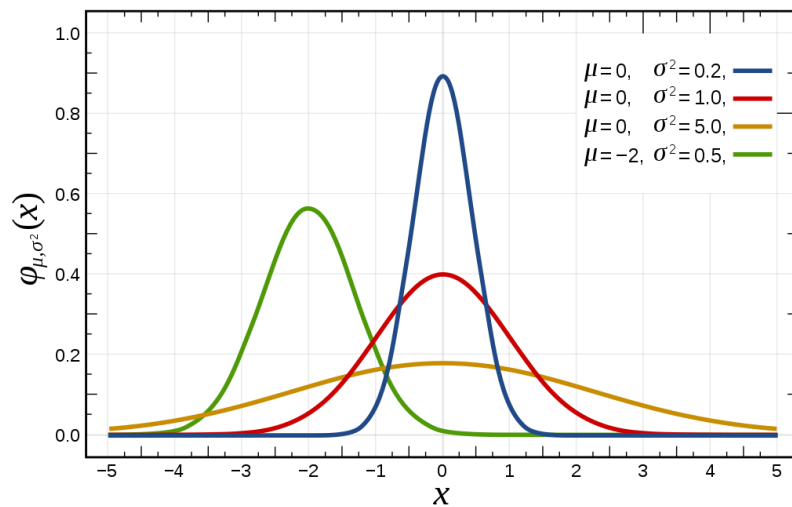


Fig. 7.3 Examples of normal distributions with different parameters [14]

The distribution is symmetric about the mean and the tails go asymptotically to (plus and minus) infinity without ever touching the axis. As with any continuous probability distribution, the area under the curve equals 1.

There is a more special normal distribution, which is called the Standard Normal Distribution. It has a mean of 0 and a standard deviation of 1. It is very useful in determining the position of certain values on any other normal distribution.

The area under the normal curve follows the 68 -95-99.7 rule, where approximately 68% of the area is within one standard deviation from the mean (one on each side), approximately 95% of the area is within two standard deviations from the

mean (on each side), and approximately 99.7% of the area is within three standard deviations from the mean.

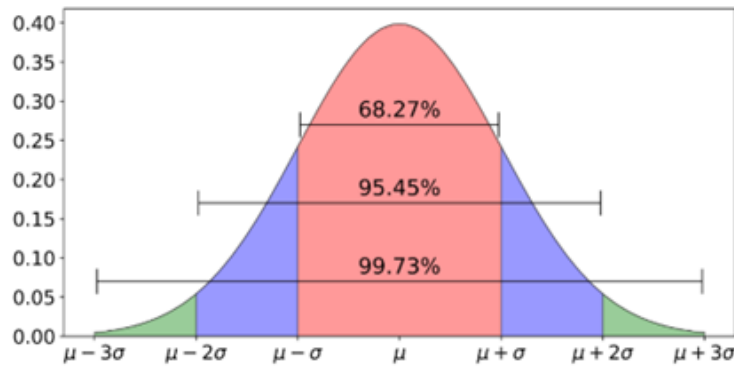


Fig. 7.4. The 68 -95-99.7 rule [15].

7.3. Student Distribution

The Student's t-distribution is a continuous probability distribution primarily used to estimate the confidence interval of a population mean, particularly when the population variance is unknown and is estimated from the sample's standard deviation. This distribution is named after William Gosset, who worked at the Guinness brewery and developed this distribution to analyze beer quality. Because he worked at a brewery and was not allowed to publish his work under his real name, he used the pseudonym "Student". Nowadays, the Student's t-distribution is used in many fields, such as market research, medicine, social sciences and engineering. It is important to understand this distribution in order to be able to interpret and analyze data collected through experiments and surveys.

The distribution is also symmetric about the mean, but has thicker tails, which means that values that are further from the mean have higher probabilities than their equivalent on the normal distribution (Figure 7.5).

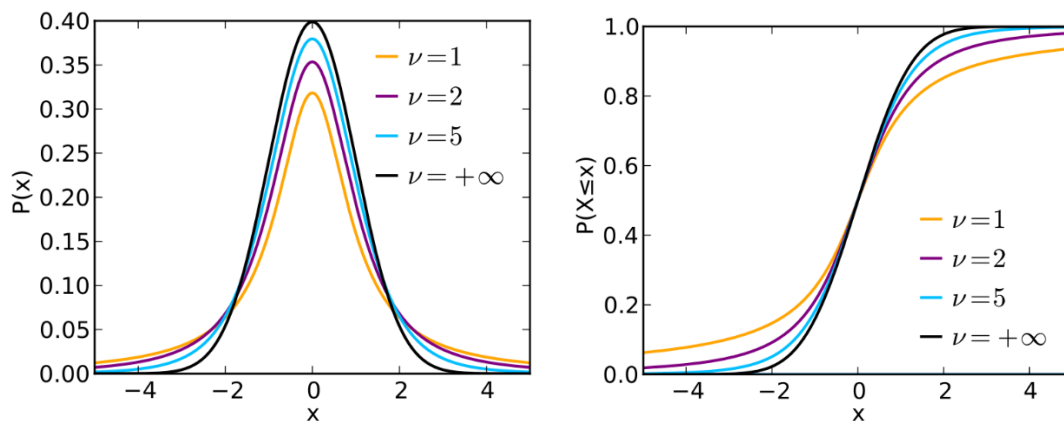


Fig. 7.5. Probability Density Function (left) and Cumulative Distribution Function (right) for the Student's t-distribution [16]

An important parameter is the number of degrees of freedom (ν):

$$\nu = n - 1$$

where n is the number of values in the sample.

Figure 7.5 shows how degrees of freedom affect the shape of the distribution. For an infinite number of degrees of freedom, the Student's t-distribution becomes the normal distribution.

7.4. Chi-square distribution

The Chi-square distribution (χ^2) is a continuous probability distribution commonly used in statistical hypothesis testing, analysis of experimental data, and in constructing confidence intervals for population variance. This distribution is derived from the standard normal distribution and is defined by the number of degrees of freedom. The Chi-square distribution is characterized by its variance and is influenced by the number of degrees of freedom. In general, the higher the number of degrees of freedom, the closer the distribution will be to a normal distribution. Therefore, the Chi-square distribution is an important distribution in many statistical applications.

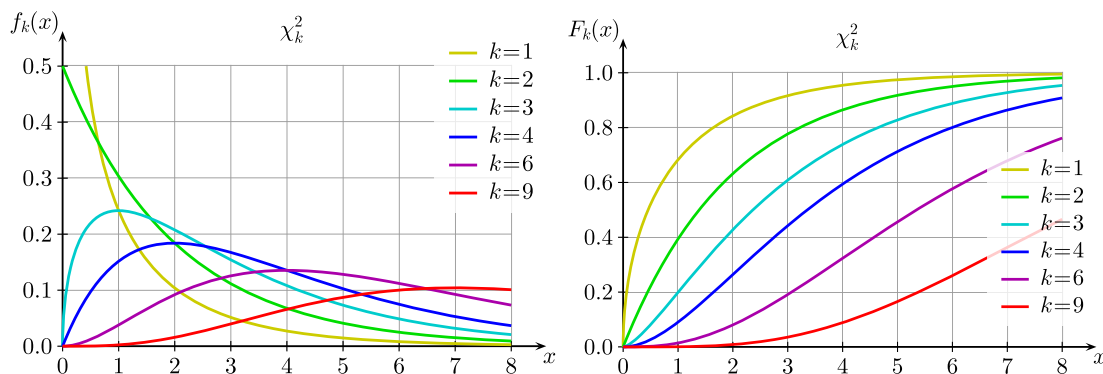


Fig. 7.6. Probability Density Function (left) and Cumulative Distribution Function (right) for the Chi-square distribution [17].

The Chi-square distribution is different from the other two continuous distributions discussed above in that it is not symmetric and does not have negative values (Figure 7.6).

7.5. Knowledge check

1. Uniform distribution assigns:
 - a. Different probabilities of all outcomes.
 - b. Same probability for all outcomes.
 - c. Increasing probabilities for consecutive results.
2. Normal distribution is characterized by:
 - a. A single peak at the mean and symmetry around the mean.
 - b. A curve with two peaks at the extremes.
 - c. Evenly distributed probabilities.
3. Which of the following statements is true for the normal distribution?
 - a. It is tilted to the right.
 - b. It is symmetrical around its mean.
 - c. It has a uniform height.
4. Student distribution is used when:
 - a. The sample size is large and the population variance is known.
 - b. The sample size is small and the population variance is unknown.
 - c. The population follows an even distribution.
5. As the degrees of freedom increase for the Student distribution, the distribution:
 - a. It becomes wider and flatter.
 - b. Approaching a normal distribution.
 - c. It remains constant, regardless of degrees of freedom.
6. A Chi-square distribution is not symmetric but has a longer tail towards:
 - a. Left.
 - b. Right.
 - c. Both left and right, depending on the situation.
7. Chi-square distribution is mainly used for:
 - a. Parameter estimation of other distributions.
 - b. Testing hypotheses on population dispersal.
 - c. Hypothesis testing of population means.

8. Normal distribution is fully described by:

- a. Average and median.
- b. Average and variance.
- c. Variance and asymmetry.

9. Uniform continuous distribution is often used to model:

- a. Variables with a known range and equal probability for any value in the range.
- b. Variables with a natural asymmetry.
- c. Variables that cluster around a central value.

10. The shape of the Chi-square distribution depends on:

- a. Sample mean.
- b. Number of attempts.
- c. Degrees of freedom.

Correct answers

1. b. Same probability for all outcomes.
2. a. A single peak at the mean and symmetry around the mean.
3. b. It is symmetrical about its mean.
4. b. The sample size is small and the population variance is unknown.
5. b. Approaching a normal distribution.
6. b. Right.
7. b. Test hypotheses about population dispersal.
8. b. Mean and variance.
9. a. Variables with a known interval and equal probability for any value in the interval.
10. c. Degrees of freedom.

8. Estimation

The concepts of a *population* and a *sample* are often used in statistics. A **population** is a complete set of items or events that are relevant to a particular research or analysis. For example, if we are interested in studying the height of adults in Romania, the population would consist of all adults in the country.

Typically, populations are very large, often comprising thousands or millions of elements, and thus collecting data from the entire population is often impractical due to the cost and effort involved. Therefore, we usually work with a subset of the population called a **sample**. The sample must be representative of the population so that we can make statistical inferences.

There are different sampling methods that can be used to collect a representative sample. Two common methods are as follows:

- **Random Sampling:** This is the simplest and most straightforward way to collect a sample. Each member of the population has an equal chance of being selected into the sample, with items being chosen at random.
- **Stratified Sampling:** The population is divided into several subgroups, or "strata" and then a random sample is drawn from each stratum. This method is useful when the population is heterogeneous, and we want to ensure that the sample reflects this diversity.

There are differences in characteristics between population and sample, but sometimes the only way to study a population is to study one or more samples taken from a population.

Once we have a sample, we can make a **point estimate**, i.e., determine a single number that serves as a "best" estimate for a population parameter such as the mean. For example, calculating the average height in our sample allows us to use this number as a point estimate of the average height in the entire population.

However, any point estimate comes with a degree of uncertainty. We can quantify this uncertainty by using a **confidence interval estimate instead**. In this case we determine a range of values within which the population parameter is likely to lie. For example, we can say that we are 95% sure that the average height of adults in Romania is between 165 cm and 175 cm. In this example, the range of 165 cm to 175 cm represents the confidence interval, and 95% is the confidence level.

Since estimation using a confidence interval is more prevalent, we will focus on it in the subsequent sections. In this course, we will focus on estimating two crucial population parameters: the *mean* and the *variance* (and consequently the standard

deviation). In Table 8.1. are notations that we have used throughout this course and that we will use to distinguish between population parameters and sample statistics.

Table 8.1. Notations used for population and sample parameters

Parameter	Population	Sample
Media	μ	\bar{x}
Standard deviation	σ	s
Variance	σ^2	s^2

From a population we can extract several samples (Figure 8.1). Due to the central limit theorem most values are in the neighborhood of the population mean (μ) and then the samples will also have values, and therefore means ($\bar{x}_1 - \bar{x}_4$), in the neighborhood of the population mean. Of course, there is also the probability of getting samples with very large values or very small values, but these are rarer because the outliers themselves are rarer.

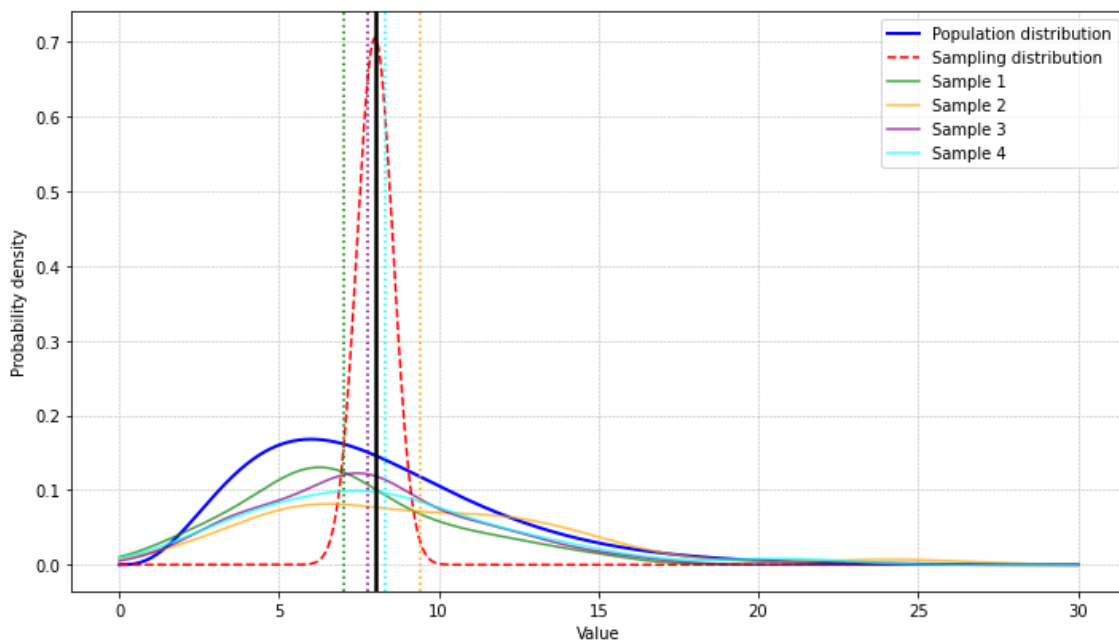


Fig. 8.1. Example of the distribution of a population and several samples

If we take a very large number of samples, and draw the distribution of their means, we get a normal distribution called **the sampling distribution**. This distribution has the property that its mean will be identical to the mean of the population from which we sampled, regardless of the shape of the population distribution.

When we estimate the parameter of a population, we determine some limits (one or two) of a range in which we assume with a certain probability that the estimated

parameter lies. The range is termed the **confidence interval** (Figure 8.2). It can have one or two bounds: the one on the left is the **lower bound**, and the one on the right is the **upper bound** of the confidence interval. The probability that the parameter we estimate lies between these bounds is called the **confidence level** and what lies outside the bounds is called the **risk**. We will denote the risk with the Greek letter α . Since the area under the curve of a distribution is a probability, it will be equal to 1. As a result, the confidence level will be all the probability from which we remove the risk ($1 - \alpha$).

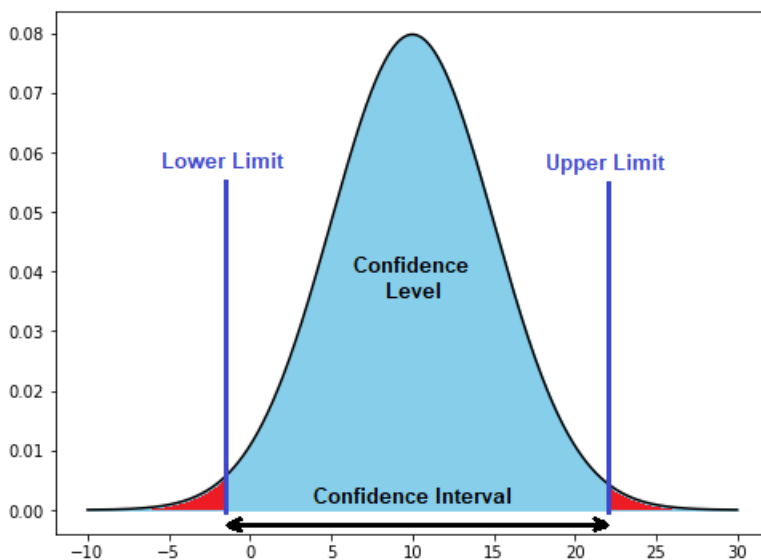


Fig. 8.2. Estimation with confidence interval (bilateral risk)

Since we can have one or two limits for the confidence interval, we can have the following types of risk (Figure 8.3):

- Right Unilateral Risk (RUD)
- Unilateral Left Risk (RUS)
- Bilateral Symmetric Risk (BSR)
- Risk Bilateral Asymmetric (RBA)

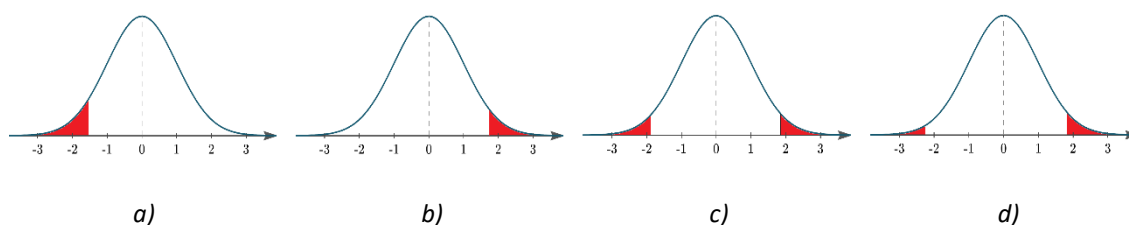


Fig. 8.3. Types of risk: a) Left Unilateral Risk; b) Right Unilateral Risk; c) Symmetric Bilateral Risk; d) Asymmetric Bilateral Risk

Once these terms are defined, we can move on to estimating the population parameters: mean and variance. An important element in estimating the mean is whether the population variance is known. As a result, we will consider these two cases when we talk about the estimation of the mean.

8.1. Estimating the mean (known population variance)

When population variance is known, estimating the population mean becomes somewhat straightforward. In such cases, one of the most used techniques is to use the Normal (z) distribution to construct confidence intervals and test hypotheses.

The formula for calculating the confidence interval when the population variance is known is as follows:

$$CI = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

CI - confidence interval limits

z - is the z-score, which can be looked up in the standard normal distribution table and corresponds to the desired confidence level (usually 90%, 95% or 99%).

σ - is the population standard deviation (considered known in this case)

n - is the sample size.

Let's see how it applies to a concrete example. A company produces several thousand parts every day. We are interested in, say, the average outside diameter of these parts. Since measuring all the parts would not be feasible, an employee takes a sample of $n = 100$ parts. He then calculates the average of the sample and gets $\bar{x} = 10.25$. Since the production process is known, the population standard deviation (and hence the variance) for this process is known $\sigma = 0.1$. Let's say we want to estimate the mean with a 95% confidence level, symmetric bilateral risk. The risk is $\alpha = 5\%$.

If the two limits of the confidence interval are x_1 and x_2 , then we can say that the mean (μ) lies between x_1 and x_2 with 95% probability and write:

$$P(x_1 < \mu < x_2) = 0.95$$

We also know that the risk is $\alpha = 5\%$ and is bilaterally symmetric, we will have half the risk ($\alpha/2 = 2.5\%$) below the lower bound of the confidence interval and the other half above the upper bound ($\alpha/2 = 2.5\%$).

To find x_1 and x_2 , we will need to use a special instance of the normal distribution called **the Standard Normal Distribution** (Figure 8.4). This distribution has mean value 0 and standard deviation equal to 1. The values represented on the axis are known as **z-scores**.

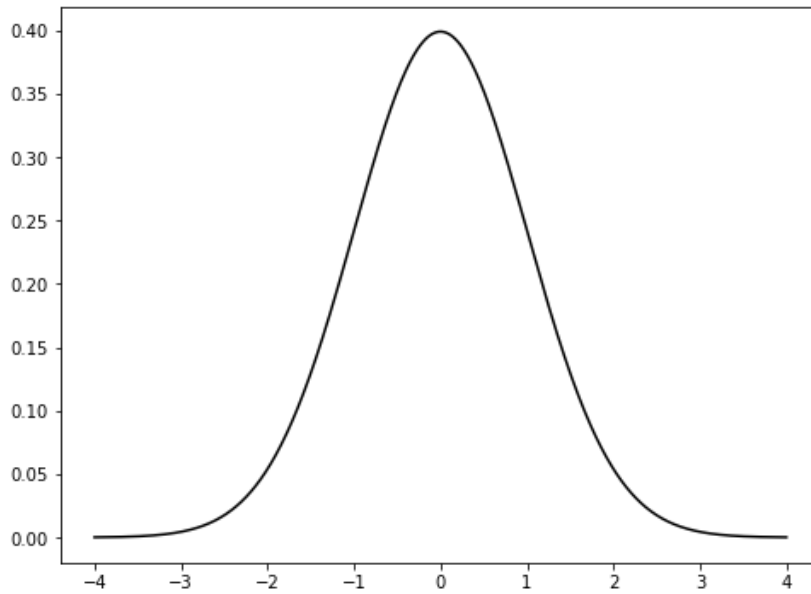


Fig. 8.4. Standard Normal Distribution

The z-scores that divide the Standard Normal Distribution in the same proportions as in our case (2.5% - 95% - 2.5%) are equivalent to the x_1 and x_2 values on our sample distribution.

The z-scores are read from a table. The table contains the z values "broken" into two parts: the first part up to the first decimal place is found in the first column of the table, and the second part corresponding to the second decimal place is found in the first row of the table (Figure 8.5). The corresponding probability ranges associated with each z-value are in the body of the table.



Fig. 8.5. Table of z-scores

To read from the Z Table, we must first identify the value in the table's body that is closest to the probability we seek. We move horizontally along the row of the identified value and ascertain the first part of the z-value in the first column. Then we go from the value of the area found upwards until we reach the first row of the table and read the second part of the value of z. For example, in Figure 8.6, we look for the value 0.0250 corresponding to the 2.5% risk on the left. Once we find it, we go horizontally down the row to the left and find the value -1.9 corresponding to the first

part of z. From the value 0.0250 we then go up the column and find the value 0.06 corresponding to the second part of the value of z. Combining the first part (-1.9) with the second part (0.06) we get the value of z = -1.96.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250

Fig. 8.6. Reading z-scores from the table

The full table is in Annex 1. You will notice that there are actually two tables: one for the risk on the left where z takes negative values and one for the risk on the right where z takes positive values. One thing to bear in mind when reading this table is that the area corresponding to the probability, we are looking for is defined between $-\infty$ and z. This means that the area increases as z increases. When z is 0, since the normal distribution is symmetric, the area will be exactly 50% (0.5000). As a result, to find the value of z corresponding to the risk to the right of 2.5%, we will need to find the area complementary to 2.5%, i.e. 97.5%. Looking at the table we will see that z will have a value of 1.96 for a risk of 2.5% on the right. This result should not be surprising given that the normal distribution is symmetric, and the risks are also symmetric.

Substituting the values obtained from the table together with the other known values in the formula for calculating the confidence interval limits, we obtain:

$$x_1 = 10.25 - 1.96 \frac{0.1}{\sqrt{100}} = 10.2304$$

$$x_2 = 10.25 + 1.96 \frac{0.1}{\sqrt{100}} = 10.2696$$

As a result, we can say with 95% confidence that the population average is between 10.2304 and 10.2696:

$$P(10.2304 < \mu < 10.2696) = 0.95$$

This example covers the case of symmetric risk. For one-sided risks you only need to determine one of the two limits (either left or right) and for asymmetric bilateral risk the calculation is done similarly, taking each limit in turn.

To apply this estimation method, we make some assumptions about our sample and its distribution:

- **Sample distribution is normal:** For a large sample size (usually $n \geq 30$), the central limit theorem says that the distribution of the sample mean is approximately normal. For smaller samples, this method only applies if the underlying population is normal.
- **Population variance is known:** This method assumes that the population variance is known, which is rarely the case in real-world scenarios.
- **Sampling was done randomly:** The method assumes that the sample is obtained by using a random sampling method.
- **Independence of observations:** The observations in the sample must be independent of each other.

8.2. Estimating the mean (population variance unknown)

If the population variance is unknown, the sample is not normally distributed. The distribution we will use in this case is called the Student distribution. Since the standard deviation of the population is unknown, we will use the standard deviation of the sample instead, as this is the best approximation we have at hand.

The problem has a formulation and solution similar to that of estimation when we know the population variance. We start by defining the confidence level:

$$P(x_1 < \mu < x_2) = 1 - \alpha$$

The formula we will use to determine the confidence interval limits will be:

$$CI = \bar{x} \pm t \frac{s}{\sqrt{n}}$$

where:

CI - confidence interval limits

t - is the t-score, which can be looked up in the Student distribution table and corresponds to the desired confidence level (usually 90%, 95% or 99%).

s - is the standard deviation of the sample

n - is the sample size.

