Advanced Statistical Methods and Optimization

Prof. Dr. Tim Weber

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
[1] 2
```

1 Basic Concepts

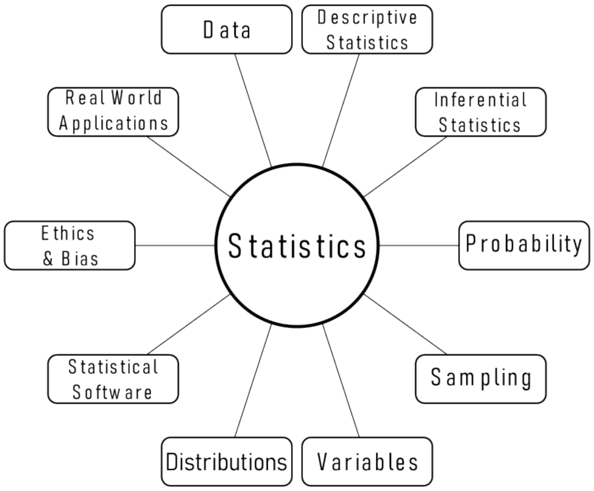


Figure 1.1: The necessary statistical ingredients.

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

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

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

3. **Inferential Statistics:** Inferential statistics is about making predictions, inferences, or decisions about a population based on a sample of data. This involves hypothesis testing, confidence intervals, and regression analysis, among other techniques.
4. **Probability:** Probability theory is the foundation of statistics. It deals with uncertainty and randomness. We use probability to describe the likelihood of events occurring in various situations, which is essential for making statistical inferences.
5. **Sampling:** In most cases, it's impractical to collect data from an entire population. Instead, we often work with samples, which are smaller subsets of the population. The process of selecting and analyzing samples is a critical aspect of statistical analysis.
6. **Variables:** Variables are characteristics or attributes that can vary from one individual or item to another. They can be categorized as dependent (response) or independent (predictor) variables, depending on their role in a statistical analysis.
7. **Distributions:** A probability distribution describes the possible values of a variable and their associated probabilities. Common distributions include the normal distribution, binomial distribution, and Poisson distribution, among others.
8. **Statistical Software:** In the modern era, statistical analysis is often conducted using specialized software packages like R, Python (with libraries like NumPy and Pandas), SPSS, or Excel. These tools facilitate data manipulation, visualization, and complex statistical calculations.
9. **Ethics and Bias:** It's essential to consider ethical principles in statistical analysis, including issues related to data privacy, confidentiality, and the potential for bias in data collection and interpretation.
10. **Real-World Applications:** Statistics has a wide range of applications, from medical research to marketing, finance, and social sciences. It helps us make informed decisions and draw meaningful insights from data in various fields.

1.1 Probability

1.1.1 Overview

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

1.1.2 What is Probability?

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

1.1.3 Basic Probability Terminology

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

1.1.4 Probability Notation

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

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

1.1.5 The Fundamental Principles of Probability

There are two fundamental principles of probability:

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

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

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

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

1.1.6 Example: Rolling a Fair Six-Sided Die

Consider rolling a fair six-sided die.

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

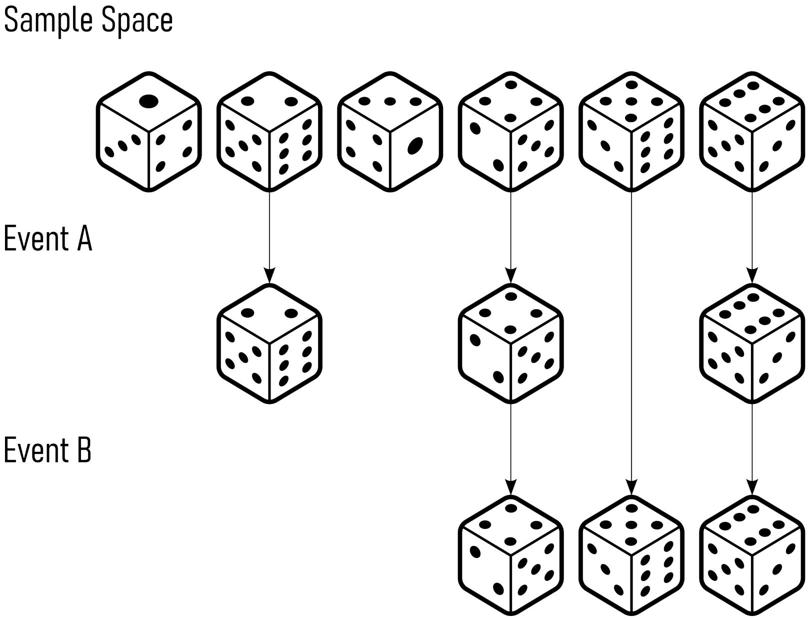


Figure 1.2: This example's sample space, as well as event A and event B.

Calculation of some probabilities:

- Probability of Event A (P(A)): $P(A) = \frac{\text{Number of outcomes in A}}{\text{Total number of outcomes in S}} = \frac{3}{6} = \frac{1}{2}$
- Probability of Event B (P(B)): $P(B) = \frac{3}{6} = \frac{1}{2}$
- Probability of both A and B (P(A and B)): 4 and 6 satisfy both A and B, $P(A \; and \; B) = \frac{2}{6} = \frac{1}{3}$

This example demonstrates the fundamental principles of probability and how they can be applied to real-world situations.

In the subsequent chapters of this course, more advanced concepts in probability theory will be explored, including conditional probability, random variables, probability distributions, and statistical inference.

1.1.7 Population Definition

In the field of statistics, a population refers to the entire group or collection of individuals, objects, or events under study. For example, when conducting a survey to examine the average income of households in a country, the population encompasses all the households within that country.

1.1.8 Random Sampling

To study a population, statisticians often employ random sampling techniques as it may not be feasible or practical to gather data from every member of the population. Probability principles come into play during the selection of a random sample, ensuring that the sampling process adheres to well-defined rules to represent the population adequately.

1.1.9 Sampling Distribution

Once a random sample is acquired, probability theory becomes instrumental in analyzing and characterizing various sample statistics, such as the sample mean and variance. The sampling distribution serves as a probability distribution, encompassing all possible values of sample statistics attainable through random sampling.

1.1.10 Inferential Statistics

Probability plays a pivotal role in drawing conclusions about the population based on the data obtained from the sample. Statistical methodologies, including hypothesis testing and constructing confidence intervals, rely on probability theory to assess the likelihood of specific outcomes and quantify the associated uncertainty.

1.1.11 Estimation

Probability is a key element in parameter estimation. For instance, when estimating population parameters like the mean or variance based on a sample, statisticians employ probability distributions such as the t-distribution or chi-squared distribution to create confidence intervals and determine margins of error.

1.1.12 Hypothesis Testing

In hypothesis testing, probability is employed to ascertain whether observed differences or associations in sample data hold statistical significance. Probability calculations aid in evaluating whether the observed results are likely to be a product of random chance or if they genuinely reflect characteristics of the population.

1.1.13 Generalization

The primary objective of statistical analysis is to generalize findings from the sample to the entire population. Probability allows for quantifying the likelihood that the characteristics observed in the sample accurately represent the entire population, while also considering the inherent uncertainty associated with the sampling process.

1.1.14 Probability in action - The Galton Board

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

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

1. **Initial Release**: At the top of the Galton board, a ball or particle is released. This ball can take one of two paths at each peg, either to the left or to the right. The decision at each peg is determined by chance, such as the flip of a coin or the roll of a die. This represents a random event.

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

2. **Multiple Trials**: As the ball progresses downward, it encounters several pegs, each of which randomly directs it either left or right. The ball continues to bounce off pegs until it reaches the bottom.

3. **Accumulation**: Over multiple trials or runs of the Galton board, you will notice that the balls accumulate in a pattern at the bottom. This pattern forms a bell-shaped curve, which is the hallmark of a normal distribution.

4. **Normal Distribution**: The accumulation of balls at the bottom resembles the shape of a normal distribution curve. This means that the majority of balls will tend to accumulate in the center, forming the peak of the curve, while fewer balls will accumulate at the extreme left and right sides.

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

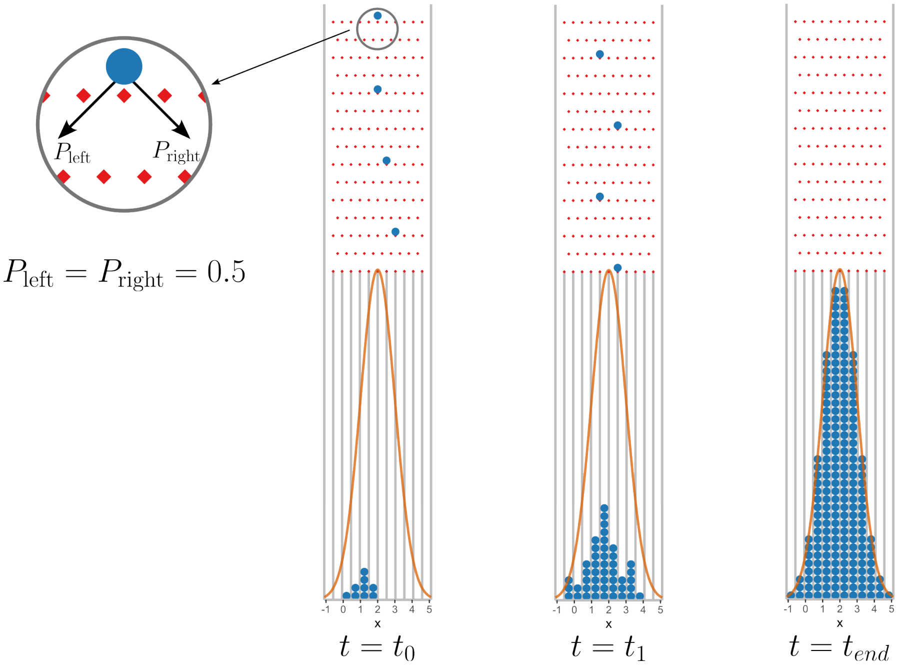


Figure 1.3: A Galton board in action.

1.2 Population

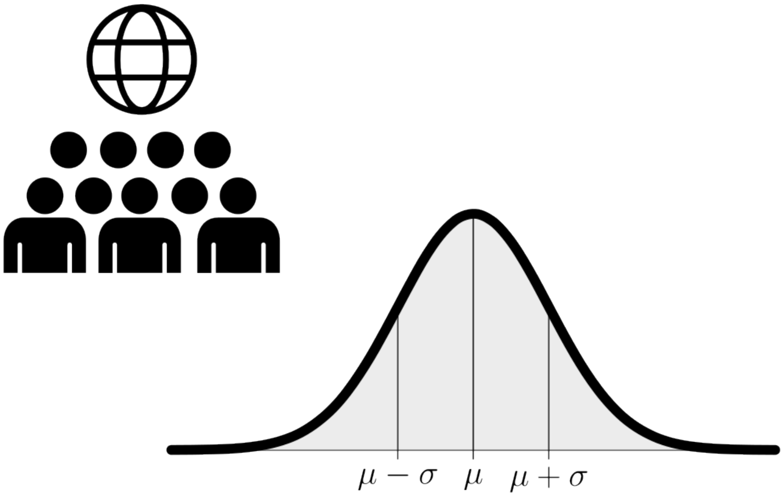


Figure 1.4: An example for a population.

In statistics, a population is the complete set of individuals, items, or data points that are the subject of a study. Understanding populations and how to work with them is fundamental in statistical analysis, as it forms the basis for making meaningful inferences and drawing conclusions about the broader group being studied. It is the complete collection of all elements that share a common characteristic or feature and is of interest to the researcher. The population can vary widely depending on the research question or problem at hand. A populations *true mean* is depicted with μ_0 and the variance is depicted with σ_0^2.

1.3 Sample

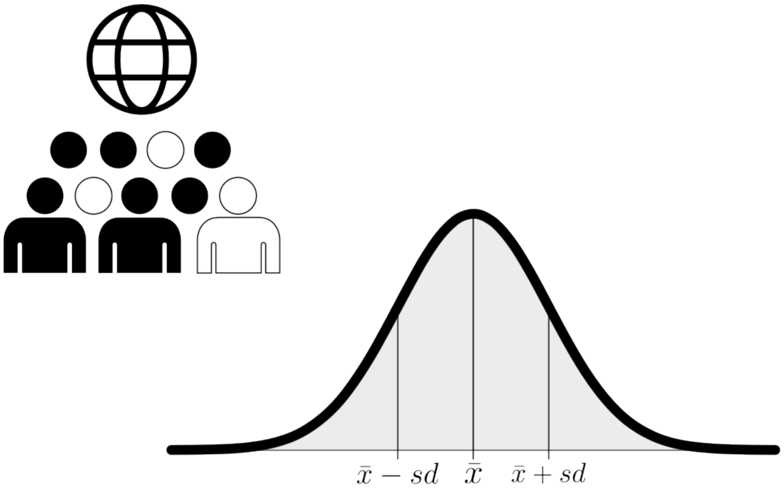


Figure 1.5: A sample drawn from the population.

