1st Edition

Dice with Data

A Journey Through Probability and Statistics

İKSAD Publishing House

Seçil Yalaz
Özge Kuran

DICING WITH DATA
A Journey Through Probability and Statistics

Seçil YALAZ
Özge KURAN

DOI: https://dx.doi.org/10.5281/zenodo.10412942

İKSAD Publishing House

Institution of Economic Development and Social Researches Publications®
(The Licence Number of Publicator: 2014/31220)
TÜRKİYE TR: +90 342 606 06 75
USA: +1 631 685 0 853
E mail: iksadyayinevi@gmail.com
www.iksadyayinevi.com

ISBN: 978-625-367-517-2
Cover Design: Kıvanç GÜNEY
November / 2023
Ankara / Türkiye
Size: 21x29,7 cm

Assoc. Prof. Dr. Seçil Yalaz and Assoc. Prof. Dr. Özge Kuran
Dicle University, Faculty of Science, Department of Statistics
Diyarbakır, TÜRKİYE

CONTENTS

CHAPTER 2
STATISTICS

PREFACE

In this book, information is provided on the solutions to the fundamental topics of probability and statistics through real-life examples. As it is known, probability is a fundamental concept that permeates various aspects of our daily lives, playing a crucial role in decision-making, risk assessment, and prediction. At its core, probability provides a mathematical framework for quantifying uncertainty and randomness, allowing us to make informed choices in the face of unpredictability. From finance and insurance to healthcare, weather forecasting, and even gaming, understanding and applying probability is instrumental in navigating the uncertainties inherent in the real world.

In finance, for instance, probability is a cornerstone of risk management. Investors and financial analysts utilize probability models to estimate the likelihood of different market outcomes, helping them make strategic decisions and optimize investment portfolios. Probability is also integral in the realm of insurance, where actuaries rely on statistical models to assess the likelihood of various events, enabling the development of policies that strike a balance between coverage and affordability.

In healthcare, probability plays a crucial role in diagnostic and treatment decisions. Medical professionals often use probabilistic models to evaluate the likelihood of a particular diagnosis or to assess the effectiveness of different treatment options. Additionally, probability is essential in epidemiology, helping to predict the spread of diseases and inform public health strategies.

The importance of probability extends to everyday situations as well. Weather forecasts, for example, rely on probabilistic models to predict the likelihood of various weather conditions. In games of chance and gambling, probability governs the odds of different outcomes, influencing decisions and strategies.

Statistics is also a powerful and indispensable tool that underlies the fabric of decision-making and understanding in our everyday lives. In a world inundated with information and variability, statistics empowers us to make sense of complex data, draw meaningful insights, and derive reliable conclusions. Its applications are diverse and far-reaching, influencing fields as varied as science, business, healthcare, and social sciences.

In business and economics, statistics plays a pivotal role in market research, forecasting, and strategic planning. Through the analysis of historical data and the application of statistical models, businesses can make informed decisions, optimize processes, and anticipate trends, thereby gaining a competitive edge in dynamic markets.

In healthcare, statistics guides medical research and informs public health policies. Clinical trials, epidemiological studies, and outcome analyses rely on statistical methods to draw valid conclusions about the effectiveness of treatments, the prevalence of diseases, and the impact of interventions.

In scientific research, statistics serves as the backbone of experimentation and hypothesis testing. Scientists use statistical tools to analyze experimental data, determine the reliability of results, and infer the significance of their findings. Whether in physics, biology, or social sciences, statistical methods provide a rigorous framework for drawing conclusions from empirical evidence.

Furthermore, statistics is pivotal in public policy and governance. Policymakers use statistical data to assess the effectiveness of social programs, allocate resources, and address societal challenges. It helps

us understand demographic trends, economic indicators, and social disparities, enabling evidence-based decision-making that can lead to more equitable and efficient policies.

Beyond these examples, statistics is deeply embedded in our daily lives, from opinion polls and sports analytics to quality control in manufacturing. Its ability to distill patterns from data and quantify uncertainty ensures that decisions are not made in the dark but are instead grounded in evidence and rational analysis.

R, a powerful and open-source statistical programming language, has been used in obtaining certain graphs. R has become an indispensable asset for professionals and researchers working in diverse fields where data analysis is paramount with its extensive range of statistical tools, data visualization capabilities, and a vibrant community of users and contributors.

The importance of R in probability and statistics lies in its ability to seamlessly translate theoretical concepts into practical applications. As a statistical computing language, R enables the implementation of complex statistical models, hypothesis testing, and probability simulations with relative ease. Its extensive collection of packages, libraries, and functions dedicated to statistical analysis makes it a versatile tool for tackling a wide array of analytical challenges.

R's robust statistical modeling capabilities empower users to explore, analyze, and interpret complex datasets. From basic descriptive statistics to advanced predictive modeling, R provides a comprehensive suite of functions for understanding the underlying patterns and relationships within data. Whether in finance, biology, social sciences, or any other field, R's flexibility makes it adaptable to the unique needs of various domains.

The collaborative nature of R is also a key factor in its significance. With a vast and engaged community of statisticians, data scientists, and researchers, R benefits from continuous development and improvement. Users can share code, contribute to packages, and collectively enhance the capabilities of R, fostering a dynamic ecosystem that evolves with the ever-changing landscape of statistical methodologies.

Moreover, R excels in data visualization, allowing users to create compelling graphs, charts, and plots that facilitate the communication of statistical findings. The graphing capabilities of the R programming language stand out as a cornerstone in the domain of probability and statistics. R's robust and versatile graphing tools empower researchers, analysts, and statisticians to visually explore, communicate, and interpret complex datasets with precision and clarity. From basic exploratory data analysis to the creation of sophisticated visualizations, R offers an extensive suite of packages and functions that make it a go-to choice for professionals in a variety of fields.

In probability and statistics, effective data visualization is paramount, as it provides an intuitive and accessible means of conveying complex information. R excels in this regard, offering a plethora of plotting functions and libraries, with the acclaimed ''ggplot2'' package leading the way. This package, known for its declarative syntax and flexibility, allows users to create a wide range of high-quality graphs, including scatter plots, histograms, box plots, and intricate multi-faceted visualizations. The integration of packages like ''ggplot2'' provides a high level of customization and aesthetics, enhancing the clarity and interpretability of statistical results.

R's graphing capabilities extend beyond basic plotting, offering advanced tools for visualizing statistical models and outcomes. With capabilities for visualizing probability distributions, density plots, and statistical models, R aids in communicating the nuances of data analysis to both technical

and non-technical audiences. This visual storytelling aspect is crucial in presenting findings, supporting decision-making processes, and fostering a deeper understanding of complex statistical concepts.

Furthermore, R's ability to integrate seamlessly with LaTeX, HTML, and other document formats enhances its utility in academic and professional settings. This facilitates the creation of publication-ready graphics and reports, streamlining the process of sharing statistical insights with peers, stakeholders, or the broader scientific community.

The collaborative nature of R's graphing community ensures a constant influx of innovative techniques and improvements. As statisticians and data scientists contribute to and expand the realm of graphing capabilities in R, users benefit from a continually evolving toolset that stays at the forefront of modern data visualization.

In essence, probability provides a systematic and quantitative framework for dealing with uncertainty, making it an indispensable tool in fields ranging from science and technology to business and everyday decision-making. While its real-life applications underscore its significance in helping individuals and organizations navigate an inherently uncertain world, the real-life importance of statistics lies in its capacity to transform raw data into meaningful information, providing a lens through which we can better understand the world, make informed choices, and address the complex challenges that define our contemporary existence. The importance of R in probability and statistics lies in its capacity to democratize data analysis, making advanced statistical techniques accessible to a broad audience. Whether you are a seasoned statistician or a beginner exploring the intricacies of probability, R stands as a reliable and efficient tool that empowers users to extract meaningful insights from data and advance the frontiers of statistical knowledge. R's graphing capabilities also play a pivotal role in probability and statistics, offering a rich and adaptable toolkit for exploring and communicating complex data patterns. Whether it's for exploratory data analysis, hypothesis testing, or model interpretation, R's graphing capabilities enhance the accessibility and impact of statistical insights in a diverse array of applications.

CHAPTER 1

PROBABILITY

1. PROBABILITY

Probability theory, a relatively recent addition to the realm of mathematics, is concerned with determining the likelihood of events occurring. It involves the mathematical calculation of the chance that a particular event might transpire and finds widespread application in various everyday scenarios. These include weather forecasting, sports strategies, insurance options, games, recreational activities, and business decision-making. In the realm of life, probability theory plays a crucial role in risk management and financial market trading. It is essential for individuals to comprehend how probability assessments are formulated and their impact on decision-making processes. Notably, major insurance corporations structure their entire business strategies around probability. Another significant area where probability theory is applied in daily life is in assessing reliability. The ability to comprehend and estimate the likelihood of different outcomes relative to one another holds significant importance in practical, real-world situations.

Italian mathematician Girolamo Cardano initiated the exploration of probability theory in the 16th century. Initially, the focus was on improving the odds in gambling scenarios. In 1650, the French gambler Chevalier de Mere sought advice from mathematician Blaise Pascal in Paris regarding issues arising in certain games of chance, particularly how to distribute stakes when a gambling game was interrupted. This led Pascal to engage in correspondence with his friend Pierre Fermat, a lawyer and amateur mathematician, as they delved into the study of probability.

The formalization of probability is credited to mathematicians Pierre de Fermat (1601 - 1665) and Blaise Pascal, prompted by Chevalier de Méré's request for assistance in navigating gambling scenarios. In the course of addressing his inquiries, they laid the groundwork for the establishment of the laws of probability. In 1933, Russian mathematician Andrey Kolmogorov systematically developed probability theory from foundational axioms, akin to Euclid's treatment of geometry, providing the basis for the modern understanding of probability. Kolmogorov’s comprehensive work, titled ‘’The Foundations of Probability Theory’’, is accessible in an English translation.

Probability theory finds application in a diverse array of fields, including the physical and biological sciences, economics, politics, sports, quality control, and production planning. Its relevance spans from quantum physics to medicine and industry. The inception of queueing theory, integral to the telecommunications industry, dates back to 1909 when Danish engineer Agner Krarup Erlang published the first research paper on the subject while working at the Copenhagen Telephone Exchange. Over the past century, this theory has extended its applicability to areas such as modeling car traffic and analyzing queues in local supermarkets.

The combination of statistics and probability plays a pivotal role in predicting the behavior of the global stock market. In the 1960s, American mathematician Edward Oakley Thorp developed and implemented hedge fund techniques, showcasing the practical use of probability in financial markets. Individual investors make decisions to invest in the stock market based on a favorable probability that the value of shares will rise. However, such investments carry inherent risks, exemplified by historical stock market crashes like the Wall Street crash of 1929 leading to the Great Depression, the Black Monday crash of 1987, and the Global Financial Crisis of 2008-2009.

In the twenty first century, probability plays a crucial role in regulating traffic flow on highways, managing telephone interchanges, and optimizing computer processors. It is instrumental in determining the genetic composition of individuals or populations, understanding the energy states of subatomic particles, estimating the propagation of rumors, and forecasting the returns on risky investments.

Initially, only a select few owned smartphones, but in a remarkably short span, it appears that virtually everyone possesses one. For instance, the surge in smartphone usage demonstrates an exponential growth pattern. Governments utilize probabilistic approaches in environmental regulation, often referred to as pathway analysis. An illustrative instance is the evaluation of how the perceived likelihood of a widespread conflict in the Middle East influences oil prices, with subsequent ripple effects on the broader economy.

We frequently encounter situations involving uncertainty, but not all events share the same level of uncertainty. While it's uncertain whether the next Nuri Ceylan Bilge film will be a blockbuster, or if it will snow in Adana city next winter, these events don't intuitively seem equally probable. To address this, we can place events on a scale ranging from impossible to certain, assigning a number between 0 and 1 to represent their position on this continuum. This numerical value is referred to as the probability of the event.

The language of probability is ingrained in our daily conversations:

- What's the likelihood of rain tomorrow during my commute to school? Should I bring my umbrella?
- How probable do you think it is that you'll pass your driving test on the next try? Should I schedule more driving lessons?
- What's the probability of winning the game this afternoon? Should we start with our best team?
- Can I be certain of reaching school on time if I take the bus instead of walking? Which mode of transportation should I use to ensure punctuality for my crucial exam?

Uncertainty and randomness manifest in various aspects of our lives. Decision-making in realms such as business, investment, agriculture, industry, healthcare, and daily life relies on expectations and predictions. A solid understanding of probability enables us to navigate uncertainties, grasp risks, and make informed decisions about the future.