As an example we will take the case of a sample of $n=5$ values, the sample mean $\bar{x} = 7.5$ the standard deviation of the sample $s = 1.5$ and asymmetric bilateral risk with one risk left $\alpha_1 = 2\%$ and right risk $\alpha_1 = 5\%$. We obtain the confidence level by

subtracting from 100% the total risk value (2%+5%=7%) i.e. 93%. As before, we need to find t-scores from a table to determine the confidence interval bounds (Appendix 2).

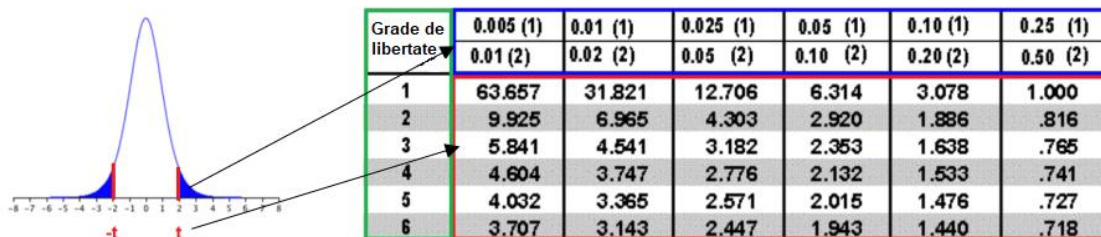


Fig. 8.7. Table of t-scores

In the t-table, the area corresponding to the risk is written in the first row of the table (for unilateral risk) and in the second row of the table (for symmetric bilateral risk). The t-value corresponding to the chosen risk is in the body of the table. The first column contains **the degrees of freedom** which are determined with the formula:

$$v = n - 1$$

We read the t-score at the intersection of the column that has the value corresponding to the risk and the row with the corresponding degrees of freedom.

In our example, we have 5 values, so we will have 4 degrees of freedom. On row 4 we will look for the values in the columns corresponding to the two risks. Since we will consider each end of the range separately, we will look on the first row of the table to read the risk values. In the table we do not have the value of 2% and we will choose the closest value, i.e. 2.5% (0.025). At the intersection of the row with 4 degrees of freedom and the column with the 2.5% risk we have the t-score value equal to 2.776. Since we are talking about left risk and the mean of the distribution is 0, we take the value with minus $t = -2.776$. For the right risk, we have the value of 5% (0.05) and the same 4 degrees of freedom, so the value of t is $t = 2.132$.

Once the t-score values have been determined, we input them into the formula:

$$x_1 = 7.5 - 2.776 \frac{1.5}{\sqrt{5}} = 5.638$$

$$x_2 = 7.5 + 2.776 \frac{1.5}{\sqrt{5}} = 9.362$$

As a result, we can say that the average population in this case is between 5,638 and 9,362 with a probability of 93%.

As with z-value estimation, there are a number of assumptions we make when estimating the mean and we do not know the variance:

- **Random sampling:** the sample must be randomly selected from the population.

- **Independence:** Observations must be independent of each other. This assumption is often met by random sampling.
- **Normality:** The sample should come from a normally distributed population.
- **Sample size:** While the t-distribution is especially useful for small samples, the sample size should not be extremely small. A common minimum is at least 5.
- **Measurement scale:** data should at least be interval or ratio scale, as these types of data allow meaningful interpretation of the arithmetic mean.
- **No outliers:** outliers can distort the mean and affect the spread, thus influencing both the estimated mean and the spread. Outliers should be carefully examined and managed accordingly.
- **Symmetry:** the t-test is more sensitive to departure from normality when the distribution is skewed. In such cases, non-parametric tests may be more appropriate.

The assumptions mentioned above should be tested before making the estimate. Otherwise, the results obtained may be biased.

8.3. Estimating population variance

Sometimes we might want to estimate the spread of population values. Variance is one of the indicators that characterize the spread of values.

Variance estimation can be useful in quality control where stability of variation is often as important as mean stability. A product may meet quality requirements on average, but if the variation is large, many individual items will be defective. In finance, variance (or standard deviation) is a measure of the volatility or risk of an investment. Knowing variance helps to test assumptions and build confidence intervals for other parameters.

To estimate the population variance, we need to know the number of values in the sample (n) and the variance (s^2) or standard deviation (s) of the sample. The way of working is similar to that of the mean estimation.

We define our problem as follows:

$$P(x_1 < \sigma^2 < x_2) = 1 - \alpha$$

Which means that σ^2 is between x_1 and x_2 , with a probability of $1 - \alpha$.

The formula we will use in this case is:

$$CI = \left[(n - 1) \frac{s^2}{\chi_{superior}^2}, (n - 1) \frac{s^2}{\chi_{inferior}^2} \right]$$

Score values χ^2 we need to find x_1 and x_2 are taken from the distribution table χ^2 (Chi-square). An extract from the table χ^2 is shown in Figure 8.8 and can be found in full in Appendix 3.

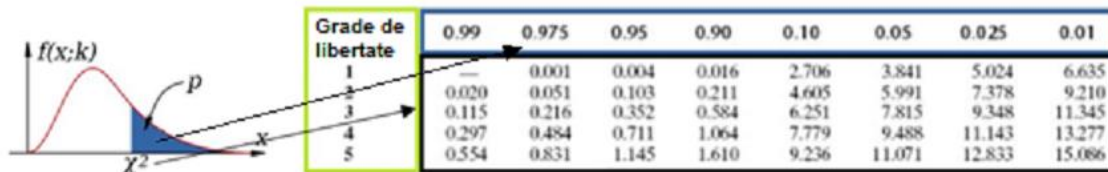


Fig. 8.8. Reading the Chi-square score table

The probability (area under the curve) is found in the first row of the table and the values χ^2 are in the body of the table. In the first column we again have the degrees of freedom as in the t-score table. χ^2 is located at the intersection of the column with the probability corresponding to the risk and the row corresponding to the degrees of freedom. When reading the table we have to take into account a few points. The risk area will increase from $+\infty$ to the left. Also note that for the lower value of the confidence interval we will use the upper value of χ^2 and for the upper value of the χ^2 lower range. This is because χ^2 is at the denominator of the fraction.

So a 1% risk to the right is in the 0.99 column, while a 1% risk to the left is in the 0.01 column.

Consider an example where we want to estimate the population variance with a symmetric bilateral risk of 10% knowing that the standard deviation of the sample is $s = 0.1$ and we have $n = 5$ values. From the bilateral risk of 10% we deduce two things: the confidence level is 90% and each risk has a value of 5% (left and right). From the table χ^2 on the row with 4 degrees of freedom and its 0.05 column we get the value $\chi^2_{superior} = 9.488$ and for $\chi^2_{inferior} = 0.711$ from its column 0.95. Plugging the data into the formula, we get:

$$CI = \left[(5 - 1) \frac{0.1^2}{9.488}, (5 - 1) \frac{0.1^2}{0.711} \right]$$

From which it follows $CI = [0.0042, 0.0562]$ indicating that the population variance in this case is in this range with a probability of 90%.

In order to estimate the variance correctly, the following assumptions must first be satisfied:

- **Random sampling:** The sample must be randomly selected from the population. This ensures that the sample is representative of the population, making the estimation more accurate.
- **Sample size:** Ideally, the sample size should be large enough to allow a robust estimation. Although there is no simple rule for what constitutes a "large"

sample, a common recommendation is at least 30 observations for the central limit theorem to come into play. However, the chi-square method for variance estimation is still applicable to smaller samples if the data are normally distributed.

- **Normality:** the data should be approximately normally distributed, especially if the sample size is small.
- **Independence of observations:** The values in the sample must be independent of each other. This assumption is typically satisfied through the random sampling process. If the data are dependent (e.g., time series data), alternative methods may be more appropriate for variance estimation.
- **Numerical data:** the data should be at interval or ratio level, as these are the types of data where it makes sense to discuss variance.
- **No outliers:** outliers can have a disproportionate effect on the variance, thus changing the estimate. The data should be checked for outliers, which should be handled appropriately before estimation.

Violation of these assumptions can lead to biased or misleading results. It is therefore important either to validate the assumptions as far as possible before proceeding with the analysis, or to use statistical techniques that are robust to violations of these assumptions.

We can summarize the three cases in Table 8.2.

Table 8.2. Estimation of population parameters (summary)

Estimate	Known parameters	Distribution	Statistic	Calculation formula
Population mean	$n, \bar{x}, \sigma^2, \sigma, \alpha$	Normal	z	$CI = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$
	$n, \bar{x}, s^2, s, \alpha$	Student	t	$CI = \bar{x} \pm t \frac{s}{\sqrt{n}}$
Population variance	n, s^2, s, α	Chi-square	χ^2	$CI = (n - 1) \frac{s^2}{\sigma^2}$

Estimating population parameters helps us in the decision-making process by providing additional information about our population. Even if the values obtained are not exact, they can still help us make decisions by giving us some threshold values with a certain probability.

Normal distribution is a practical tool for a variety of disciplines. Its prevalence in natural phenomena makes it essential for the social sciences in standardising tests and scores, allowing us to compare different data sets on a common scale.

Student distribution plays a critical role when dealing with small sample sizes, a common scenario in experimental and clinical studies where obtaining large samples may be impractical or impossible. It allows researchers to make inferences about population means with some level of confidence despite having limited data. This is particularly valuable in fields such as psychology and medicine, where individual variability is high and sample sizes are often limited.

The chi-square distribution plays a key role in the analysis of categorical data. It is the cornerstone of the chi-square independence test, which allows researchers to determine the relationship between categorical variables. This is extremely useful in fields such as genetics to assess the association of genetic traits, in marketing to assess consumer preferences, and in ecology to study the distribution of species.

These distributions are not just mathematical abstractions; they are the lenses through which we view and make sense of the world. They allow statisticians to draw meaningful conclusions from the data, test hypotheses and ultimately contribute to the advancement of knowledge in various fields. As such, understanding their properties and applications is important for any statistician, scientist or researcher who wants to make informed decisions based on data.

8.4. Knowledge check

1. Which parameter is estimated by the sample mean?
 - a. Median population
 - b. Population mode
 - c. Average population

2. The central limit theorem is important in estimation because:
 - a. Allows the sample mean to be used as a point estimate for the population mean.
 - b. It states that the distribution of sample means will be normally distributed, regardless of sample size.
 - c. Ensures that the sample variance is an unbiased estimator of the population variance.

3. When the population variance is unknown and the sample size is small, which distribution should be used for estimation?
 - a. Normal distribution
 - b. Binomial distribution
 - c. Student Distribution

4. What is the purpose of constructing a confidence interval?
 - a. To provide a range of values that is likely to contain the population parameter.
 - b. To accurately determine the population parameter.
 - c. To test a hypothesis about the population parameter.

5. The width of a confidence interval for estimating the mean of a population will:
 - a. Increases as sample size increases.
 - b. It decreases as the sample variance decreases.
 - c. It remains constant regardless of sample size.

6. The point estimate for population variance is:
 - a. Range of variation of the sample data.
 - b. Standard deviation of the sample.
 - c. Sample variance.

7. Which of the following is an example of a point estimate?
 - a. Sample mean
 - b. Confidence interval
 - c. Hypothesis testing

8. Confidence level (e.g. 95%) in the context of a confidence interval refers to:
 - a. Percentage of sample data falling within the range.
 - b. Probability that the interval includes the true population parameter.
 - c. Proportion of the population from which the sample was drawn.

9. If you wanted to estimate the average population with a higher degree of confidence, what would you do?
 - a. Increase the sample size.
 - b. Decrease the width of the confidence interval.
 - c. Use a higher alpha level.

10. To estimate the mean of a population when the variance is known, we use:
 - a. Normal distribution
 - b. Binomial distribution
 - c. Student Distribution

Correct answers

1. c. Average population
2. a. Allows use of the sample mean as a point estimate for the population mean.
3. c. Student Distribution
4. a. To provide a range of values that is likely to contain the population parameter.
5. b. Decreases as sample variance decreases.
6. c. Sample variance.
7. a. Sample mean
8. b. Probability that the interval includes the true population parameter.
9. a. Increase the sample size.
10. a. Normal distribution

9. Statistical process control

Statistical Process Control (SPC) is a method of monitoring, controlling and improving processes through statistical analysis. It comprises six tools used to identify problems and their causes for better process understanding and monitoring:

- Histogram
- Pareto diagram
- Dot diagram
- Control charts
- Cause-effect diagram
- Process diagram

Below we explore each of these tools, some in greater detail and others more briefly.

9.1. The histogram

Histograms are used to visualize the distribution of continuous data. By understanding the data distribution, we can draw conclusions about the process. The histogram has been discussed in previous sections and we will not dwell on it. Figure 9.1 illustrates a histogram of the diameters of parts manufactured in a factory.

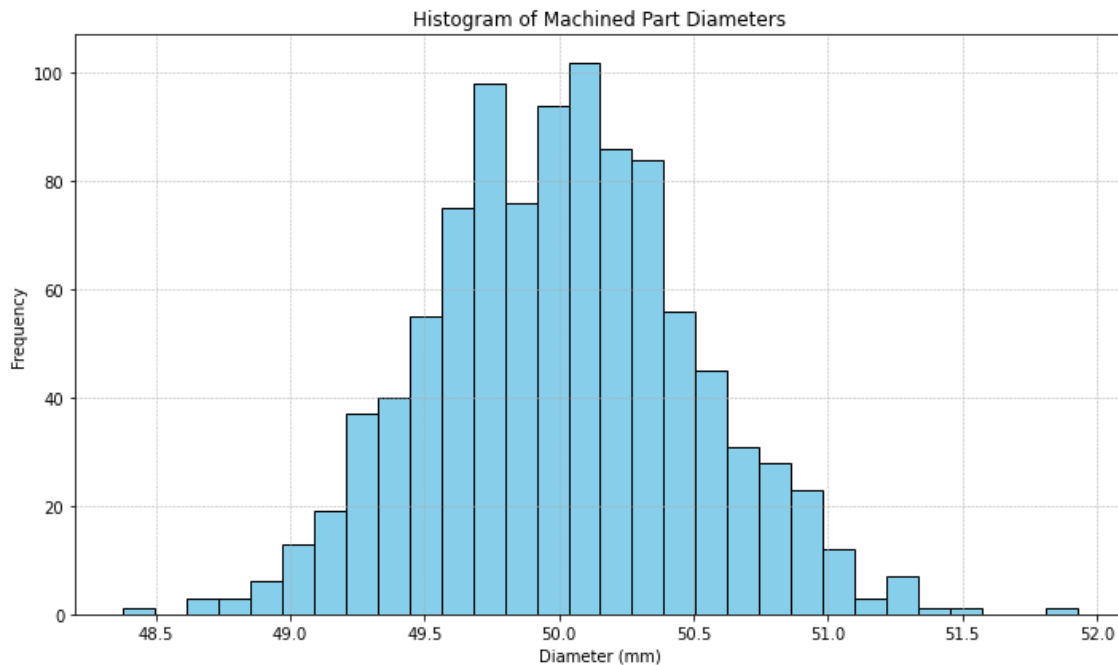


Fig. 9.1. An example of a histogram

With the histogram we can see the shape of the data distribution and whether it is close to a desired distribution, such as Normal.

9.2. Pareto chart

The main purpose of the Pareto chart is to highlight the most important factors in a data set. The Pareto chart, developed by Vilfredo Pareto in the 1800s, is based on the Pareto principle. This, often referred to as the 80/20 rule, is a principle that stating that 80% of the outcomes (or outputs) are the result of 20% of all the causes (or inputs) of a given event. In other words, in many situations, a small number of causes will produce the majority of outcomes or effects. For example: 80% of a company's profits might come from 20% of its customers; 80% of complaints might come from 20% of customers; 80% of software errors might be caused by 20% of known problems. The values (80/20) are not strict proportions but are indicative and illustrate disproportionality between inputs and outputs. The Pareto principle is a powerful tool that can help increase efficiency and effectiveness. By understanding which factors (the essential few) contribute to most outputs, efforts and resources can be concentrated on those critical areas, leading to better results and better use of resources.

In quality control, the Pareto chart is used to visualize this principle, helping to prioritize problems or causes that need to be addressed. By addressing the few major causes that lead to most problems, organizations can achieve significant improvements with relatively limited effort.

The Pareto chart (Figure 9.2) contains both bars and a line graph. Individual values are represented in descending order by bars, while the cumulative total is represented by the line.

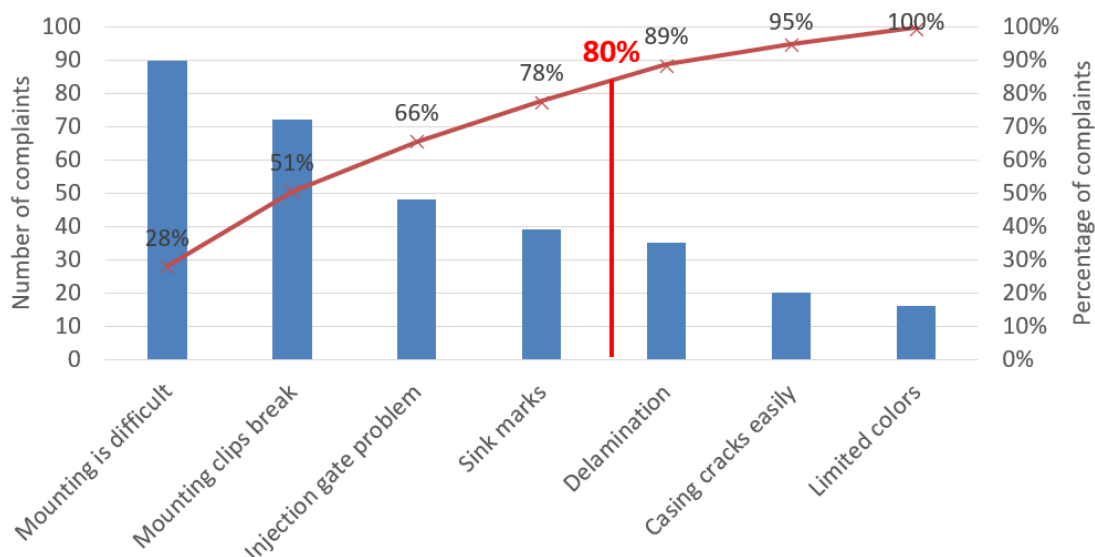


Fig. 9.2. An example of a Pareto chart

Creating a Pareto chart involves some systematic steps to ensure that it effectively visualizes the most important factors in a data set:

1. **Identification of categories:** Decide which categories or factors you want to analyze. These could be types of defects, sources of complaints, or any other classifications relevant to the issue at hand.
2. **Data collection:** Collect data for each category. This could involve counting defects of each type, counting complaints for each category, etc.
3. **Category ordering:** Arrange the categories in descending order by frequency.
4. **Calculate cumulative totals and percentages:** For each category, calculate the cumulative total and the cumulative percentage.
5. **Plot the graph:** Plot the categories on the horizontal axis, starting from the left with the category with the highest frequency. Plot the frequency for each category as bars. The height of the bar represents the frequency of that category. Using a second vertical axis on the right, plot the cumulative percentage as a line graph. This line should start at the top of the first bar and end at 100% on the last column.
6. **Drawing reference lines (optional):** You can add a reference line to the 80% mark on the cumulative percentage axis to easily identify the few vital categories that contribute to 80% of the effect.
7. **Graph analysis:** The leftmost columns (and the categories they represent) are usually the most significant contributions to the problem or situation. The rightmost bars represent categories that have less effect but may still need attention in specific contexts.
8. **Taking action:** Use the information in the Pareto chart to prioritize actions. Focus on the categories that have the most significant impact. Addressing problems in these areas can lead to the most substantial improvements. Be careful, this prioritization is done strictly in terms of frequency of occurrence and not the importance of the effect. Some causes, although rare, can have significant effects on the problem you are trying to solve. Use your judgement in prioritizing your actions.
9. **Review and update the chart:** Over time, as actions are taken and processes change, the distribution of issues by category may also change. It is important to review and update the Pareto chart periodically to reflect the current situation and to ensure that efforts remain focused on critical areas.

The categories that make up 80% of the total are the most common and should be addressed. In the example above, the first four categories make up about 80%.

9.3. Scatter plot

A scatter plot is a graphical representation in which each observation in the data set is represented as a point. The position of each point is determined by the value of

two variables: one variable determines the position on the x-axis and the other variable determines the position on the y-axis.

The purpose of the diagram is to visualize and compare the relationship or correlation between two quantitative variables. It can also be used to identify patterns, trends, clusters or outliers in the data.

The diagram is composed of (Figure 9.3):

- X-axis (horizontal axis): represents the values of a variable.
- Y-axis (vertical axis): represents the values of the second variable.
- Points: Each point on the graph represents a single observation in the dataset.

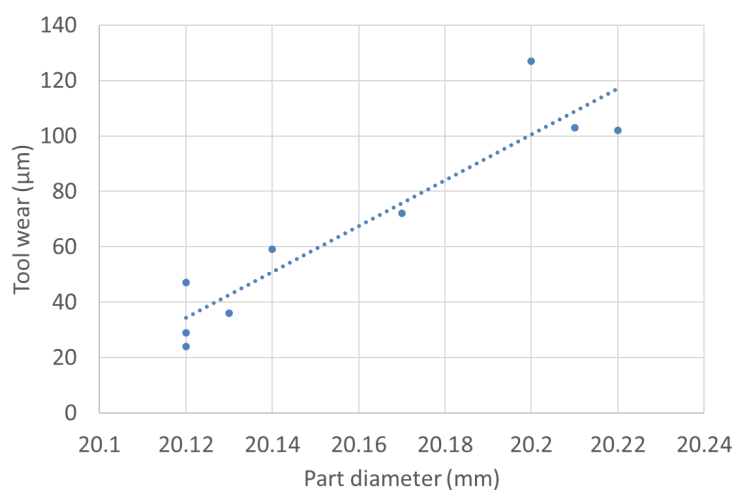


Fig. 9.3. An example of a dot plot

In the example in Figure 9.3, the X-axis represents the diameter of a part expressed in millimeters and on the vertical axis the wear of the tool used to obtain the part in micrometers. Each point represents a pair of diameter-wear values collected for each part. By representing several pairs of values, we can see that as the diameter of the part increases, the tool wear increases. As a result, the points illustrate an increasing trend visible on the graph.

The following steps must be taken to create such a chart:

1. Choose the variables you want to analyze.
2. Plot the values of a variable on the x-axis.
3. Plot the values of the second variable on the y-axis.
4. For each observation in the data set, mark a point where the x and y values intersect.

Dot plots, although simple, can help us understand relationships between two variables, identify trends and identify outliers.

9.4. Control charts

Production processes can be subject to different variations due to fluctuations in raw material quality, variations in machine settings, human error and more. If these variations are not monitored and controlled, they can lead to a lower quality product or service. Control charts help us identify and correct these variations before they become serious problems.

Control charts are widely used in quality engineering and process management to monitor processes and identify significant deviations from an established standard. Control charts can be used to measure multiple parameters such as mean or median and range or standard deviation. They help to identify variations in the process and help to distinguish between natural fluctuations and anomalies that require corrective action. They allow a proactive approach to quality control, minimizing defects and reducing waste.

9.4.1. Process capability

Control charts can also be used to determine whether a process can meet specified requirements. In this case we are talking about process capability, which is the ability of a process or machine to operate in a way that meets quality specifications. To determine process capability, we need to know:

- **LSL** - lower specification limit
- **USL** - upper specification limit
- **s** - standard deviation of the process

Specification limits can be tolerance limits or control limits of a process. The formula for process capability when using tolerance limits is:

$$Cp = \frac{UTL - LTL}{6s}$$

Where:

UTL - upper tolerance limit

LTL - lower tolerance limit

s - standard deviation of the data collected from the process

Another indicator of capability is the process capability index, Cpk. To determine it we need to calculate the distance to the mean of the \bar{x} and then the upper (Cps) and lower (Cpi) capability:

$$Cps = \frac{UTL - \bar{x}}{3s}$$

$$C_{pi} = \frac{\bar{x} - LTL}{3s}$$

From this we derive Cpk with the formula:

$$C_{pk} = \text{Min}(C_{pi}, C_{ps})$$

The higher the process capability, the better the process. A process with a capability index greater than 1.33 is considered capable, while one with a capability less than 1 is considered incapable. A process with a capability between 1 and 1.33 is borderline capable and needs some adjustments to become capable. Some companies use stricter limits than 1.33.

9.4.2. Elements of a control chart

A control chart monitors the evolution of the mean (or other central trend parameter) and the spread of the data. The target is placed on the central line, and the limits are on either side of it (upper and lower control limits). The control limits are usually placed within 3 standard deviations of each side of the mean. Warning limits can also be placed (Figure 9.4.). If warning limits are added, they are placed at 90% of the distance between the center line and the control limits.