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

1.4 Descriptive Statistics

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

1. Measures of Central Tendency:
    - Mean (average)
    - Median (middle value)
    - Mode (most frequent value)
2. Measures of Variability or Dispersion:
    - Range (difference between the maximum and minimum values)
    - Variance (average of the squared differences from the mean)
    - Standard Deviation (square root of the variance)
3. Frequency Distributions:
    - Histograms
    - Density plots
    - Frequency tables
    - Bar charts
4. Summary Statistics:
    - Percentiles
    - Quantiles

1.4.1 Histogram

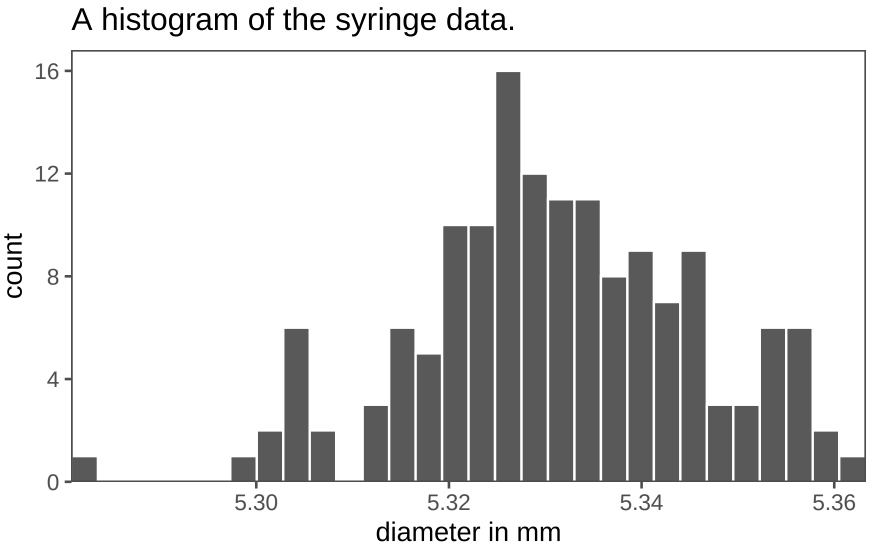


Figure 1.6: An example for descriptive statistics (histogramm)

An example for descriptive statistics is shown in Figure 1.6 as a histogram. It shows data from a company that produces pharmaceutical syringes, taken from Ramalho (2021). During the production of those syringes, the so called *barrel diameter* is a critical parameter to the function of the syringe and therefore of special interest for the Quality Control.

A histogram as shown in Figure 1.6 shows the data of 150 measurements during the QC. On the `x-axis` the *barrel diameter* is shown, while the count of each *binned* diameter is shown on the `y-axis`. The binning and of data is a crucial parameter for such a plot, because it already changes the appearance and width of the bars. Binning is a trade-off between visibility and readability.

1.4.2 Density plot

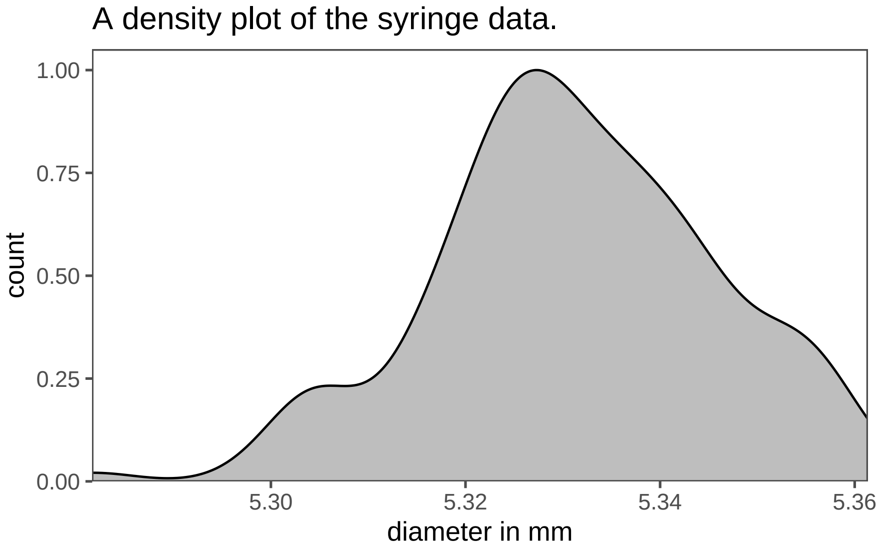


Figure 1.7: An example for a density plot for the syringe data (barrel diameter).

Density plots are another way of displaying the statistical distribution of an underlying dataset. The biggest strength of those plots is, that no binning is necessary in order to show the data. The limitation of this kind of plot is the interpretability. An example of a density plot for the syringe data is shown in Figure 1.7. On the `x-axis` the syringe barrel diameter is shown (as in a histogram). The `y-axis` in contrast does not display the count of a binned category, but rather the Probability Density Function for the specific diameter. The grey area under the density curve depicts the probability of a syringe diameter to appear in the data. The complete area under the curve equals to 1 meaning that a certain diameter is sure to appear in the data.

1.4.3 Boxplot

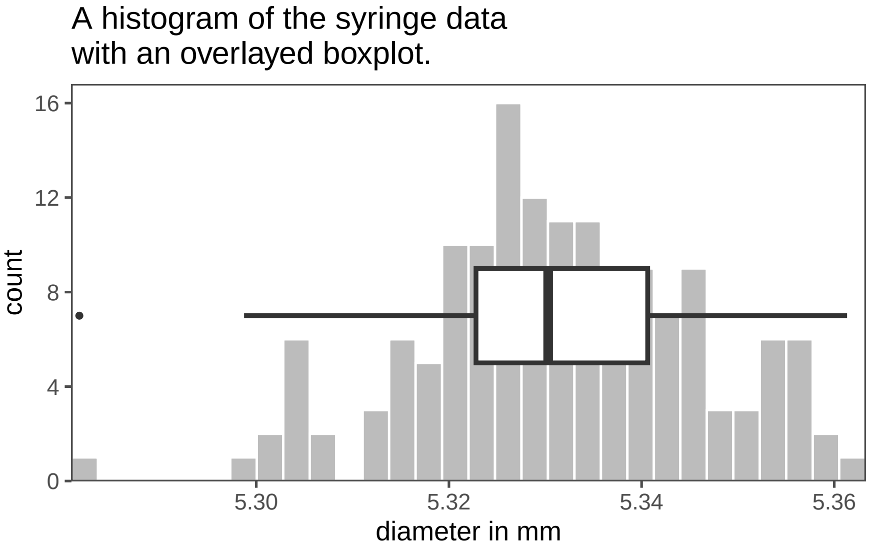


Figure 1.8: A boxplot of the same syringe data combined with the according histogram.

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

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

The syringe data in boxplot form is shown in Figure 1.8 as an overlay of the histogram plot before. Boxplots are useful for quickly understanding the central tendency, spread, and presence of outliers in a dataset, making them a valuable tool in data analysis and visualization.

1.4.4 Average, Standard deviation and Range

Very popular measures of central tendency include the *average* (mean) and the *standard deviation* (variance) of a dataset. The computed mean from an actual dataset is depicted with $\bar{x}$ and calculated via (1.3).

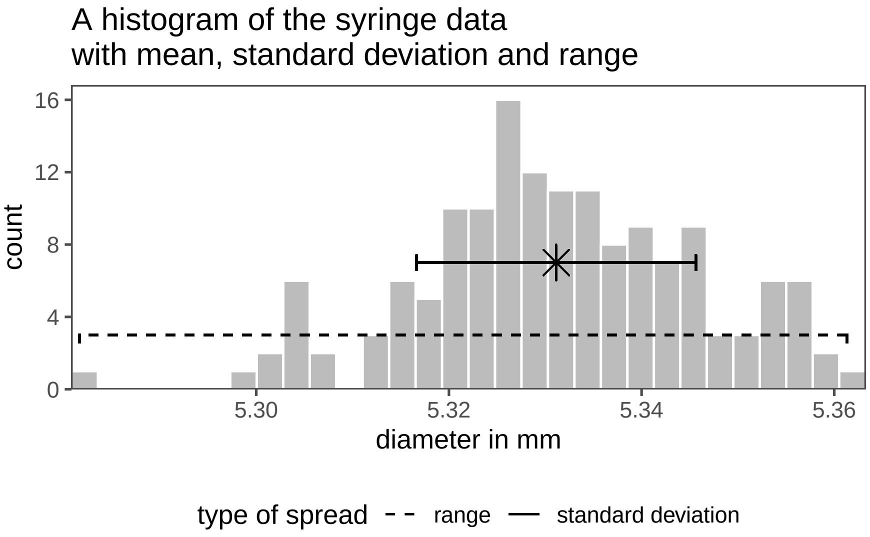


Figure 1.9: A histogram of the syringe data with mean, standard deviation and range.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.3}$$

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

The computed *standard deviation* from an actual dataset is depicted with sd and calculated via (1.4).

$$sd = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.4}$$

The *standard deviation* can therefore be explained as the square root of the sum of all differences of each individual datapoints to the mean of a dataset divided by the number of datapoints. It is not to be confused with the true variance σ_0^2 of a population. The variance of a dataset can be calculatd via (1.5).

$$\sigma = sd^2 \tag{1.5}$$

The *range* from an actual dataset is depicted with r and calculated via (1.6).

$$r = \max(x_i) - \min(x_i) \tag{1.6}$$

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

1.5 Visualizing Groups

1.5.1 Boxplots

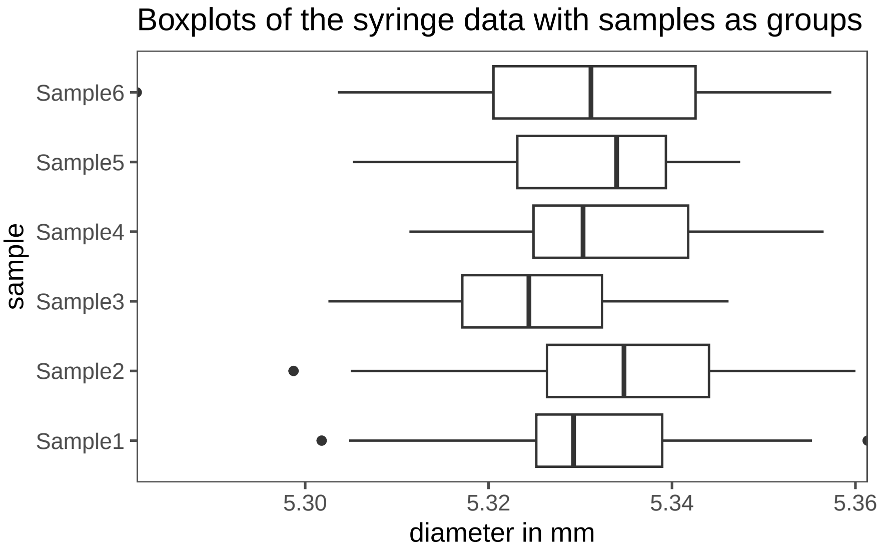


Figure 1.10: Boxplots of the syringe data with the samples as groups.

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

If the underlying distribution of the data is unknown, a good start to visualize groups within data is usually a boxplot as shown in Figure 1.10. The syringe data from Ramalho (2021) contains six different groups, one for every sample drawn. Each sample consists of 25 observations in total. On the `x-axis` the *diameter* in mm is shown, the `y-axis` depicts the sample number. The `boxplots` are then drawn as described above (median, 25^{th} and 75^{th} percentile box, 5^{th} and 95^{th} whisker). The 25^{th} and 75^{th} percentile box is also known as the Interquartile Range

1.5.2 Mean and standard deviation plots

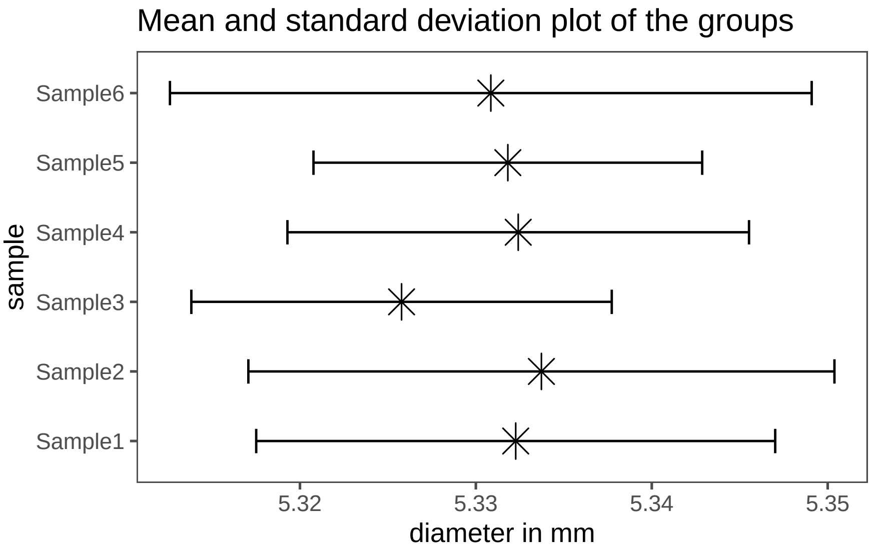


Figure 1.11: Mean and standard deviation plots of the groups in the dataset.