In reality, predicting future events with absolute certainty is impossible. Assessing the likelihood or probability of events occurring, represented by a numerical value between 0 and 1, is crucial. This number is termed probability. For probability to be valuable, it must extend beyond merely reflecting past experiences. It should also enable us to predict future probabilities and use these predictions to make well-informed decisions.

Probability involves the exploration of randomness and uncertainty. In everyday language, we use the term 'random' to refer to things that are unpredictable. Random events, while not perfectly predictable in the short term, exhibit long-term regularities that can be described and quantified using probability. In contrast, haphazard events lack consistent long-term patterns.

It's crucial to differentiate between random and haphazard (or chaos). While both may seem similar initially because their outcomes cannot be anticipated with certainty, random events demonstrate long-term predictability, whereas haphazard events do not.

Consider the example of tossing an unbiased coin to observe the number of heads. When the coin is thrown, there are only two possible outcomes: heads or tails.

Data are collected by observing either uncontrolled natural events or controlled situations in a laboratory, and the term experiment is used to describe both methods of data collection. Examples of experiments include throwing a coin, rolling a dice and noting the number on the top surface, counting cars at a traffic light when it turns green, measuring daily rainfall in a specific area, and so on.

In the realm of mathematics, a description of a random phenomenon is termed a probability model. For instance, when tossing a coin, the outcome cannot be known in advance, but we can state that the outcome will be either heads or tails. Assuming the coin is balanced, we believe each outcome has a probability of 0.50. This description includes a list of possible outcomes and the probability assigned to each outcome, forming the basis of a probability model.

It's essential to recognize that the proportion of heads in a small number of tosses may deviate from the probability. Probability models describe long-term trends. While the outcome of a single coin toss is unpredictable, in repeated throws, the percentage of times the coin lands on heads tends to stabilize at a limit of 50%. Therefore, although the result of an individual coin toss is unpredictable, the long-term average behavior is foreseeable, justifying the consideration of the outcome of tossing a fair coin as a random event.

1.1. Theoretical and Experimental Probability

Forecasts based on theoretical probability are the most dependable predictions, relying on observable and measurable physical relationships that remain constant. These predictions encompass activities like coin flips, spinners, and number cubes.

Practical estimation of the probability of an event is achieved through repetitive experimentation, known as theoretical (empirical) or experimental probability. For instance, consider a factory producing electrical components, some of which are known to be defective. Determining the likelihood of a randomly chosen component, such as component A, being faulty involves applying the principles of probability.

Probability, as a field of study, facilitates comprehension of the uncertainty associated with events. A probability experiment, or simply an experiment, entails something with an uncertain outcome. It is a systematic process for gathering data about events. In a random experiment, uncertainty exists regarding which of multiple potential outcomes will transpire.

The sample space (denoted as S) in a random experiment represents the set of all possible outcomes. Alternatively, any letter may be used to denote the sample space. For example, for a single coin toss, the sample space is $S = \{heads, tails\}$ or simply $\{h, t\}$.

An event is a collection of outcomes in the sample space possessing a specific characteristic. It is an outcome observed in a single trial of the experiment, such as getting tails in the coin-tossing experiment, rolling a 4 with a six-sided die, selecting a Queen from a deck of cards, or obtaining an odd number when spinning a numbered spinner.

For an experiment involving throwing a die and observing the number on the top face, the simple events are $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$. The set of all these simple events constitutes the sample space, represented as $S = \{1, 2, 3, 4, 5, 6\}$.

The probability of an event occurring can be expressed verbally, but a more useful representation is a numerical value between 0 and 1. On this scale, 0 denotes an impossible event, and 1 signifies an event certain to happen. An impossible event has a 0% chance and is assigned a probability of 0, while a certain event has a 100% chance and is assigned a probability of 1. All other events can be assigned a probability between 0 and 1. The accompanying number line illustrates how different probabilities can be interpreted.

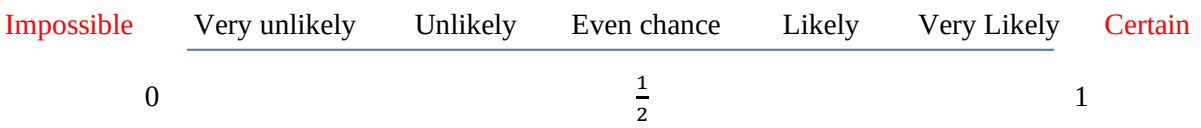


Figure 1: Probability line

The probability of an event is always within the range of 0 to 1 and can be expressed as a decimal, fraction, or percentage. We denote the probability of an event A occurring as $P(A)$, where $0 \leq P(A) \leq 1$.

There are two primary methods for determining the probability of an event:

1. *Theoretical Approach to Probability:* This involves establishing probabilities based on what we anticipate happening, often utilizing arguments of symmetry.
2. *Experimental Approach to Probability:* This entails determining probabilities by observing the outcomes of an experiment.

1.1.1. *Theoretical approach*

Theoretical probability, as the name implies, involves the theoretical aspects of probability, providing outcomes of events based on mathematical principles and reasoning. It offers insights into what should happen in an ideal scenario without the need for actual experiments.

The application of theoretical probability proves particularly valuable in situations where conducting practical experiments is impractical, such as in satellite launches. Predictions, considered reasonable guesses about future occurrences, should ideally be grounded in facts and probability.

In certain situations, it is feasible to predict probability before conducting an experiment, referred to as theoretical probability. To accomplish this, it is essential to list all possible outcomes, known as the sample space. The theory of equally likely outcomes emphasizes symmetry and the indistinguishability of outcomes. If an experiment has n equally likely outcomes with no preference, the probability of each outcome is $\frac{100\%}{n}$ or $\frac{1}{n}$. For instance, in the case of a balanced coin, the probability of it landing heads is equal to it landing tails, both being $\frac{100\%}{2}$ or 50%. Similarly, for a fair six-sided die, the probability of landing on any specific number, like 1, is $\frac{100\%}{6}$ or $\frac{1}{6}$.

Probabilities in the theory of equally likely outcomes range from 0% to 100%. For events with multiple possible outcomes, the chance of the event occurring is determined by the number of ways it can happen divided by the total possible outcomes. For example, the chance of a die showing an even number is the number of even outcomes (2, 4, or 6) divided by the total possible outcomes (1, 2, 3, 4, 5, 6).

Probability, ranging from 0% to 100%, is synonymous with chance. The theoretical approach to probability is rooted in events where the likelihood of each event is known. In the case of a fair six-

sided die, with a sample space of $\{1, 2, 3, 4, 5, 6\}$, each outcome is equally likely, resulting in a theoretical probability of 1 in 6 or $\frac{1}{6}$ for any particular outcome.

Sample spaces, representing all possible outcomes, can be illustrated through lists, 2-dimensional grids, tree diagrams, tables, or Venn diagrams. The notation $n(S)$ indicates the number of elements in the sample space. For example, in rolling a fair six-sided die, $n(S)$ equals 6, while for flipping a fair coin, $n(S)$ is 2. If event A is defined as rolling a 6, then $n(A)$ is 1, indicating one occurrence of the desired outcome in the sample space. The probability of event A occurring is denoted as $P(A)$ and calculated as $\left(\frac{n(A)}{n(S)}\right)$, resulting in $\frac{1}{6}$ for rolling a 6 with a fair six-sided die.

If a sample space consists of n outcomes that are equally likely when the experiment is conducted once, then each individual outcome has a probability of $\frac{1}{n}$ of occurring.

Let's consider the event of spinning a prime number with an octagonal spinner (Figure 2).

Figure 2: Octagonal spinner

Among the 8 possible outcomes, the event corresponds to the numbers 2, 3, 5, and 7. Therefore, the probability of rolling a prime number is 4 in 8, expressed as $\frac{4}{8}$ or simplified as $\frac{1}{2}$, which equals 0.5. The notation $\frac{4}{8}$ is read as ‘’4 chances in 8’’.

When outcomes are equally likely, the probability (P) of an event A occurring is given by the formula:

$$P(A) = \frac{the\ number\ of\ members\ of\ the\ event\ A}{the\ total\ number\ of\ possible\ outcomes} = \frac{\text{number of outcomes corresponding to A}}{\text{number of outcomes in the sample space}} = \frac{n(A)}{n(S)}.$$

In a general sense, the theoretical probability of any event A is expressed as $P(A) = \frac{n(A)}{n(S)}$, where $n(A)$ is the count of ways event A can occur, and $n(S)$ is the total number of possible outcomes.

While it might seem evident that the two definitions (experimental and theoretical) are equivalent, proving this equivalence can be intricate. The law of large numbers is often studied to understand this concept better.

For instance, in a lottery with two possible outcomes (win or not win), it doesn’t imply a probability of one-half for winning, as the outcomes may not be equally likely. Mistakes in probability often stem from this type of misconception.

If a six-sided die is rolled six times, the expectation is to get a 6 once because the probability in each roll is $\frac{1}{6}$. For 12 rolls, the anticipated number of sixes is $12 \times \frac{1}{6} = 2$. However, when considering a smaller number of rolls, like 10, the expected number of sixes would be $10 \times \frac{1}{6} = 1\frac{2}{3}$. While it is not possible to have a fraction of a outcome like $1\frac{2}{3}$ sixes, this concept needs to be reassessed when exploring probability distributions.

1.1.2. *Empirical approach*

Empirical probability, also known as experimental probability, is another method alongside theoretical probability for determining the likelihood of a random event occurring.

In experiments involving chance, the following terms are used to describe the process and results:

- The number of trials represents the total instances the experiment is repeated.
- Outcomes refer to the various possible results for a single trial of the experiment.
- The frequency of a specific outcome is the count of times that particular outcome is observed.
- The relative frequency of an outcome is the frequency expressed as a fraction or percentage of the total number of trials.

For instance, in a scenario where a small plastic cone was thrown into the air 279 times, landing on its side 183 times and on its base 96 times, the terms used are:

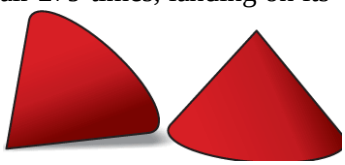


Figure 3: Plastic cone

- The number of trials: 279
- Outcomes: Side and Base (Figure 3: Plastic cone)
- Frequencies of Side and Base: 183 and 96, respectively
- Relative frequencies of Side and Base: 0.656 and 0.344, respectively.

In the absence of additional data, the relative frequency of each event serves as the best estimate of the probability of that event occurring.

Empirical probability is equated to relative frequency, representing the probability derived from experimentation. In frequency theory, probability is viewed as the limit of the relative frequency of an event occurring in repeated trials. Relative frequencies always fall between 0% and 100%. According to the frequency theory of probability, stating "the probability that A occurs" as p% means that with repeated, independent, and essentially identical experiments, the percentage of the time A occurs will approach p. For example, asserting a 50% chance that a coin lands heads means that in multiple, independent throws, the ratio of heads to total throws converges to a limiting value of 50% as the number of throws increases. As this ratio always falls between 0% and 100%, the probability, when it exists, must be within this range.

Relative frequency denotes the ratio of the number of occurrences of an event to the total number of opportunities for it to occur within the same timeframe.

In experimental probability, the likelihood of an event is determined by the outcomes of a series of trials. As the number of trials increases, the probability converges toward the actual probability of the event happening. This method is particularly valuable when outcomes are not equally probable, requiring estimation.

Consider the example of a factory producing various electrical components, some of which may be faulty. To ascertain the probability of a specific component (e.g., component A) being faulty, testing is conducted. Testing all components is impractical, so a process of testing a subset and calculating the relative frequency of faults is employed. If the second component is not faulty, we could conclude the probability of a component being faulty is $\frac{1}{2}$ since (so far) half (one out of two) of all components are faulty.

Continuing this process a number of times and calculating the ratio:

$$\frac{\text{the number of faulty components}}{\text{the number of components tested}} = \text{the relative frequency of the component being faulty}$$

Similar to the dice investigation, as the number of components tested grows, the relative frequency approaches the actual probability of a component being faulty. Relative frequency serves as an estimate of probability, with larger numbers of trials leading to closer alignment between relative frequency and probability.

Applying the concept of equally likely outcomes, the probability of an event (A) is the number of favorable outcomes in A divided by the total number of outcomes in the sample space (S). This relationship is expressed as $P(A) = \frac{n(A)}{n(S)}$, where $n(A)$ represents the number of outcomes in A, and $n(S)$ represents the total number of outcomes.

If we perform an experiment a number of times, then the experimental probability of an event occurring is

$$\text{experimental probability} = \text{relative frequency of event} = \frac{\text{number of times event occurs}}{\text{number of trials}}$$

In the study of expectation, the equation

$$\text{number of times event occurs} = \text{experimental probability} \times \text{number of trials}$$

is utilized. Conversely, if there are n trials, and an event has a probability p of occurring in each trial, then the expected number of occurrences is np.