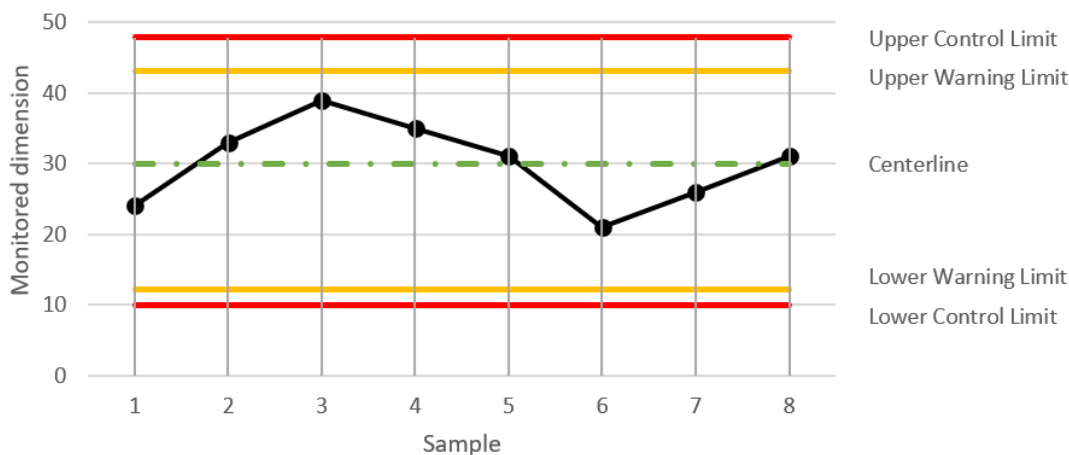


Fig. 9.4. Elements of a control chart

Each standard deviation from the mean divides the graph into 3 symmetric areas: A, B and C. Zone A is the furthest from the mean and zone C is the central zone (Figure 9.5.). These zones will help us to identify possible problems in the process.

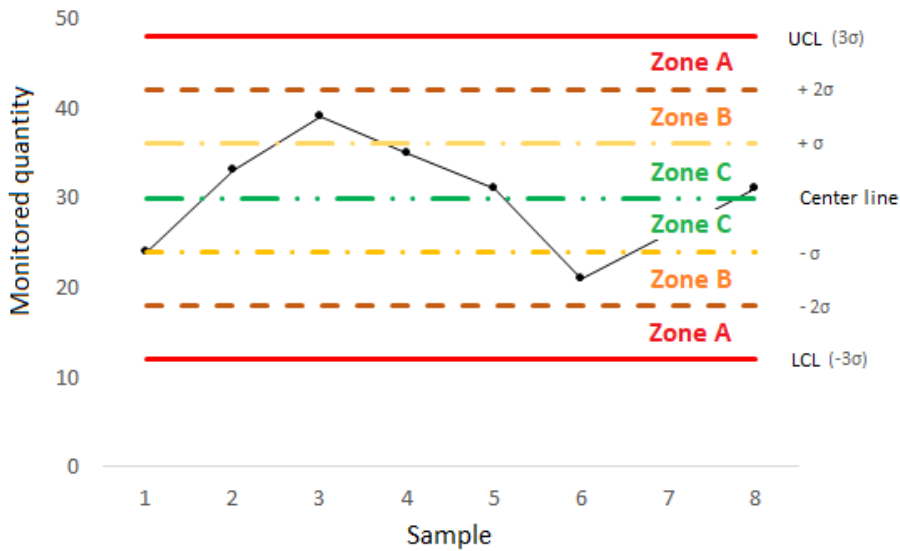
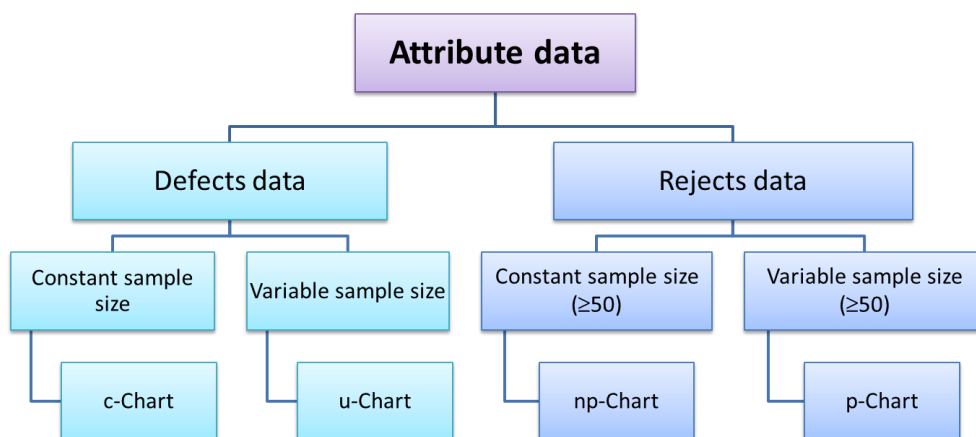


Fig. 9.5. Elements of a typical control chart and the three zones

In order to use control charts correctly, we must have a stable process that does not vary much over time, with normally distributed data, and control limits must fall on either side of the center line.

9.4.3. Types of control charts

There are different control charts depending on the type of data and sample size. Figure 9.6 shows the different types of control charts.



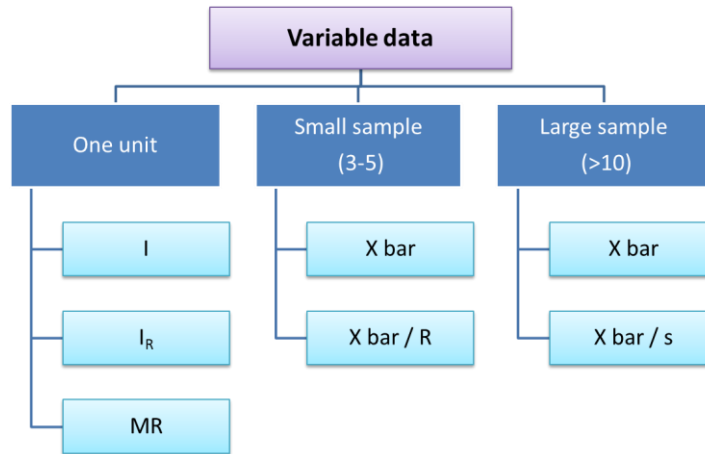


Fig. 9.6. Types of Control Charts by Data Type and Sample Size

We can have continuous (or variable) data and discrete (or attribute) data. For continuous data we have I/MR, X-Bar/R and X-Bar/S diagrams and for attribute data we have c, u, np and p diagrams. I/MR charts use individual values, while X-Bar charts are used when we have groups of samples. When the subgroup is small, we use the range (R) to characterize the spread and when the sample size is larger, we use the standard deviation (s).

Attribute control charts use discrete (countable) data. In quality control analysis, this countable data falls into one of two categories:

- **Defects** - represents the number of non-conformities of an item, such as a part. There is no limit to the number of possible defects. Defect charts count the number of defects in the inspection unit.
- **Rejects** - where the entire item is deemed not to conform to product specifications. Each item can have only one number associated with it: 1 or 0. Scrap charts count the number of scraps in a subgroup.

Depending on the sample size we can have different cards. The **c-type** card is used for fault monitoring and the sample size is constant. If the sample is of variable size, we will use type **u** cards. In the case of rejects, for samples of constant size we will use type **n** card and for samples of variable size, type **np** card.

C-charts

C-charts are used to monitor the number of faults in a given unit over time. This type of chart is particularly useful when we are interested in the total number of defects per unit rather than the proportion of defects. For example, you might use a C-chart to monitor the number of scratches on painted machine parts or the number of broken seals in batches of canned food.

As this chart represents the number of defects, the data in such a chart cannot be negative. This is why the lower control limit (LIC) is often zero. The C-type chart usually

assumes that defects follow a Poisson distribution. This assumption is important for the correct interpretation of control limits and other statistical properties. Because it uses count data, a Type C chart may be more sensitive to changes in the process that lead to an increase in the number of defects per unit, making it valuable for early detection of quality problems.

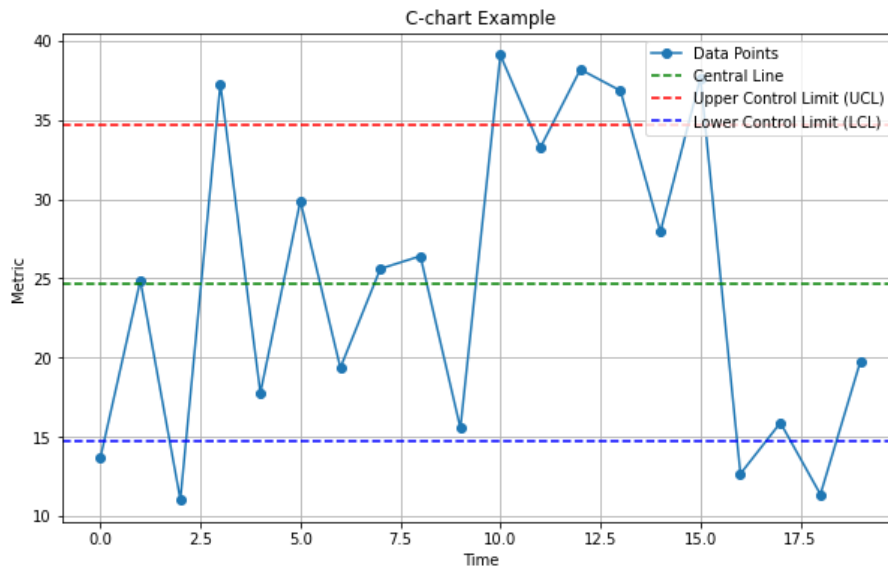


Fig. 9.7. An example of a c-chart

The chart consists of the following elements:

1. **Points** representing data collected: These represent the number of defects per unit in the process. For example, if you are monitoring the quality of machine painting jobs, each point might represent the number of scratches on a single machine.
2. **Center line**: This line represents the average number of defects across all units. It serves as the baseline for assessing process stability.
3. **Upper Control Limit (UCL)**: This line is calculated based on the average number of defects and represents the upper threshold for what is considered normal variation in the process [18], [19]:

$$LSC = \bar{c} + 3\sqrt{\bar{c}}$$

where:

$$\bar{c} = c/m$$

and

c- number of defects

m - number of samples

- Lower Control Limit (LCL):** Similarly, this line is calculated based on the average number of defects and represents the lower threshold for what is considered normal variation. [18], [19]:

$$LIC = \bar{c} - 3\sqrt{\bar{c}}$$

Since we cannot have negative results, this limit is considered 0 if the result of the calculation is negative:

$$LIC = \max(0, \bar{c} - 3\sqrt{\bar{c}})$$

- Time axis:** The horizontal axis represents the sequence in which data are collected, which is usually time-based.

u-Chart

A u-chart, where 'u' stands for 'unit', is an attribute control chart that displays how the frequency of defects, or non-conformances, changes over time for a process or system. The number of defects is collected for the opportunity area in each subgroup. The opportunity area can be either a group of items or just an individual item on which the defect count is performed. The u-chart is an indicator of the consistency and predictability of the level of defects in the process. A u-chart is appropriate when the area of opportunity for a defect varies from subgroup to subgroup

The u-chart chart (Figure 9.8) is best suited for monitoring the unit defect rate when the sample size varies. In contrast to the c-chart, which focuses on the total number of defects, the u-chart focuses on the defect rate, allowing a standardized comparison between samples of different sizes. It is particularly useful in scenarios where production volumes or batch sizes fluctuate.

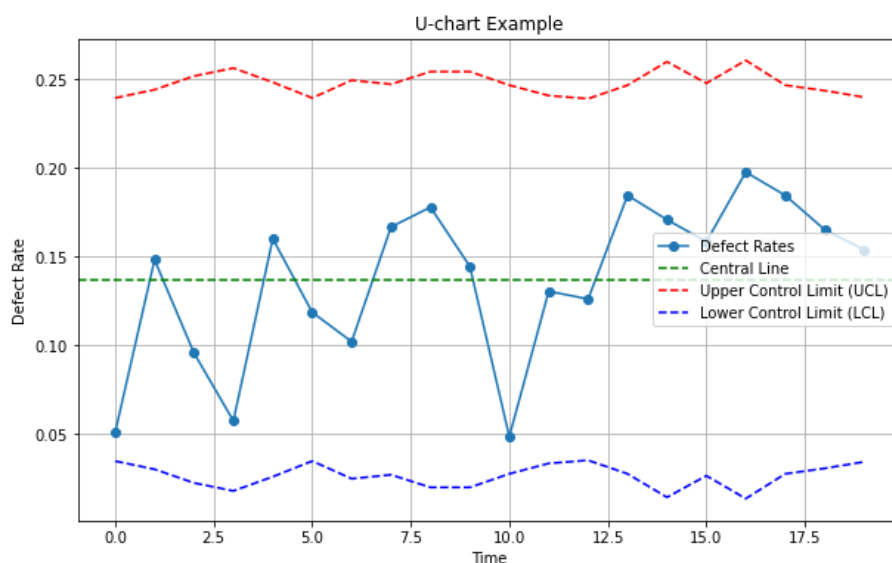


Fig. 9.8. An example of a u-Chart

np-Chart

The np-Chart (Figure 9.9) is designed to monitor the number of defective items in a sample of constant size. It is an indicator of the consistency and predictability of the level of defects in a process. The chart contains the same elements as the u-Chart, such as the center line and control limits, except that each point represents the number of defects in a sample, not in a unit.

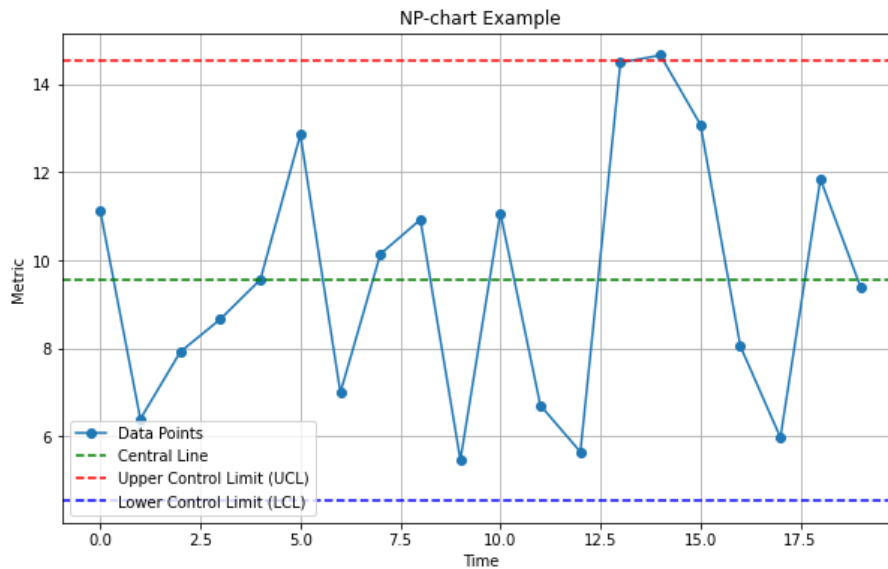


Fig. 9.9. An example of an np card

Imagine a scenario where you check the quality of tablets produced in a batch of 500. The np chart will show the number of defective tablets in each batch, helping you to identify any quality issues.

p-Chart

The p-Chart (Figure 9.10) is used to monitor the proportion of defective items in a sample. It is used when the subgroup size varies, and the chart displays the proportion or fraction of rejected items rather than the number rejected.

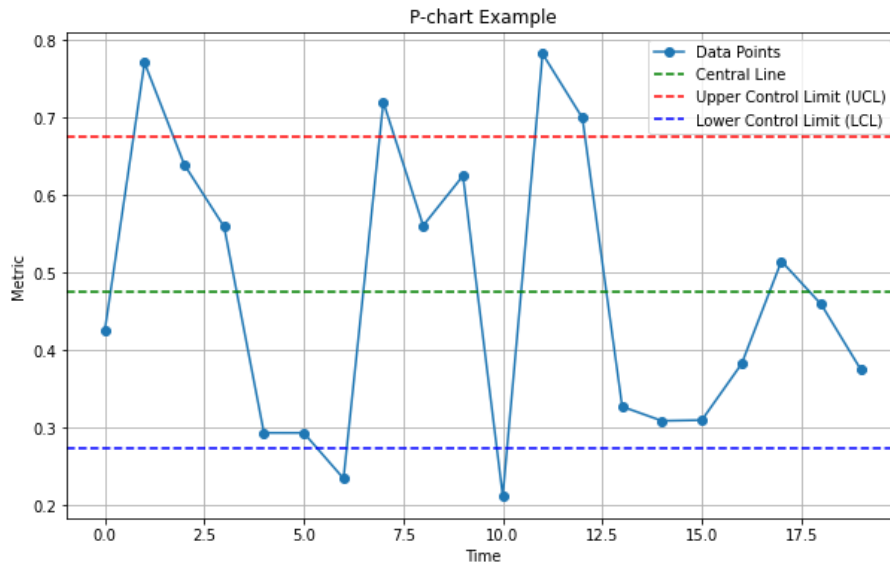


Fig. 9.10. An example of a p-Chart

p-Charts are focused on proportions or percentages, rather than the total number of defects as in C-charts.

These graphs often assume a binomial distribution of the data, making them suitable for large samples and proportions that are not extremely small or large.

I-Chart

The I-Chart (Figure 9.11) is used to monitor the variation of continuous, individual values over time. In this chart each point is an observation, such as the duration of an operation in minutes.

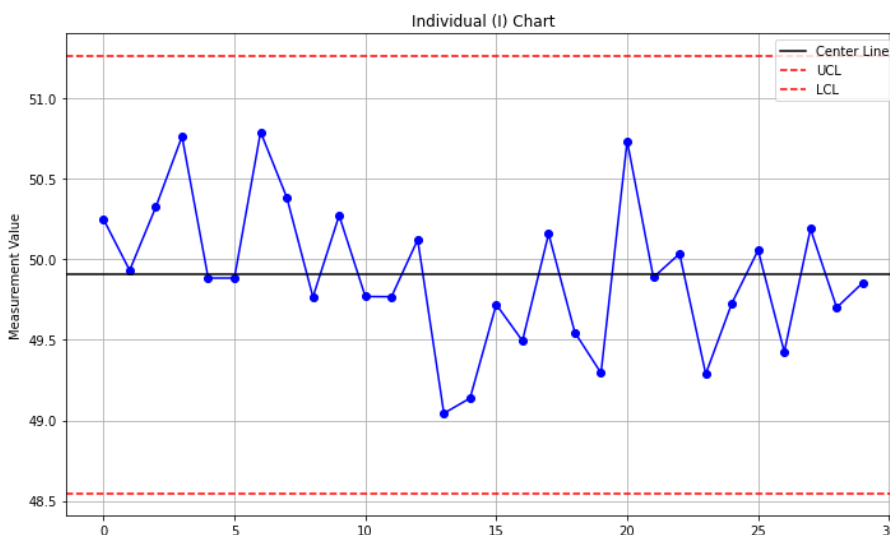


Fig. 9.11. An example of an I-Chart

This is a continuous value card and, unlike attribute data cards, will also have non-integer values. For example, we can have the diameter of a part of 23.57 mm. Because each point represents an individual value, the chart is sensitive to small changes in the

process unlike other charts that use averaging. This type of graph is used when our data follows or approaches a normal distribution. It is often used alongside a chart that tracks variation stability (such as the MR-Chart).

X-Bar chart

The X-Bar chart (Figure 9.12) is used to monitor the evolution over time of the mean of some values in a subgroup. Each point on the chart represents a sample (or subgroup) mean. Depending on the sample size, this can be used in conjunction with the R or S chart. The chart shows how consistent and predictable a process is.

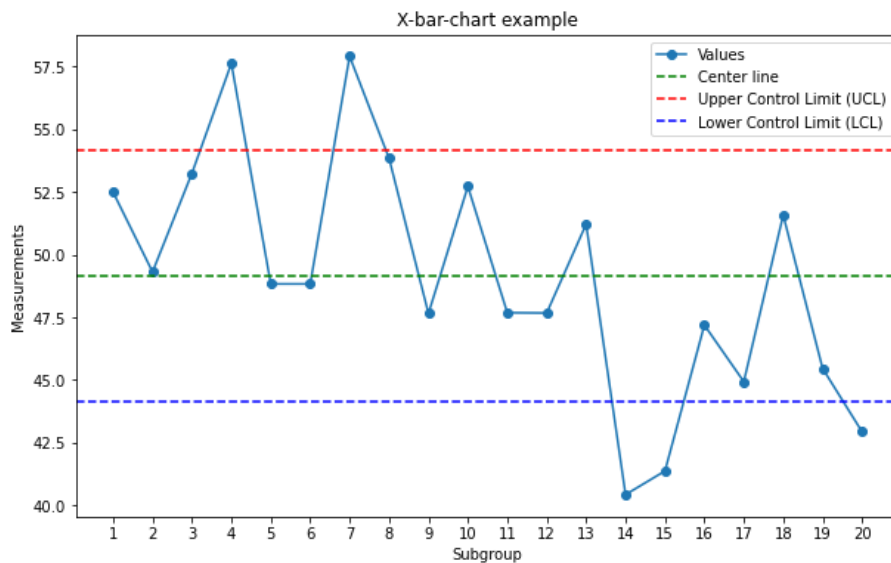


Fig. 9.12. An example of an X-Bar chart

MR-Chart

The MR (Moving-Range) chart is used to monitor data variation (Figure 9.13). It is used in conjunction with the I-Chart. Each value on the chart represents the difference between two consecutive observations. Since the difference represents the range, and this is applied successively to each pair of values, thus "moving" on the graph from left to right, the chart is called Moving-Range.

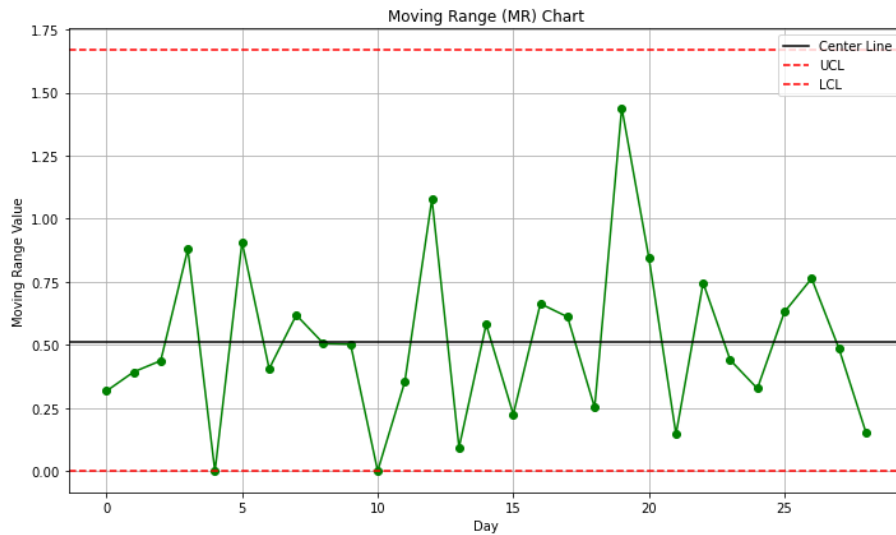


Fig. 9.13. An example of an MR chart

R-Chart

The R-Chart (Figure 9.14) is used in the analysis of the variation of data in a group of samples using the range. It is often used in conjunction with the X-bar chart. The sample size is usually small (less than 5 values). Each value in the chart represents the range of the subgroup of samples.

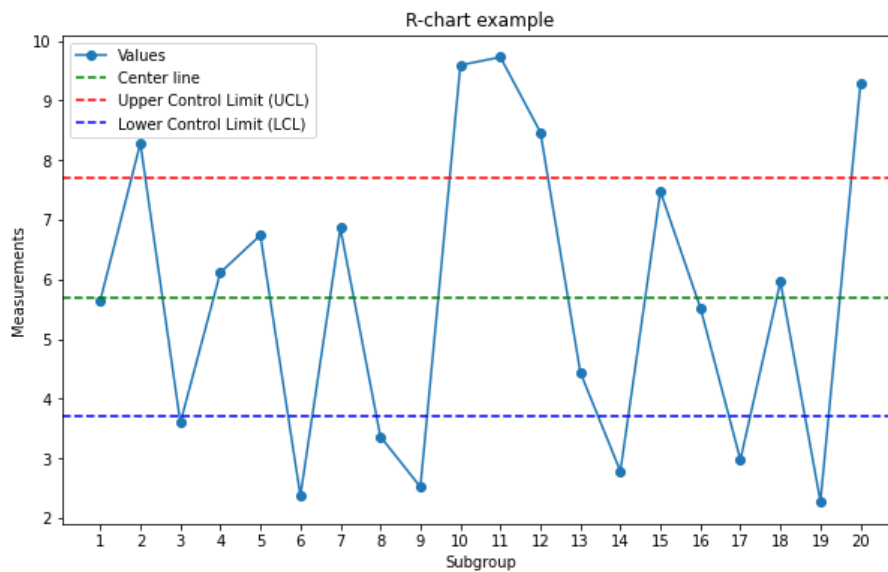


Fig. 9.14. An example of an R-Chart

Type S-chart

The S-chart (Figure 9.15) is used like the R chart to monitor data variation but is used when the subgroup size is larger (more than 5 values in a subgroup). For each subgroup, the standard deviation of the values in the subgroup is calculated and plotted on the graph. It can be used in conjunction with the X-bar chart.