If the data follows a normal distribution, showing the mean and standard deviation for each group is also very common. For the syringe dataset, this is shown in Figure 1.11. The plot follows the same logic as for the boxplots (`x-axis-data`, `y-axis-data`), but the data itself shows the mean with a ×-symbol, as the length of the horizontal errorbars accords to $\bar{x} \pm sd(x)$.

1.5.3 Half-half plots

Boxplots and mean-and-standard-deviation plots sometimes hide some details within the data, that may be of interest or simply important. Half-half plots, as shown in shown in Figure 1.12, incorporate different plot mechanisms. The left half shows a violin plot, which outlines the underlying distribution of the data using the PDF. This is very similar to a density plot. The right half shows the original data points and give the user a visible clue about the sample size in the data size. Note that the `y`-position of the points is jittered to counter *overplotting*. Details can be found in Tiedemann (2022).

1.5.4 Ridgeline plots

Figure 1.13 shows so called *ridgeline* plots as explained in Wilke (2022). They are in essence density plots that use the `y-axis` to differentiate between the groups. On the `x-axis` the density of the underlying dataset is shown. More info on the creation of these plots and graphics

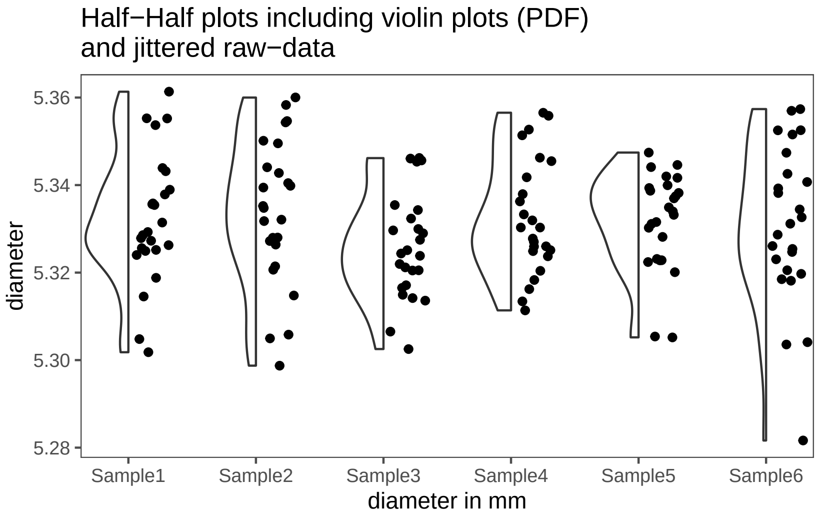


Figure 1.12: Half-half plots that incooperate different types of plots

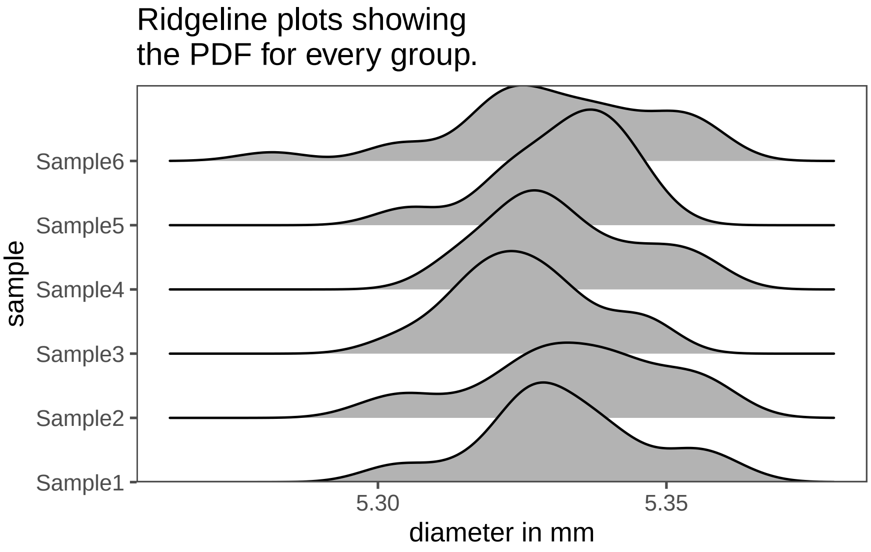


Figure 1.13: Ridgeline plots for distributions within groups.

is available in Wickham (2016) as well as "The R Graph Gallery – Help and Inspiration for r Charts" (2022).

1.6 The drive shaft exercise

1.6.1 Introduction

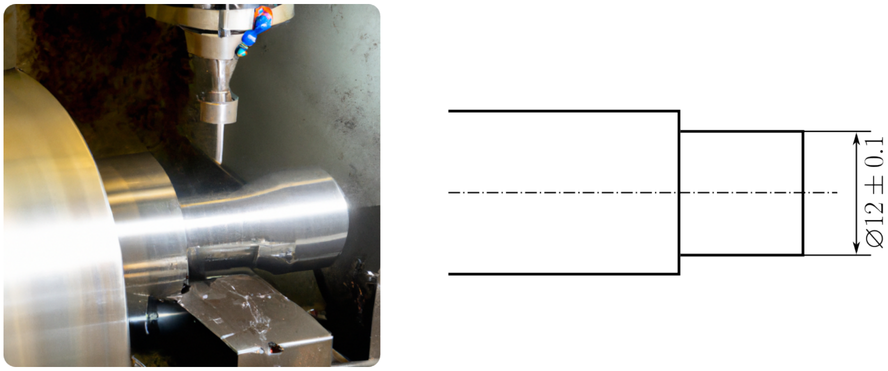


Figure 1.14: The drive shaft specification.

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

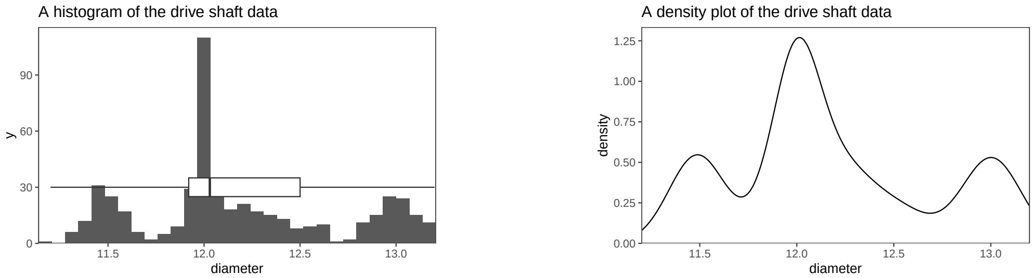
1. Material Selection: Quality steel or aluminum alloys are chosen based on the specific application and requirements.
2. Cutting and Machining: The selected material is cut and machined to achieve the desired shape and size. Precision machining is crucial for balance and performance.
3. Welding or Assembly: Multiple sections may be welded or assembled to achieve the required length. Proper welding techniques are used to maintain structural integrity.
4. Balancing: Balancing is critical to minimize vibrations and ensure smooth operation. Counterweights are added or mass distribution is adjusted.
5. Surface Treatment: Drive shafts are often coated or treated for corrosion resistance and durability. Common treatments include painting, plating, or applying protective coatings.
6. Quality Control: Rigorous quality control measures are employed to meet specific standards and tolerances. This includes dimensional checks, material testing, and defect inspections.

7. Packaging and Distribution: Once quality control is passed, drive shafts are packaged and prepared for distribution to manufacturers of vehicles or machinery.

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

The end diameter of the drive shaft shall be $\emptyset 12 \pm 0.1mm$ (see Figure 1.14). This example will haunt us the rest of this lecture.

1.6.2 Visualizing all the Data



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

Figure 1.15: The raw data of the measured drive shaft diameter.

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

Table 1.1: The summary table of the drive shaft data

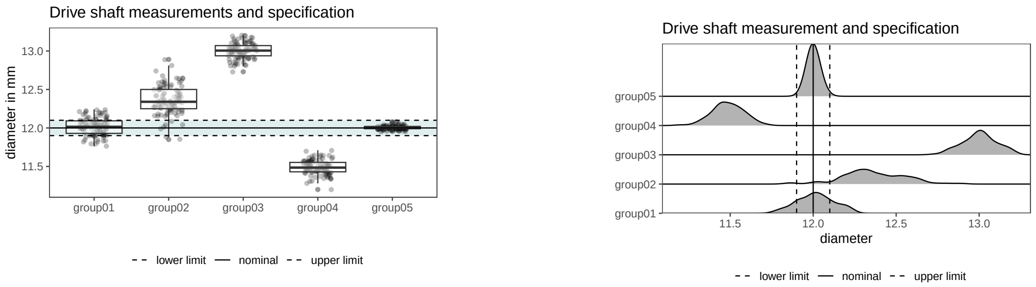
Variable	**N = 500**[1]
diameter	12.17 (0.51), 12.03 (0.58)

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

In Figure 1.15 the data and the distribution thereof is visualized using different modalities. The complete `drive shaft data` is shown as a histogram (Figure 1.15a) and as a density plot

(Figure 1.15b). A single boxplot is plotted over the histogram data in Figure 1.15a, providing a link to Table 1.1 (median and IQR). One important conclusion may be draw from those plots already: There may be more than one dataset hidden inside the data. We will explore this possibility further.

1.6.3 Visualizing groups within the data



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

(b) The groups visualized as ridgeline plots

Figure 1.16: The raw data of the measured drive shaft diameter.

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

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

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

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

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

First in Figure 1.16a the raw data points are shown as points with overlayed boxplots. On the `x-axis` the groups are depicted, while the Parameter of Interest (in this case the *end diameter* of the drive shaft) is shown on the `y-axis`. Because we are interested how the manufactured drive shafts behave with respect to the specification limit, the `nominal` value as well as the `uppper` and the `lower` specification limit is also shown in the plot as horizontal lines.

In Figure 1.16b the data is shown as ridgeline density plots. On the `x-axis` the diameter is depiected, while the `y-axis` shows two types of data. First, the groups $1 \dots 5$ are shown. For the individual groups, the probability is depicted as a line, therefore indicating which values are most probable in the given group. Again, because we are interested how the manufactured drive shafts behave .w.r.t the specification limit, the `nominal` value as well as the `uppper` and the `lower` specification limit is also shown in the plot as vertical lines.

2 Statistical Distributions

2.1 Types of data

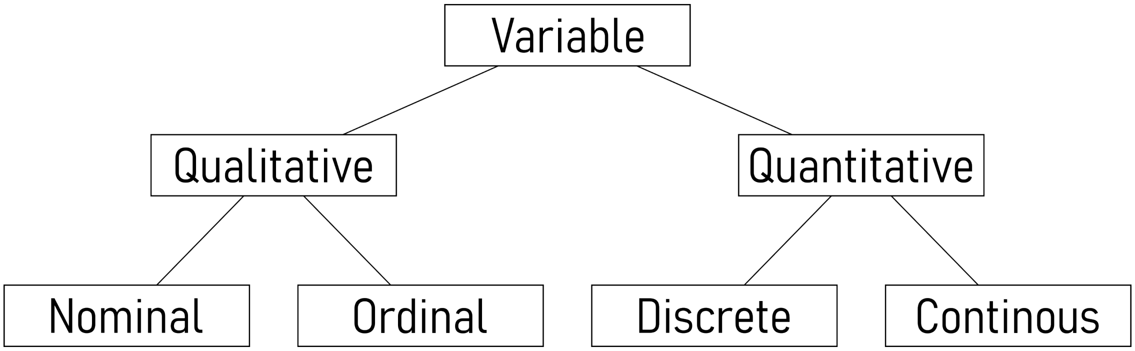


Figure 2.1: Data can be classified as different types.

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

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

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

2.1.1 Nominal Data

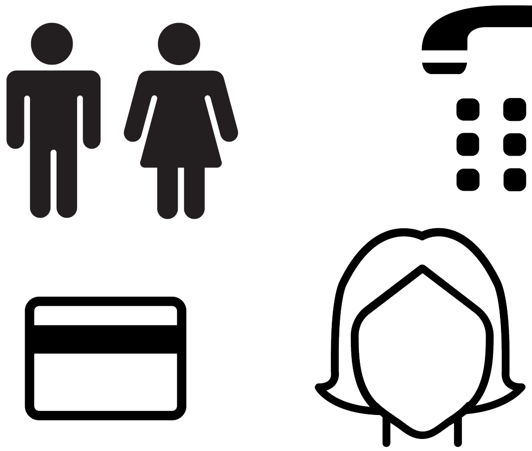


Figure 2.2: Some example for nominal data.

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

2.1.2 Ordinal Data

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

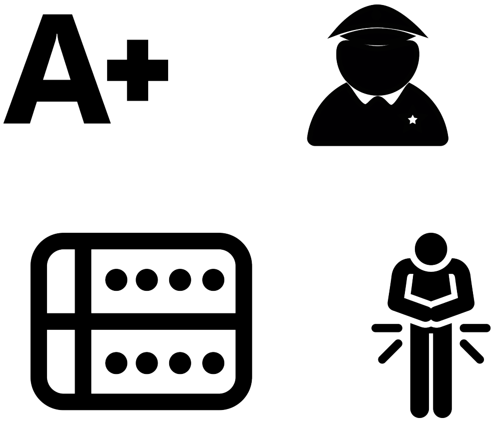


Figure 2.3: Some example for ordinal data.