The distinction between theoretical probability and empirical probability is outlined in the table below:

Theoretical Probability	**Empirical Probability**
Theoretical probability can be defined as the theory behind probability.	Empirical probability or experimental probability is the probability calculated on historic data.
No experiments are conducted to find the theoretical probability. Instead, it depicts what is expected to happen.	Empirical probability is the result of an experiment.
It is predicted by using logical reasoning and having knowledge of the situation.	It is determined by repeating experiments and observing the various outcomes.
The theoretical probability formula is given below: $\text{Theoretical Probability} = \frac{\text{Number of favorable outcomes}}{\text{Number of possible outcomes}}$.	The empirical probability formula is as follows: $\text{Empirical Probability} = \frac{\text{Number of times an event occurs}}{\text{Total number of trials}}$.

Theoretical Probability	Empirical Probability
Example: A fair coin is tossed. Probability of getting a head and a tail, $P(Head) = 1/2 = 0.5$; $P(Tail) = 1/2 = 0.5$.	Example: On tossing a fair coin 15 times it is noted that a head appears 5 times and a tail appears 3 times. $P(Head) = 5/15 = 0.33$, $P(Tail) = 3/15 = 0.2$.

While experimental probability converges to theoretical probability with a large number of trials, studying theoretical probability is essential when conducting numerous real-life experiments is either impractical or costly. In cases like predicting satellite launch failures, assumptions are made, and theoretical probability is calculated based on these assumptions, proving its significance in applications where extensive experimentation is not feasible.

Think about conducting an experiment like flipping a coin or picking a card from a deck. This process can be repeated numerous times to enhance accuracy. However, in scenarios such as determining the probability of a satellite launch failure, conducting multiple launch experiments isn't viable or practical. In these instances, it becomes essential to establish specific assumptions. Theoretical probability is then calculated based on these assumptions, proving valuable in situations where actual experimentation is unfeasible, particularly in various applications.

1.2. Probability and Set Theory

Set theory serves as a fundamental framework for the entire field of mathematics. The language employed in probability closely aligns with that of set theory, allowing logical statements to be construed as assertions about sets. This correspondence facilitates the introduction of a systematic approach to framing probability problems.

Considering that an event represents an outcome or a set of outcomes from a random experiment, it is also conceivable to view the event as a subset of the sample space or as an amalgamation of individual outcomes. Utilizing certain tools from set theory, it can be beneficial to visualize an experiment. The connections between set theory and probability are notably striking, and recognizing this relationship proves advantageous. A particularly useful visual aid in this context is the Venn diagram. Consequently, outcomes in the sample space and events can be articulated using set notation and depicted with the assistance of a Venn diagram.

1.2.1. *Drawing Venn diagrams, complements, intersections and unions*

To establish an axiomatic probability system, a grasp of fundamental set theory and probability rules is crucial.

When dealing with information about the number of individuals in intersecting groups, Venn diagrams serve as a valuable tool for visualizing the data. Beyond depicting frequencies, Venn diagrams can also represent probabilities.

It is customary to assign the unique number or probability to each region within the diagram. This is often achieved by labeling the intersection of all groups with an unknown and then proceeding outward. The total for regions joined together is not labeled. To draw a Venn diagram, begin by

outlining a rectangle that encompasses all the items under consideration. This rectangle is referred to as the universal set, denoted by S, as it contains every item.

For instance, if there are 100 students in a year group, you denote $n(S) = 100$, and the rectangle symbolizes these 100 students. Within the year group, if 36 students study art, the probability that a randomly chosen student studies art is represented as $P(A)$, with $P(A) = \frac{36}{100} = \frac{9}{25}$.

The region outside of set A (but still within the sample space S) signifies the complement of set A. This is everything other than the event happening. For example, the complement of rolling a 6 on a die is rolling not a 6, so, 1, 2, 3, 4, or 5. Two events are complementary if exactly one of them must occur. If A is an event, then A' is the complementary event of A, or ''not A''.

The law of the excluded middle asserts that an event either happens or it does not, serving as a fundamental axiom of standard logic. However, fuzzy logic, an alternative logical system, allows for events to be in a state of ''maybe happening''. This concept finds applications in various real-world scenarios and is also philosophically illustrated in Schrödinger's Cat, a physics problem that explores the idea of uncertainty.

This represents A', the complement of set A, indicating students who don't study art. Thus, $n(A') = n(S) - n(A)$. In this case, $n(A') = 100 - 36 = 64$.

To find the probability that a randomly chosen student doesn't study art, we need $P(A')$, which is calculated as $P(A') = \frac{n(A')}{n(S)} = \frac{64}{100} = \frac{16}{25}$.

Observing $P(A) + P(A')$, we note that $P(A') + P(A) = \frac{9}{25} + \frac{16}{25} = 1$. In general, for an event A and its complement A', $P(A') + P(A) = 1$, and $P(A') = 1 - P(A)$.

In everyday language, the term ''or'' can be unclear. Stating, ''Everyone at the party is a student or a teacher'', typically doesn't imply that each person could be both. However, in a game scenario, saying, ''I win if I get a black number or an even number'', suggests winning with a black even number. In probability, we employ ''or'' in this latter sense, where ''A or B'' signifies that A, B, or both events could occur. The reason mathematicians prefer precise language is that everyday expressions often lack clarity. For instance, saying I play football or basketball might be misconstrued by some to mean I don't play both, introducing ambiguity that mathematicians strive to avoid.

We now delve into a more detailed examination of compound events involving multiple occurrences within our sample space. This complexity may arise due to multiple processes within the experiment or the interest in various properties of the outcome.

Now, considering compound events with two events A and B, denoted as $A \cap B$ ('A intersection B''), if 40 students study biology and 12 study both art and biology, we can represent this on a Venn diagram by adding region B.

The intersection of A and B, $A \cap B$, includes outcomes common to both, and since there were $100 - 24 - 28 - 12 = 36$ students not studying art or biology, these go outside the circles.

For example, the probability a student chosen at random studies both art and biology, \(P(A \cap B)\), is calculated as $P(A \cap B) = \frac{\mathrm{n}(A \cap B)}{n(S)} = \frac{12}{100} = \frac{3}{25}$.

The probability a student chosen at random doesn't study biology but studies art, $P(A \cap B')$, is $P(A \cap B') = \frac{\mathrm{n}(A \cap B')}{n(S)} = \frac{24}{100} = \frac{6}{25}$.

$A' \cap B'$ represents the intersection of students not studying art and not studying biology, where $n(A' \cap B') = 36$.

When the combined probability of events A or B is not simply the sum of their individual probabilities, we introduce the concept of finding the probability of A or B when they are not mutually exclusive. To address this, we define another operation.

The union of two events, A and B, denoted by the symbol $A \cup B$, represents the event containing all outcomes belonging to A, or B, or both. The occurrence of either A or B or both is expressed as $A \cup B$ and read as ''A union B''.

For instance, the probability that a randomly chosen student studies either art or biology is denoted as $P(A \cup B)$. In a Venn diagram, the shaded area corresponding to $P(A \cup B)$ is determined as $n(A \cup B) = 24 + 12 + 28 = 64$.. Thus, $P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{64}{100} = \frac{16}{25}$.

Now, let's consider $A \cup B'$, which represents all students studying art or not studying biology. The count is given by $n(A \cup B') = 24 + 12 + 36 = 72$, resulting in $P(A \cup B') = \frac{n(A \cup B')}{n(S)} = \frac{72}{100} = \frac{18}{25}$.

Several properties characterize set operations, including basic consequences of the definitions. Examples include

$$A \cup B = B \cup A,$$
$$(A')' = A,$$
$$A \cap S = A,$$
$$A \cup S = S,$$
$$A \cap A' = \emptyset,$$
$$A \cup A' = S,$$

where $\emptyset$ is the empty set.

In mathematics, ''and'' corresponds to union $\cup$, ''or'' corresponds to intersection $\cap$, and these operations have specific properties. For instance, De Morgan's laws state

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

and

$$(A \cap B)' = A' \cup B'.$$

Finally, the distributive laws are expressed as

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

These concepts help in understanding and calculating probabilities when events are not mutually exclusive.

1.2.2. *Probability rules*

Irrespective of the theory to which we adhere, the principles of probability remain applicable.

Rule 1:

Probability, denoted as $P(A)$, is always a number between 0 and 1. This means $0 \leq P(A) \leq 1$. If the probability is 0, the event never occurs, and if it's 1, the event always occurs. For example, when rolling a standard die, getting the number 9 is impossible ($P(9) = 0$), and the probability of getting any integer between 1 and 6 is 1.

Even if you believe an event has a minimal likelihood, negative probability does not exist.

Rule 2:

The sum of the probabilities for all possible outcomes must equal 1 because these outcomes are mutually exclusive, collectively covering the entire sample space. For instance, if someone provides the probabilities for randomly selecting a grade 1, 2, 3, or 4 student in a high school as 0.24, 0.24, 0.25, and 0.19, respectively, the total exceeds 1 (0.92). Similarly, if the reported probabilities are 0.24, 0.28, 0.25, and 0.26, the total surpasses 1 (1.03).

Regardless of how likely an event may seem, it is essential to note that a probability cannot exceed 1.

Rule 3:

The addition rule asserts that for any two events, A and B, the probability of their union, $P(A \cup B)$, can be calculated using the formula $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In simpler terms, this equation signifies that the probability of either A or B or both occurring is equal to the sum of the probabilities of A and B, minus the probability of both A and B.

In this context, both the individual probabilities $P(A)$ and $P(B)$ encompass the probability of the joint occurrence of A and B, denoted as $P(A \cap B)$. To avoid double counting the joint probability in the calculation of $P(A \cup B)$, one instance of this joint probability is subtracted.

An alternative interpretation of this formula is: ''To determine the number of ways of obtaining either A or B or both, count the ways of obtaining A and add to that the ways of obtaining B. However, since we have counted the ways of obtaining both A and B twice, we need to compensate by subtracting it''.

This probability addition rule extends to the scenario of mutually exclusive events. For two mutually exclusive events A and B, where $P(A \cap B) = 0$, the addition law simplifies to $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$. This simplification leads to Rule 4 in probability calculations.

Rule 4:

If two events share no common outcomes, the probability of either event occurring is the sum of their individual probabilities. Events with no shared outcomes, and therefore cannot happen simultaneously, are termed disjoint events or mutually exclusive events. Mutually exclusive events are those whose outcomes cannot co-occur.

For instance, among 100 students, 28 are found to study chemistry. Since it is impossible to study both chemistry and art simultaneously due to conflicting class schedules, the events A (studying art) and C (studying chemistry) are considered mutually exclusive.

In general, for two mutually exclusive events A and B, it follows that the probability of their intersection, $P(A \cap B)$, is 0. The addition rule for mutually exclusive events is expressed as $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$.

As an illustration, consider the scenario of tossing three coins. The events of obtaining exactly two heads or exactly two tails are disjoint, resulting in the probability of getting exactly two heads or two tails being calculated as $\frac{3}{8}+\frac{3}{8}=\frac{6}{8}=\frac{3}{4}$.

Furthermore, the addition of probabilities is always permissible because outcomes are inherently disjoint. A trial cannot yield two different outcomes simultaneously. This provides a means to verify the legitimacy of assigned probabilities.

In Venn diagrams, mutually exclusive events do not overlap.

Rule 5:

If the probability of receiving a 7 on your exam is 0.2, then the probability of not receiving a 7 on the exam is the complement of that event and is denoted by A'. The probability of A' (not receiving a 7) is calculated as $P(A')=1-P(A)$, where $P(A)$ is the probability of receiving a 7. Alternatively, the probability of receiving a 7 ($P(A)$) is expressed as $1-P(A')$.

In other words, the probability of an event and the probability of its complement add up to 1, as they cover all possible outcomes: $P(A)+P(A')=1$. This relationship is fundamental in probability theory and provides a useful way to calculate one probability when the other is known.

1.2.3. *Equally likely outcomes*

In certain situations, we can assume that individual outcomes are equally probable due to the experiment's balance. For example, tossing a balanced coin results in heads or tails being equally likely, each with a 50% probability. Similarly, rolling a standard balanced die yields numbers 1 to 6 as equally likely, each with a probability of $\frac{1}{6}$.