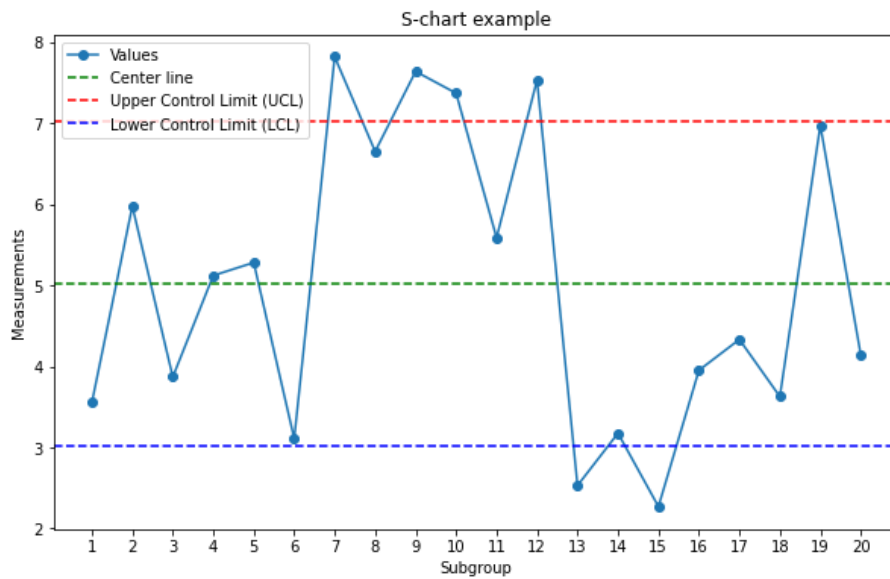


Fig. 9.15. An example of an S-Chart

9.4.4. Interpretation of a control chart

Control charts are used to identify potential problems in the process, with the aim of avoiding major process rejects or malfunctions. In interpreting control charts there are a few rules that help us in identifying potential problems in the process we are monitoring. Table 9.1 summarizes these rules.

Table 9.1. Troubleshooting a process with the control chart

Rule	Description
1. Points outside the boundaries	One or more points are out of bounds
2. Zone A test	2 out of 3 consecutive points are in Zone A or further
3. Zone B test	4 out of 5 consecutive points are in zone B or further
4. Zone C test	7 or more consecutive points are on one side of the average (in Zone C or beyond)
5. Trend	7 consecutive points are trending up or down
6. Mixing	8 consecutive points without a point in Zone C
7. Layering	15 consecutive points in Zone C
8. Supra-control	14 consecutive alternating points

If we have bridging outside the control limits (Figure 9.16) this indicates a problem with the monitored process. This can happen due to large variations from normal parameters and may be due to misconfiguration of production equipment, measurement errors or even the presence of a new employee. Another indicator of a large variation is if 2 out of 3 consecutive points are in zone A or beyond (Figure 9.16).

This is called the Zone A test and may indicate the omission of a production step, equipment failure, etc.

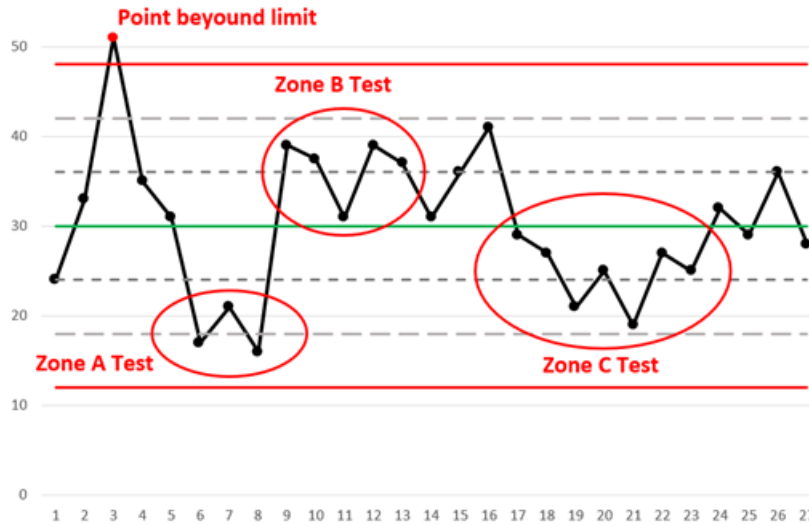


Fig. 9.16. Rule 1-4

If we observe 4 out of 5 consecutive points in Zone B or beyond, or 7 or more consecutive points are on one side of the mean (in Zone C or beyond), these may indicate small to medium variations from the nominal value. These represent the tests for Zone B and C respectively (Figure 9.16). Possible causes may be a change in work instructions, measuring devices or even material.

When 7 consecutive points on the chart have an up or down trend (Figure 9.17) this may indicate the effects of a gradually evolving phenomenon such as tool wear or tool heating.

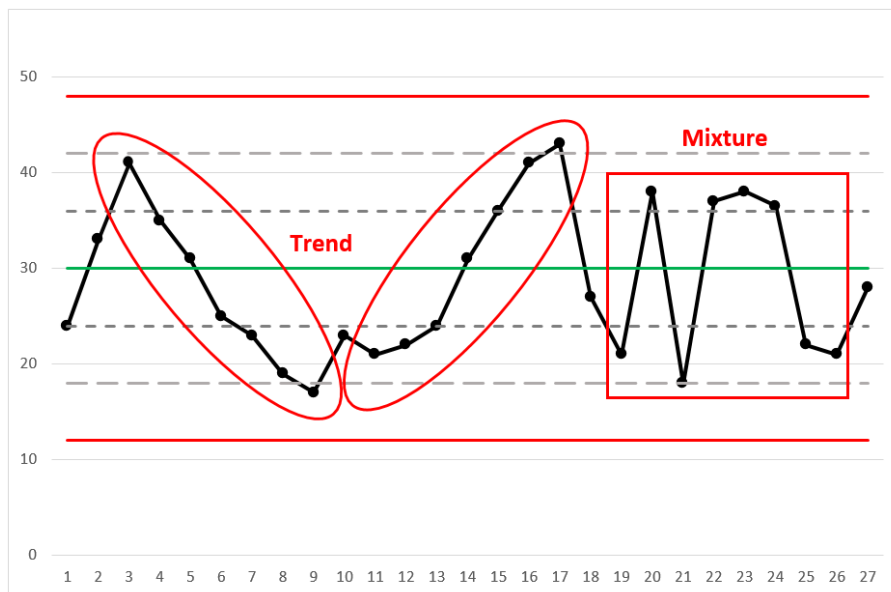


Fig. 9.17. Rule 5-6

The presence of 8 consecutive points with no points in area C indicates mixing of values from two distinct processes such as from two different machines or from the use of two different materials (Figure 9.17). The same causes can be identified when we have 15 consecutive points in area C. This phenomenon is called stratification (Figure 9.18).

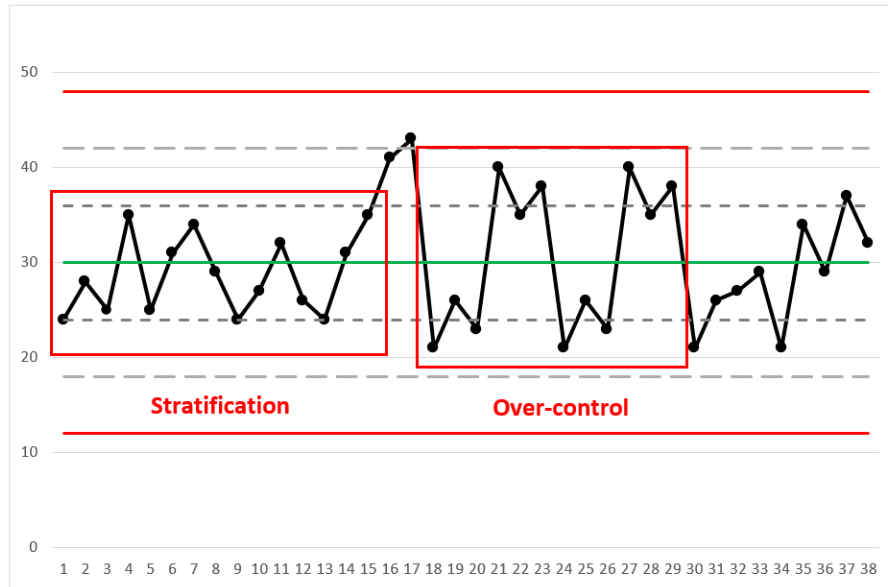


Fig. 9.18. Rule 7-8

When we have 14 consecutive alternating points or there are some repetitive patterns it may be an indication of overcontrol caused by operator manipulation of the data or the alternating use of more than one material.

The possible causes briefly described for the above rules are shown in Table 9.2

Table 9.2. Possible causes of rules observed in the control chart

Description of the event	Rules	Possible causes
Large variations from the average	1, 2	New employee; wrong configuration; measurement error; production step skipped; step not completed; power failure; faulty equipment
Small variations from the average	3, 4	Change in material; change in work instructions; different measuring device; different work shift; improvement of worker skills; change in maintenance schedule; change in installation procedure
Trends	5	Tool wear; thermal effects (cooling, heating)

Mixing	6	The existence of several processes (shifts, machines, materials)
Layering	7	The existence of several processes (shifts, machines, materials)
Overcontrol	8	Manipulation of data by the operator; Alternative use of more than one material

Whatever patterns are identified with the control chart, a first step is to stop the process being monitored and identify the causes. Once the causes are identified, the process can be resumed.

9.5. Cause-effect diagram

The cause-effect diagram, also known as the Ishikawa diagram or "fish bone" diagram, is used to determine the root cause for a particular effect or problem (Figure 9.19). Each root cause is linked to the backbone of the diagram. The common root causes for a technical process are: machine, method, measurement, materials, people (manpower) and environment, but these are adaptable to the process under analysis. Each root cause has sub-causes which in turn may have other sub-causes. This tool helps to visualize how different systems interact with each other and what root cause might exist for a particular effect.

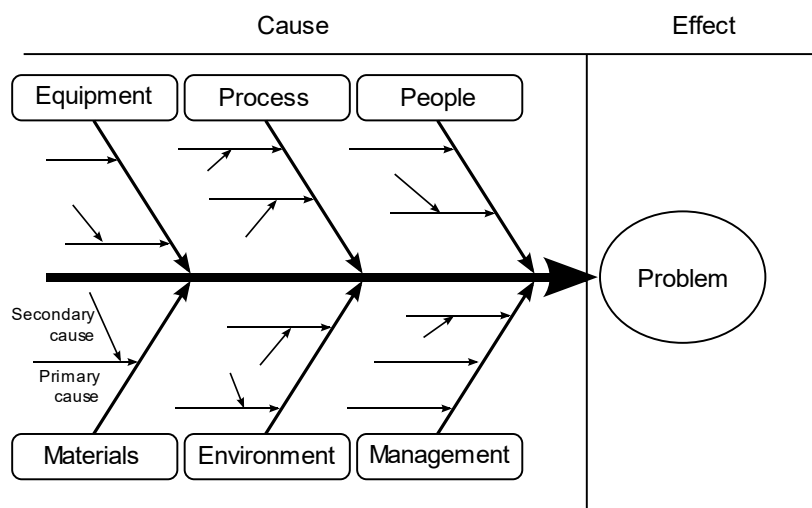


Fig. 9.19. An example of a cause-effect diagram [20]

For example, we can use the Ishikawa diagram to determine why our products have a higher percentage of defects than the accepted limit. By analyzing the production process from each perspective (Machine/Equipment, Materials, Process, Environment, Personnel and Management) we can identify primary, secondary and even tertiary causes leading to the high defect rate.

The diagram does not aim to identify a single cause but gives an overview of potential causes and their influence on the observed effect.

9.6. Process Diagrams

Flowchart is a tool that helps to visualize the steps of a process. It helps to better understand the process and to see how its different parts interact (Figure 9.20).

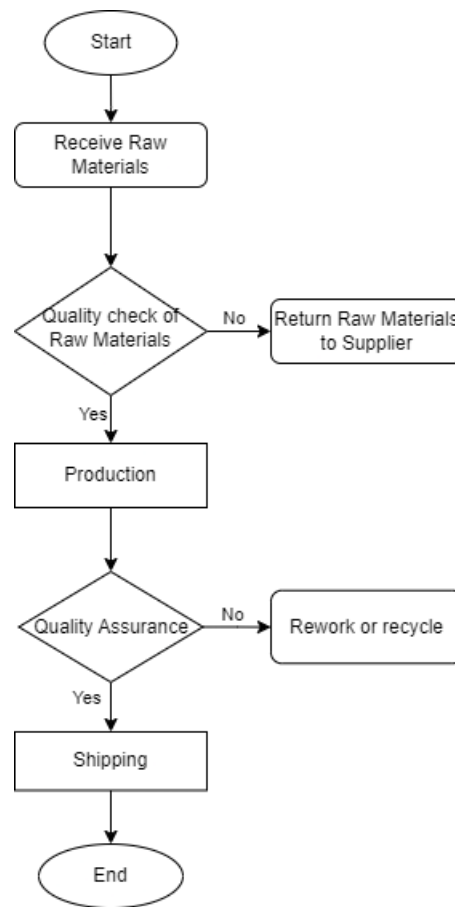


Fig. 9.20. An example of a process flow diagram

Each process can in turn be divided into sub-processes which can have their own diagrams. The diagram must be simple and easy to understand, otherwise it does not fulfil its purpose of giving an overview of the process.

9.7. Knowledge check

1. What is a control chart?
 - a. A musical instrument
 - b. A graph used to track the performance of a process
 - c. A geographical map
 - d. A type of fishing chart
2. What is the main purpose of a control chart?
 - a. Improving data aesthetics
 - b. Process quality monitoring and control
 - c. Creating animations
 - d. None of these
3. What is the center line (CL) in a control chart?
 - a. The control limit of the chart
 - b. The lower control limit
 - c. The process mean
 - d. The process variation
4. What type of data is most suitable for a P-chart?
 - a. Date continued
 - b. Attribute data
 - c. Categorical dates
 - d. Complex dates
5. In what situation is an X-bar chart useful?
 - a. When tracing defects on a part
 - b. When tracking subgroup averages
 - c. When tracking variation between subgroups
 - d. All the above
6. What are LSC and LIC?
 - a. Central Working Unit
 - b. Central Sustainability Limit and Innovation Limit Considered
 - c. Upper Control Limit and Lower Control Limit
 - d. None of these
7. Which of the following is an indicator of an unstable process?
 - a. Random points throughout the chart
 - b. Points centered around the center line
 - c. Seven or more consecutive points above or below the center line
 - d. All points are between the control limits

8. What type of chart is used to monitor the number of defects per unit?
- p-Chart
 - c-Chart
 - X-bar card
 - R-chart
9. What type of card is used to monitor process variation?
- p-Chart
 - X-bar Chart
 - u-Chart
 - S-Chart
10. Which of the following statements is true for an I-MR chart?
- Uses subgroup averages
 - Uses individual data
 - It is only used for categorical data
 - None of these

Correct Answers

1. b. A graph used to track the performance of a process
2. b. Process quality monitoring and control
3. c. The process mean
4. b. Attribute data
5. b. When tracking subgroup averages
6. c. Upper Control Limit and Lower Control Limit
7. c. Seven or more consecutive points above or below the center line
8. b. C-Chart
9. d. S-Chart
10. b. Use individual data

10. Correlation and regression

This chapter will guide you through the basic principles of correlation and regression which are powerful statistical tools, illustrates their application with practical examples and lays the foundations for more advanced statistical modelling. Whether you are predicting economic trends, analyzing scientific data or simply trying to understand the world a little better, an understanding of correlation and regression is indispensable.

Correlation provides a quantifiable measure of how variables change together. We will discuss the correlation coefficient, a statistic that captures the degree to which two variables are linearly related. This concept is essential for many applications: in finance, it aids in portfolio diversification, and in healthcare, it assists in identifying disease risk factors.

Linear regression helps us model the relationship between an independent variable and a dependent variable. Regression analysis not only allows us to describe the association but also to make predictions, control for confounding variables and even infer causality under the right conditions.

10.1. Correlation

The Pearson correlation coefficient is used to determine the strength and direction of a linear relationship between two continuous variables [21]. Specifically, the test uses a coefficient called the **Pearson correlation coefficient**, denoted as r . This coefficient's value ranges from -1, indicating a perfect negative linear relationship, to +1, indicating a perfect positive linear relationship. A value of 0 (zero) indicates that there is no relationship between the two variables. The correlation can be visualized using a dot plot (Figure 10.1).

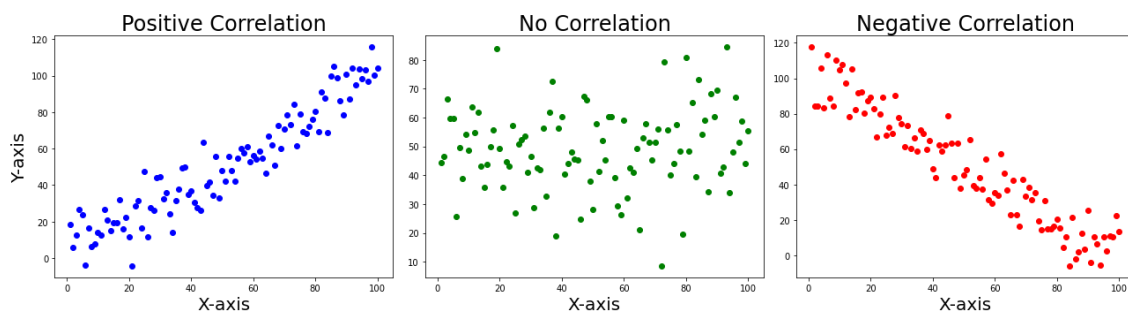


Fig. 10.1. Types of correlations

In the case of a positive correlation, the points are arranged in ascending order. The less the points are spread out and tend to be arranged on a straight line, the closer the coefficient r is to 1. Similarly if the points are arranged in decreasing order, the correlation is negative. If we cannot distinguish an increasing or decreasing trend, then we have no correlation.

To calculate the correlation coefficient correctly, some preconditions must be met:

- The measurement level must be interval or ratio for both variables
- Variables must be approximately normally distributed
- The association between the two variables should be linear
- Data should be free of outliers

For example, we could investigate whether there is a link (or correlation) between the height and weight of some people in a focus group. In this case, we have two variables measured on a ratio scale. If there are no disturbing factors, height and weight are usually normally distributed in a population. One way we can increase our chances of having normally distributed data is to have larger sample sizes (>30). Weight and height are usually linearly associated. Taller people tend to have higher weight, and shorter people lower weight. Of course, this is not a strict rule and we may have exceptions. Outliers can be addressed using various methods.

The formula for calculating the correlation coefficient r is:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where:

x_i - the values of the first variable

y_i - the values of the second variable

\bar{x} - average of the values of the first variable

\bar{y} - average of the values of the second variable

There are several types of correlation coefficients (Spearman, Kendal, etc.) but the Pearson correlation is the most commonly used. If one or more of the preconditions are violated then other types of correlation are used. For example, if the data are not normally distributed or if we are working with ordinal data, we can use the Spearman correlation coefficient which makes no assumptions about the distribution of the data and uses ranks for the calculation instead of absolute values.

It is very important to understand that the existence of correlation does not imply causality. In other words, if two variables are correlated, it does not mean that one

determines the other. There may be other hidden factors causing the correlation between the variables under analysis. For example, if there are more admissions of people with sunstroke during the summer on the one hand and a higher consumption of ice cream on the other, although we can say that the two variables are correlated, we cannot say that if we consume more ice cream we get sunstroke.

10.2. Linear regression

Simple linear regression enables us to predict the value of one variable based on another [22]. The variable we predict is called the dependent or predicted variable and the variable used for prediction is called the independent or predictor variable. We use linear regression when the relationship between the two variables is linear

We can use regression to:

- determine whether the linear relationship between two variables is statistically significant,
- calculate how much of the variation in the dependent variable is explained by the independent variable,
- to understand the direction and nature of the relationship
- predict values of the dependent variables based on different values of the independent variable.

The basic idea is to get a line that best fits our data. The best-fitting line is the one that minimizes the distances between the points and the regression line. The distance between each point and the regression line is called the total prediction error.

The equation of the regression line is:

$$Y = a * X + b$$

where:

Y - dependent variable

X - independent variable

a, b - parameters of the regression line

The prediction error is calculated by summing the squares of the differences between the observed values (from the data set) and the predicted values:

$$\Delta = \sum (y - \hat{y})^2$$

where:

y - observed values

\hat{y} - predicted values

The graphical representation of a regression line for two variables X and Y is shown in Figure 10.2.

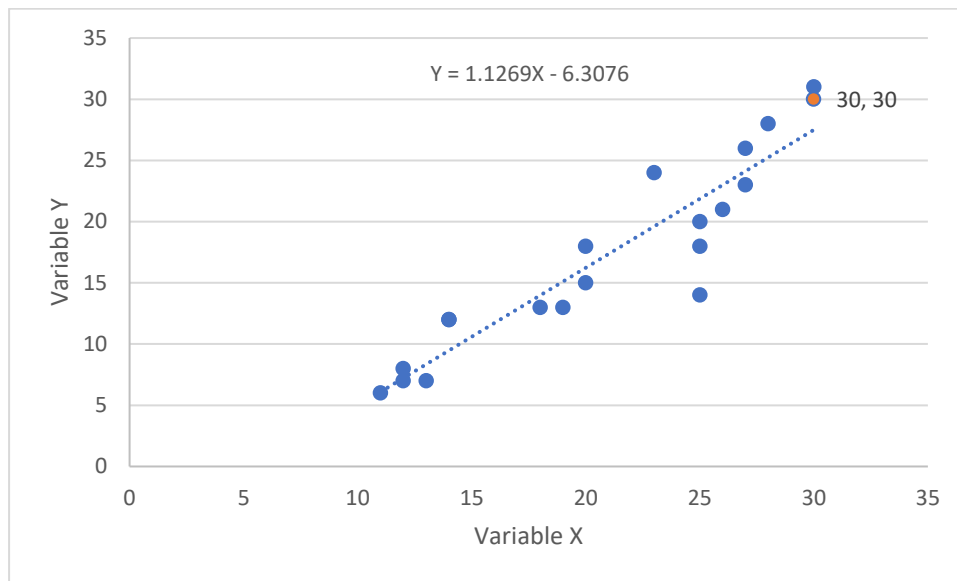


Fig. 10.2 Regression line for two variables X and Y.

Each observation is represented by a point that has the values of variables X and Y respectively. For example, the point marked in red has the values X=30 and Y=30. The best-fitting line for these data has the equation:

$$Y = 1.1269 * X - 6.3076$$

This is called the regression model. In this model, the coefficient of X is called the regression slope and tells us the average change in Y for a one unit change in X. The free coefficient represents the value of Y when X is 0 and is called the free term.

The quality of the model obtained can be assessed using the coefficient of determination (R^2). This is a statistical measure of how close the data are to the regression line found. R^2 is the percentage of the variation in the dependent (response) variable explained by the linear model:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

This indicator takes values between 0 and 1, where 0 indicates that the model does not explain the variability of the response data around its mean and 1 indicates that the model explains all the variability of the response data around its mean. In general, the higher the R^2 , the better the model fits the data. A quality model has high coefficient values (0.8 - 0.9) but the quality threshold varies by domain, application, etc.

Linear regression is based on several key assumptions, which are necessary to obtain reliable results:

1. **Linearity:** The relationship between the independent variables and the dependent variable is linear. This can be checked by using dot plots or by assessing the correlation coefficient.
2. **Independence:** Observations are independent of each other. This is a key assumption in regression models that assumes that observations of the dependent variable are collected without any related influence.
3. **Homoscedasticity:** The errors have a constant variance at each level of an independent variable. In other words, the variance or "spread" of errors should be approximately the same at all levels of the independent variables (homoscedasticity), as opposed to having a variance that increases or decreases with the adjusted values (heteroscedasticity).
4. **Residuals normality:** Residuals in the model should be approximately normally distributed. This assumption is not necessary for estimating the coefficients themselves, since the least squares method that estimates the coefficients is non-parametric. However, normality is necessary for the proper construction of confidence intervals and hypothesis tests.
5. **No or little multicollinearity:** Multicollinearity occurs when independent variables are highly correlated with each other. This can make it difficult to determine the individual effect of each independent variable on the dependent variable due to redundancy of information. It is preferable to have independent variables that are not highly correlated (coefficients greater than 0.8-0.9).
6. **No endogeneity:** regressors (X) must not be correlated with the error term (ϵ). Endogeneity may occur due to omitted variable bias, measurement error, or some type of simultaneous causality between independent and dependent variables.
7. **Residuals are independent of predictors:** residuals are uncorrelated with predictor variables. This ensures that the predictors have a consistent and unbiased influence on the dependent variable Y .

If these assumptions are not met, the results of the regression analysis may be unreliable or invalid. It is important to test these assumptions and apply remedial measures as appropriate, which may include transforming variables, adding variables to the model or using alternative estimation techniques.