categories indicate a level of agreement, but the differences between them may not be uniform or measurable.

2.1.3 Discrete Data

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

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

2.1.4 Continous Data

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

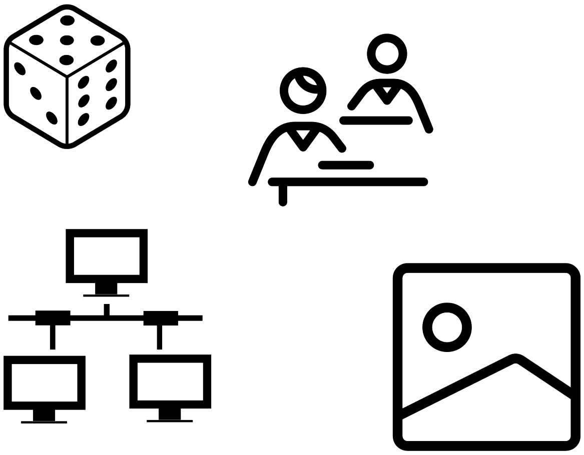


Figure 2.4: Some example for discrete data.

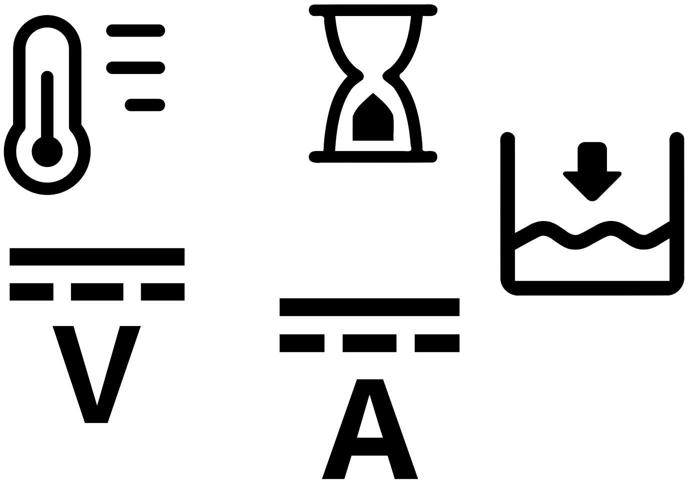


Figure 2.5: Some example for continous data.

2.2 The Normal Distribution

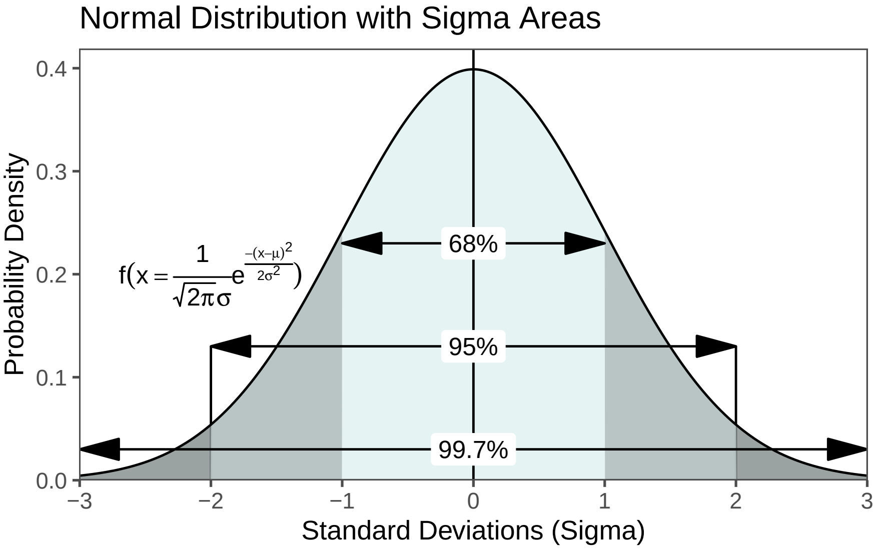


Figure 2.6: The standarized normal distribution

The normal distribution is a fundamental statistical concept that holds immense significance in the realms of engineering and production. It is often referred to as the Gaussian distribution or the bell curve, is a mathematical model that describes the distribution of data in various natural and human-made phenomena, see Johnson (1994). It forms a symmetrical curve when plotted, is centered around a mean (μ_0) and balanced on both sides (Figure 2.6). The spread or dispersion of the data points is characterized by σ_0^2. Those two parameters completley define the normal distribution. A remarkable property of the normal distribution is the empirical rule, which states that approximately 68% of the data falls within one standard deviation from the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations (Figure 2.6). The existence of the normal distribution in the real world is a result of the combination of several factors, including the principles of statistics and probability, the Central Limit Theorem, and the behavior of random processes in nature and society.

2.2.1 Central Limit Theorem (CLT)

The primary reason for the existence of the normal distribution in many real-world datasets is the Central Limit Theorem (Taboga 2017). The CLT states that when you take a large enough number of random samples from any population, the distribution of the sample means will tend to follow a normal distribution, even if the original population distribution is not normal. This means that the normal distribution emerges as a statistical consequence of aggregating random data points. This is shown in Figure 2.7.

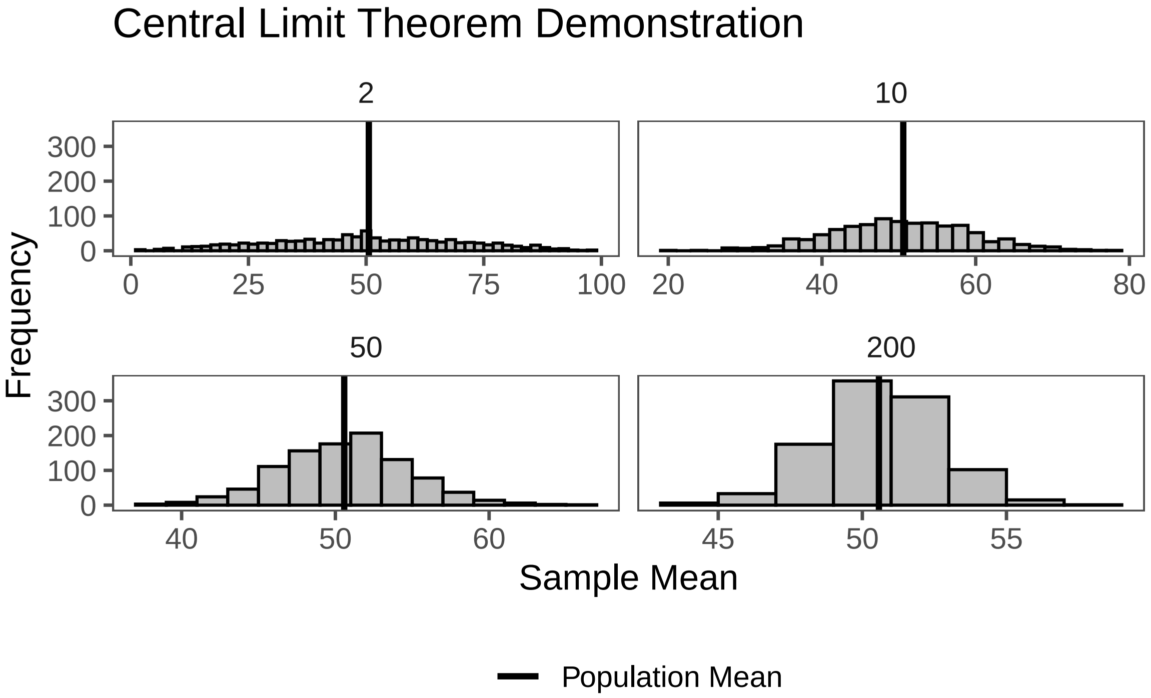


Figure 2.7: The central limit theorem in action.

From $n = 10000$ uniformly disitrbuted data points (the *population*) ($min = 1, max = 100$) either $2, 10, 50$ or 200 samples are taken randomly (the *samples*). For each of the samples the mean is calculated, resulting in 1000 mean values for each $(2, 10, 50$ or $200)$ sample size. In Figure 2.7 the results from this numerical study are shown. The larger the sample size, the closer the mean calculated $\bar{x}$is to the population mean (μ_0). The effect is especially large on the standard deviation, resulting in a smaller standard deviation the larger the sample size is.

2.2.2 Randomness and Independence

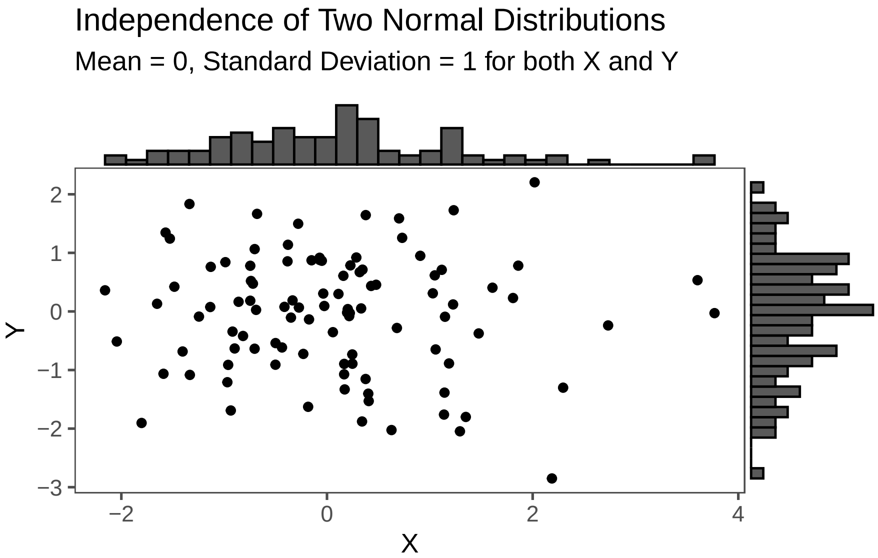


Figure 2.8: Two random and independent samples drawn from a normal distribution. They do not show any indepedence.

In nature and society, many processes involve a large number of random, independent, and additive factors. When these factors combine, their individual effects tend to follow a normal distribution, as predicted by the CLT. This principle is observed in various contexts, such as the behavior of particles in a gas (Brownian motion), the genetics of traits in populations, or the variability in the heights and weights of individuals in a population.

2.2.3 Law of Large Numbers

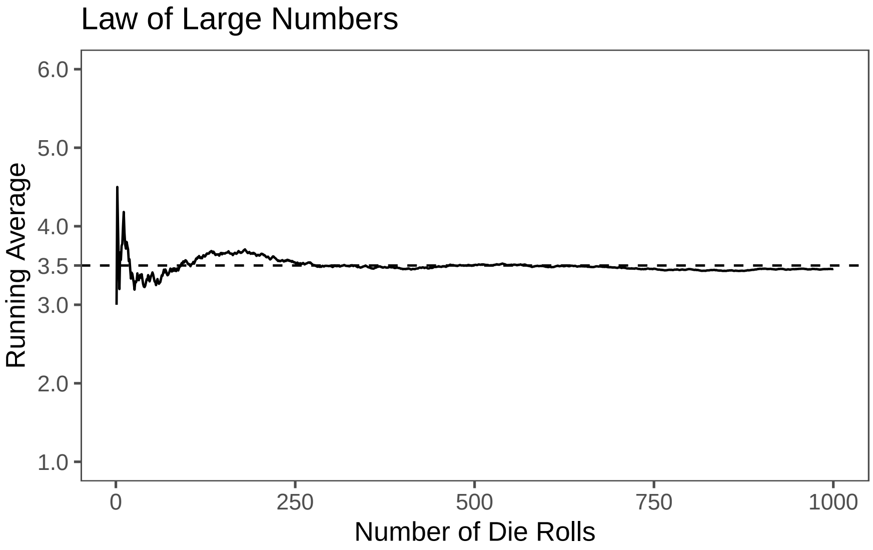


Figure 2.9: The Law of Large Numbers in Action with die rolls as an example.

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

The LLN ar work is shown in Figure 2.9. A fair six-sided die is rolled 1000 times and the running average of the roll results after each roll is calculated. The resulting line plot shows how the running average approaches the expected value of 3.5, which is the average of all possible outcomes of the die. The line in the plot represents the running average It fluctuates at the beginning but gradually converges toward the expected value of 3.5. To emphasize this convergence, a dashed line indicating the theoretical expected value which is essentially the expected value applied to each roll. This visualization demonstrates the Law of Large Numbers, which states that as the number of trials or rolls increases, the *sample mean* (running average in this case) approaches the *population mean* (expected value) with greater accuracy, showing the predictability and stability of random processes over a large number of observations.