Assuming all digits are equally likely, the probabilities are depicted in Table 1:

First digit	0	1	2	3	4	5	6	7	8	9
Probability	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 1: All digits are equally likely to happen

$$P(A) = 0.1$$
$$P(B) = P(6) + P(7) + P(8) + P(9) = 4 \times 0.1 = 0.4$$
$$P(C) = P(1) + P(3) + P(5) + P(7) + P(9) = 5 \times 0.1 = 0.5$$

Additionally, according to the complement rule, the probability that the first digit is not 1 is given by:

$$P(A') = 1 - P(A) = 1 - 0.1 = 0.9.$$

1.2.4. *Tree diagrams, tables, and grids*

Venn diagrams alone don't provide formulas for intersection or union. Another method, tree diagrams, is valuable for this purpose. When multiple events occur, either sequentially or simultaneously, a tree diagram lists all potential outcomes. It initiates with branches for all possible outcomes of one event, and from each branch, it enumerates outcomes for the next event. Probabilities for each branch are recorded along the way.

Probability situations may involve either with replacement or without replacement, and there exists a distinction in the calculated probabilities for each scenario. The creation of a list detailing possible

outcomes and associated probabilities may not always be the most efficient way to represent probabilities. Probability tree diagrams offer a visual representation for situations where multiple events occur simultaneously. Before revisiting the earlier problem, let's explore two additional examples.

Consider a scenario with a large group of objects. If we randomly select one object and examine it for specific features, this process is termed sampling. If the object is returned to the group before making another selection, it is known as sampling with replacement. On the other hand, if the chosen object is set aside, it is termed sampling without replacement.

Sampling is frequently employed in industrial process quality control, where certain inspection processes prevent the return of the object to the larger group. Examples include determining the hardness of a chocolate, checking the yolk content of an egg, or assessing the correctness of an object's construction, which may involve disassembly.

Let's consider a box containing 3 red, 2 blue, and 1 yellow marble. When sampling two marbles, we can do so either with replacement or without replacement before the second draw. Notably:

- with replacement (independent events), $P(two\ reds) = \frac{3}{6} \times \frac{3}{6} = \frac{1}{4}$,
- without replacement (dependent events), $P(two\ reds) = \frac{3}{6} \times \frac{2}{5} = \frac{1}{5}$.

Tree diagrams offer an advantage over the sample space method, handling more than two events and scenarios where outcomes aren't equally likely.

In cases involving multiple operations, the sample space can still be listed, but illustrating it on a 2-dimensional grid or using a tree diagram is often more efficient. For example, a tree diagram can illustrate possible outcomes when spinning a spinner three times. To highlight the event of obtaining blue twice, represented as B for blue and W for white, outcomes include $BBB, BEW, BWB, BWW, WEB, WBW, WWB, WWW$.

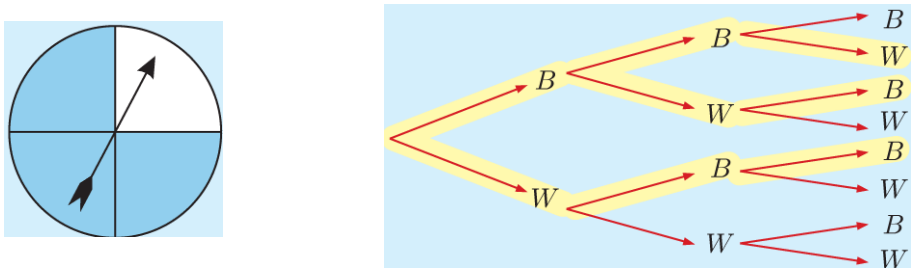


Figure 4: Spinner and outcomes

In a study to examine patients' blood types, the process can be conceptualized as a two-stage experiment. Initially, we determine the blood type, followed by classifying the Rh factor as either positive or negative. The individual outcomes in this experiment can be enumerated using a tree diagram, which proves to be highly effective and beneficial for resolving probability-related issues. The sample space for this experiment encompasses the set {A+, A-, B+, B-, AB+, AB-, O+, O-}, and these outcomes can be derived from the results listed in the last column. Alternatively, the same individual events can be organized in a probability table, as depicted in Table 2.

	Blood type			
Rh factor	A	B	AB	O
Positive	A+	B+	AB+	O+
Negative	A-	B-	AB-	O-

Table 2: Blood types probability table

Two-way tables, also known as contingency tables, are tabulations that juxtapose two categorical variables. For instance, individuals leaving a recently introduced attraction at a theme park were surveyed about their preference for or against the ride. The gathered responses are condensed in the provided table. In this illustration, we will explore how these tables can be employed to approximate probabilities. To facilitate this analysis, we enhance the table by incorporating totals for each row and column.

	Child	**Adult**
Liked the ride	55	28
Disliked the ride	17	30

Table 3: Responses for new ride

Utilize this table to approximate the likelihood that a person chosen at random who experienced the ride ‘’liked the ride, is a child and disliked the ride, is an adult or disliked the ride’’.

We expand the table to incorporate overall counts for each row and column.

	Child	**Adult**	**Total**
Liked the ride	55	28	83
Disliked the ride	17	30	47
Total	72	58	130

Table 4: Responses of children and adults for new ride

83 out of the 130 people surveyed liked the ride.

$$P(liked\ the\ ride) = \frac{83}{130} = 0.638$$

17 of the 130 people surveyed are children who disliked the ride.

$$P(child\ and\ disliked\ the\ ride) = \frac{17}{130} = 0.131$$

$28 + 30 + 17 = 75$ of the 130 people are adults or people who disliked the ride.

$$P(adults\ or\ disliked\ the\ ride) = \frac{75}{130} = 0.577$$

Two-dimensional grids are valuable tools for visualizing two-stage or sequential probability models. For instance, when rolling a fair six-sided die twice, a two-dimensional grid and the count of possible events are used to calculate probabilities. For example:

- Probability of at least one roll showing a 6: Count the points in the column corresponding to 6 on the first roll and the row corresponding to 6 on the second roll, avoiding double-counting the corner point.
- Probability of the numbers on both rolls being the same: Count the points on the diagonal.
- Probability that the first roll shows a number larger than the second roll: Pick the points below the diagonal. Thus, $P(first\ number > second\ number) = \frac{15}{36}$.

Two-dimensional grids serve as effective visual representations of sample spaces, allowing us to visually assess favorable outcomes and subsequently calculate probabilities.

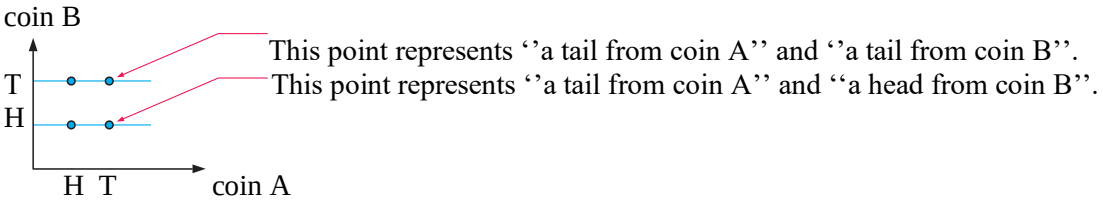


Figure 5: Two-dimensional grids

There are four members of the sample space.

1.3. Geometric Probability

Geometric probability within statistics pertains to situations where there is an equal likelihood of success and failure, the outcomes are independent, and there is no specified maximum number of trials for conducting the experiment.

In probability problem-solving, scenarios may arise where the number of possible outcomes is infinite. In such instances, it proves beneficial to convert the probability problem into a geometric one, wherein the notion of infinite intervals can be visually represented through lines, segments, and enclosed figures. Geometric probability, by definition, is a method that illustrates the concept of infinite intervals within a measurable figure of length, area, or volume. This approach can be applied in one-dimensional, two-dimensional, or three-dimensional contexts. The key point is that when it is established that the population space of the experiment involves an infinite number of outcomes or when dealing with continuous data, geometric probability offers a means to determine those probabilities.

Consider a scenario where events are construed as lengths. For example, if Seçkin needs to attend a business meeting scheduled between 5 pm and 6 pm and he arrives at 5:45 pm, what is the probability that he will arrive on time? Time, being a form of continuous data with infinite possibilities, accounts for minutes, seconds, and even microseconds. To address this, fractions of time are considered in the potential outcomes. This problem can be represented using a number line, as illustrated below:

Figure 6: Line represents the time intervals

In the Figure 6, the crimson portion of the segment denotes the unfavorable time, encompassing all moments before 5:45 pm, while the green segment represents the accepted time range between 5:45 pm and 6 pm. Dividing these segments into equal intervals offers a more visually comprehensible representation of probability. Assuming each segment has a unit length of 1, making the entire segment 4 units, and the desired outcome interval has a length of 1, the geometric probability of Seçkin arriving on time is calculated as follows:

$$Probability = \frac{desired\ lengt}{total\ lengt} = \frac{1}{4} = 0.25.$$

Therefore, the probability of Jacob arriving at the meeting on time is 25%.

In scenarios where events are interpreted as areas in the plane, consider randomly shooting at a circular target. The probability of hitting the central part is given by $P = (\pi(R/4)\text{^}2)/(\pi R\text{^}2) = 1/16$. This calculation is derived from the area of the small circle over the area of the entire target.

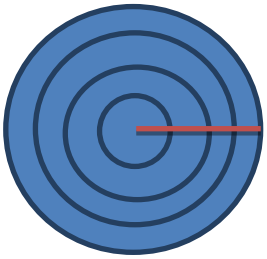


Figure 7: Circular target (red line=R)

Beren and Eda plan to meet at the museum between 12:00 and 13:00. The first person to arrive will wait 15 minutes, and if the second person doesn't show up, the first person will leave. Assuming random arrival times, the probability of them meeting is determined by the conditions $|x - y| \leq 15$ and $x \leq 60, y \leq 60$, where x and y represent Lydia and Rania's arrival times in minutes after 12:00, respectively. Geometrically, the central region in the diagram illustrates the arrival times that facilitate their meeting.

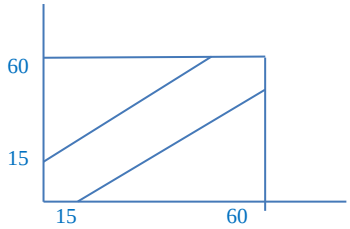


Figure 8: Area representation

The area for each triangle is $\frac{1}{2}bh = \frac{1}{2}(45)^2$ making the center area 60^2- 45^2. Therefore, the probability they meet is $\frac{60^2 - 45^2}{60^2} = \frac{7}{16}$.

1.4. Probability Calculation for Equally Likely Outcomes Using Counting Principles

Counting principles can provide a method for calculating probabilities in problems that would be particularly challenging through other means. When dealing with a random arrangement of objects, a sample space is formed, allowing the calculation of the probability of a specific condition A. In experiments where all outcomes are equally likely, the theoretical probability of an event A is expressed as $P(A) = \frac{n(A)}{n(S)}$, where $n(A)$ represents the number of outcomes comprising event A, and $n(S)$ is the total number of outcomes in the sample space. The focus here is on the computation of $n(A)$ and $n(S)$, and such calculations involve counting principles.

A combination of r objects selected from n objects constitutes a subset of the set of n objects. For instance, consider the letters $ABCDE$. There are 10 combinations of 3 letters chosen from these 5: $ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE$. The general rule states that the number of combinations or subsets of r objects out of n objects is given by the binomial coefficient:

$$^nC_r = \binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

Hence, the number of ways to choose 3 letters from $ABCDE$ is:

$$^5C_3 = \binom{5}{3} = \frac{5!}{2!\,3!} = 10$$

It's worth noting that $^8C_5 = {}^8C_3$.

1.5. Dependent and Independent Events and Conditional Probability

1.5.1. *Dependent events*

Consider a scenario where a hat contains 5 red and 3 blue tickets. A ticket is randomly chosen, its color is noted, and then set aside without being returned to the hat. Subsequently, a second ticket is randomly selected. What is the probability that the second ticket is red?

If the first ticket was red, then $P(second\ is\ red) = \frac{4}{7} = \frac{4 \text{ reds remaining}}{7 \text{ to choose from}}$.

If the first ticket was blue, $P(second\ is\ red) = \frac{5}{7} = \frac{5 \text{ reds remaining}}{7 \text{ to choose from}}$.

The probability of the second ticket being red is contingent on the color of the first ticket. Hence, we are dealing with dependent events.

Events are considered dependent when the occurrence of one event affects the occurrence of other events. This dependency arises when events are not independent. Some straightforward examples of dependent events include

- getting into a traffic accident, which depends on driving or riding in a vehicle,
- receiving a parking ticket, which is more likely if you park your vehicle illegally,
- winning a lottery, where the odds increase with the number of tickets purchased.

For dependent events A and B, the probability $P(A \cap B)$ is determined by $P(A) \times P(B\ given\ that\ A\ has\ occurred)$.

In general, when an experiment involves sampling:

- without replacement, we have dependent events,
- with replacement, we have independent events.