10.3. Knowledge check

1. What is the slope in a simple linear regression equation?
 - a. Predicted value of Y when X is zero
 - b. Average change in Y for a one unit change in X
 - c. Correlation between X and Y
2. In regression analysis, what does R-squared (R^2) represent?
 - a. The proportion of variation in the dependent variable that can be predicted by the independent variable.
 - b. Variance of residues.
 - c. The correlation coefficient between X and Y.
3. Which of the following indicates the strongest relationship between two variables in a linear regression model?
 - a. $R^2 = -0.8$
 - b. $R^2 = 0$
 - c. $R^2 = 0.9$
4. What is the main purpose of using regression analysis?
 - a. To describe the association between variables.
 - b. To predict the value of a dependent variable based on the value of at least one independent variable.
 - c. To prove the cause-effect relationship between variables.
5. Which of the following assumptions is NOT required for linear regression analysis?
 - a. Homoscedasticity
 - b. Normal distribution of variables
 - c. Independence of errors
6. Which method is commonly used to find the best-fit line in a simple linear regression?
 - a. Least squares estimation
 - b. Maximum likelihood estimation
 - c. Estimating the mode
7. What does a correlation coefficient of zero mean?
 - a. Perfect positive correlation
 - b. Perfect negative correlation
 - c. No correlation
8. If the dot plot of two variables forms a perfectly straight, downward sloping line, what is the correlation coefficient?
 - a. 0

- b. 1
 - c. -1
9. Which of the following is a true statement about correlation?
- a. Correlation implies causation.
 - b. Correlation measures the strength and direction of a linear relationship between two variables.
 - c. Correlation can only be found in linear models.
10. Under what conditions should the Pearson correlation coefficient not be used?
- a. When the relationship is non-linear.
 - b. When variables are on different scales.
 - c. When the dataset contains outliers.

Correct answers

1. b. Average change in Y for a one unit change in X
2. a. The proportion of variation in the dependent variable that can be predicted by the independent variable.
3. c. $R^2 = 0.9$
4. b. Predicting the value of a dependent variable from the value of at least one independent variable.
5. b. Normal distribution of variables
6. a. Least squares estimation
7. c. No correlation
8. c. -1
9. b. Correlation measures the strength and direction of a linear relationship between two variables.
10. a. When the relationship is non-linear.

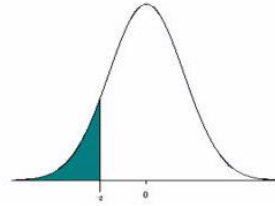
11. References

- [1] Cambridge Dictionary, "data," Cambridge Dictionary. Accessed: Jul. 24, 2022. [Online]. Available: <https://dictionary.cambridge.org/dictionary/english/data>
- [2] C. Zins, "Conceptual approaches for defining data, information, and knowledge," *J. Am. Soc. Inf. Sci. Technol*, vol. 58, no. 4, pp. 479-493, 2007, doi: 10.1002/asi.20508.
- [3] R. L. Ackoff, "From data to wisdom," *J. Appl. Syst. Anal.*, vol. 16, pp. 3-9, 1989.
- [4] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *J. Inf. Sci.*, vol. 33, no. 2, pp. 163-180, Apr. 2007, doi: 10.1177/0165551506070706.
- [5] D. Chaffey and S. Wood, *Business Information Management: Improving Performance Using Information Systems by Chaffey, Dave, Wood, Steve (2004) Paperback*.
- [6] "Definition of VARIABLE." Accessed: Aug. 10, 2022. [Online]. Available from: <https://www.merriam-webster.com/dictionary/variable>
- [7] O.Theobald, *Statistics for Absolute Beginners: A Plain English Introduction*. Amazon Digital Services LLC - KDP Print US, 2017.
- [8] C. Mckay, *Probability and Statistics*. Scientific e-Resources, 2019.
- [9] Incnis Mrsi, "Discrete probability distribution illustration." Wikimedia. Accessed: May 15, 2023. [Online]. Available from: [https://commons.wikimedia.org/wiki/File:Discrete probability distribution illustration.svg](https://commons.wikimedia.org/wiki/File:Discrete_probability_distribution_illustration.svg)
- [10] "Binomial distribution," *Wikipedia*. May 06, 2023. Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/w/index.php?title=Binomial_distribution&oldid=1153481251
- [11] "Hypergeometric distribution," *Wikipedia*. May 15, 2023. Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/w/index.php?title=Hypergeometric_distribution&oldid=1154916959
- [12] "Continuous uniform distribution," *Wikipedia*. Mar. 13, 2023. Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/w/index.php?title=Continuous_uniform_distribution&oldid=1144473074
- [13] "1.3.6.6.1. Normal Distribution." Accessed: May 15, 2023. [Online]. Available from: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm>
- [14] "Normal distribution - Wikipedia." Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/wiki/Normal_distribution
- [15] "Empirical Rule (68-95-99.7) Explained | Built In." Accessed: May 15, 2023. [Online]. Available from: <https://builtin.com/data-science/empirical-rule>
- [16] "Student's *t*-distribution," *Wikipedia*. May 09, 2023. Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/w/index.php?title=Student%27s_t-distribution&oldid=1153912628
- [17] "Chi-squared distribution - Wikipedia." Accessed: May 15, 2023. [Online]. Available from: https://en.wikipedia.org/wiki/Chi-squared_distribution
- [18] L. A. Doty, *Statistical Process Control*, 2nd edition. New York: Industrial Press, Inc. 1996.
- [19] S. Glen, "C Chart: Definition, Formulas," Statistics How To. Accessed: Oct. 10, 2023. [Online]. Available from: <https://www.statisticshowto.com/c-chart/>

- [20] F. at de.wikipedia, *Ishikawa fishbone-type cause-and-effect diagram*. 2008. Accessed: Oct. 25, 2023. [Online]. Available at: https://commons.wikimedia.org/wiki/File:Ishikawa_Fishbone_Diagram.svg
- [21] "Pearson's Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example." Accessed: Oct. 25, 2023. [Online]. Available from: <https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php>
- [22] "Linear Regression Analysis in SPSS Statistics - Procedure, assumptions and reporting the output." Accessed: Oct. 25, 2023. [Online]. Available from: <https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php>

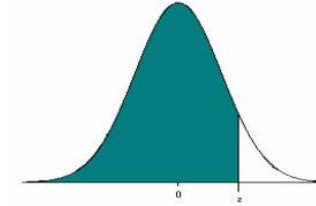
Annex 1 - Table for normal distribution (z-values)

Table of Standard Normal Probabilities for Negative Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

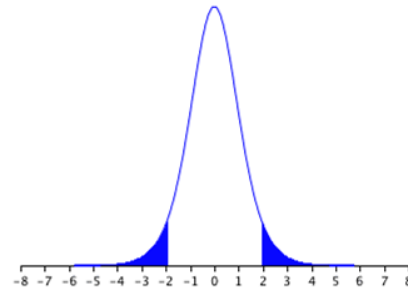
Table of Standard Normal Probabilities for Positive Z-scores



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Note that the probabilities given in this table represent the area to the LEFT of the z-score.
The area to the RIGHT of a z-score = 1 – the area to the LEFT of the z-score

Appendix 2 - Student distribution table (t-values)

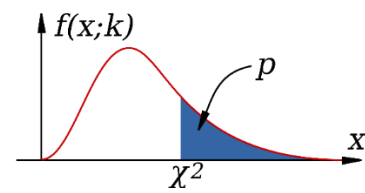


Significance level = α

Degrees of Freedom	0.005 (1)	0.01 (1)	0.025 (1)	0.05 (1)	0.10 (1)	0.25 (1)
	0.01 (2)	0.02 (2)	0.05 (2)	0.10 (2)	0.20 (2)	0.50 (2)
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.878	2.485	2.060	1.708	1.316	.684
26	2.779	2.479	2.056	1.706	1.315	.684
27	2.771	2.473	2.052	1.703	1.314	.684
28	2.763	2.467	2.048	1.701	1.313	.683
29	2.756	2.462	2.045	1.699	1.311	.683
Large	2.575	2.327	1.960	1.645	1.282	.675

(1) - unilateral; (2) - bilateral symmetric

Appendix 3 - Chi-square distribution table (χ^2 values)

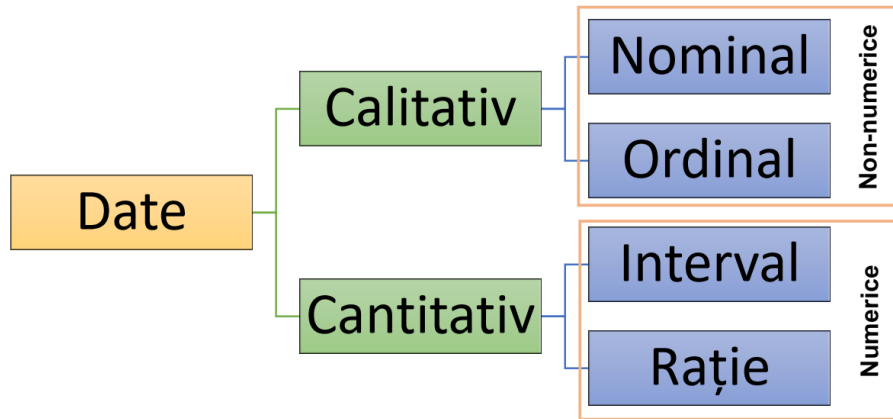


Chi-Square (χ^2) Distribution								
Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Annex 4 - Summary of concepts

Descriptive statistics

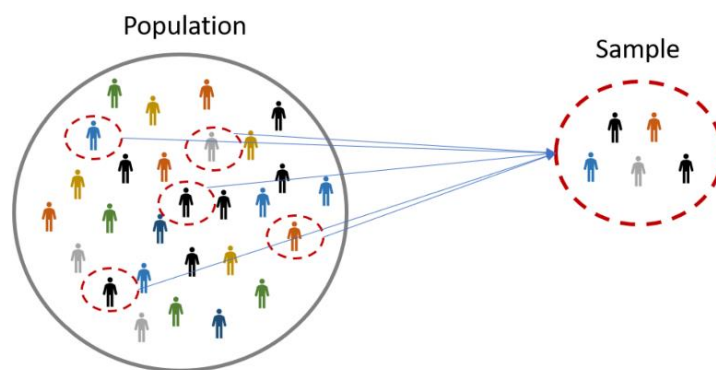
Measurement levels



Frequency

Absolute	Relative
Number of items	Proportion (percentage) of elements $\text{Relative frequency} = \frac{\text{absolute frequency}}{\text{total number of elements}}$

Population and sample



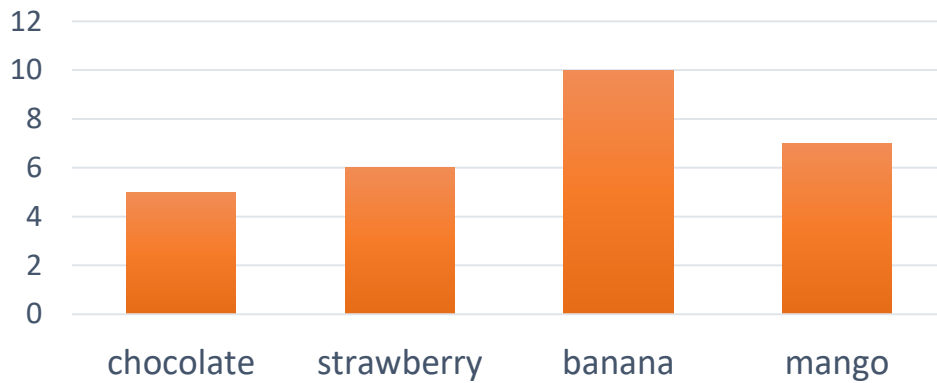
<p>Population parameters</p> <p>μ - average population</p> <p>σ^2 - population variance</p> <p>σ - population standard deviation</p>	<p>Sample characteristics:</p> <p>\bar{x} - sample mean</p> <p>s^2 - sample variance</p> <p>s - standard deviation of the sample</p>
--	--

Indicators

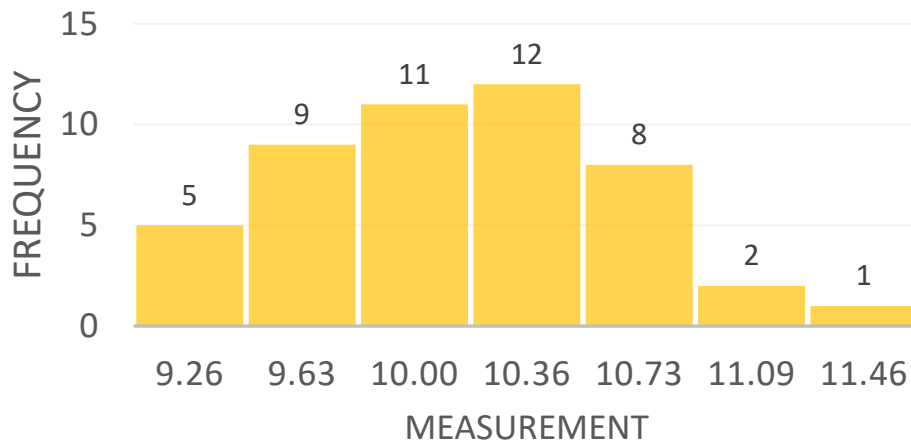
Measure	Indicator	Form	
		Population	Sample
Central tendency	Arithmetic mean	$\mu = \frac{\sum_{i=1}^n x_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
	Median	Value in the middle of the ordered string	
	Modal	Category/range with highest frequency	
	Mean Square	$M_p = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$	
	Geometric mean	$M_g = \sqrt[n]{\prod_{i=1}^n x_i}$	
	Harmonic mean	$M_a = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$	
	Central value	$x_c = \frac{Max - Min}{2}$	
Spread	Minim	Lowest value in the string	
	Maxim	Highest value in the string	
	Range	$R = x_{max} - x_{min}$	
	Variance	$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$
	Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n}}$	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$
Form	Asymmetry	$g_1 = \frac{\sum_{i=1}^n (x_i - \mu)^3}{\sigma^3}$	
	Flattening	$k = \frac{\sum_{i=1}^n (x_i - \mu)^4}{(\sum_{i=1}^n (x_i - \mu)^2)^2}$	

Visualizations

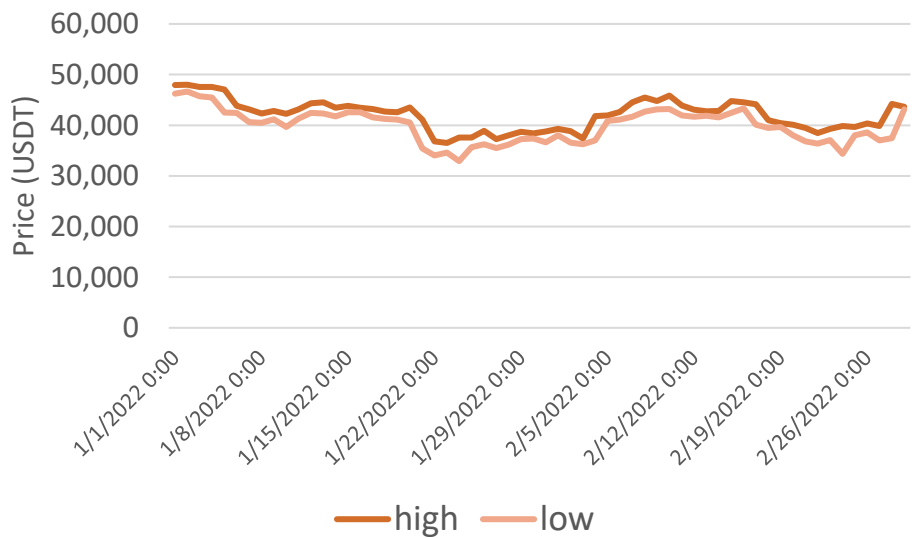
Column chart



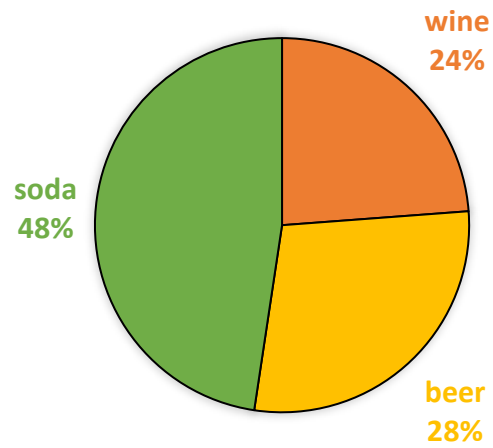
Histogram



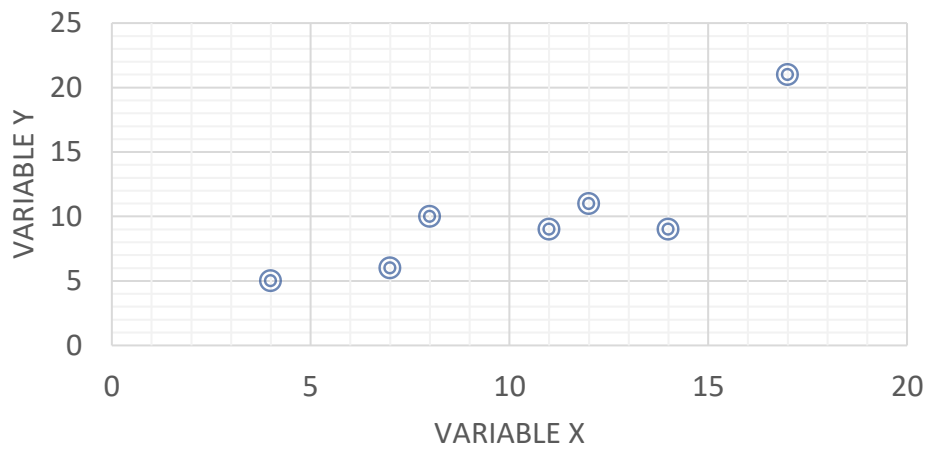
Line chart



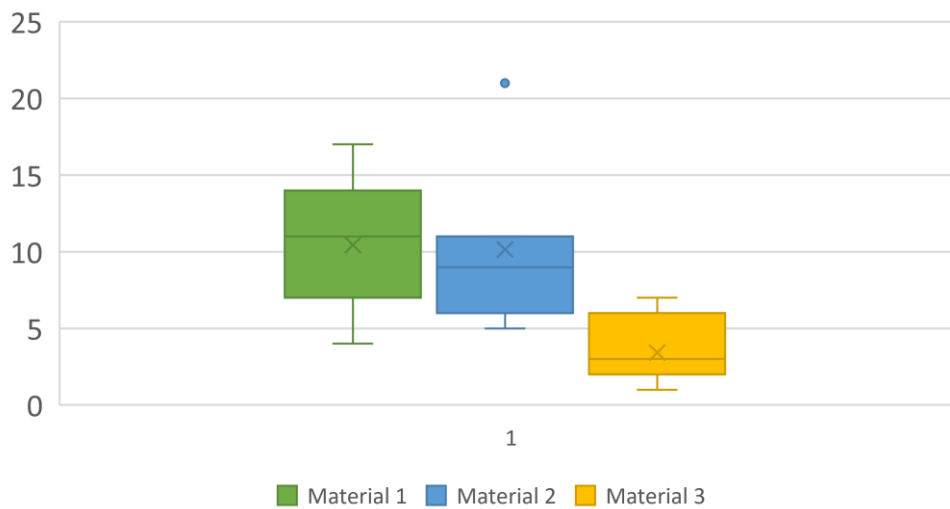
PIE CHART



Scatter plot



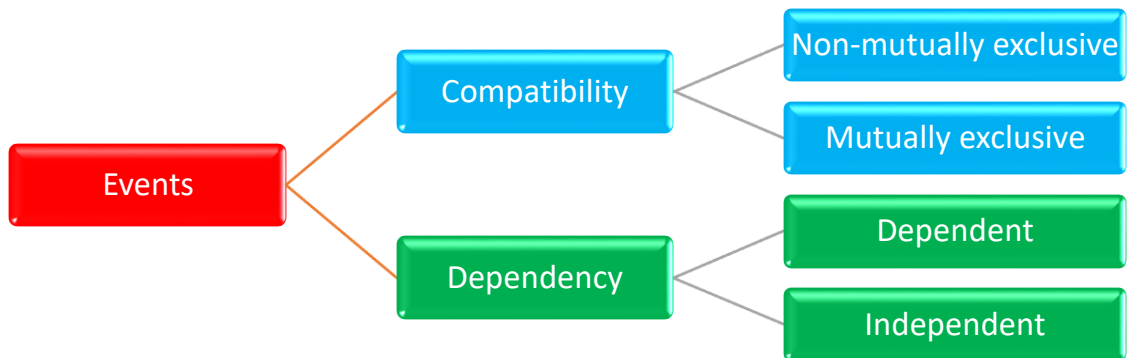
Box plot



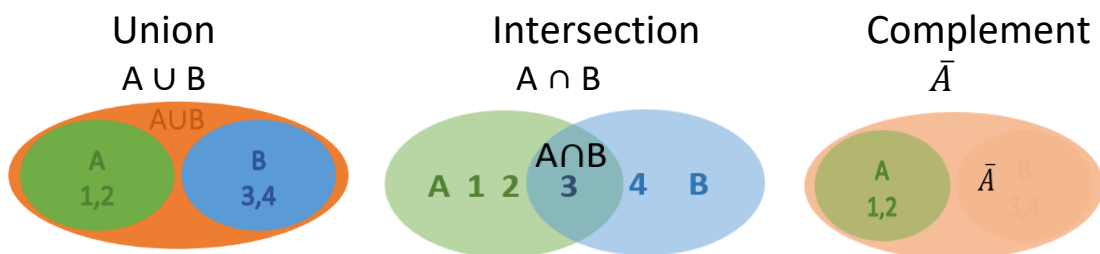
Probabilities

Types of events

- Impossible: \emptyset
- Random: $A = \{1, 2, 3\}$
- Sure: E



Event operations



Probability

$$P(X) = \frac{\text{no. of favorable events}}{\text{total no. of equally likely events}}$$

Conditional probability: $P(B | A)$ <- probability of B, given that A occurred

Operation	Type of event	Probability
Union	Mutually exclusive	$P(A \cup B) = P(A) + P(B)$
	Non-mutually exclusive and independent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B)$
	Non-mutually exclusive and dependent	$P(A \cup B) = P(A) + P(B) - P(A) * P(B A)$
Intersection	Non-mutually exclusive and independent	$P(A \cap B) = P(A) * P(B)$
	Non-mutually exclusive and dependent	$P(A \cap B) = P(A) * P(B A)$

Laws of probability

Independence of events

$$P(B|A) = P(B)$$

General addition rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Multiplication rule

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n)$$


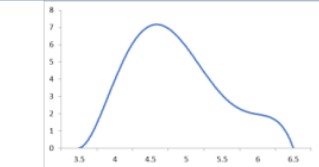
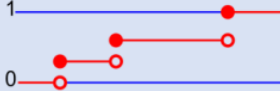
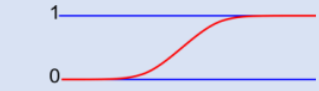
Law of total probability

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

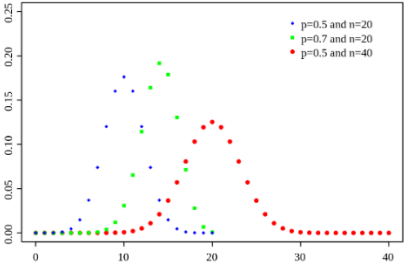
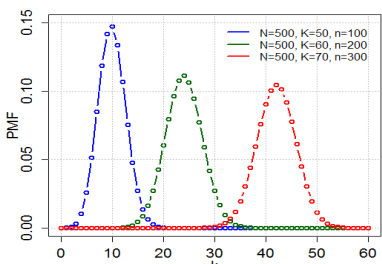
Bayes' rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Random variables

	Discrete Random Variables	Continuous Random Variables
Probabilities can be written in a table	Yes	No
We can determine the probability for a certain value	Yes	No
Probability Distribution Function Name	Probability Mass Function	Probability Density Function
Graphical representation		
Cumulative Distribution Function		