2.2.4 The drive shaft exercise - Normal Distribution

In Figure 2.10 the `drive shaft data` is shown for each group in a histogram. As an overlay, the respective *normal distribution* (with the groups $\bar{x}, sd$) is overlayed. If the data is normally

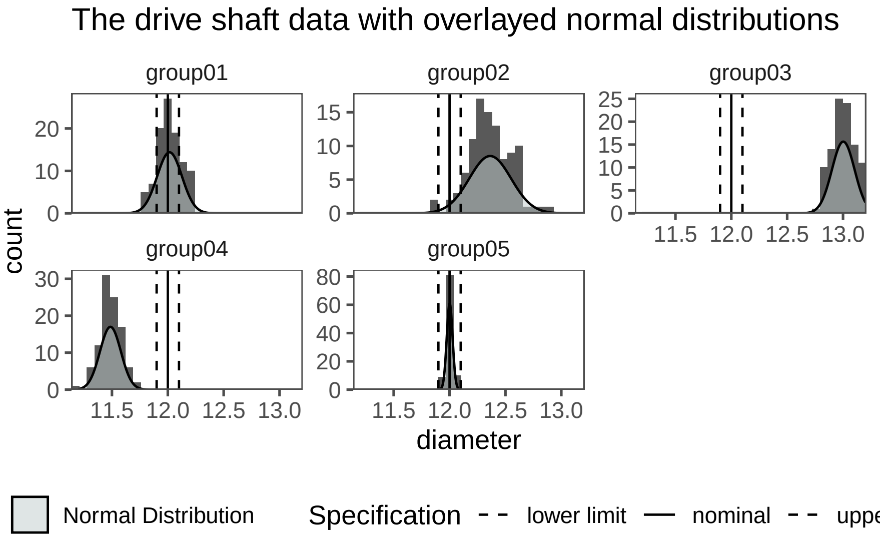


Figure 2.10: The drive shaft data with the respective normal distributions.

distributed, is a different question.

2.3 Z - Standardization

The Z-standardization, also known as standard score or z-score, is a common statistical technique used to transform data into a standard normal distribution with a mean of 0 and a standard deviation of 1 (Taboga 2017). This transformation is useful for comparing and analyzing data that have different scales and units (2.1).

$$Z = \frac{x_i - \bar{x}}{sd} \tag{2.1}$$

How the z-score can be applied is shown in Figure 2.11 and Figure 2.12. The data for group `X` and group `Y` may be measured in different units (Figure 2.11). To answer the question, which of the values $x_i (i = 1 \dots 5)$ is more probable, the single data points are transformed to the respective z-score using (2.1). In Figure 2.12, the z-scores for both groups are plotted against each other. The perfect correlation of the datapoints shows, that for every x_i the same probability applies. Thus, the datapoints are comparable.

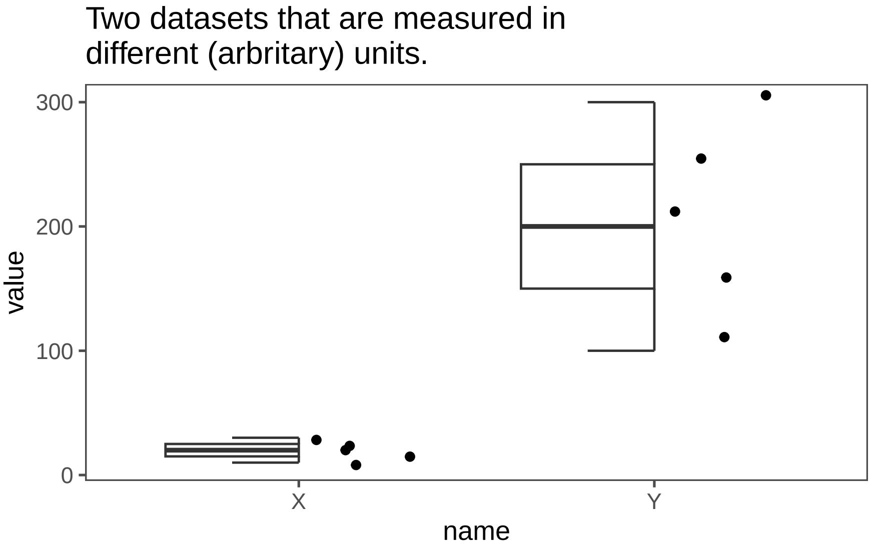


Figure 2.11: The original data of group X and group Y

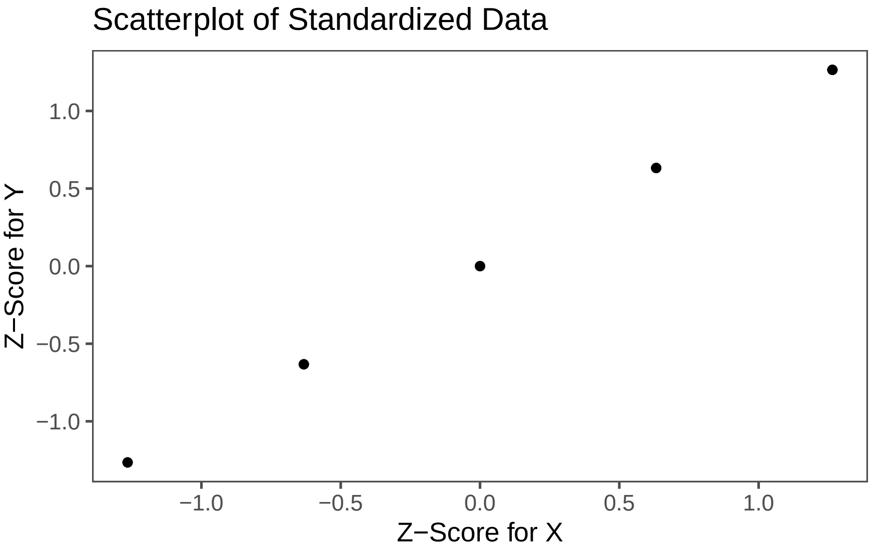


Figure 2.12: The correlation of the z-score shows, that every point x_i is equally probable

2.3.1 The drive shaft exercise - Z-Standardization

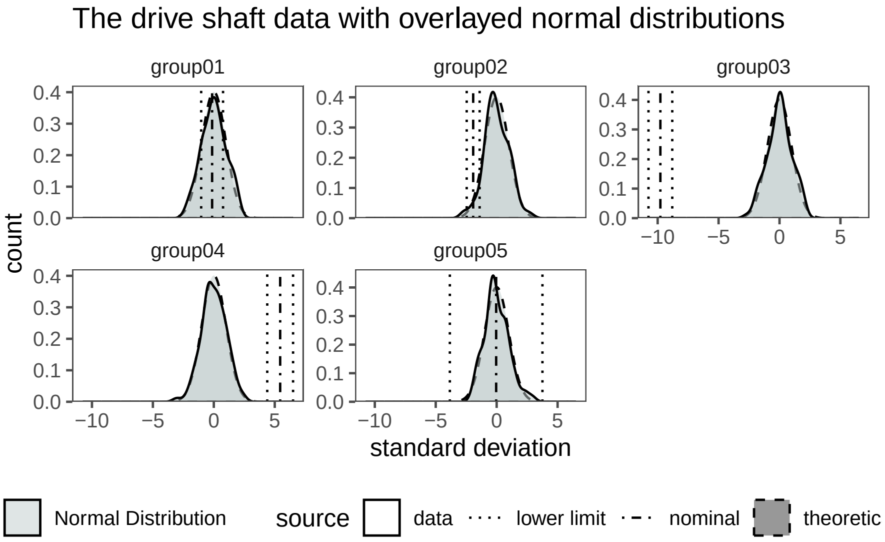


Figure 2.13: The standardized data of the drive shaft data.

In Figure 2.13 the standardized drive shaft data is shown. The mean of the data ($\bar{x}$) is now centered at 0 and the standard deviation is 1. For this case, the specification limits have also been transferred to the respective z-score (even though they can not be interpreted as such anymore). For every x_i the probability to be within a normal distribution is now known. When comparing this to the transferred specification limits, it is clear to see that for `group01` "most" of the data points are within the limits in contrast to `group03` where none of the data points lies within the specification limits. When looking at `group03` we see, that the *nominal* specification limit is -9.78 standard deviations away from the centered mean of the datapoints. The probability of a data point being located there is $6.8605273 \times 10^{-23}$ which does not sound an awful lot. We will dwelve more into such investigation in another chapter, but this is a first step in the direction of inferential statistics.

2.4 Probability Density Function (PDF)

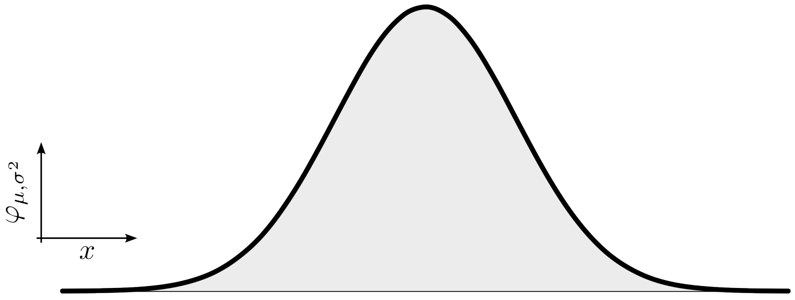


Figure 2.14: A visual represenstation of the PDF for the normal distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{2.2}$$

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

2.5 Cumulative Density Function (CDF)

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

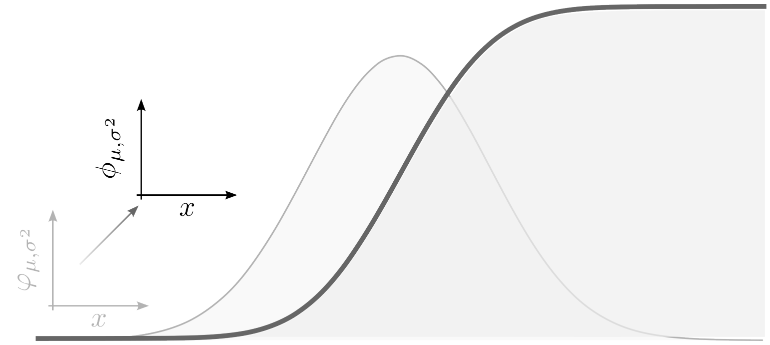


Figure 2.15: A visual represenstation of the CDF for the normal distribution.

$$\begin{aligned}
z &= \frac{x-\mu}{\sigma} \\
\varphi(x) &= \frac{1}{2\pi} e^{\frac{-z^2}{2}} && (2.3) \\
\phi(x) &= \int \frac{1}{2\pi} e^{\frac{-x^2}{2}}\, dx && (2.4) \\
\lim_{x\to\infty} \phi(x) &= 1 \\
\lim_{x\to-\infty} \phi(x) &= 0
\end{aligned}$$

2.6 Likelihood and Probability

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

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

In summary, while both likelihood and probability deal with the chance of events occurring, likelihood is often used in the context of comparing different *hypotheses or models* based

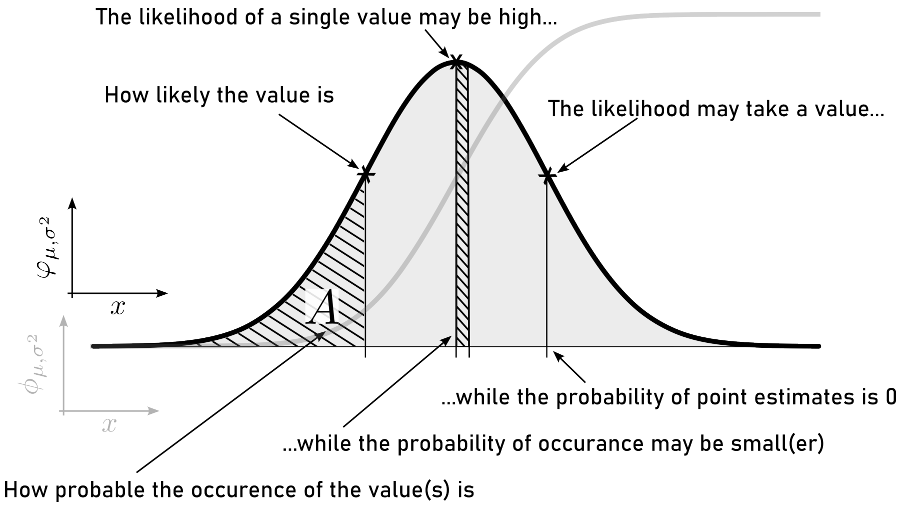
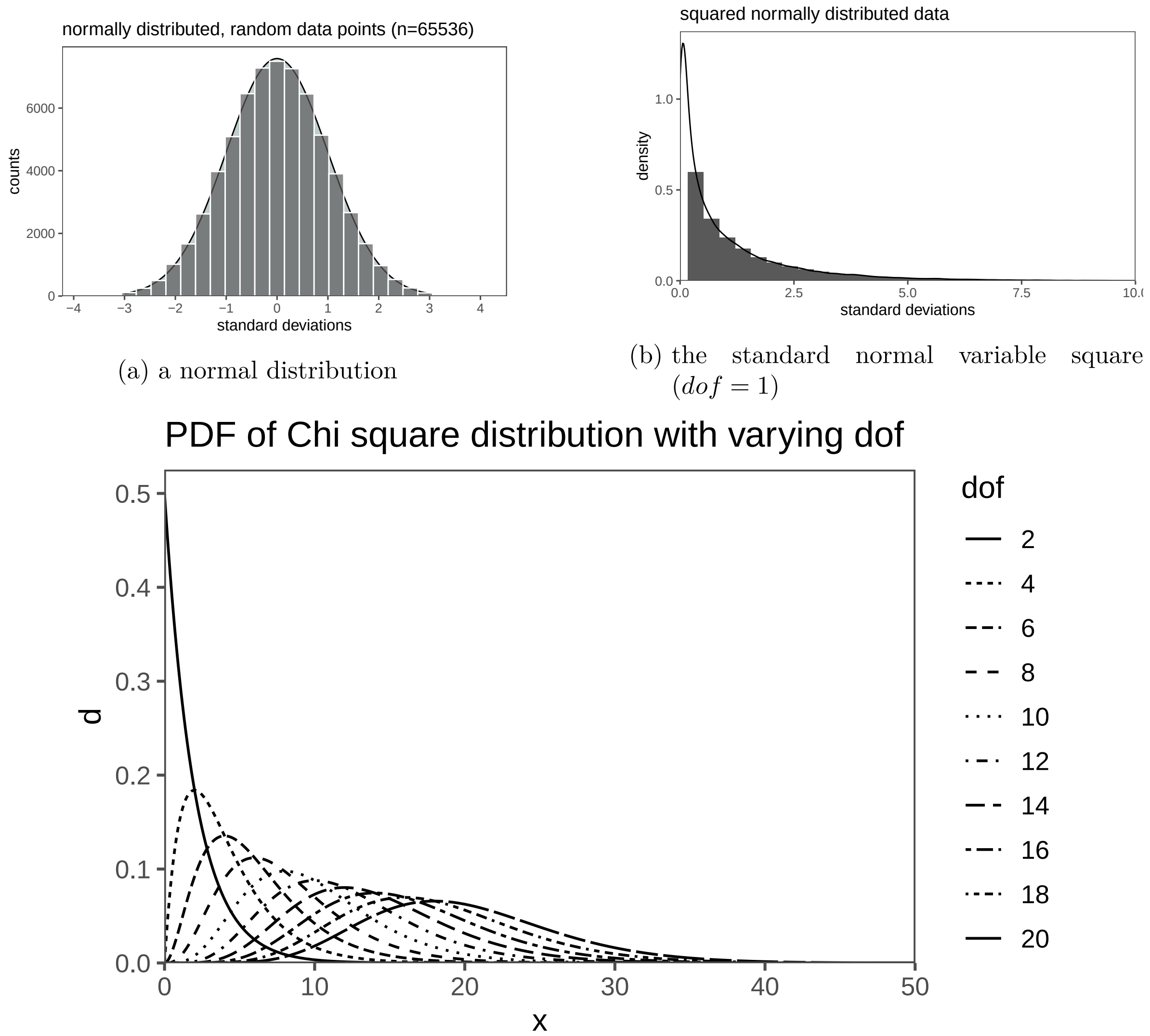


Figure 2.16: The subtle difference between likelihood and probability.