It’s important to note that not all scenarios involving dependent events are related to sampling. For instance, the event of a student walking to school today may be dependent on the event of rain occurring today.

1.5.2. *Independence and independent events*

In the mid-1600s, mathematicians Blaise Pascal, Pierre de Fermat, and Antoine Gombaud grappled with a simple gambling problem: determining the likelihood of rolling at least one six on four throws of one die compared to rolling at least one double six on 24 throws with two dice.

While we can assess the intersection and union of two events A and B if we have knowledge of $P(A\ given\ that\ B\ has\ occurred)$ or $P(B\ given\ that\ A\ has\ occurred)$, finding this information can be challenging. However, there is a significant exception when the two events are independent,

meaning they do not influence each other. Events A and B are considered independent if the occurrence of one event does not impact the likelihood of the other occurring.

Examples of independent events include

- owning a dog and having an aunt named Nalan
- taking a cab home and finding your favorite movie on cable.

In such cases, knowing that event B has occurred does not affect the probability of event A occurring, and vice versa.

For independent events A and B, the product rule or multiplication rule for independent events is as follows:

$$P(A \cap B) = P(A \; given \; that \; B \; has \; occurred) \times P(B) = P(A) \times P(B)$$

This rule serves as a useful tool for checking the independence of events. Two events are independent $\Leftrightarrow P(A \cap B) = P(A) \times P(B)$ or $P(A \; given \; that \; B \; has \; occurred) = P(A)$ ($\Leftrightarrow$ means ‘’if and only if’’). If this condition is not met, the events are considered dependent.

Consider a scenario in a large school where 55% of students are male, and it’s known that the percentage of cyclists among males and females is the same at 22%. The probability of selecting a male cyclist when randomly choosing a student from this population can be calculated using the multiplication rule for independent events. Since the proportion of cyclists is the same in both groups (males and females), cycling and gender are independent. Therefore, the chance of picking a male student is 55%, and from that 55%, we know that 22% are cyclists. Thus, the probability of selecting a male cyclist is calculated as $0.22 \times 0.55 = 12.1\%$.

For independent events, $P(A|B)$ = P(A), and it follows that:

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

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

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

Some useful results related to independent events:

1. $P(A) = P(A \cap B) + P(A \cap B')$
2. $P(B) = P(A \cap B) + P(A' \cap B)$
3. $P(A' \cap B') = 1 - P(A \cup B)$

It is crucial not to confuse independence with disjoint events. Disjoint events mean that if one event occurs, the other does not, while independent events mean that knowledge of one event does not influence the occurrence of the other.

1.5.3. *Conditional probability*

In probability, conditioning involves introducing new constraints on the outcomes of an experiment, adjusting probabilities to consider updated information.

Conditional probability stands in contrast to unconditional probability, which denotes the likelihood of an event occurring without regard to whether other events have transpired or any additional conditions

are present. Unconditional probability, also referred to as marginal probability, quantifies the chance of an event happening without incorporating information from preceding or external events. As this probability remains unchanged by new information, it is considered constant.

Consider estimating the probability that a randomly selected individual is a millionaire. Would this estimate change if you were informed that they live in a mansion? The inclusion of additional information leads to changes in probabilities. In this case, $P(millionaire)$ significantly differs from $P(millionaire|lives\ in\ a\ mansion)$..

One effective approach for determining conditional probabilities involves restricting the sample space. Initially, we enumerate all equally likely possibilities before receiving any information, and subsequently eliminate possibilities that the provided information rules out.

Venn diagrams are particularly helpful when contemplating conditional probability, allowing us to exclude irrelevant portions of the diagram based on the given information.

Key points about conditional probability include:

- Conditional probability pertains to the likelihood of one outcome occurring given that another event has also transpired.
- It is often expressed as the probability of B given A, denoted as $P(B|A)$., where the probability of B is contingent on the occurrence of A.
- Conditional probability is distinct from unconditional probability.
- Probabilities are categorized as conditional, marginal, or joint.
- Bayes' theorem is a mathematical formula used for calculating conditional probability.

To compute the probability that an event A occurs given that another event B has occurred, we use the formula:

$$P(A\ given\ that\ B\ has\ occurred) = \frac{number\ of\ \text{events that both } A \text{ and } B \text{ occur}}{number\ of\ \text{events that } B \text{ occur}}$$

For events A and B, denoted as $A|B$, the conditional probability is expressed as:

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

When outcomes in each event are equally likely, the formula can be simplified to:

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

Thus, the conditional probability formula is given by:

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

The conditional probability of event A given event B, denoted as $P(A|B)$, represents the likelihood of A occurring with the knowledge that B has taken place. If A is an event with probability $P(A) = p \neq 0$, and $A \cap B = \emptyset$ (A and B are disjoint), learning that B occurred implies that A did not occur,

resulting in a revised probability of A being zero ($P(A|B) = 0$). Conversely, if $A \cap B = B$ (implying B is a subset of A), learning that B occurred implies that A must have occurred, leading to a revised probability of A being 100% ($P(A|B) = 1$).

This concept underscores the idea that the probability assigned to an event can change when information about another event is known, forming the basis of understanding conditional probability.

Consider a scenario where you are playing cards, and your opponent is about to give you a card. Initially, the probability of receiving a queen from a well-shuffled deck of 52 cards $P(queen) = 4/52 = 1/13$. However, if you already have one queen in your hand among five cards, the probability of receiving another queen, given this information, becomes $P(queen|1\ queen\ in\ hand) = 3/47 \neq \frac{1}{13}$. Thus, knowing about the presence of one queen in your hand alters the probability of the next card being a queen.

Applying this idea to a real-life example, let's consider a student applying for admission to a university with hopes of receiving an academic scholarship. The school accepts 10% of every 1000 applicants and awards scholarships to 0.2% of the 500 students accepted. If 50% of scholarship recipients also receive stipends, the chance of a student being accepted, receiving a scholarship, and getting a stipend is calculated as 0.1% (0.1x0.02x0.5).

In interpreting questions, distinguishing between conditional probability and combined probability is crucial. For instance, if asked about the probability that a boy with green eyes also has brown hair, it might be tempting to find $P(green\ eye \cap brown\ hair)$. However, since having green eyes is already given, the correct approach is to find $P(green\ eye|brown\ hair)$.

Key Probability Definitions:

Conditional Probability, $P(A|B)$: The probability of event A occurring given that event B has occurred. For example, given that you drew a red card, what's the probability that it's a four ($P(four|red) = 2/26 = 1/13$). So out of the 26 red cards (given a red card), there are two fours so $1/13$

Marginal Probability: The probability of an event occurring in isolation ($P(A)$), considered as unconditional probability. It is not conditioned on another event. Example: the probability that a card drawn is red ($P(red) = 0.5$). Another example, the probability that a card drawn is a 4 ($P(four) = 1/13$).

Joint Probability, $P(A \cap B)$: The probability of both events A and B occurring simultaneously, denoted as the intersection of the events. The probability of the intersection of A and B may be written $P(A \cap B)$. Example, the probability that a card is a four and red $P(four\ and\ red) = 2/52 = 1/26$.

1.5.4. *Bayes' theorem*

Bayes' theorem, named after the 18th-century British mathematician Thomas Bayes, is a mathematical formula used to determine conditional probability. This theorem offers a mechanism to update existing predictions or theories by incorporating new evidence. In the realm of finance, Bayes' theorem finds application in assessing the risk associated with lending money to potential borrowers. It is also referred to as Bayes' Rule or Bayes' Law and serves as the cornerstone of Bayesian statistics. This set of probability rules enables the adjustment of predictions based on received information, leading to more accurate and dynamic estimates.

For instance, consider a test for a rare medical disease that boasts 99% accuracy, meaning $P(positive\ result|you\ have\ the\ disease) = 0.99$ and $P(negative\ result|you\ don'thave\ the\ disease) = 0.99$. If a positive result is obtained, the common intuition might suggest a 99% likelihood of having the disease. However, this might not be accurate. Bayes' theorem addresses the probability of interest, namely $P(you\ have\ the\ disease|positive\ result)$, by relating it to $P(positive\ result|you\ have\ the\ disease)$.

Starting with the formula for conditional probability $P(A|B) = \frac{P(A\cap B)}{P(B)}$, Bayes' theorem emerges when replacing $P(A \cap B)$ with $P(A)\times P(B|A)$:

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

The denominator, $P(B)$, is often challenging to calculate and can be determined using a tree diagram, accounting for scenarios where B occurs after A has happened or after A has not happened. This leads to the expression:

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

Substituting this back into Bayes' theorem, the formula becomes:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(A)P(B|A) + P(A')P(B|A')}$$

Applying Bayes' theorem to the disease detection example, where $A = positive\ result$ and $B = you\ have\ the\ disease$, then the required probability is $P(B|A)$ and the information given tells us that $P(A|B) = 0.99$ and $P(A'|B') = 0.99$, implying that $P(A|B') = 0.01$.

Hence:

$$P(you\ have\ the\ disease|positive\ result) = \frac{0.99P(B)}{0.99P(B) + 0.01(1 - P(B))}.$$

The calculation depends on the incidence of the disease in the population denoted by $P(B)$. If $P(B)$ is small, the resulting probability of having the disease given a positive result can vary significantly.

For example, if $P(B) = \frac{1}{100}$ (so 1% of the population have the disease), the required probability is 50%, but if $P(B) = \frac{1}{1000}$ it is under 1%. We can make sense of this by arguing 'the disease is so rare that a positive test is far more likely to be a faulty test result of a healthy person than an accurate result of a diseased person.

Bayes' theorem can be extended to situations with more than two outcomes, denoted as B_1, B_2 and B_3. The total probability of A occurring is expressed as:

$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)$.

This leads to the second version of Bayes' theorem:

$$P(B_i|A) = \frac{P(B_i) \times P(A|B_i)}{P(B_1) \times P(A|B_1) + P(B_2) \times P(A|B_2) + \cdots + P(B_n) \times P(A|B_n)}.$$

It is crucial to note that outcomes B_1, B_2 up to B_n and including B_n must encompass all possible outcomes.

1.6. Random Variables

In statistics, the focus often shifts to determining the mean or standard deviation of collected data. Nevertheless, in practical situations, the ability to forecast these quantities in advance proves valuable. While predicting the outcome of a single event in a random scenario, such as the result of one die roll, is impossible, it's interesting to note that with a sufficient number of events, the average can be predicted with considerable accuracy.

Due to the inherent randomness in these variables, exact predictions for their values in the next measurement are challenging. However, it's frequently feasible to identify potential values the variable might assume and assign a probability to each possible value.

In previous probability chapter, events were often described using words. Nonetheless, when feasible, numerical representations are more convenient. This chapter extends the probability concepts to model the random variation or distribution of numerical variables.

Variables, defined as characteristics that vary over time or for different objects in probability, will have numerical values that change based on the outcomes of the experiment. For instance, if we are counting the number of smartphones owned by families in a city, the variable of interest, X, can assume values like 0, 1, 2, 3, and so on, contingent upon the random experiment's outcome. This classification as a random variable emphasizes the variability in its values.

In probability experiments, the primary interest often lies in the numerical quantity derived from the outcomes rather than the detailed outcome specifics. Consider tossing two dice in games, where the focus might be on their sum rather than the individual values on each die. A sample space, where each point is equally likely, can be represented in Table 5, consisting of 36 ordered pairs (a, b), with ''a''denoting the number on the first die and ''b'' representing the number on the second die. In this context, the random variable X can signify the sum of the numbers, and the resulting values of X are presented in the table.

(1. l); x=2	*(2, l); x=3*	*(3, l); x=4*	*(4, l); x=5*	*(5, l); x=6*	*(6, l); x=7*
(1, 2); x=3	*(2, 2); x=4*	*(3, 2); x=5*	*(4, 2); x=6*	*(5, 2); x=7*	*(6, 2); x=8*
(1, 3); x=4	*(2, 3); x=5*	*(3, 3); x=6*	*(4, 3); x=7*	*(5, 3); x=8*	*(6, 3); x=9*
(1. 4); x=5	*(2, 4); x=6*	*(3, 4); x=7*	*(4, 4); x=8*	*(5, 4); x=9*	*(6, 4); x=10*
(1. 5); x=6	*(2, 5); x=7*	*(3, 5); x=8*	*(4, 5); x=9*	*(5, 5); x=1 0*	*(6, 5); x=11*
(1, 6); x=7	*(2, 6); x=8*	*(3, 6); x=9*	*(4, 6); x=10*	*(5, 6); x=11*	*(6, 6); x=12*

Table 5: Sample space and the values of the random variable X in the two-dice experiment

Events can be more precisely and succinctly defined using the random variable X. For instance, the event of obtaining a sum greater than or equal to 5 but less than 9 can be replaced by the inequality $5 \leq x < 9$.