Discrete distributions

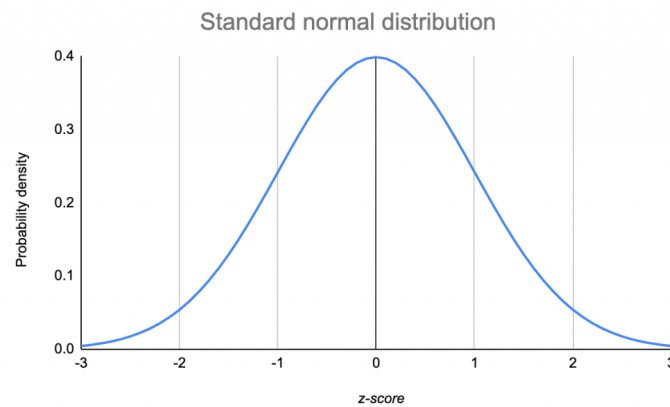
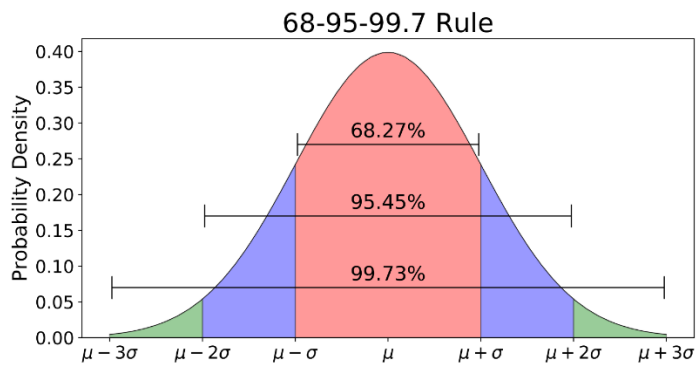
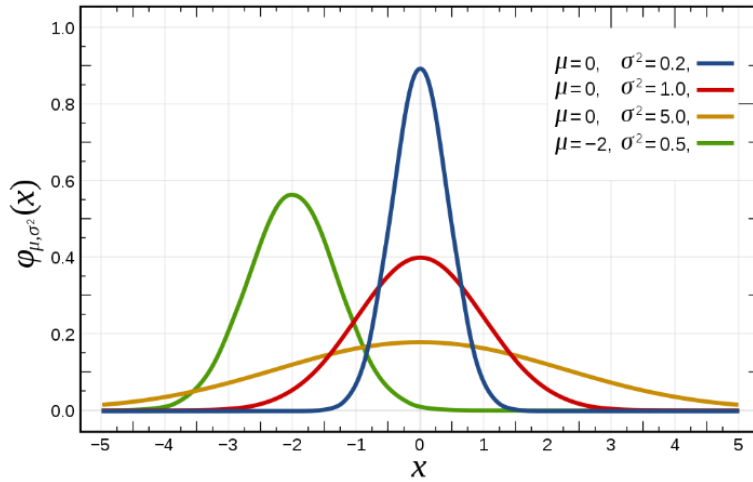
<i>Binomial distribution</i>	<i>Hypergeometric distribution</i>
The number of successes in extracting a sample of size n from a population of size N , <u>putting the part in place each time.</u> <p>p - probability of success q - probability of failure n - number of drawings Note:</p> $C_n^k = \frac{n!}{(n-k)! * k!}$	The number of successes in extracting a sample of size m from a population of size n , <u>without putting the part in place each time.</u> <p>p - probability of success q - probability of failure m - number of drawings a - number of successes ($a=n*p$) b - number of failures ($b=n*q$)</p>
$P(X = k) = C_n^k * p^k * q^{n-k}$	$P(X = k) = \frac{C_a^k * C_b^{m-k}}{C_n^m}$
$F(k) = P(X \leq k) = \sum C_n^k p^k q^{n-k}$	$F(x) = P(X \leq k) = \frac{1}{C_n^m} \sum_{i=0}^k C_a^i * C_b^{m-i}$
	

Continuous distributions

Normal distribution

Parameters: mean (μ) and standard deviation (σ)

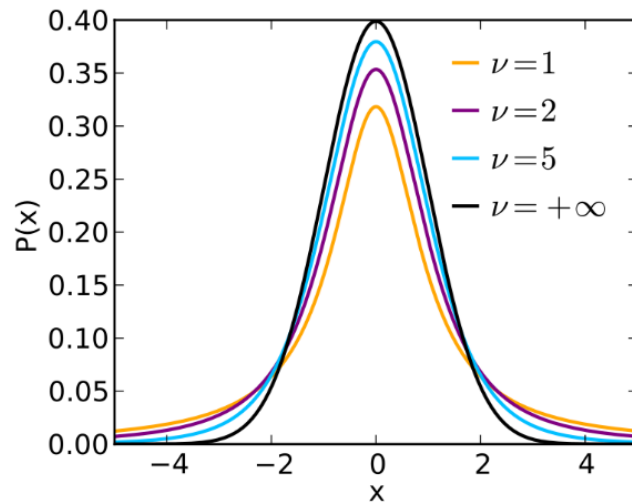
Symmetrical around the mean



Student Distribution

Parameter $\nu = n-1$

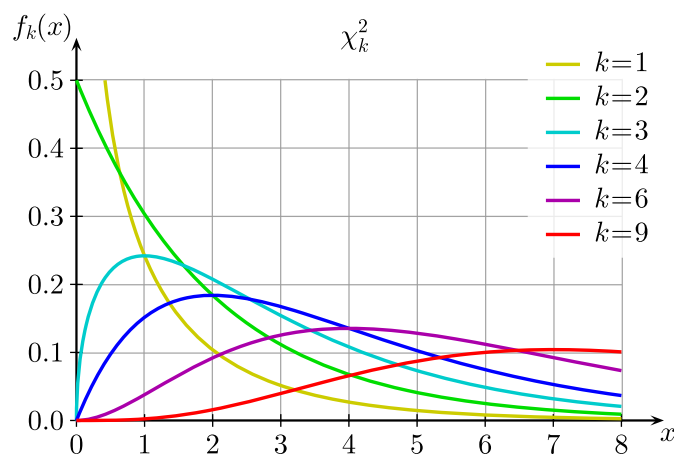
Symmetrical around the mean



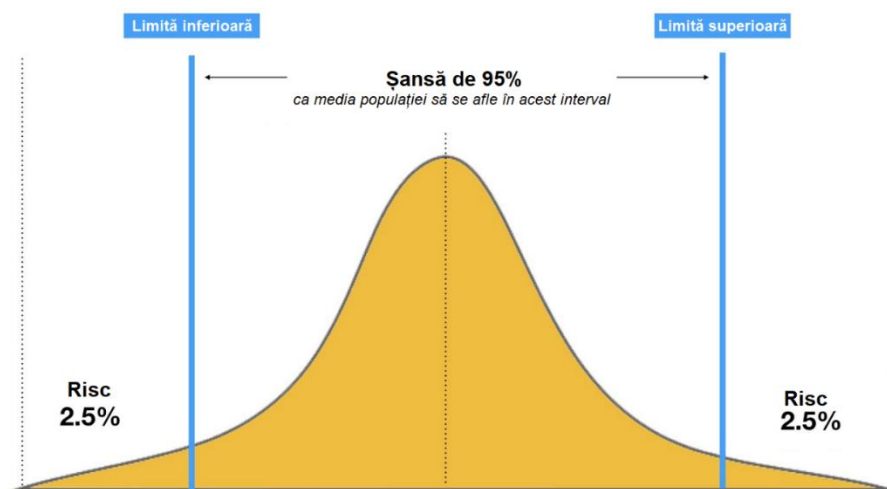
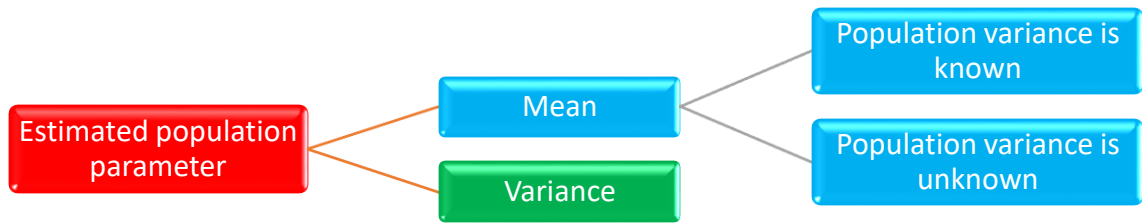
Chi-square distribution (χ^2)

Parameter $k = n-1$ (degrees of freedom)

Asymmetric (degrees of freedom) and strictly positive



Estimate



Types of risk

Unilateral Left Risk (ULR)	Unilateral Right Risk (URR)	Symmetrical Bilateral Risk (SBR)	Asymmetrical Bilateral Risk (ABR)

Estimate

Estimating the mean		Estimating the variance																																																																																																																																																																																																																																																																																																																																																																																		
Variance is known	Variance not known																																																																																																																																																																																																																																																																																																																																																																																			
Normal distribution (z values)	Student distribution (t values)	Chi-squared distribution (χ^2)																																																																																																																																																																																																																																																																																																																																																																																		
$\bar{x} - z_{\alpha_{lft}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha_{rgt}} \frac{\sigma}{\sqrt{n}}$	$\bar{x} - t_{\alpha_{lft}} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha_{rgt}} \frac{s}{\sqrt{n}}$	$(n-1) \frac{s^2}{\chi_{1-\alpha}^2} < \sigma^2 < (n-1) \frac{s^2}{\chi_{\alpha}^2}$																																																																																																																																																																																																																																																																																																																																																																																		
<table border="1"> <tr><th>z</th><th>0.00</th><th>0.01</th><th>0.02</th><th>0.03</th></tr> <tr><td>-3.4</td><td>0.0003</td><td>0.0003</td><td>0.0003</td><td>0.0003</td></tr> <tr><td>-3.3</td><td>0.0005</td><td>0.0005</td><td>0.0005</td><td>0.0004</td></tr> <tr><td>-3.2</td><td>0.0007</td><td>0.0007</td><td>0.0006</td><td>0.0006</td></tr> <tr><td>-3.1</td><td>0.0010</td><td>0.0009</td><td>0.0009</td><td>0.0009</td></tr> <tr><td>-3.0</td><td>0.0013</td><td>0.0013</td><td>0.0013</td><td>0.0012</td></tr> <tr><td>-2.9</td><td>0.0019</td><td>0.0018</td><td>0.0018</td><td>0.0017</td></tr> <tr><td>-2.8</td><td>0.0026</td><td>0.0025</td><td>0.0024</td><td>0.0023</td></tr> <tr><td>-2.7</td><td>0.0035</td><td>0.0034</td><td>0.0033</td><td>0.0032</td></tr> <tr><td>-2.6</td><td>0.0047</td><td>0.0045</td><td>0.0044</td><td>0.0043</td></tr> <tr><td>-2.5</td><td>0.0062</td><td>0.0060</td><td>0.0059</td><td>0.0057</td></tr> <tr><td>-2.4</td><td>0.0082</td><td>0.0080</td><td>0.0078</td><td>0.0076</td></tr> </table>	z	0.00	0.01	0.02	0.03	-3.4	0.0003	0.0003	0.0003	0.0003	-3.3	0.0005	0.0005	0.0005	0.0004	-3.2	0.0007	0.0007	0.0006	0.0006	-3.1	0.0010	0.0009	0.0009	0.0009	-3.0	0.0013	0.0013	0.0013	0.0012	-2.9	0.0019	0.0018	0.0018	0.0017	-2.8	0.0026	0.0025	0.0024	0.0023	-2.7	0.0035	0.0034	0.0033	0.0032	-2.6	0.0047	0.0045	0.0044	0.0043	-2.5	0.0062	0.0060	0.0059	0.0057	-2.4	0.0082	0.0080	0.0078	0.0076	<table border="1"> <tr><th>Degrees of freedom</th><th>0.005 (1)</th><th>0.01 (1)</th><th>0.025 (1)</th><th>0.05 (1)</th></tr> <tr><td>1</td><td>63.657</td><td>31.821</td><td>12.706</td><td>6.314</td></tr> <tr><td>2</td><td>9.925</td><td>6.965</td><td>4.303</td><td>2.920</td></tr> <tr><td>3</td><td>5.841</td><td>4.541</td><td>3.182</td><td>2.353</td></tr> <tr><td>4</td><td>4.604</td><td>3.747</td><td>2.776</td><td>2.132</td></tr> <tr><td>5</td><td>4.032</td><td>3.365</td><td>2.571</td><td>2.015</td></tr> <tr><td>6</td><td>3.707</td><td>3.143</td><td>2.447</td><td>1.943</td></tr> <tr><td>7</td><td>3.501</td><td>2.998</td><td>2.365</td><td>1.900</td></tr> <tr><td>8</td><td>3.358</td><td>2.898</td><td>2.306</td><td>1.860</td></tr> <tr><td>9</td><td>3.251</td><td>2.819</td><td>2.262</td><td>1.828</td></tr> <tr><td>10</td><td>3.173</td><td>2.751</td><td>2.228</td><td>1.796</td></tr> <tr><td>11</td><td>3.110</td><td>2.693</td><td>2.199</td><td>1.771</td></tr> <tr><td>12</td><td>3.057</td><td>2.643</td><td>2.174</td><td>1.749</td></tr> <tr><td>13</td><td>3.013</td><td>2.599</td><td>2.152</td><td>1.730</td></tr> <tr><td>14</td><td>2.976</td><td>2.561</td><td>2.132</td><td>1.713</td></tr> <tr><td>15</td><td>2.943</td><td>2.528</td><td>2.114</td><td>1.698</td></tr> <tr><td>16</td><td>2.913</td><td>2.499</td><td>2.098</td><td>1.684</td></tr> <tr><td>17</td><td>2.886</td><td>2.474</td><td>2.084</td><td>1.671</td></tr> <tr><td>18</td><td>2.861</td><td>2.451</td><td>2.071</td><td>1.659</td></tr> <tr><td>19</td><td>2.838</td><td>2.430</td><td>2.059</td><td>1.648</td></tr> <tr><td>20</td><td>2.816</td><td>2.411</td><td>2.048</td><td>1.638</td></tr> <tr><td>21</td><td>2.796</td><td>2.393</td><td>2.038</td><td>1.629</td></tr> <tr><td>22</td><td>2.777</td><td>2.377</td><td>2.029</td><td>1.620</td></tr> <tr><td>23</td><td>2.759</td><td>2.362</td><td>2.021</td><td>1.612</td></tr> <tr><td>24</td><td>2.742</td><td>2.348</td><td>2.013</td><td>1.604</td></tr> <tr><td>25</td><td>2.727</td><td>2.335</td><td>2.006</td><td>1.597</td></tr> <tr><td>26</td><td>2.712</td><td>2.323</td><td>2.000</td><td>1.590</td></tr> <tr><td>27</td><td>2.698</td><td>2.311</td><td>1.994</td><td>1.584</td></tr> <tr><td>28</td><td>2.685</td><td>2.299</td><td>1.989</td><td>1.578</td></tr> <tr><td>29</td><td>2.673</td><td>2.288</td><td>1.984</td><td>1.573</td></tr> <tr><td>30</td><td>2.661</td><td>2.278</td><td>1.980</td><td>1.568</td></tr> </table>	Degrees of freedom	0.005 (1)	0.01 (1)	0.025 (1)	0.05 (1)	1	63.657	31.821	12.706	6.314	2	9.925	6.965	4.303	2.920	3	5.841	4.541	3.182	2.353	4	4.604	3.747	2.776	2.132	5	4.032	3.365	2.571	2.015	6	3.707	3.143	2.447	1.943	7	3.501	2.998	2.365	1.900	8	3.358	2.898	2.306	1.860	9	3.251	2.819	2.262	1.828	10	3.173	2.751	2.228	1.796	11	3.110	2.693	2.199	1.771	12	3.057	2.643	2.174	1.749	13	3.013	2.599	2.152	1.730	14	2.976	2.561	2.132	1.713	15	2.943	2.528	2.114	1.698	16	2.913	2.499	2.098	1.684	17	2.886	2.474	2.084	1.671	18	2.861	2.451	2.071	1.659	19	2.838	2.430	2.059	1.648	20	2.816	2.411	2.048	1.638	21	2.796	2.393	2.038	1.629	22	2.777	2.377	2.029	1.620	23	2.759	2.362	2.021	1.612	24	2.742	2.348	2.013	1.604	25	2.727	2.335	2.006	1.597	26	2.712	2.323	2.000	1.590	27	2.698	2.311	1.994	1.584	28	2.685	2.299	1.989	1.578	29	2.673	2.288	1.984	1.573	30	2.661	2.278	1.980	1.568	<table border="1"> <tr><th>Degrees of Freedom</th><th>0.99</th><th>0.975</th><th>0.95</th><th>0.90</th></tr> <tr><td>1</td><td>—</td><td>0.001</td><td>0.004</td><td>0.016</td></tr> <tr><td>2</td><td>0.020</td><td>0.051</td><td>0.103</td><td>0.211</td></tr> <tr><td>3</td><td>0.115</td><td>0.216</td><td>0.322</td><td>0.584</td></tr> <tr><td>4</td><td>0.297</td><td>0.484</td><td>0.711</td><td>1.064</td></tr> <tr><td>5</td><td>0.554</td><td>0.831</td><td>1.145</td><td>1.610</td></tr> <tr><td>6</td><td>0.872</td><td>1.237</td><td>1.635</td><td>2.204</td></tr> <tr><td>7</td><td>1.239</td><td>1.690</td><td>2.167</td><td>2.833</td></tr> <tr><td>8</td><td>1.646</td><td>2.180</td><td>2.733</td><td>3.490</td></tr> <tr><td>9</td><td>2.099</td><td>2.700</td><td>3.325</td><td>4.168</td></tr> <tr><td>10</td><td>2.599</td><td>3.247</td><td>3.940</td><td>4.865</td></tr> <tr><td>11</td><td>3.101</td><td>3.816</td><td>4.576</td><td>5.578</td></tr> <tr><td>12</td><td>3.599</td><td>4.388</td><td>5.226</td><td>6.300</td></tr> <tr><td>13</td><td>4.099</td><td>4.963</td><td>5.892</td><td>7.032</td></tr> <tr><td>14</td><td>4.599</td><td>5.541</td><td>6.576</td><td>7.779</td></tr> <tr><td>15</td><td>5.099</td><td>6.121</td><td>7.276</td><td>8.541</td></tr> <tr><td>16</td><td>5.599</td><td>6.703</td><td>7.992</td><td>9.309</td></tr> <tr><td>17</td><td>6.099</td><td>7.287</td><td>8.724</td><td>10.083</td></tr> <tr><td>18</td><td>6.599</td><td>7.873</td><td>9.472</td><td>10.863</td></tr> <tr><td>19</td><td>7.099</td><td>8.461</td><td>10.236</td><td>11.649</td></tr> <tr><td>20</td><td>7.599</td><td>9.051</td><td>11.016</td><td>12.441</td></tr> <tr><td>21</td><td>8.099</td><td>9.643</td><td>11.812</td><td>13.239</td></tr> <tr><td>22</td><td>8.599</td><td>10.237</td><td>12.624</td><td>14.043</td></tr> <tr><td>23</td><td>9.099</td><td>10.833</td><td>13.452</td><td>14.853</td></tr> <tr><td>24</td><td>9.599</td><td>11.431</td><td>14.296</td><td>15.669</td></tr> <tr><td>25</td><td>10.099</td><td>12.031</td><td>15.156</td><td>16.491</td></tr> <tr><td>26</td><td>10.599</td><td>12.633</td><td>16.032</td><td>17.319</td></tr> <tr><td>27</td><td>11.099</td><td>13.237</td><td>16.924</td><td>18.153</td></tr> <tr><td>28</td><td>11.599</td><td>13.843</td><td>17.832</td><td>18.993</td></tr> <tr><td>29</td><td>12.099</td><td>14.451</td><td>18.756</td><td>19.839</td></tr> <tr><td>30</td><td>12.599</td><td>15.061</td><td>19.696</td><td>20.691</td></tr> </table>	Degrees of Freedom	0.99	0.975	0.95	0.90	1	—	0.001	0.004	0.016	2	0.020	0.051	0.103	0.211	3	0.115	0.216	0.322	0.584	4	0.297	0.484	0.711	1.064	5	0.554	0.831	1.145	1.610	6	0.872	1.237	1.635	2.204	7	1.239	1.690	2.167	2.833	8	1.646	2.180	2.733	3.490	9	2.099	2.700	3.325	4.168	10	2.599	3.247	3.940	4.865	11	3.101	3.816	4.576	5.578	12	3.599	4.388	5.226	6.300	13	4.099	4.963	5.892	7.032	14	4.599	5.541	6.576	7.779	15	5.099	6.121	7.276	8.541	16	5.599	6.703	7.992	9.309	17	6.099	7.287	8.724	10.083	18	6.599	7.873	9.472	10.863	19	7.099	8.461	10.236	11.649	20	7.599	9.051	11.016	12.441	21	8.099	9.643	11.812	13.239	22	8.599	10.237	12.624	14.043	23	9.099	10.833	13.452	14.853	24	9.599	11.431	14.296	15.669	25	10.099	12.031	15.156	16.491	26	10.599	12.633	16.032	17.319	27	11.099	13.237	16.924	18.153	28	11.599	13.843	17.832	18.993	29	12.099	14.451	18.756	19.839	30	12.599	15.061	19.696	20.691
z	0.00	0.01	0.02	0.03																																																																																																																																																																																																																																																																																																																																																																																
-3.4	0.0003	0.0003	0.0003	0.0003																																																																																																																																																																																																																																																																																																																																																																																
-3.3	0.0005	0.0005	0.0005	0.0004																																																																																																																																																																																																																																																																																																																																																																																
-3.2	0.0007	0.0007	0.0006	0.0006																																																																																																																																																																																																																																																																																																																																																																																
-3.1	0.0010	0.0009	0.0009	0.0009																																																																																																																																																																																																																																																																																																																																																																																
-3.0	0.0013	0.0013	0.0013	0.0012																																																																																																																																																																																																																																																																																																																																																																																
-2.9	0.0019	0.0018	0.0018	0.0017																																																																																																																																																																																																																																																																																																																																																																																
-2.8	0.0026	0.0025	0.0024	0.0023																																																																																																																																																																																																																																																																																																																																																																																
-2.7	0.0035	0.0034	0.0033	0.0032																																																																																																																																																																																																																																																																																																																																																																																
-2.6	0.0047	0.0045	0.0044	0.0043																																																																																																																																																																																																																																																																																																																																																																																
-2.5	0.0062	0.0060	0.0059	0.0057																																																																																																																																																																																																																																																																																																																																																																																
-2.4	0.0082	0.0080	0.0078	0.0076																																																																																																																																																																																																																																																																																																																																																																																
Degrees of freedom	0.005 (1)	0.01 (1)	0.025 (1)	0.05 (1)																																																																																																																																																																																																																																																																																																																																																																																
1	63.657	31.821	12.706	6.314																																																																																																																																																																																																																																																																																																																																																																																
2	9.925	6.965	4.303	2.920																																																																																																																																																																																																																																																																																																																																																																																
3	5.841	4.541	3.182	2.353																																																																																																																																																																																																																																																																																																																																																																																
4	4.604	3.747	2.776	2.132																																																																																																																																																																																																																																																																																																																																																																																
5	4.032	3.365	2.571	2.015																																																																																																																																																																																																																																																																																																																																																																																
6	3.707	3.143	2.447	1.943																																																																																																																																																																																																																																																																																																																																																																																
7	3.501	2.998	2.365	1.900																																																																																																																																																																																																																																																																																																																																																																																
8	3.358	2.898	2.306	1.860																																																																																																																																																																																																																																																																																																																																																																																
9	3.251	2.819	2.262	1.828																																																																																																																																																																																																																																																																																																																																																																																
10	3.173	2.751	2.228	1.796																																																																																																																																																																																																																																																																																																																																																																																
11	3.110	2.693	2.199	1.771																																																																																																																																																																																																																																																																																																																																																																																
12	3.057	2.643	2.174	1.749																																																																																																																																																																																																																																																																																																																																																																																
13	3.013	2.599	2.152	1.730																																																																																																																																																																																																																																																																																																																																																																																
14	2.976	2.561	2.132	1.713																																																																																																																																																																																																																																																																																																																																																																																
15	2.943	2.528	2.114	1.698																																																																																																																																																																																																																																																																																																																																																																																
16	2.913	2.499	2.098	1.684																																																																																																																																																																																																																																																																																																																																																																																
17	2.886	2.474	2.084	1.671																																																																																																																																																																																																																																																																																																																																																																																
18	2.861	2.451	2.071	1.659																																																																																																																																																																																																																																																																																																																																																																																
19	2.838	2.430	2.059	1.648																																																																																																																																																																																																																																																																																																																																																																																
20	2.816	2.411	2.048	1.638																																																																																																																																																																																																																																																																																																																																																																																
21	2.796	2.393	2.038	1.629																																																																																																																																																																																																																																																																																																																																																																																
22	2.777	2.377	2.029	1.620																																																																																																																																																																																																																																																																																																																																																																																
23	2.759	2.362	2.021	1.612																																																																																																																																																																																																																																																																																																																																																																																
24	2.742	2.348	2.013	1.604																																																																																																																																																																																																																																																																																																																																																																																
25	2.727	2.335	2.006	1.597																																																																																																																																																																																																																																																																																																																																																																																
26	2.712	2.323	2.000	1.590																																																																																																																																																																																																																																																																																																																																																																																
27	2.698	2.311	1.994	1.584																																																																																																																																																																																																																																																																																																																																																																																
28	2.685	2.299	1.989	1.578																																																																																																																																																																																																																																																																																																																																																																																
29	2.673	2.288	1.984	1.573																																																																																																																																																																																																																																																																																																																																																																																
30	2.661	2.278	1.980	1.568																																																																																																																																																																																																																																																																																																																																																																																
Degrees of Freedom	0.99	0.975	0.95	0.90																																																																																																																																																																																																																																																																																																																																																																																
1	—	0.001	0.004	0.016																																																																																																																																																																																																																																																																																																																																																																																
2	0.020	0.051	0.103	0.211																																																																																																																																																																																																																																																																																																																																																																																
3	0.115	0.216	0.322	0.584																																																																																																																																																																																																																																																																																																																																																																																
4	0.297	0.484	0.711	1.064																																																																																																																																																																																																																																																																																																																																																																																
5	0.554	0.831	1.145	1.610																																																																																																																																																																																																																																																																																																																																																																																
6	0.872	1.237	1.635	2.204																																																																																																																																																																																																																																																																																																																																																																																
7	1.239	1.690	2.167	2.833																																																																																																																																																																																																																																																																																																																																																																																
8	1.646	2.180	2.733	3.490																																																																																																																																																																																																																																																																																																																																																																																
9	2.099	2.700	3.325	4.168																																																																																																																																																																																																																																																																																																																																																																																
10	2.599	3.247	3.940	4.865																																																																																																																																																																																																																																																																																																																																																																																
11	3.101	3.816	4.576	5.578																																																																																																																																																																																																																																																																																																																																																																																
12	3.599	4.388	5.226	6.300																																																																																																																																																																																																																																																																																																																																																																																
13	4.099	4.963	5.892	7.032																																																																																																																																																																																																																																																																																																																																																																																
14	4.599	5.541	6.576	7.779																																																																																																																																																																																																																																																																																																																																																																																
15	5.099	6.121	7.276	8.541																																																																																																																																																																																																																																																																																																																																																																																
16	5.599	6.703	7.992	9.309																																																																																																																																																																																																																																																																																																																																																																																
17	6.099	7.287	8.724	10.083																																																																																																																																																																																																																																																																																																																																																																																
18	6.599	7.873	9.472	10.863																																																																																																																																																																																																																																																																																																																																																																																
19	7.099	8.461	10.236	11.649																																																																																																																																																																																																																																																																																																																																																																																
20	7.599	9.051	11.016	12.441																																																																																																																																																																																																																																																																																																																																																																																
21	8.099	9.643	11.812	13.239																																																																																																																																																																																																																																																																																																																																																																																
22	8.599	10.237	12.624	14.043																																																																																																																																																																																																																																																																																																																																																																																
23	9.099	10.833	13.452	14.853																																																																																																																																																																																																																																																																																																																																																																																
24	9.599	11.431	14.296	15.669																																																																																																																																																																																																																																																																																																																																																																																
25	10.099	12.031	15.156	16.491																																																																																																																																																																																																																																																																																																																																																																																
26	10.599	12.633	16.032	17.319																																																																																																																																																																																																																																																																																																																																																																																
27	11.099	13.237	16.924	18.153																																																																																																																																																																																																																																																																																																																																																																																
28	11.599	13.843	17.832	18.993																																																																																																																																																																																																																																																																																																																																																																																
29	12.099	14.451	18.756	19.839																																																																																																																																																																																																																																																																																																																																																																																
30	12.599	15.061	19.696	20.691																																																																																																																																																																																																																																																																																																																																																																																
<p>$\sigma=1.5$ $\bar{x}=10.3$ $n=30$</p> <p>Estimate μ with a unilateral left risk of 4%. From the table $z = -1.75$</p> $10.3 - 1.75 \frac{1.5}{\sqrt{30}} = 9.82$ <p>Result: μ is greater than 9.82 with a probability of 96%. There is a risk of 4% that μ is smaller than 9.82</p>	<p>$s=1.2$ $\bar{x}=5.1$ $n=20$</p> <p>Estimate μ with a unilateral right risk of 2.5%. Degrees of freedom $v = 20-1 = 19$ From the table $t = 2.093$</p> $5.1 + 2.093 \frac{1.2}{\sqrt{20}} = 5.66$ <p>Result: μ is smaller than 5.66 with a probability of 97.5%. There is a risk of 2.5% that μ is greater than 5.66</p>	<p>$s=2, \bar{x}=7, n=25$ Estimate σ^2 with a bilateral symmetrical risk of 10%. Degrees of freedom $v = 25-1 = 24$ Left risk $(1-\alpha/2) = 5\%$ Right risk $(\alpha/2) = 5\%$ From the table: $\chi_{1-\alpha/2}^2 = 13.848$ $\chi_{\alpha/2}^2 = 36.415$</p> <p><i>Left_limit</i> $= 24 \frac{4}{36.415} = 2.63$</p> <p><i>Right_limit</i> $= 24 \frac{4}{13.848} = 6.93$</p> <p>Result: σ^2 is between 2.63 and 6.93 with a probability of 90%. There is a risk of 10% that σ^2 is outside this interval.</p>																																																																																																																																																																																																																																																																																																																																																																																		