on *observed data*, while probability is more broadly used to quantify the chances of *events happening* in *general*.

2.7 Chi^2 - Distribution



(a) a normal distribution

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

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

Figure 2.17: What a χ^2 distribution reprepresents and how it relates to a the normal distribution.

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

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k} \tag{2.5}$$

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

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

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

2.7.1 The drive shaft exercise - Chi^2 Distribution

In Figure 2.18 the squared standad deviation for every datapoint (from the stanardized data) is shown as a histogram for every group with an overlayed (and scaled) density plot. In the background of every group the theoretical χ^2-distribution with $dof = 1$ is plotted to visually compare the empirical distribution of the datapoints to the theorectial.

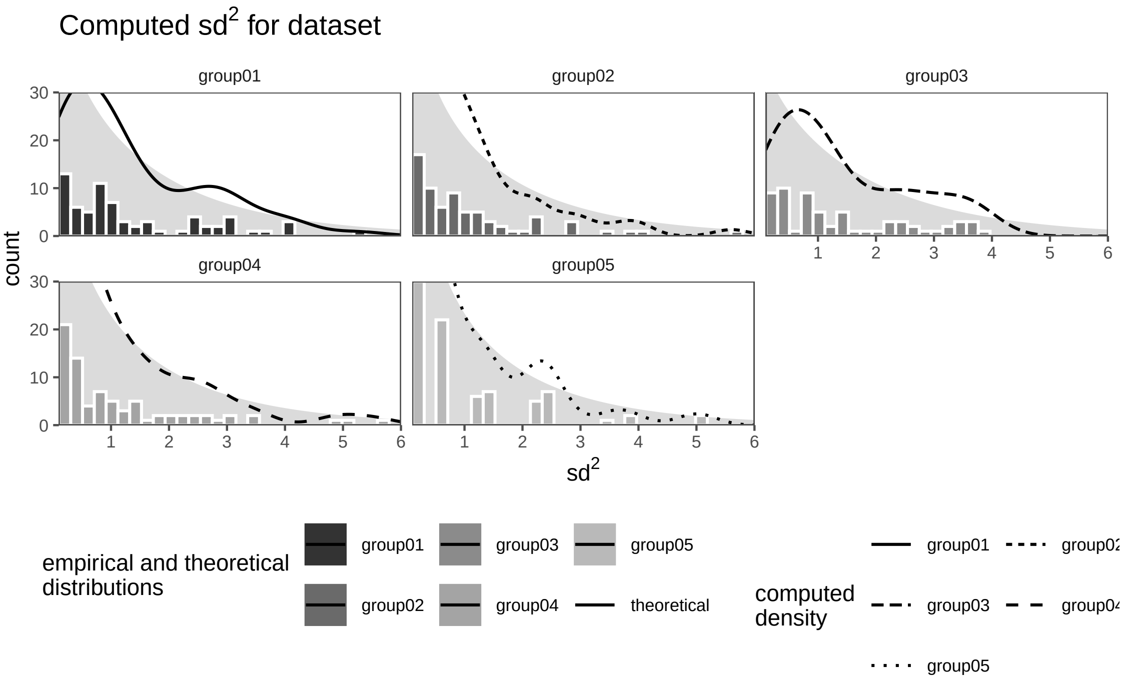


Figure 2.18: The χ^2 disitribution of the drive shaft data.

2.8 t - Distribution

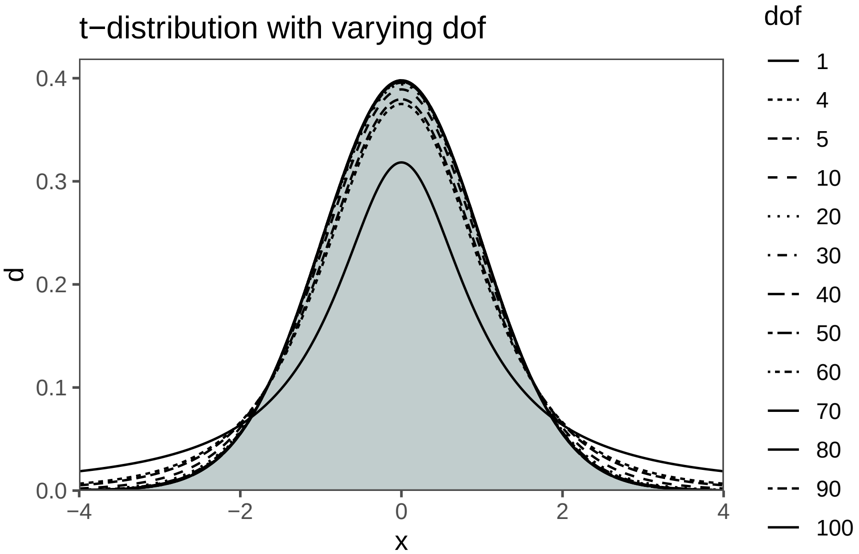


Figure 2.19: PDF of t-distribution with varying dof

The t-distribution, also known as the Student's t-distribution (Student 1908), is a probability distribution that plays a significant role in statistics[1]. It is a symmetric distribution with a bell-shaped curve, similar to the normal distribution, but with heavier tails. The key significance of the t-distribution lies in its application to inferential statistics, particularly in hypothesis testing and confidence interval estimation.

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

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

3. **Degrees of Freedom:** The shape of the t-distribution is determined by a parameter called degrees of freedom (df). As the df increases, the t-distribution approaches the normal distribution. When df is small, the tails of the t-distribution are fatter, allowing for greater uncertainty in estimates.

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

Statisticians found that if they took samples of a constant size from a normal population, computed a statistic called a *t-score* for each sample, and put those into a relative frequency distribution, the distribution would be the same for samples of the same size drawn from any normal population. The shape of this sampling distribution of t's varies somewhat as sample size varies, but for any n, it is always the same. For example, for samples of 5, 90% of the samples have t-scores between -1.943 and $+1.943$, while for samples of 15, 90% have t-scores between ± 1.761. The bigger the samples, the narrower the range of scores that covers any particular proportion of the samples (2.9) (Note the similarity to (2.1)). Since the *t-score* is computed for every x_i the resulting sampling distribution is called the *t-disitribution*.

$$t_i = \frac{x_i - \mu_o}{sd/\sqrt{n}} \tag{2.6}$$

In Figure 2.19 it is shown, that with increasing dof (in this case *sample size*), the *t-distribution* approximates a normal distribution (gray area). Figure 2.19 also shows an example of the *t-distribution* in action. Of all possible samples with 9 dof 0.025 ($2\frac{1}{2}$%) of those samples would have t-scores greater than 2.262, and .975 (97.5%) would have t-scores less than 2.262. The advantage of the *t-score* and *t-distribution* is clearly visible. All these values can be computed from sampled data, the population can remain *estimated* (2.9).

2.8.1 The drive shaft exercise - t-Distribution

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

Suppose we have some data (maybe the drive shaft exercise?) with which calculations can be done. First, the mean $\bar{x}$ and sd is calculated according to (1.3) and (1.4). After this, the confidence level (we will get to this later in more detail) is specified. A value of 95% is a common choice of cl.

$$ci = 0.95 \quad \text{(for a 95\% confidence level)} \tag{2.7}$$

Then the Standard Error (SE) is calculated using (2.8), which takes the sd and n of a sample into account (notice, how we did not use any population estimation?).

$$SE = \frac{sd}{\sqrt{n}} \tag{2.8}$$

In the next step, the critical *t-score* is calculated using the cl as shown in (2.9). *qt* in this case returns the value of the inverse cumulative function of the t-distribution given a certain random variable (or datapoint x_i) and $n-1$ dof. Think of it as an automated look up in long statistical tables.

$$t_{score} = qt\left(\frac{1-ci}{2}, df = n-1\right) \tag{2.9}$$

With this, the *margin of error* can be calculated using the SE and the *t-score* as shown in (2.10).

$$margin\ of\ error = t_{score} \times SE \tag{2.10}$$

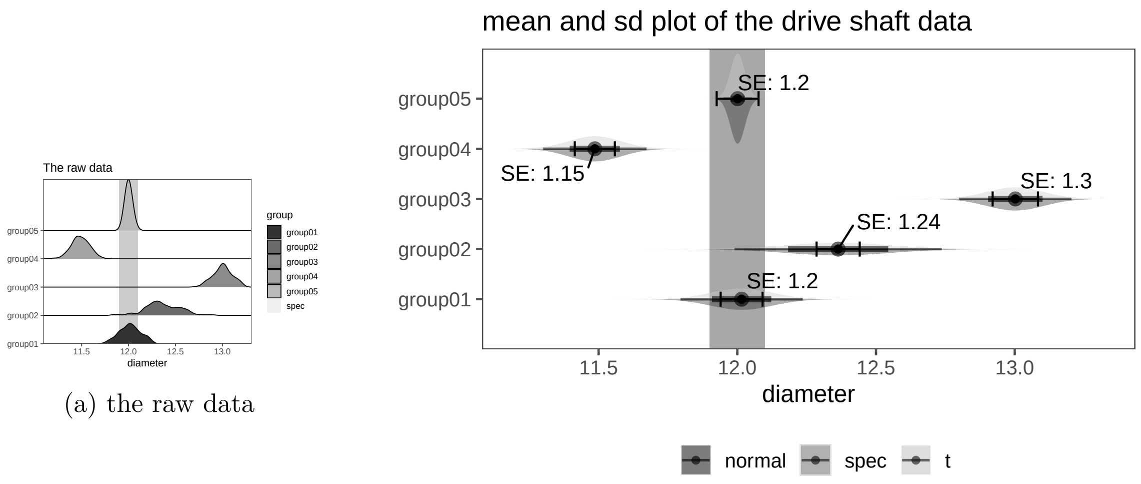
In the last step the Confidence Interval is calculated for the `lower` and the `upper` bound with (2.11) and (2.12).