Between 2008 and 2010, Paul the Octopus accurately predicted the outcomes of 12 out of 14 football matches by choosing one of two boxes marked with the flags of the national teams playing. This results in an impressive success rate of nearly 86%. To assess the credibility of this success, it's crucial to determine the probability of such an outcome occurring purely by chance.

Paul is not the only renowned predicting animal; Punxsutawney Phil, a groundhog featured in the film Groundhog Day, has been forecasting the weather for more than 130 years. The legend is that if he sees his shadow on February 2, six more weeks of winter are expected; no shadow indicates an early spring. Phil has a 39% success rate. Evaluating whether this success rate is low requires understanding the expected success rate. This chapter explores scenarios like these, aiming to ascertain the probability of events that are essentially outcomes of pure chance.

A random variable is a quantity whose value is contingent upon chance, such as the outcome when rolling a die. If the probabilities associated with each possible value are known, mathematical calculations become feasible. Represented by a capital letter like X, a random variable has measurable values denoted by lowercase letters. For instance, if X represents the numbers resulting from a dice throw, $x = 2$ represents the outcome of rolling a 2.

Random variables represent in numerical form the potential outcomes of a random experiment. Various examples include:

- The number of calls received by a household on a Sunday night.
- The number of available beds in hotels in a small city.
- The number of customers a salesperson contacts in a working day.
- The length of a plastic bar produced by a specific machine.
- The weight of newborn babies in a university hospital.

Random variables can be either discrete or continuous. A random variable is considered discrete if its possible values are isolated points on the number line, indicating a countable number of potential values. For example, consider the number of coin flips until the head side appears, with possible values $x = 1, 2, 3$, and so on. Despite the potential infinity of values, this variable remains discrete.

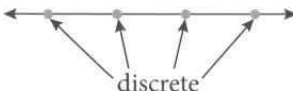


Figure 9: discrete variables

A discrete variable is characterized by having distinct and separate values, typically integer values. For instance, in a class of n students, a variable might represent the number of left-handed students, taking values like $0, n$, or any whole number in between. However, it cannot take non-integer values like 1.5, 21.28, or negative numbers.

The specific number of left-handed students in a class of 25 can differ from one class to another, making it a variable.

A discrete random variable X encompasses a set of distinct and specific possible values. For example, X could represent:

- The number of wickets a bowler takes in a cricket innings, with values 0, 1, 2, and so on, up to 10.

- The number of defective light bulbs in a purchase order of 50, with values ranging from 0 to 50.
- The number of houses in a suburb equipped with a power safety switch.
- The annual sales of new bicycles by a bicycle store.
- The number of defective light bulbs in a purchase order for a city store.

To determine the value of a discrete random variable, a counting process is necessary.

On the other hand, a continuous variable can take any value within an entire interval on the number line. Consider the time it takes a student at your school to eat lunch, which could vary anywhere from zero to the length of the lunch period at your school.

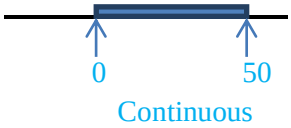


Figure 10: Continuous variables

A continuous random variable, denoted as X, has the capability to assume any value within a specific interval on the number line. For instance, X could represent:

- The heights of men, falling within the interval $50\ cm < X < 250\ cm$.
- The volume of water in a tank, which might lie in the interval $0\ m^3 < X < 100\ m^3$.

To ascertain the value of a continuous random variable, a measurement process is required.

1.6.1. Discrete Random Variables

In the preceding section, we explored various ways of assigning probabilities to events. In this section, our focus will be on understanding the emergence of discrete random variables and their distributions.

These variables and their distributions can arise from different scenarios, such as

- conducting experiments with repeated trials until a pattern emerges,
- leveraging symmetry principles (like the $\frac{1}{2}$ chance of a coin landing heads or tails or $\frac{1}{6}$ chance of an ordinary die),
- evaluating the form of tennis players to predict their outcomes in upcoming matches.

For any random variable, there exists a corresponding probability distribution that outlines the likelihood of the variable taking specific values. The probability that the variable X equals a particular value x is denoted as $P(X = x)$.

The collection of all values in the sample space of a random variable, along with their associated probabilities, is termed the probability distribution or probability mass function of the variable. This information is often best presented in a table.

The probability distribution of a discrete random variable can be represented in various forms, including a table, graphical representation, or as a function known as a probability distribution function or probability mass function $(P(x) = P(X = x))$.

The domain of the probability mass function encompasses the set of possible values of the variable, while the range comprises the values in the probability distribution of the variable.

The probability distribution for a discrete random variable is also referred to as the probability mass function, and sometimes as the probability distribution function. For each possible value x of the random variable X, the probability mass function specifies the probability of observing that particular value during the experiment.

For example, if X represents the number of sixes obtained when a dice is rolled 3 times, we would therefore write $P(X = x)$ to represent "the probability that the number of sixes is x" where x can take the values 0, 1, 2 and 3.

Consider the following crucial facts for any random variable X:

- $0 \leq P(X = x) \leq 1$, indicating that a probability will always be a value between 0 and 1 (inclusive).
- $\sum P(X = x) = 1$, signifying that the sum of all probabilities in a probability distribution must always equal 1. This fact is especially helpful when complete information about the probabilities is not available.

If X is a random variable with possible values $\{x_1, x_2, x_3, \cdots, x_n\}$ and corresponding probabilities $\{p_1, p_2, p_3, \cdots, p_n\}$ such that $P(X = x_i) = p_i$ for $i = 1, \cdots, n$, then:

- $0 \leq p_i \leq 1$ for all $i = 1, \cdots, n$,
- $\sum_{i=1}^{n} p_i = p_1 + p_2 + p_3 + \cdots + p_n = 1$,
- $\{p_1, \cdots, p_n\}$ describes the probability distribution of X.

For example, suppose X is the number of heads obtained when 2 coins are tossed. The possible values for X are $\{0, 1, 2\}$ with corresponding probabilities $\left\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right\}$. Here, we observe that $0 \leq p_i \leq 1$ for each value of i, and the probabilities add up to 1.

Figure 11: 1 Turkish Lira coin

$\boldsymbol{x}$	0	1	2
$\boldsymbol{P(X = x)}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Table 6: Probability distribution function of X

We can depict this probability distribution either in the form of a table or a graph. Another method for representing the probability distribution is through a histogram, where each bar corresponds to the probability associated with a specific value of x. The values of x inherently represent mutually exclusive events. The summation of $P(x)$ across all values of X is tantamount to adding the probabilities of all simple events in the sample space, resulting in a total of 1. This outcome can be extended to all probability distributions.

When dealing with discrete random variables, it is common to examine intervals of values that the variable may assume. Caution is required regarding whether the endpoints of an interval are inclusive. For instance, consider the following intervals:

Notation	*Statement*
$P(X = 3)$	the probability that *X* equals 3
$P(X < 3)$	the probability that *X* is less than 3
$P(X \leq 3)$	the probability that *X* is at most 3
$P(X > 3)$	the probability that *X* is more than 3
$P(X \geq 3)$	the probability that *X* is at least 3
$P(3 < X < 7)$	the probability that *X* is between 3 and 7
$P(3 \leq X \leq 7)$	the probability that *X* is at least 3 but no more than 7
$P(3 < X \leq 7)$	the probability that *X* is more than 3 but no more than 7
$P(3 \leq X < 7)$	the probability that *X* is at least 3 but less than 7

Table 7: Random variable statements

For a given value x of the random variable X, there is often an interest in calculating the probability that the observed value of X is no more than x. This leads to the concept of the cumulative distribution function.

The notation in the cumulative distribution function, also referred to as the cumulative probability function, indicates that the summation is performed over all possible values of y that are less than or equal to x. The choice of the variable name (y in this case) is arbitrary, and any letter can be utilized.

The cumulative distribution function of a random variable X expresses the probability that X does not surpass the value x as a function of x. In other words:

$$F(x) = P(X \leq x) = \sum_{y;y \leq x} p(y)$$

In many instances, the cumulative distribution is employed to determine individual probabilities:

$$P(X = x) = P(X \leq x) - P(X < x).$$

This property is particularly valuable when examining the binomial and Poisson distributions.

Here is an example of the frequency distribution for 25 families in İstanbul surveyed in a marketing study to record the number of liters of milk consumed during a specific week. Table 8 provides the liters consumed, rounded to the nearest liter, along with the corresponding relative frequency. One interpretation of probability is that it signifies the long-term relative frequency of the event.

Number of litres to the nearest litre	**Relative frequency**
0	0.08
1	0.20
2	0.36
3	0.20
4	0.12
5	0.04

Table 8: Number of litres of milk consumed by families duringa particular week

In the context of milk consumption, the cumulative distribution function can be represented as shown in Table 9. For instance, $F(3) = 0.84$ signifies the probability of a family consuming up to 3 liters of milk. This outcome is obtained by summing the probabilities associated with $x = 0, 1, 2,$ and 3.

X	0	1	2	3	4	5
$F(x)$	0.08	0.28	0.64	0.84	0.96	1.00

Table 9: Cumulative distribution function for milk consumption

If a discrete random variable X has possible values $x_1, x_2, x_3, \cdots, x_n$ and each value has an equal probability of $\frac{1}{n}$, then X is considered a uniform discrete random variable.

An illustrative example of a uniform discrete random variable is the outcome X when rolling a fair six-sided die. In this case, the potential values for X are 1, 2, 3, 4, 5, and 6, with each value having a probability of $\frac{1}{6}$.

In contrast, if two dice are rolled, the sum of the resulting numbers Y is not a uniform discrete random variable.

1.6.1.1. *The mode and median*

A fundamental inquiry regarding a random variable is the value it is most likely to assume, referred to as the mode. In the given example, the random variable X has a mode of 1, indicating that the most probable outcome is obtaining a head once. It's noteworthy that a random variable may lack a mode, as is the case with the outcomes of a fair die where each result is equally probable. Alternatively, a random variable can have multiple modes, and if the highest probability corresponds to two outcomes, it is termed bimodal.

For a discrete probability distribution, the mode represents the most frequently occurring value of the variable. Specifically, the mode is the data value x_i associated with the highest probability p_i.

The median of the distribution corresponds to the 50th percentile. In the context of a cumulative probability distribution, the median, denoted as m, is the smallest X value for which $P(X \leq m)$ exceeds $\frac{1}{2} = 0.5$. If such a value m exists, the median is the mean of this value and the next largest X. If the possible values $\{x_1, x_2, x_3, \cdots, x_n\}$ are listed in ascending order, the median is the value x_j when the cumulative sum $p_1 + p_2 + \cdots + p_j$ reaches 0.5.

To determine the median of a discrete random variable, the defining property of the median is applied, ensuring that half of the data falls below it. In probability distribution terms, this implies that the median m satisfies $P(X \leq m)$. If a value m meets this condition, the median is the average of m and the next largest X value. Cumulative probabilities, such as $P(X \leq m)$, provide insights into the likelihood of being less than or equal to a specific value. In the example of milk consumption, $P(X \leq 2)$ is 0.64, indicating that the median is 2. If the distribution were as per Table 8, where $P(X \leq 2)$ is precisely 0.5, the median would be the average of 2 and 3, resulting in 2.5.

1.6.1.2. *Expectation*

The expectation of a random variable represents the mean outcome when the variable is measured an infinite number of times, providing an average value for the random variable.

The probability distribution for a random variable resembles the relative frequency distribution, with the distinction that the former serves as a model for the entire population, while the latter characterizes a sample of measurements. Similar to how the mean and standard deviation offer measures for the center and spread of sample data, analogous measures can be computed to describe the center and spread of the population.

The population mean, also known as the expected value of the random variable X, gauges the average value of X within the population. It denotes the value anticipated on average when the experiment is repeated an infinite number of times. The expected value is determined using a straightforward formula, which can be clarified through an example.

Consider the milk consumption example with the probability distribution outlined in Table 8. If a large number of families, say 100000, is chosen, the expected outcomes would intuitively involve observing 8000 families consuming no milk, 20000 consuming 1 liter, and the remaining consuming: 36000, 20000, 12000, and 4000 liters, respectively. The average (mean) value of X is calculated as the sum of all measurements divided by n, resulting in 2.2 liters:

$$\frac{\text{Sum of all measurements}}{\text{n}} = \frac{0 \times 8000 + 1 \times 20000 + 2 \times 36\,000 + 3 \times 20\,000 + 4 \times 12000 + 5 \times 4000}{100000}$$
$$= 0 \times 0.08 + 1 \times 0.20 + 2 \times 0.36 + 3 \times 0.20 + 4 \times 0.12 + 5 \times 0.04$$
$$= 0 \times p(0) + 1 \times p(1) + 2 \times p(2) + 3 \times p(3) + 4 \times p(4) + 5 \times p(5) = 2.2$$

That is, we expect to see families, on average, consuming 2.2 litres of milk. This does not mean that we know what a family will consume, but we can say what we expect to happen.