Statistical Process Control (SPC)

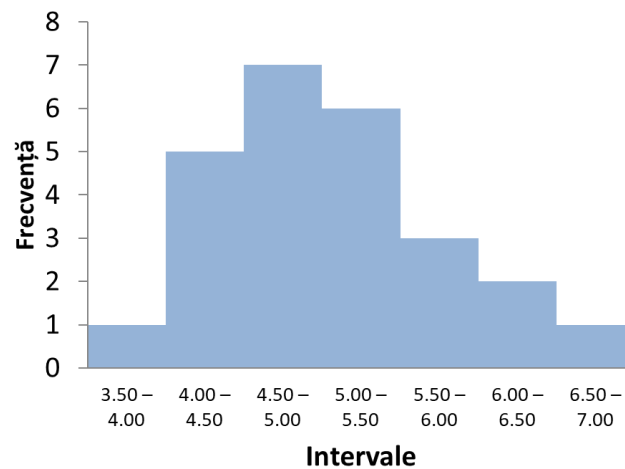
SPC tools

- Histogram
- Pareto diagram
- Dot diagram
- Control charts
- Cause-effect diagram
- Process diagram

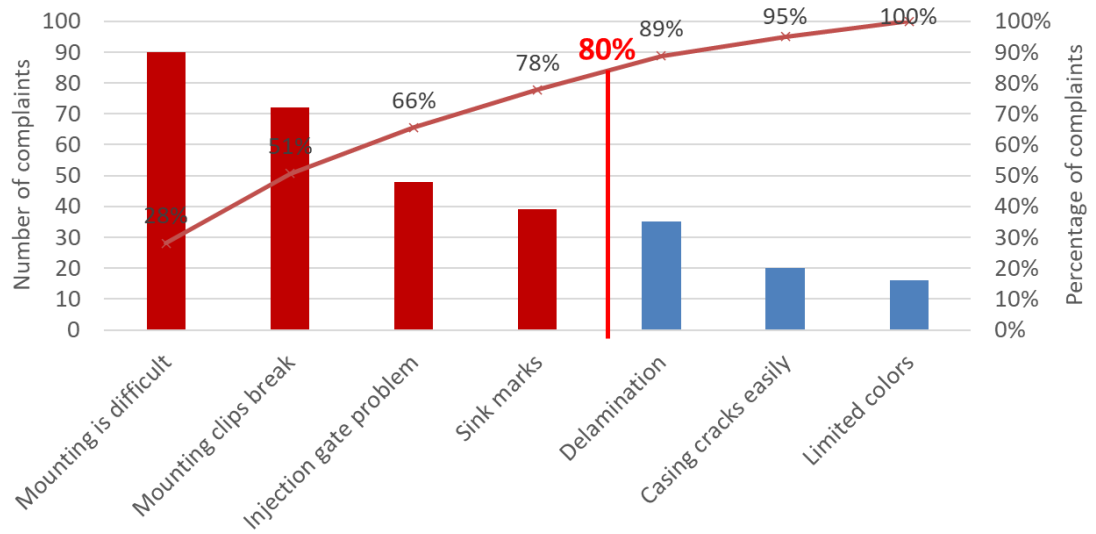
Histogram

Steps in constructing a histogram:

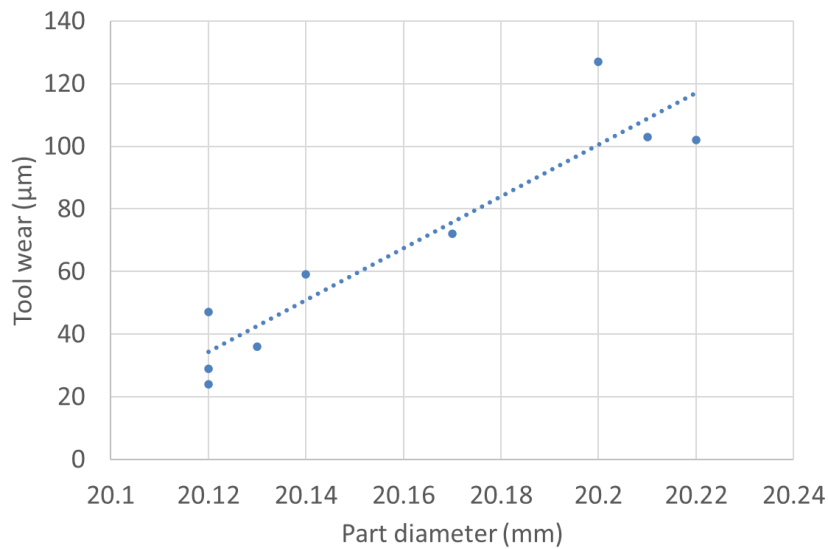
1. Determine the minimum and maximum values and calculate the range
2. Divide the range into the set number of intervals to find the interval length
3. Use the interval length to determine the interval ends for each interval
4. Count how many values in the string are in each range
5. Use a column chart to view the number of values in each range.



Pareto diagram

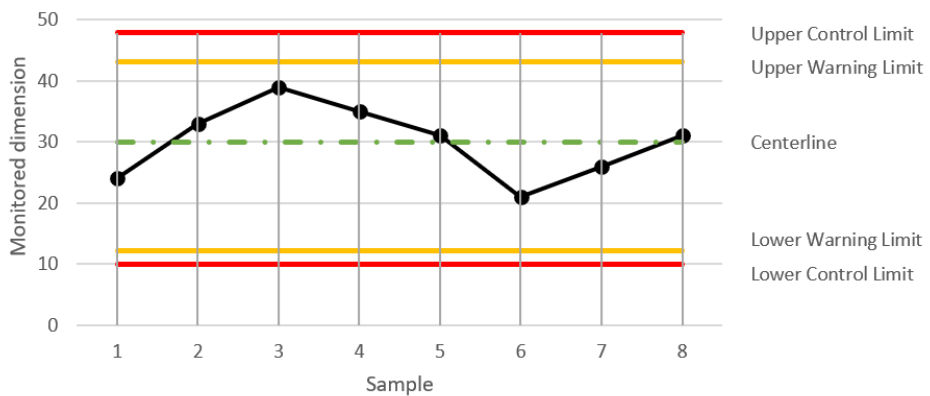


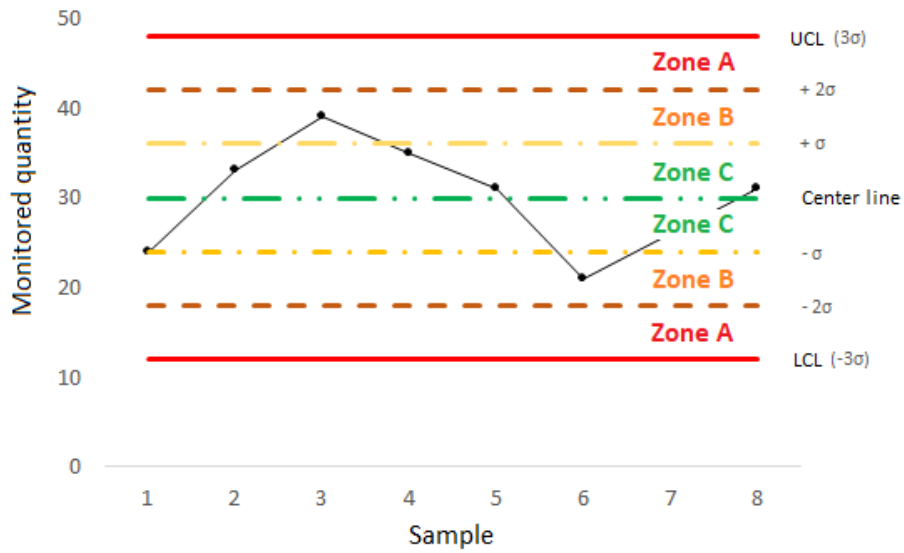
Scatter plot



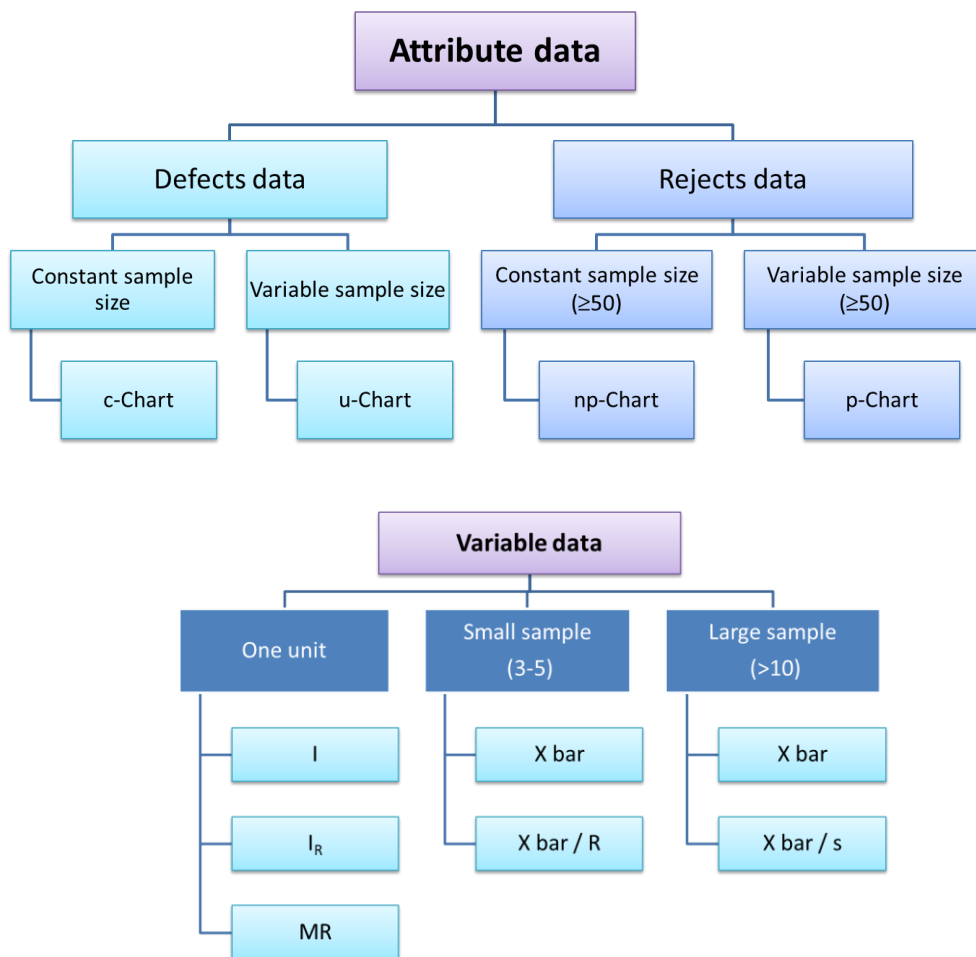
Control charts

Elements of a control chart



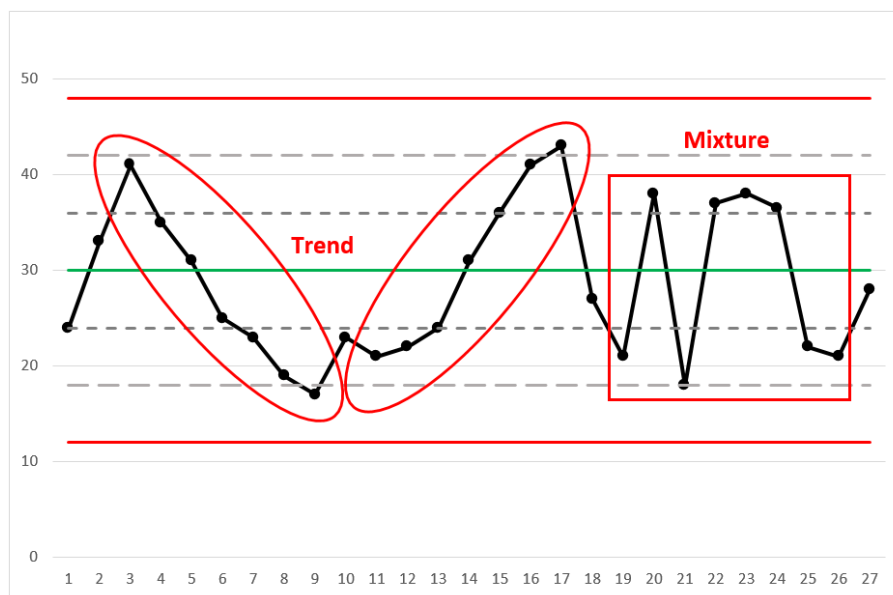
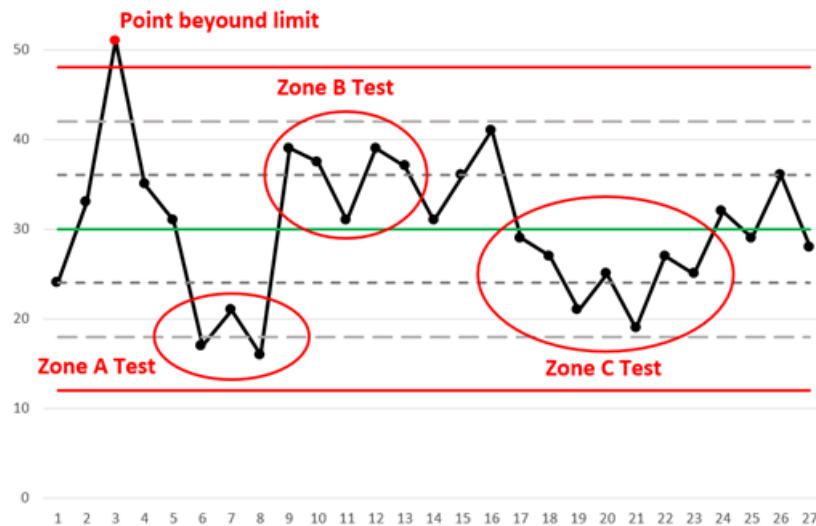


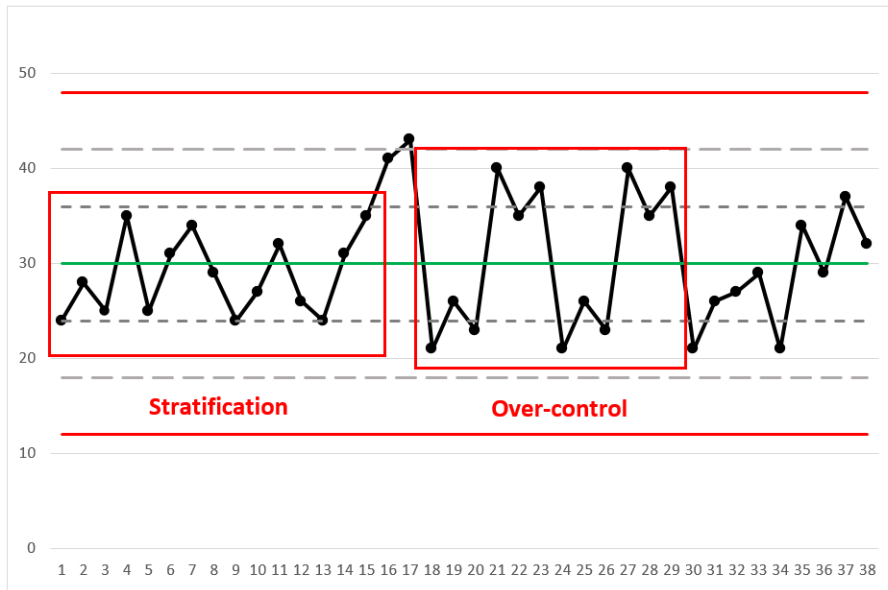
Types of control chart



How to identify problems with the process

Rule	Description
1. Points outside the boundaries	One or more points are out of bounds
2. Zone A test	2 out of 3 consecutive points are in Zone A or further
3. Zone B test	4 out of 5 consecutive points are in zone B or further
4. Zone C test	7 or more consecutive points are on one side of the average (in Zone C or beyond)
5. Trend	7 consecutive points are trending up or down
6. Mixing	8 consecutive points without a point in Zone C
7. Layering	15 consecutive points in Zone C
8. Supra-control	14 consecutive alternating points

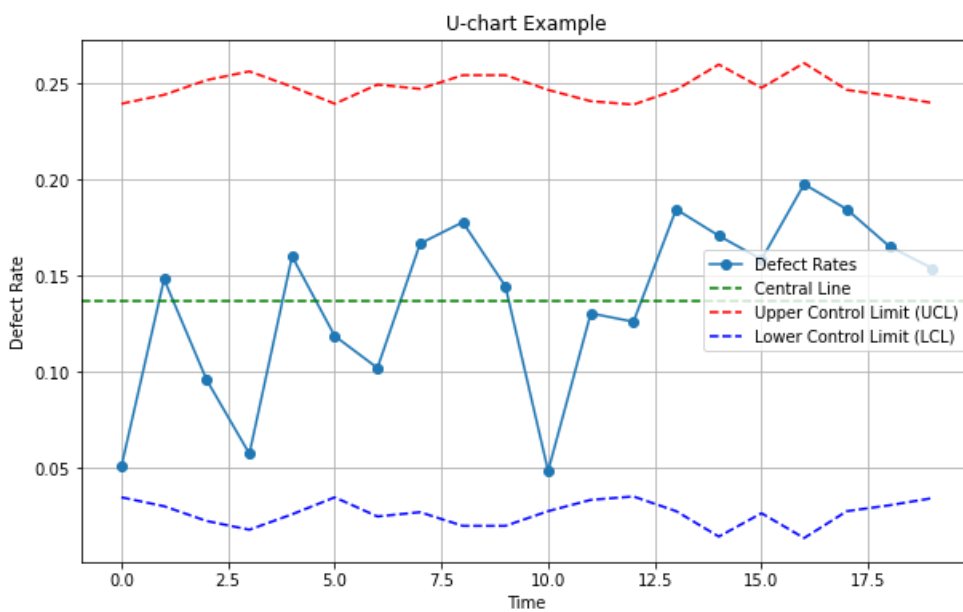
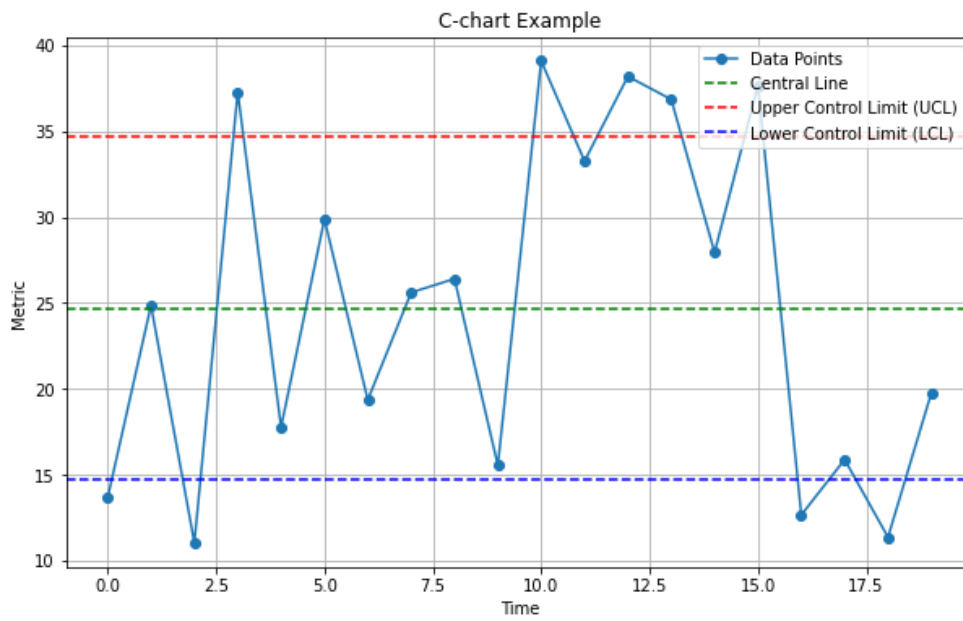