$$lo = \bar{x} - margin\ of\ error \tag{2.11}$$
$$hi = \bar{x} + margin\ of\ error \tag{2.12}$$

It all looks and feels very similar to using the normal disitrbution. Why this is the case, is shown in Figure 2.20. In Figure 2.20a the raw dataset is shown with the underlayed specification limits for the manufacturing of the drive shaft. For some groups the judgement if the drive shaft is wihtin specification is quite clear (`group 1`, `group 2` and `group 5`). For the other groups, this can not be done so easily. For the drive shaft data, we of course now some population data, therefore the *normal distribution* can be compared to the *t-distribution*. This is done in Figure 2.20b. On the `x-axis` the diameter is shown, the `y-axis` depicts the groups (as before). The distribution on top of the estimated parameters is the population (normal distribution), the distribution on the bottom follow a *t-distribution*. With $n > 30$ (as for this dataset), the difference between disitrbution is very small, further showcasing the use of the *t-distribution* (also see Figure 2.19 for comparison).

2.9 F - Statistics

F-statistics, also known as the *F-test* or *F-ratio*, is a statistical measure used in analysis of variance and regression analysis (Taboga 2017). It assesses the ratio of two variances, indicating the extent to which the variability between groups or models is greater than the



(a) the raw data

(b) normal disitribution, t-distribution and confidence intervalls using the t-distribution

Figure 2.20: The drive shaft data and the application of the t-Distribution

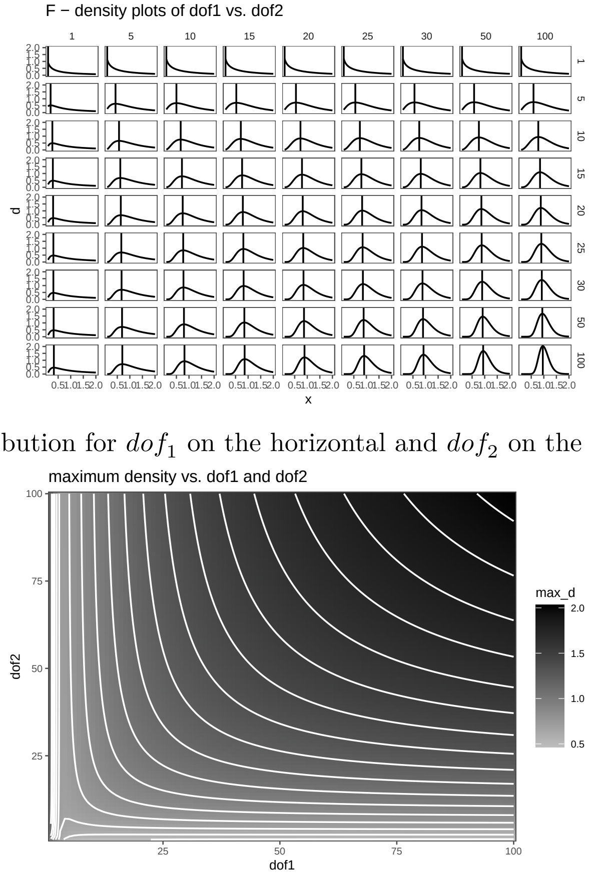
variability within those groups or models. The *F-statistic* plays a crucial role in hypothesis testing and model comparison.

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

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

2. **Regression Analysis:** F-statistics are used in regression analysis to assess the overall significance of a regression model. Specifically, in multiple linear regression, it helps determine whether the model, which includes multiple predictor variables, is better at explaining the variance in the response variable compared to a model with no predictors. It tests the null hypothesis that all coefficients of the model are equal to zero.

The degrees of freedom in an *F-distribution* refer to the two sets of numbers that determine the shape and properties of the distribution (Figure 2.21).



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

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

Figure 2.21: The influence of dof_1 and dof_2 on the density in the F-disitribution

Numerator Degrees of Freedom (dof_1): The numerator degrees of freedom, often denoted as dof_1, is associated with the variability between groups or models in statistical analyses (Figure 2.21a - horizontal axis). In the context of ANOVA, it represents the dof associated with the differences among group means. In regression analysis, it is related to the number of predictors or coefficients being tested simultaneously.

Denominator Degrees of Freedom (dof_2): The denominator degrees of freedom, often denoted as dof_2, is associated with the variability within groups or models (Figure 2.21b - vertical axis). In ANOVA, it represents the degrees of freedom associated with the variability within each group. In regression analysis, it is related to the error or residual degrees of freedom, indicating the remaining variability not explained by the model.

The F-distribution is used to compare two variances: one from the numerator and the other from the denominator. The F-statistic, calculated as the ratio of these variances, follows an F-distribution (2.13).

$$f(x; dof_1, dof_2) = \frac{\Gamma\left(\frac{dof_1+dof_2}{2}\right)}{\Gamma\left(\frac{dof_1}{2}\right)\Gamma\left(\frac{dof_2}{2}\right)}\left(\frac{dof_1}{dof_2}\right)^{\frac{dof_1}{2}} \frac{x^{\frac{dof_1}{2}-1}}{\left(1+\frac{dof_1}{dof_2}x\right)^{\frac{dof_1+dof_2}{2}}} \tag{2.13}$$

$$F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n} \tag{2.14}$$

In practical terms: A higher numerator degrees of freedom (dof_1) suggests that there are more groups or predictors being compared, which may result in larger F-statistic values. A higher denominator degrees of freedom (dof_2) implies that there is more data within each group or model, which may lead to smaller F-statistic values. The F-distribution is right-skewed and always positive. It has different shapes depending on the values of dof_1 and dof_2 (Figure 2.21b). The exact shape is determined by these degrees of freedom and cannot be altered by changing sample sizes or data values (Figure 2.21b). Researchers use F-distributions to conduct hypothesis tests, such as F-tests in ANOVA and F-tests in regression, to determine if there are significant differences between groups or if a regression model is statistically significant.

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

2.10 Interconnections

1. Normal Distribution The **Normal Distribution** is characterized by its mean (μ) and standard deviation (σ), see Figure 2.22. It serves as the foundation for many statistical analyses.

2. Standardized Normal Distribution The **Standardized Normal Distribution**, denoted as $Z \sim N(0,1)$, is a special case of the normal distribution. It has a mean (μ) of 0 and a standard deviation (σ) of 1. It is obtained by standardizing a normal distribution variable X: $Z = \frac{X-\mu}{\sigma}$ (Figure 2.22).

3. t Distribution The **t Distribution** is related to the normal distribution and depends on degrees of freedom. As dof increases, the t-distribution approaches the standard normal distribution (Figure 2.22).

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

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

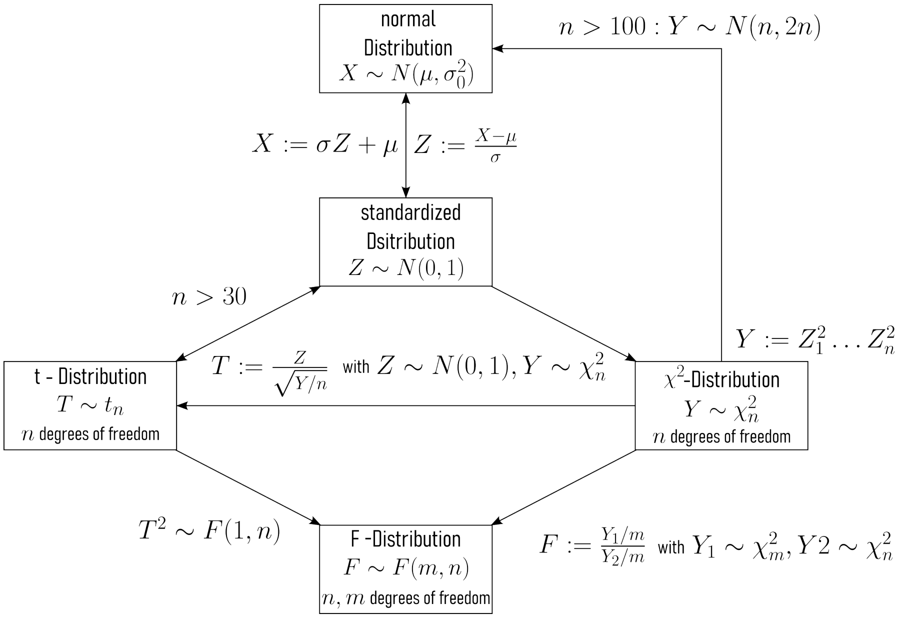

normal Distribution $X \sim N(\mu, \sigma_0^2)$
$n > 100 : Y \sim N(n, 2n)$
$X := \sigma Z + \mu$
$Z := \frac{X-\mu}{\sigma}$
standardized Dsitribution $Z \sim N(0,1)$
$n > 30$
$Y := Z_1^2 \dots Z_n^2$
t - Distribution $T \sim t_n$ n degrees of freedom
$T := \frac{Z}{\sqrt{Y/n}}$ with $Z \sim N(0,1), Y \sim \chi_n^2$
χ^2-Distribution $Y \sim \chi_n^2$ n degrees of freedom
$T^2 \sim F(1,n)$
F -Distribution $F \sim F(m,n)$ n, m degrees of freedom
$F := \frac{Y_1/m}{Y_2/m}$ with $Y_1 \sim \chi_m^2, Y2 \sim \chi_n^2$

Figure 2.22: The distributions are interconnected in several different ways.

2.11 Bionmimal Distribution

The binomial distribution is a **discrete** probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of

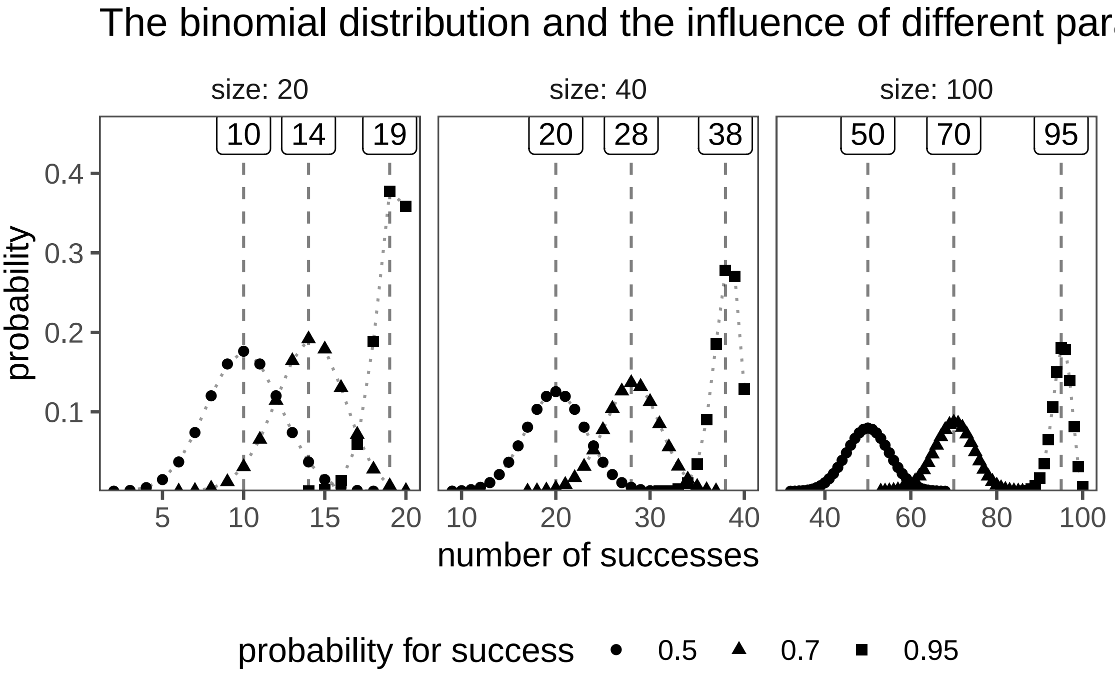


Figure 2.23: The binomial distribution

success. A Bernoulli trial, named after Swiss mathematician Jacob Bernoulli[2], is a random experiment or trial with two possible outcomes: success and failure. These outcomes are typically labeled as 1 for success and 0 for failure. The key characteristics of a Bernoulli trial are:

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

Examples of Bernoulli trials include:

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

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

2.11.1 Probability Mass Function (PMF)

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

- Definition: The PMF describes the probability distribution of a discrete random variable. It gives the probability that the random variable takes on a specific value. In other words, the PMF assigns probabilities to each possible outcome of the random variable.
- Formal Representation: For a discrete random variable X, the PMF is denoted as P(X = x), where x represents a specific value. Mathematically, the PMF is defined as: $P(X = x) = \text{probability that } X \text{ takes the value } x$

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

- Properties: The probabilities associated with all hypothetical values must be non-negative and sum up to 1. Thinking of probability as “mass” helps avoid mistakes, as the total probability for all possible outcomes is conserved (similar to how physical mass is conserved).
- Comparison with Probability Density Function (PDF): A PMF is specific to *discrete* random variables, while a PDF is associated with continuous random variables. Unlike a PDF, which requires integration over an interval, the PMF **directly** provides probabilities for individual values.
- Mode: The value of the random variable with the largest probability mass is called the mode.
- Measure-Theoretic Formulation: The PMF can be seen as a special case of more general measure-theoretic constructions. It relates to the distribution of a random variable and the probability density function with respect to the counting measure.