The mean value of X remains consistent regardless of the number of trials, allowing us to find the mean or expected value $E(X)$ by summing each value x multiplied by its respective probability $P(X = x)$. This is akin to conducting the experiment just once, and the formula for expected value is expressed as:

$$E(X) = \sum_{x} xP(X = x)$$

For a random variable X with possible values $x_1, x_2, x_3, \cdots, x_n$ and associated probabilities $p_1, p_2, p_3, \cdots, p_n$, the expected value of X is

$$E(X) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

Expectation is essentially the mean of the underlying distribution, often denoted by μ. Importantly, the expectation of a random variable need not be a value that the variable can itself assume.

To illustrate with a practical scenario, an insurance company offers a policy that pays ৳10000 when a car is damaged beyond repair or ৳5000 for major damages (50%). The company charges ৳50 per year for this service. To assess profitability, consider a probability model for this policy, where 1 out of every 1000 cars is damaged beyond repair, and an additional 2 out of 1000 sustain serious damage.

Type of accident	**Amount paid x**	**Probability $P(X = x)$**
Total damage	10000	$\frac{1}{1000}$
Major damage	5000	$\frac{2}{1000}$
Minor or no damage	0	$\frac{997}{1000}$

Table 10: Probability table for car insurance policy

The expected payout for the insurance company can be calculated as follows:

$$\mu = E(X) = \sum xP(x) = ₺10{,}000\left(\frac{1}{1000}\right) + ₺5{,}000\left(\frac{2}{1000}\right) + ₺0\left(\frac{997}{1000}\right) = ₺20$$

This implies that, on average, the insurance company expects to pay ₺20 per insured car. Given that the company charges ₺50 for the policy, it anticipates a profit of ₺30 per car. From another perspective, if they insure 1000 cars, the company expects to pay ₺10000 for one car and ₺5000 each for two cars with major damage. This totals ₺20000 for all cars, averaging ₺20 per car.

It's crucial to note that this expected value doesn't represent the actual cost for any specific policy. Due to the random nature of events, some car owners may receive ₺10000 or ₺5000, while most others receive nothing. To account for this variability, the insurance company needs to assess a measure of this variability, known as the standard deviation.

When considering a spinner, players are awarded points based on the resulting number of spins. On average, the expected points per spin can be calculated by considering the probability of each score. For instance, if the spinner has scores of 50, 15, 10, and 5, with equal probabilities, the total score in this case would be $50 + 15 + 10 + 5 = 80$, which is an average of $\frac{80}{4} = 20$ points per spin. This means the average score per spin is $\frac{1}{4}(50 + 15 + 10 + 5) = \frac{1}{4} \times 50 + \frac{1}{4} \times 15 + \frac{1}{4} \times 10 + \frac{1}{4} \times 5 = 20$ points.

Moreover, when dealing with repeated experiments, expectations can be applied to estimate the number of occurrences of a specific event. For example, rolling a die 120 times, with each face equally likely, suggests that $\frac{1}{6}$ of these rolls, or 20 times, can be expected to result in a ‘’six’’. This expected value, denoted as np, is utilized to determine the number of expected occurrences when an event has a probability p of happening in each trial. In gambling, the concept of expected gain is pivotal, representing the expected return or payout from a game, minus the cost to play. A game is considered fair when the expected gain is zero, expressed as $E(X) = 0$, where X represents the gain from each game.

1.6.1.3. *Variance and standard deviation*

In addition to understanding the expectation and median, there is interest in quantifying how much an outcome deviates from the average. The variance of a random variable serves as a measure representing the extent of variation expected when the variable is measured an infinite number of times, reflecting how dispersed the variable is.

The variance is computed by determining the deviation from the mean, $x - \mu$, and squaring it, a process applicable to random variables as well. The population variance σ^2 and, consequently, the population standard deviation σ can be justified using similar reasoning. These measures characterize

the dispersion of the random variable values around the center, utilizing the concept of the average or expected value of squared deviations from the mean μ, or $E(X)$.

Another formula for the variance can be derived as follows:

$\sigma^2 = \sum(x-\mu)^2 P(x) = \sum x^2 P(x) - \mu^2 = \sum x^2 P(x) - [E(x)]^2 = E(x)^2 - [E(x)]^2 = \sum x^2 P(x) - [\sum xP(x)]^2$

This formula is often expressed as ‘’the mean of the squares minus the square of the mean’’.

Returning to the milk consumption example, where the expected mean value was calculated as 2.2 liters, let's tabulate the work to facilitate the manual computation of the variance.

X	$P(x)$	Deviation $(x-\mu)$	Squared deviation $(x-\mu)^2$	$(x-\mu)^2 P(x)$
0	0.08	-2.2	4.84	0.3872
1	0.20	-1.2	1.44	0.2880
⋮	⋮	⋮	⋮	⋮
5	0.04	2.8	7.84	0.3136
		Total	$\sum(x-\mu)^2 P(x)$	**1.52**

Table 11: Calculating variance for milk consumption

Hence, 1.52 litres is the variance of the milk consumption, and the standard deviation is $\sqrt{1.52} = 1.233$ litres.

1.6.2. Discrete Probability Distributions

Various everyday situations present examples of discrete random variables. Nevertheless, a subset of discrete probability distributions holds significant applicability and serves as models for numerous scenarios.

One notable type of discrete random variable is employed in sampling with replacement, and its associated probability distribution is known as the binomial probability distribution. On the other hand, when dealing with sampling without replacement, the hypergeometric probability distribution serves as the applicable model.

1.6.2.1. *The Binomial distribution*

Certain discrete probability distributions are so commonly encountered that they have been designated names and formal notations. Among these, the binomial distribution stands out as one of the most crucial. The binomial distribution arises in scenarios where a fixed number of experiments or trials are conducted, each having two possible outcomes. Typically denoted as n, the number of trials involves one outcome termed as ‘’success’’ and the other as ‘’failure’’. The probability of success is denoted as p. If the likelihood of success in a trial remains constant, and the trials are conducted independently, the binomial distribution is used to model the number of successes.

Real-life examples of situations well-suited for the binomial distribution include:

- Testing a Drug:

In the pharmaceutical field, the effectiveness of a new drug in curing a disease can be modeled using the binomial distribution. The outcomes are whether the drug successfully treats the disease or not, and similarly, the occurrence of side effects can be treated as a success or failure.

- Participating in a Lucky Draw:

Luck-based contests, such as lucky draw competitions, where the outcome is either winning or losing, align with the binomial distribution. The probability of success (winning) or failure (losing) can be expressed using this distribution.

- Estimating the Number of Fraudulent Transactions:

Banks utilize the binomial distribution to estimate the likelihood of a credit card transaction being fraudulent. By recording the total number of fraudulent transactions in a specific area, the distribution helps predict the number of future fraud cases.

- Number of Spam Emails Received:

Predicting the number of spam emails received is another application of the binomial distribution. An email can either be categorized as spam or not, making it a suitable scenario for this distribution.

- Number of Returns:

When considering product returns, the binomial distribution aids in estimating the probability of a customer returning a product within a given time frame. This assists shopkeepers in predicting the number of expected returns.

- Participating in an Election:

Election outcomes, where a participant can either win or lose, find representation through the binomial distribution. This distribution can be employed to calculate the chances of a political party winning or losing based on historical election records.

- Supporting a Sports Team:

Whether your favored sports team wins or loses in a game or event corresponds to the success or failure of the experiment. Hence, the binomial distribution is well-suited for scenarios involving sports team support and game outcomes.

Consider an experiment with two possible outcomes: success occurs if a certain event happens, and failure if the event does not occur.

Let X represent the number of occurrences of success (blues)

when spinning a given spinner once.

The random variable X signifies the total number of successes in n trials.

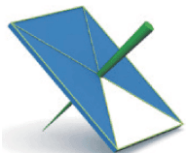


Figure 12: Spinner

If we spin the spinner n times independently and count the number of blues in each trial, it is termed a binomial experiment. In a binomial experiment, the probability of getting a blue is constant for each spin, and spins are independent of one another. The probability distribution of X is given in Table 12:

$\boldsymbol{x}$	0	1
$\boldsymbol{P(X=x)}$	$\frac{1}{4}$	$\frac{3}{4}$

Table 12: The probability distribution of X

If we repeat this experiment in a number of independent trials, we spin the spinner n times, and count the number of blues that result, we call it a binomial experiment because the probability that we get a blue is the same for each spin, and each spin is independent of every other spin.

The probability of success, denoted as p, must remain constant across all trials in a binomial experiment. Since success and failure are complementary events, the probability of failure, denoted as $1-p$, is also constant for all trials.

A binomial experiment possesses the following characteristics:

- The experiment comprises n identical trials.
- Each trial results in either success or failure.
- The probability of success on a single trial, p, remains constant throughout the entire experiment. The probability of failure, denoted as $1-p$ or q, is also constant, where $p+q=1$.
- The trials are independent.
- The variable of interest is the number of successes (x) during the n trials, where $x=0,1,2,\cdots,n$.

The outcomes of a binomial experiment, along with the corresponding probabilities, constitute a binomial distribution.

The unique parameters defining a binomial distribution are the values of n (number of trials) and p (probability of success). The notation $X \sim B(n,p)$ indicates that the variable X follows a binomial probability distribution. The symbol $\sim$ conveys the concept ‘’is distributed as’’, and concise abbreviations are often used for standard distributions.

In every binomial distribution:

- There is a fixed number (n) of trials.
- Each trial has two possible outcomes: success or failure.
- The probability of success (p) is constant from trial to trial.
- Trials are independent of each other.

The probability distribution function for the binomial random variable x is expressed as:

$$P(x \text{ successes in } n \text{ indefiendent trials}) = P(x) = {}^{n}C_{x}p^{x}(1-p)^{n-x} = \binom{n}{x}p^{x}(1-p)^{n-x}$$

$$= \binom{n}{x}p^{x}q^{n-x}, \text{ for } x = 0,1,2,\cdots,n.$$

Here, $\binom{n}{r}$ represents the number of ways r successes can be arranged among the n trials, and $p^{r}q^{n-r}$ is the probability of obtaining r successes and $n-r$ failures in a specific order.

The Galton board, also known as a quincunx or bean machine, devised by Sir Francis Galton, is a statistical experiment tool. It features an upright board with evenly spaced nails, a lower half with slots, and a funnel at the upper edge for pouring balls. Each time a ball hits a nail, it can bounce left or right with equal probability, resulting in a binomial distribution in the heights of ball heaps in the lower slots, resembling a normal or bell curve.

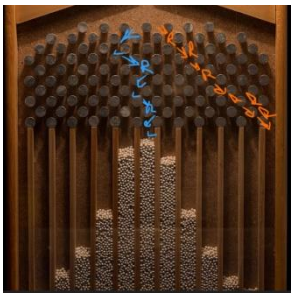


Figure 13: Bean machine[1]

Suppose a spinner possesses three blue edges and one white edge. Clearly, each spin results in either blue or white, with a $\frac{3}{4}$ chance of landing on blue and a $\frac{1}{4}$ chance of landing on white. If we designate a blue result as ‘’success’’ and a white result as ‘’failure’’, we create a binomial experiment.

Let p be the probability of obtaining blue, where $p = \frac{3}{4}$. The probability of obtaining white is $1 - p = \frac{1}{4}$. Consider twirling the spinner $n = 3$ times. Let the random variable X represent the number of successes or blue results, with X taking values 0, 1, 2, or 3.

$$P(X = 0) = P(\text{none are blue}) = P(\text{WWW}) = \left(\frac{1}{4}\right)^3 \approx 0.0156$$

$$P(X = 1) = P(1 \text{ blue and } 2 \text{ white}) = P(\text{BWW or WBW or WWB}) = 3\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^2 \approx 0.1406$$

$$P(X = 2) = P(2 \text{ blue and } 1 \text{ white}) = P(\text{BBW or BWB or WBB}) = 3\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right) \approx 0.4219$$

$$P(X = 3) = P(3 \text{ blue}) = P(\text{BBB}) = \left(\frac{3}{4}\right)^3 \approx 0.4219$$

The factor of 3 represents the number of ways of obtaining one success in three trials, denoted as $\binom{3}{1}$. This suggests that $P(X = x) = \binom{3}{x}\left(\frac{3}{4}\right)^x\left(\frac{1}{4}\right)^{3-x}$ where $x = 0, 1, 2,$ or 3.