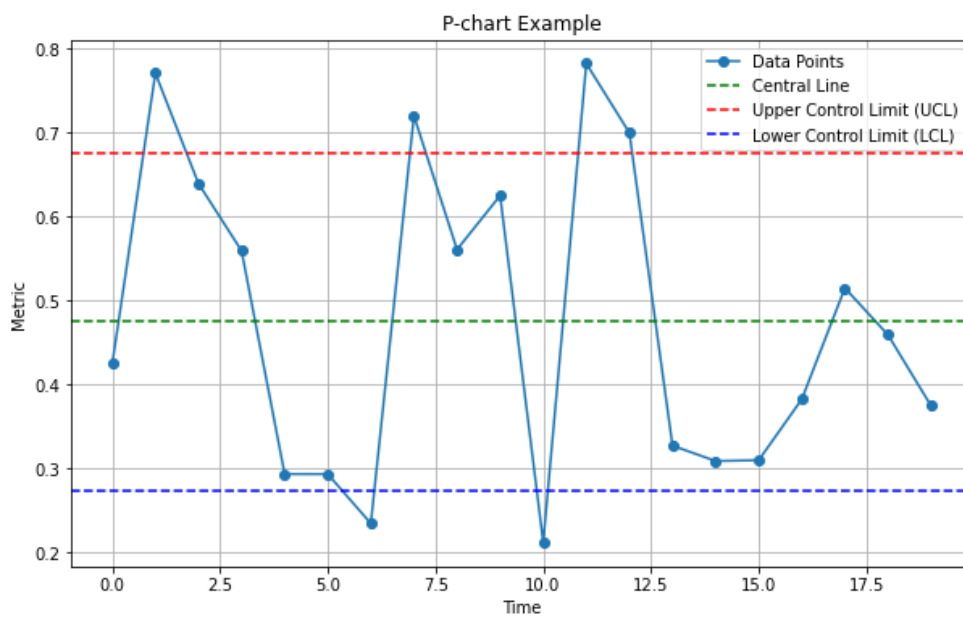
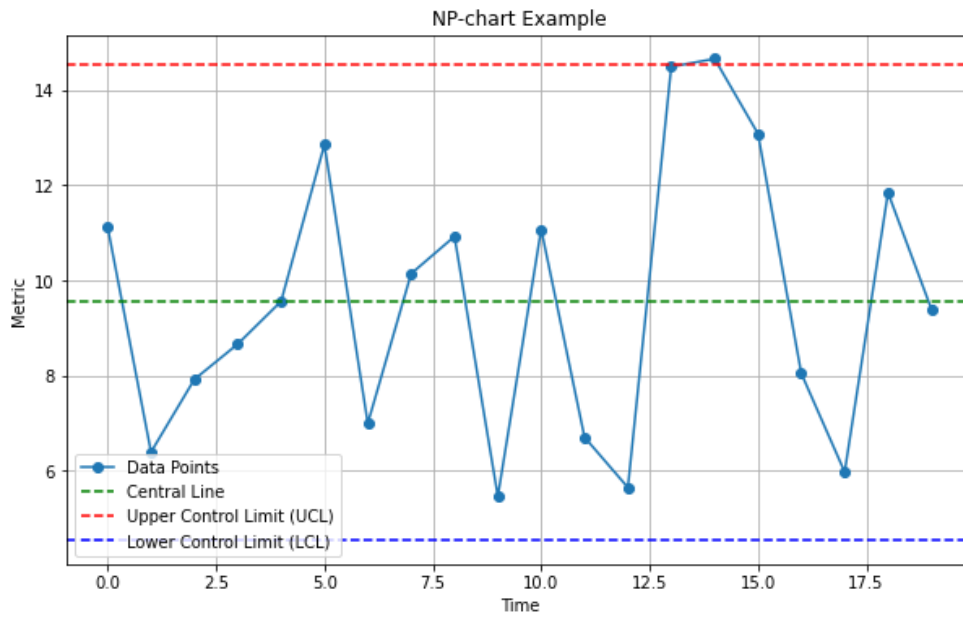


Description of the event	Rules	Possible causes
Large variations from the average	1, 2	New employee; wrong configuration; measurement error; production step skipped; step not completed; power failure; faulty equipment
Small variations from the average	3, 4	Change in material; change in work instructions; different measuring device; different work shift; improvement of worker skills; change in maintenance schedule; change in installation procedure
Trends	5	Tool wear; thermal effects (cooling, heating)
Mixing	6	The existence of several processes (shifts, machines, materials)
Layering	7	The existence of several processes (shifts, machines, materials)
Overcontrol	8	Manipulation of data by the operator; Alternative use of more than one material

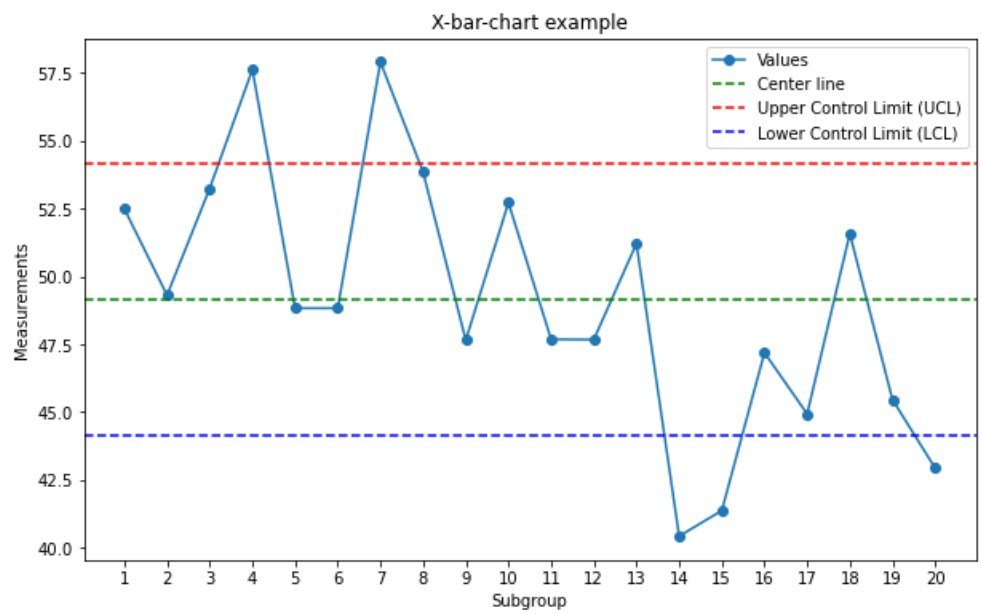
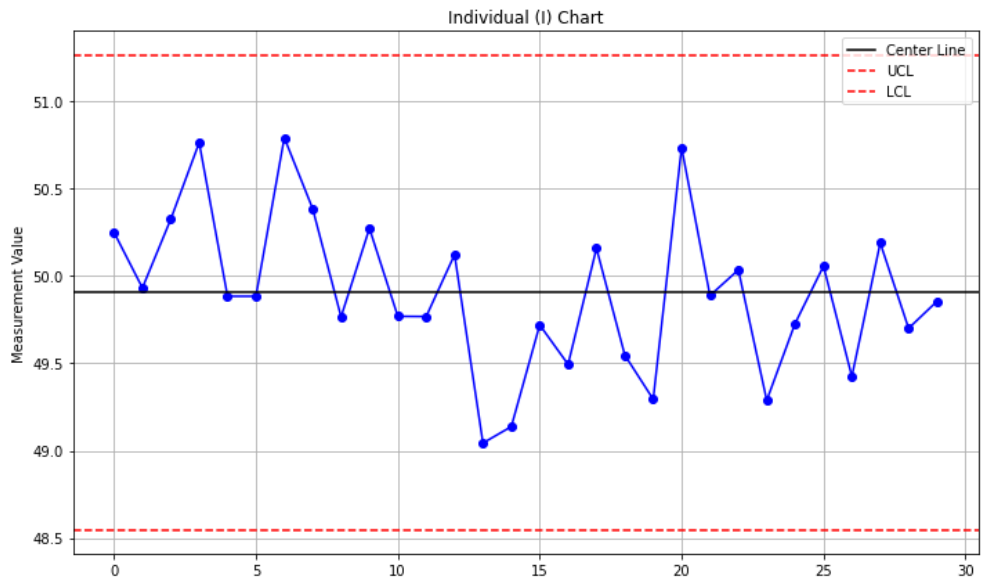
Chart types

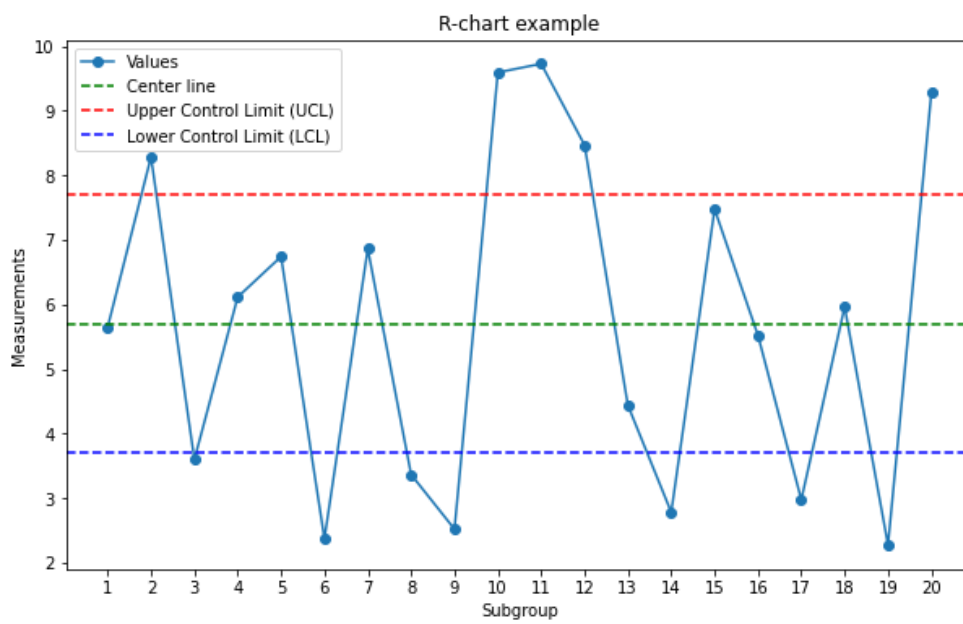
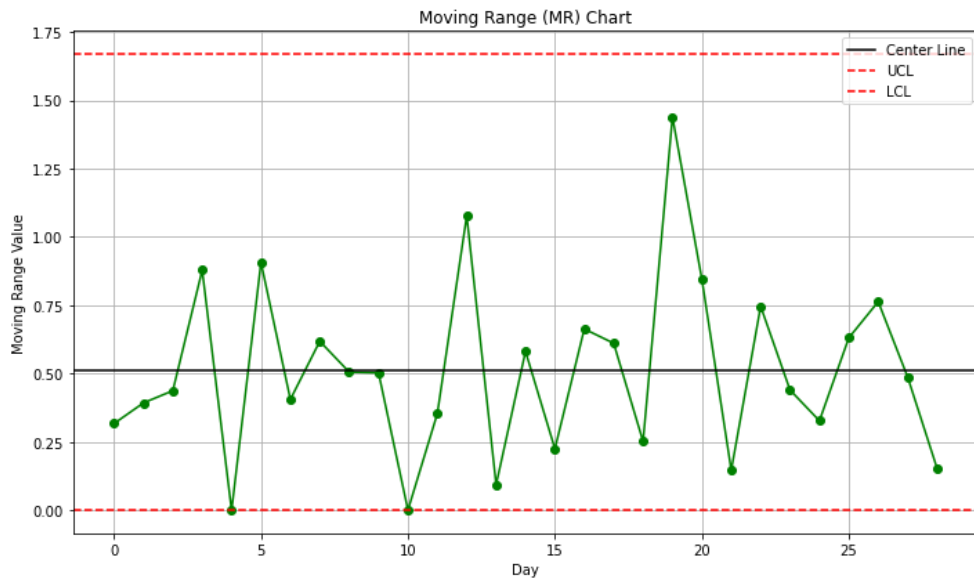
Attribute charts

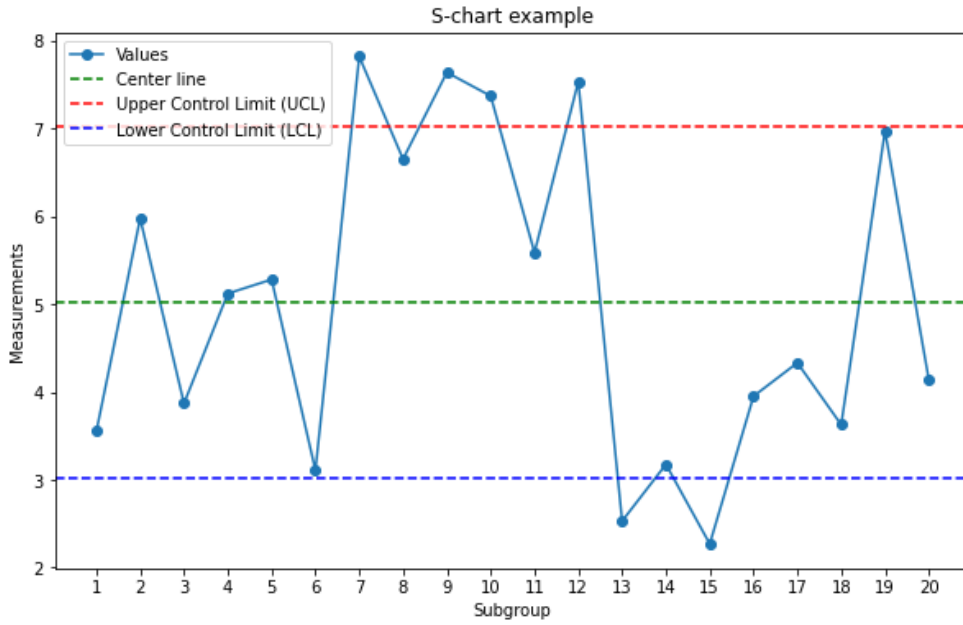




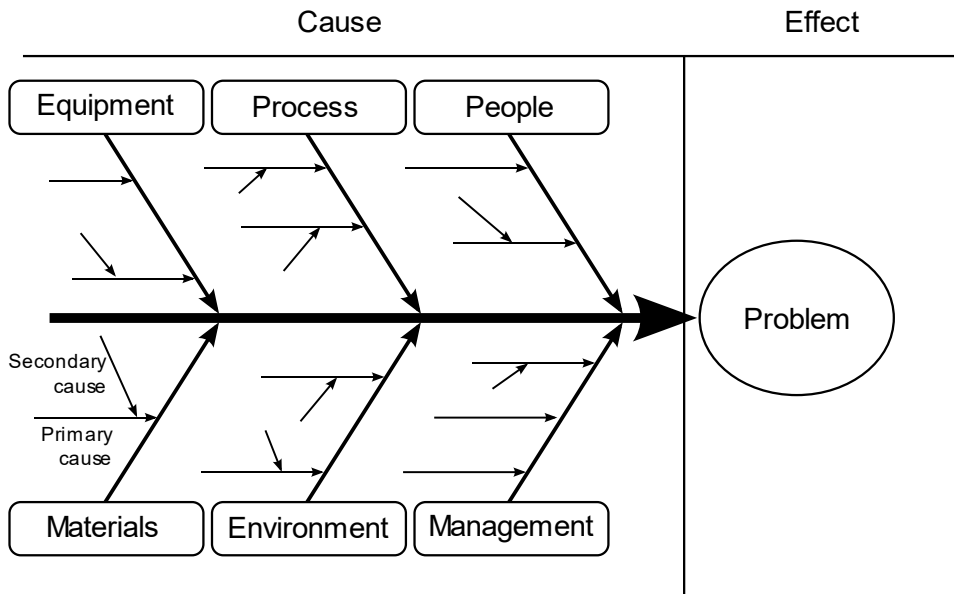
Variable charts



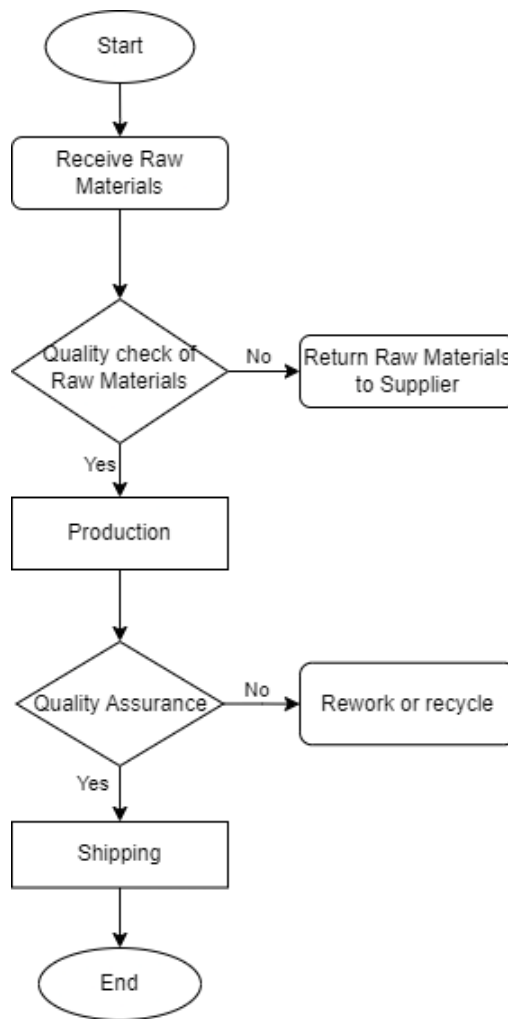




Ishikawa diagram



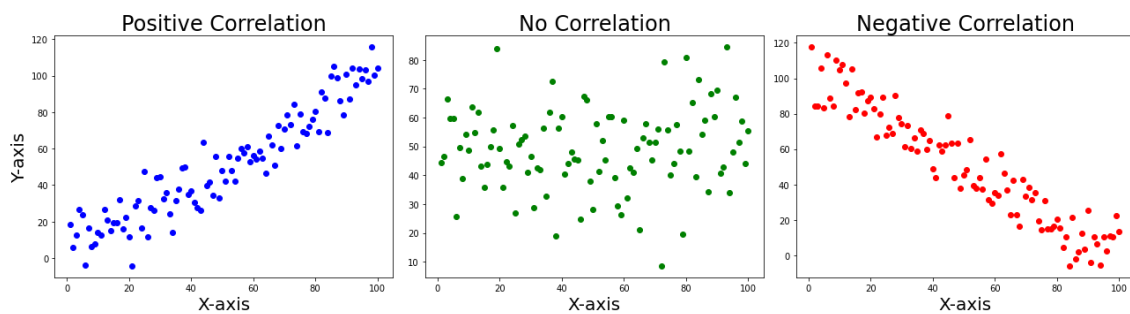
Process flow diagram



Correlation and regression

Pearson correlation coefficient

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$



Regression line equation:

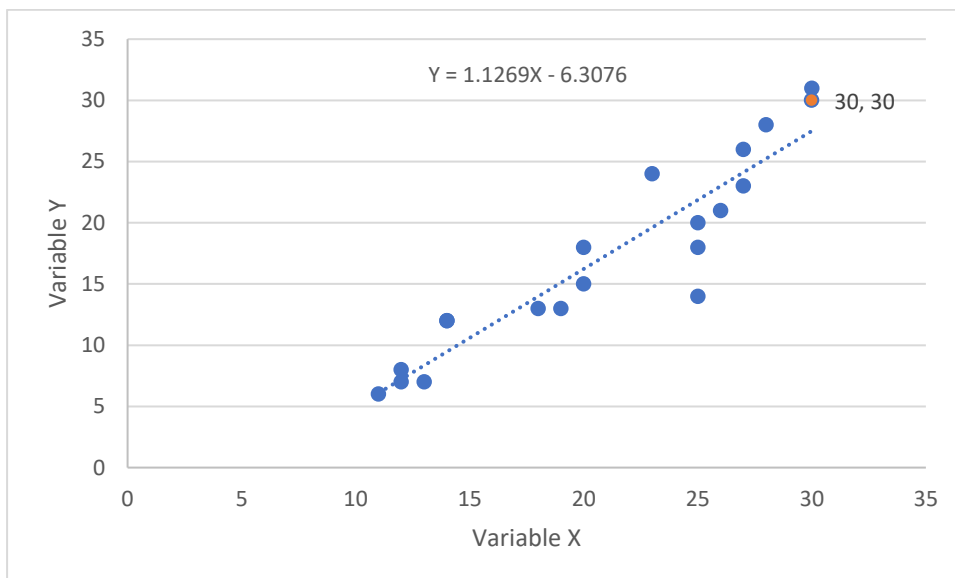
$$Y = a * X + b$$

Prediction error:

$$\Delta = \sum (y - \hat{y})^2$$

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$



Index of terms

Notation	Name / Description
a	Number of successes
ABR	Asymmetrical Bilateral Risk
ai	Absolute frequency
α (alpha)	Risk
b	Number of failures
χ² (chi squared)	Statistic for the χ ² distribution
f_i	Relative frequency
∅ (fi)	The null set / Impossible event
IQR	Interquartile Range
k (discrete distributions)	Current observation
k (Student distribution)	Degrees of freedom
m (Hypergeometric distribution)	Number of trials
M_a	Harmonic mean
M_e	Median
M_g	Geometric mean
μ	Population mean
M_o	Modal
M_p	Quadratic mean
M_x	Arithmetic mean
n (Binomial distribution)	Number of trials
n (Hypergeometric distribution)	Number of elements
v (niu)	Degrees of freedom
p	Probability of success
P(A B)	Probability of A given B
P(x)	Probability of X
q	probability of failure
Q1, Q2, Q3	First, Second, Third quartile
R	Range
r	Pearson's correlation coefficient
R²	Determination coefficient
s	Sample standard deviation
S	Sample space
s²	Sample variance
SBR	Symmetrical Bilateral Risk
σ (sigma)	Population standard deviation
σ²	Population variance
SS_{res}	Sum of squares of residuals
SS_{total}	Total sum of squares
t	Statistic for the Student distribution
ULR	Unilateral Left Risk
URR	Unilateral Right Risk
\bar{x} (x_bar)	Sample mean
x_{max}	Maximum value
x_{min}	Minimum value
z	Statistic for the Normal distribution

List of figures

Fig. 1.1. Knowledge-Information-Data Hierarchy [4], [5]	6
Fig. 1.2. Types of data and their levels of measurement	9
Fig. 1.3. Modal determination by graphical method	12
Fig. 2.1. Simple column diagram	21
Fig. 2.2. Simple diagram with horizontal bars	21
Fig. 2.3. Multi-instance column chart	22
Fig. 2.4. Stacked columns diagram	22
Fig. 2.5. Line chart	23
Fig. 2.6. Scatter plot	24
Fig. 2.7. Pie chart	25
Fig. 2.8. Age distribution in age ranges	28
Fig. 2.9. Comparison of height distributions of students from two different classes (Class A – left, class B- right)	29
Fig. Representation of events by Venn diagrams	39
Fig. 3.2. Union of two events	40
Fig. 3.3. Intersection of two events	40
Fig. 3.4. Complement of an event	41
Fig. 4.1. Law of total probability	47
Fig. 4.2. Bayes rule explained in an example	48
Fig. 5.1. Distribution of results of rolling a die 100 times	53
Fig. 5.2. Example curve for a continuous distribution	54
Fig. 5.3. Cumulative distribution function for a discrete (left) and continuous (right) function [9]	55
Fig. Probability function and cumulative distribution function for uniform distribution.	60
Fig. Distribution (left) and probability (right) functions of the binomial distribution [10].	61
Fig. Distribution function and cumulative distribution function for hypergeometric distribution [11].	64
Fig. 7.1 Distribution (left) and cumulative (right) function for continuous uniform distribution [12]	68
Fig. 7.2. Probability Density Function (left) and Cumulative Distribution Function (right) for the normal distribution [13]	69
Fig. 7.3 Examples of normal distributions with different parameters [14]	69
Fig. 7.4. The 68 -95-99.7 rule [15]	70
Fig. 7.5. Probability Density Function (left) and Cumulative Distribution Function (right) for the Student’s t-distribution [16]	70
Fig. 7.6. Probability Density Function (left) and Cumulative Distribution Function (right) for the Chi-square distribution [17]	71
Fig. 8.1. Example of the distribution of a population and several samples	76
Fig. 8.3. Types of risk: a) Left Unilateral Risk; b) Right Unilateral Risk; c) Symmetric Bilateral Risk; d) Asymmetric Bilateral Risk	77
Fig. 8.4. Standard Normal Distribution	79
Fig. 8.5. Table of z-scores	79
Fig. 8.6. Reading z-scores from the table	80
Fig. 8.7. Table of t-scores	82

Fig. 8.8. Reading the Chi-square score table 84

Fig. 9.1. An example of a histogram 90

Fig. 9.2. An example of a Pareto chart..... 91

Fig. 9.3. An example of a dot plot..... 93

Fig. 9.4. Elements of a control chart..... 95

Fig. 9.5. Elements of a typical control chart and the three zones 96

Fig. 9.6. Types of Control Charts by Data Type and Sample Size..... 97

Fig. 9.7. An example of a c-chart..... 98

Fig. 9.8. An example of a u-Chart..... 99

Fig. 9.9. An example of an np card..... 100

Fig. 9.10. An example of a p-Chart..... 101

Fig. 9.11. An example of an I-Chart..... 101

Fig. 9.12. An example of an X-Bar chart..... 102

Fig. 9.13. An example of an MR chart..... 103

Fig. 9.14. An example of an R-Chart 103

Fig. 9.15. An example of an S-Chart..... 104

Fig. 9.16. Rule 1-4..... 105

Fig. 9.17. Rule 5-6..... 105

Fig. 9.18. Rule 7-8..... 106

Fig. 9.19. An example of a cause-effect diagram [20] 107

Fig. 9.20. An example of a process flow diagram 108

Fig. 10.1. Types of correlations 112

List of tables

Table 1.1 - Frequency table (absolute, relative, cumulative ascending and descending) 10

Table 1.2. A group of students with the same average in 2 different subjects..... 13

Table 2.1. Representative statistical indicators 17

Table 2.2 Example of a table..... 19

Table 2.3. The 6 ranges and the limits for each range 27

Table 2.4. Frequency table for the 6 intervals..... 27

Table 4.1. Multiplication and addition rules according to event types..... 46

Table 5.1. Comparison between discrete and continuous variables..... 56

Table 8.1. Notations used for population and sample parameters..... 76

Table 8.2. Estimation of population parameters (summary) 85

Table 9.1. Troubleshooting a process with the control chart..... 104

Table 9.2. Possible causes of rules observed in the control chart..... 106