The PMF for the binomial distribution is given in (2.15)

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{2.15}$$

2.11.2 The drive shaft exercise - Binomial Distribution

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

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

2.12 Weibull - Distribution

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

The Weibull distribution is characterized by two parameters:

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

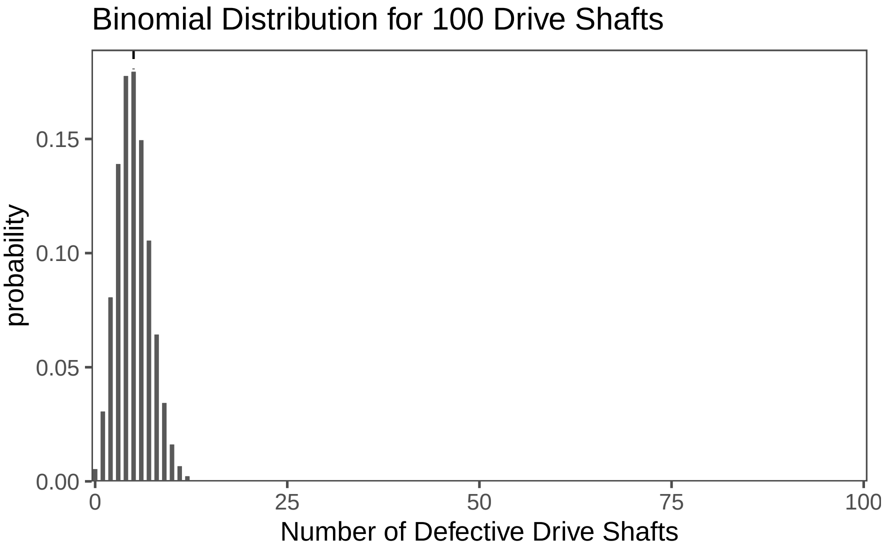


Figure 2.24: The binomial disitribution and the drive shaft exercise.

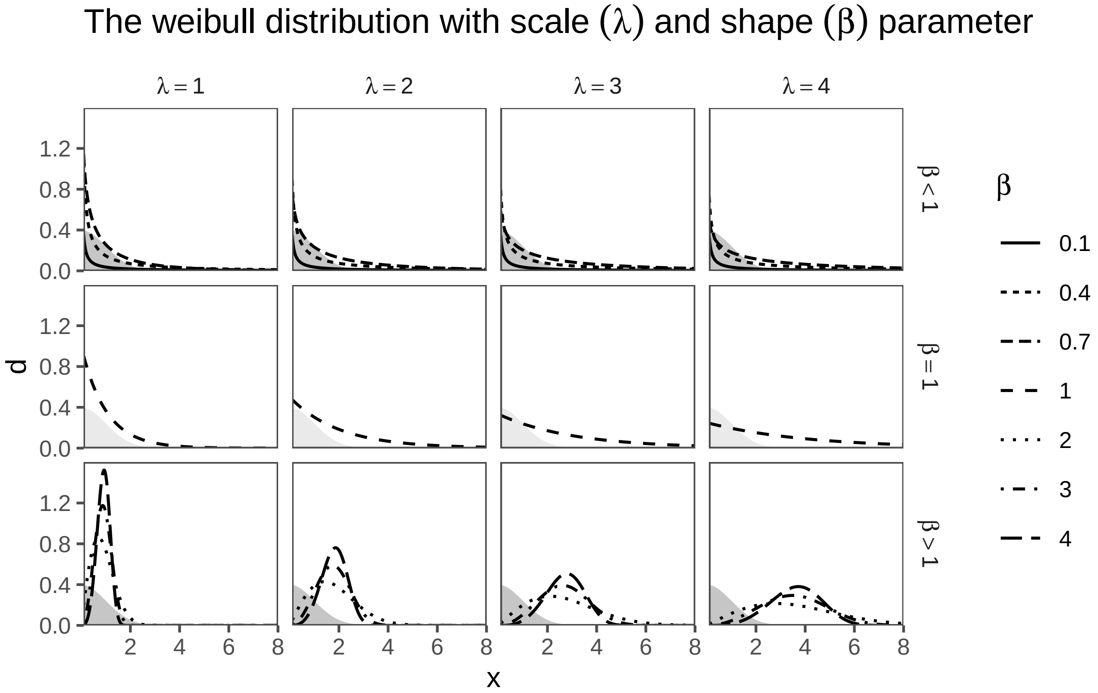


Figure 2.25: The weibull distribution and the influence of β and λ

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

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

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

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

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

2.12.1 The drive shaft exercise - Weibull distribution

The weibull distribution can be applied to estimate the probability of a part to fail after a given time. Suppose there have been $n = 100$ drive shafts produced. In order to assure that the assembled drive shaft would last during their service time, they have been tested in a test-stand that mimics the mission profile[4] of the product. This process is called *qualification* and a big part of any product development (Meyna 2023). The measured hours are shown in Figure 2.26 in a histogram of the data. On the `x-axis` the `Time to failure`is shown, while the `y-axis` shows the number of parts that failed within the time. They histogram plot is overlayed with an empirical density plot as a solid line, as well as the theoretical distribution as a dotted line (Luckily, we know the distribution parameters).

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

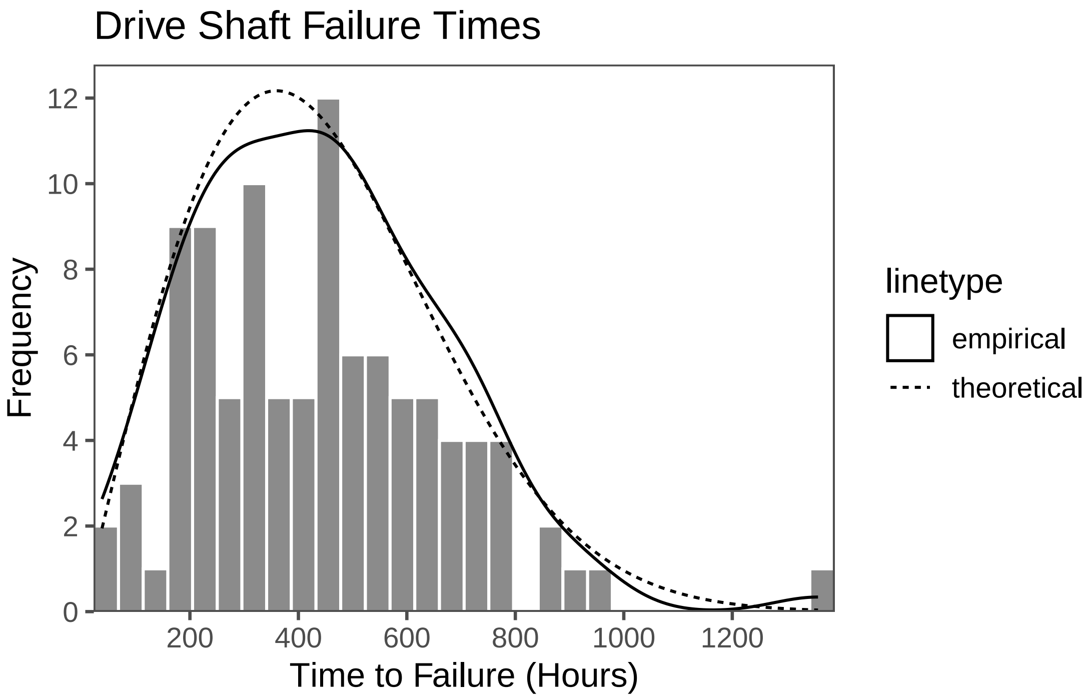


Figure 2.26: The measured hours how long the drive shafts lasted in the test stand.

2.13 Poisson - Distribution

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

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

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

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

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

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

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

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

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

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

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

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

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

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

When λ is small, the distribution is skewed to the right and is more concentrated toward zero (Figure 2.27). When λ is moderate, the distribution approaches a symmetric bell shape (Figure 2.27). When λ is large, the distribution becomes increasingly similar to a normal distribution(Figure 2.27).

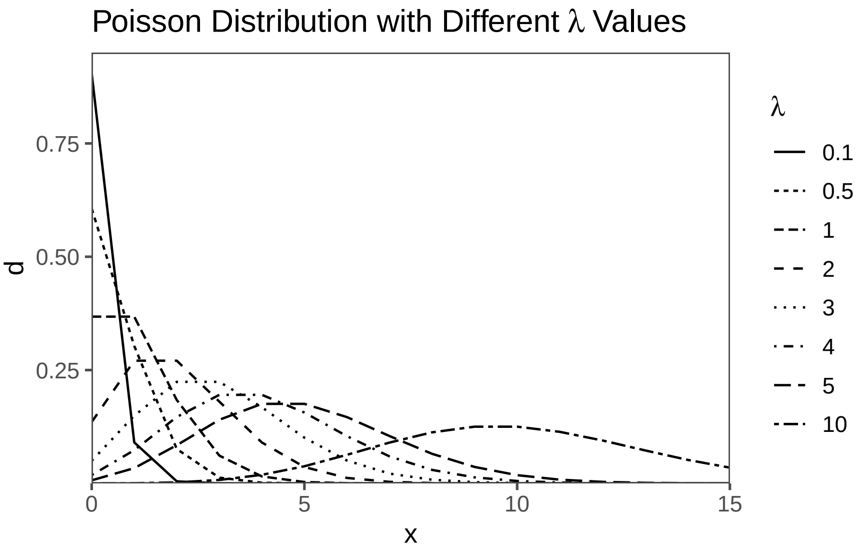


Figure 2.27: The poisson distribution with different λ values.

2.14 Gamma - Distribution

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

Key points about the gamma distribution:

1. It is often used to model the waiting times for events that occur at a constant rate, such as the time between arrivals in a Poisson process.
2. The exponential distribution is a special case of the gamma distribution when $\alpha = 1$ (Figure 2.28).
3. The gamma distribution is right-skewed for $\alpha > 1$ and left-skewed for $0 < \alpha < 1$ (Figure 2.28).
4. The mean of the gamma distribution is $\frac{\alpha}{\beta}$, and its variance is $\frac{\alpha}{\beta^2}$ (Figure 2.28).

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

The connection to other distributions:

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

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

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

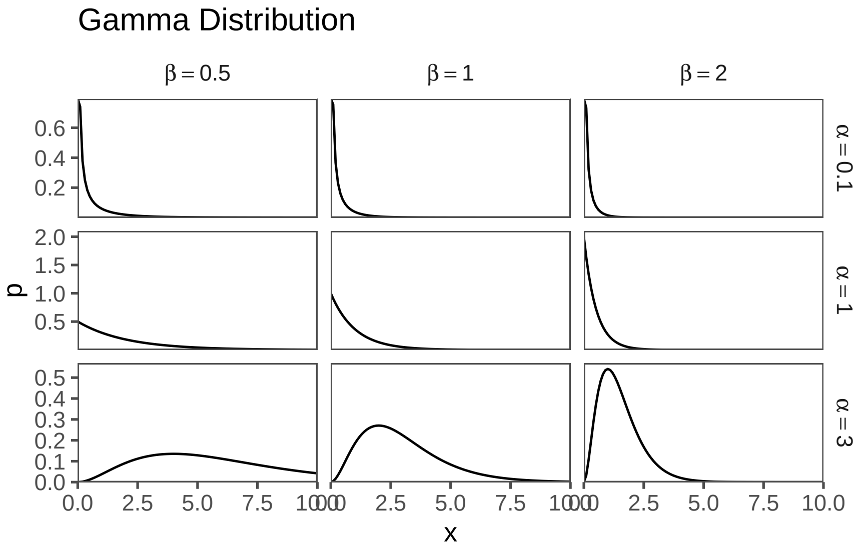


Figure 2.28: The Gamma distribution with varying α (shape) and β (scale)

References

J. Bibby, E. J. G. Pitman. 1980. “Some Basic Theory for Statistical Inference.” *The Mathematical Gazette* 64 (428): 138–38. https://doi.org/10.2307/3615104.

Johnson, Norman Lloyd. 1994. *Continuous Univariate Distributions*. Wiley.

Meyna, Arno. 2023. *Sicherheit Und Zuverlässigkeit Technischer Systeme*. Carl Hanser Verlag GmbH & Co. KG. https://doi.org/10.3139/9783446468085.fm.

Ramalho, Joao. 2021. *industRial: Data, Functions and Support Materials from the Book "industRial Data Science"*. https://CRAN.R-project.org/package=industRial.

Student. 1908. “The Probable Error of a Mean.” *Biometrika* 6 (1): 1. https://doi.org/10.2307/2331554.

Taboga, Marco. 2017. *Lectures on Probability Theory and Mathematical Statistics - 3rd Edition*. Createspace Independent Publishing Platform.

“The R Graph Gallery – Help and Inspiration for r Charts.” 2022. https://r-graph-gallery.com/.

Tiedemann, Frederik. 2022. *Gghalves: Compose Half-Half Plots Using Your Favourite Geoms*. https://CRAN.R-project.org/package=gghalves.

Weibull, Waloddi. 1951. “A Statistical Distribution Function of Wide Applicability.” *Journal of Applied Mechanics* 18 (3): 293–97. https://doi.org/10.1115/1.4010337.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wilke, Claus O. 2022. *Ggridges: Ridgeline Plots in 'Ggplot2'*. https://CRAN.R-project.org/package=ggridges.