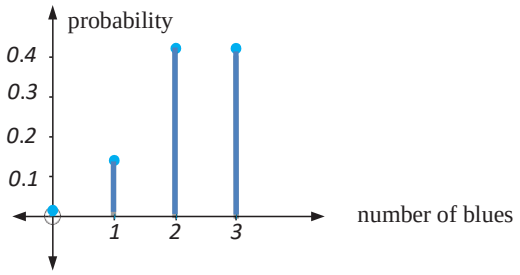


Figure 14: Probability function graph

[1] https://sanket-m-kangle.medium.com/what-is-population-sample-and-sampling-error-6298d79c3771

The total of the probabilities $(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$ equals 1.

In the case of the binomial distribution where $X \sim B(n, p)$, the expected value of X is $E(X) = \mu = np$. The variance formula for a binomial distribution with $X \sim B(n, p)$ is $Var(X) = \sigma^2 = npq = np(1 - p)$. The binomial probability function is utilized to determine the probability $P(X = k)$ of the variable taking the value k.

To calculate the probability that the variable falls within a range of values, such as $P(X \leq k)$ or $P(X \geq k)$, the binomial cumulative probability function is employed. The cumulative distribution function $F(x)$ for a random variable X indicates the probability that X does not exceed the value x, expressed as

$$F(x) = P(X \leq x) = \sum_{y;y \leq x} p(y).$$

Consequently, for the binomial distribution, the cumulative distribution function is represented as:

$$F(x) = P(X \leq x) = \sum_{y;y \leq x} p(y) = \sum_{y;y \leq x} {}^{n}C_{y} p^{y} q^{n-y} = \sum_{y;y \leq x} \binom{n}{y} p^{y} q^{n-y}.$$

Utilizing the cumulative distribution proves valuable when determining the probability of a binomial variable falling within a specific interval.

1.6.2.2. *The Poisson distribution*

When waiting for a bus, there are two possible outcomes at any given moment – it either arrives or it does not. Attempting to model this situation using a binomial distribution encounters ambiguity regarding what constitutes an individual trial. Instead, we consider a rate of success, representing the number of buses arriving in a fixed time period.

In scenarios where we are aware of the rate of events within a specific space or time, ranging from commercial contexts (such as the number of calls through a telephone exchange per minute) to biological settings (like the number of clover plants observed per square meter in a pasture), the Poisson distribution proves useful. This distribution is entirely defined by the rate of success, conventionally denoted as λ. The distribution is stated as $X \sim Po(\lambda)$, and its probability formula is

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; \quad x = 0,1,2,\cdots$$

Real-life applications of the Poisson distribution include:

- Number of Network Failures per Week:

Cell phone companies and wireless service providers use Poisson distribution to predict the probability of network failures based on historical data, enabling proactive problem-solving. This aids in enhancing their operational efficiency and customer satisfaction metrics. By leveraging the historical data on network failures within a specific timeframe in a given area, it becomes feasible to predict the likelihood of future network failures using the Poisson distribution. This predictive capability enables broadcasting organizations to proactively address potential issues, allowing them to prepare solutions in advance. Consequently, customers utilizing their services can experience minimal inconvenience.

- Number of Bankruptcies Filed per Month:

In banking, Poisson distribution aids in estimating the probability of customer bankruptcies over a given time, guiding reserve cash planning. As an example, suppose the bank's records indicate an average of four bankruptcies filed each month in a specific area. This data can be utilized to predict the likelihood of zero, one, two, or three bankruptcies being filed in the subsequent month. This predictive analysis assists bank managers in estimating the necessary reserve cash that should be readily available in the event of a specific number of bankruptcies.

- Number of Website Visitors per Day:

Predicting the number of website visitors per day is suited to Poisson distribution, assisting in maintaining website functionality. For example, if a specific website receives 100 visitors per day, the likelihood of more or fewer than 100 visitors in the subsequent day can be precalculated using the Poisson distribution. Once the probability of incoming visitors to the website is determined, the risk of a website crash can be assessed. Consequently, the site engineer takes measures to maintain optimal data uploading and downloading speeds, allocates suitable bandwidth to accommodate the expected number of visitors, and adjusts website parameters like processing capacity accordingly to prevent potential website crashes.

- Number of Arrivals at a Restaurant:

Restaurants use Poisson distribution to estimate daily customer visits, aiding in preparation and resource allocation. Suppose a specific restaurant receives a daily average of 200 visitors; in this case, the Poisson distribution can be applied to predict whether more or fewer than 200 people might visit the restaurant on the following day. This provides the owner with insights into the expected number of patrons and aids in determining the necessary quantity of raw materials for their service.

- Number of Calls per Hour at a Call Center:

Call centers apply Poisson distribution to determine staffing needs, optimizing efficiency based on expected call volumes. For example, if a call center typically handles 10 calls per hour, the Poisson formula can be applied to estimate the probability of receiving zero, one, two, three, or any other integer number of calls per hour. This information enables managers to understand the expected call volume at various hours throughout the day, facilitating the creation of an appropriate schedule for employees to follow.

- Number of Books Sold per Week:

Booksellers use Poisson distribution to predict future sales based on known data which is collected in a week, facilitating inventory planning. To achieve this goal, the individual typically reviews existing data or information on bookstore sales. Using the Poisson distribution, they calculate the probability of selling a specific number of books within a defined and fixed time duration.

- Average Number of Storms in a City:

Poisson distribution aids in predicting natural calamities, allowing for risk estimation and preventive measures. Risk assessment enables environmental engineers and scientists to implement appropriate measures to safeguard lives and reduce property damage substantially. To achieve this, they document the average occurrence of storms or other disasters in a specific area over a designated period. This recorded information serves as input for the Poisson distribution calculator, which then provides the probability of a specific number of disasters that could occur in the same area in the near future.

- Number of Emergency Calls Received Every Minute:

Poisson distribution is used to prepare for varying emergency call rates, ensuring readiness for any possible scenario. Using the average number of emergency calls received every minute, the number of emergency calls that might be received in the next hour can be found. This helps the staff be ready for every possible emergency.

When a question involves an average rate of success or events occurring at a constant rate, Poisson distribution is appropriate. If a fixed number of trials can be identified, then binomial distribution is suitable.

For Poisson distribution, the expectation is $E(X) = \lambda$, and the variance is also $Var(X) = \sigma^2 = \lambda$.

Consider an example where recordable accidents occur in a factory at an average rate of 7 every year, independently of each other. To find the probability of exactly 2 recordable accidents in a given year, the random variable X is defined as the number of accidents in a year, and the probability distribution is $X \sim Po(7)$. Using the Poisson distribution probability function, $P(X = 2) = \frac{e^{-7}7^2}{2!} = 0\ 223$.

The Poisson distribution exhibits scalability. If the number of butterflies observed on a flower in a 10-minute interval conforms to a Poisson distribution with a mean (expectation) of λ, then the number of butterflies observed on a flower in a 20-minute interval adheres to a Poisson distribution with a mean of 2λ. Likewise, the number of butterflies observed in a 5-minute interval adheres to a Poisson distribution with a mean of $\frac{\lambda}{2}$.

1.6.3. Continuous Probability Distributions

When dealing with a discrete random variable X, positive probabilities are assigned to each possible value of X, forming the probability distribution for X. The sum of all these probabilities for different values of X equals 1. In the case of discrete variables, we can visually represent the probabilities associated with various values of the random variable X using a probability histogram (relative frequency histogram). Each bar's area in this histogram corresponds to the probability of the specific value it represents.

Discrete random variables and binomial probability distributions involve cases where the random variable X can assume non-negative integer values, such as $x = 0, 1, 2, 3, 4, \cdots, n$ for some finite $n \in N$. However, many real-world variables, such as height, weight, and time, are continuous. For a continuous random variable X, x can take any real value within a reasonable domain.

Consider the following data for masses of 5 kg bags of rice.

Mass/kg	4.9	5.0	5.1	5.2
Frequency	12	16	20	14

Table 13: Frequencies for masses of 5 kg bags of rice

Not all data labeled as 5.1 kg corresponds precisely to a mass of exactly 5.1 kg. For instance, a bag with a mass of 5.1358 kg or 5.0879546 kg would fall within this category. Due to the impracticality of listing all potential actual masses and the inherent limitations in measuring mass with absolute precision, continuous data must be grouped. Consequently, discussions about the probability of a single value for a continuous random variable are not feasible; instead, we focus on the probability of the variable falling within a specified range. A convenient representation of this concept is through the

area under a graph, where no area exists above a single value, but the area above any given range can be determined.

The area under a graph is calculated through integration. The function subject to integration is termed the probability density function, denoted as $f(x)$. The key characteristic of $f(x)$ is that the area between two x values equates to the probability of the continuous random variable falling between those two values: $P(a < X < b) = \int_a^b f(x)dx$.

Given the infinite potential values X can assume, and even if a measuring device theoretically enabled precise measurement of X, measurements of X from any two population members would never be identical. Consequently, the probability of X being exactly equal to any specific value is zero. For a continuous variable X, the probability that X is exactly equal to a particular value $P(X = x) = 0$ is zero for all x.

Consider a continuous random variable X, such as height, weight, or the lifespan of a product like a TV set. Due to its continuous nature, X can take values over an interval, with an infinite number of possibilities. Therefore, it's impractical to list all possible values and their probabilities. Instead, for continuous variables, we employ a probability density function or distribution curve.

The probability density function, denoted as $f(x)$, is not a probability itself; rather, probabilities are determined by calculating areas under the probability density function curve for specific intervals. It is essential to recognize that for a continuous variable, we can only discuss the probability of an event lying in an interval. Since $P(X = x) = 0$ for all x in the case of continuous variables, a probability mass function is not applicable. Instead, we use a probability density function or distribution curve. The value of this function is not a probability; rather, probabilities are obtained by calculating areas under the curve for a given interval.

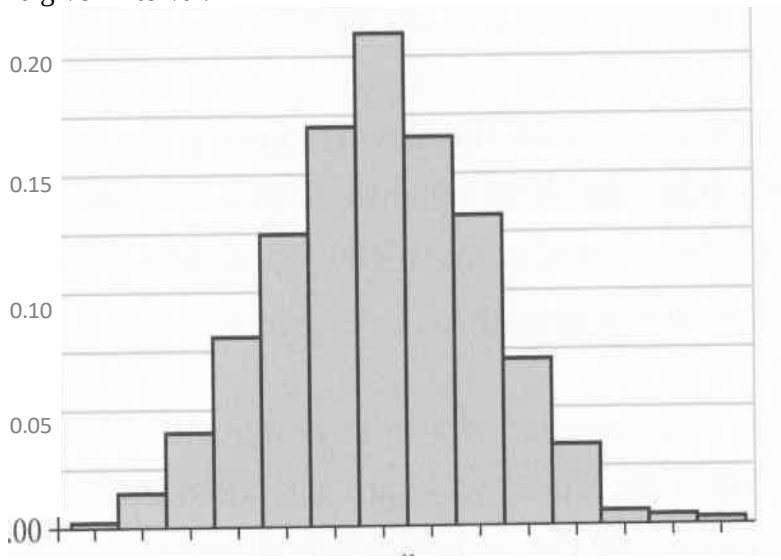


Figure 15: Relative frequency histogram

When handling continuous data, we group it into classes to create a relative frequency histogram. As more measurements are collected, narrower classes are utilized to refine the histogram. With a substantial number of measurements and narrow class widths, the relative frequency histogram increasingly resembles a smooth curve—a characteristic of continuous data. This smooth curve represents the probability density function of X, depicted by a curve $y = f(x)$. This curve ensures the total area beneath it is 1, and the area between any two points indicates the probability that x falls between those two points.

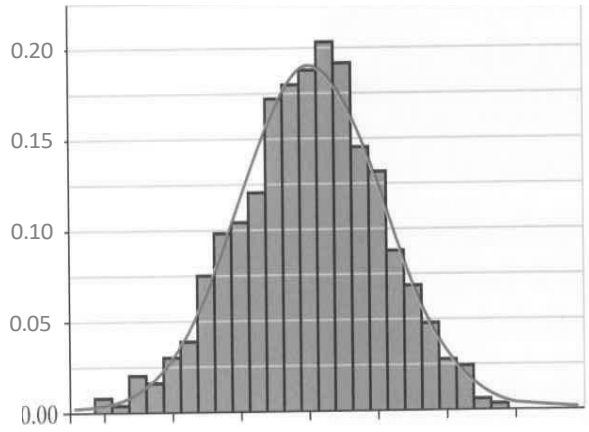


Figure 16: As the number of measurements increases, class width becomes narrower

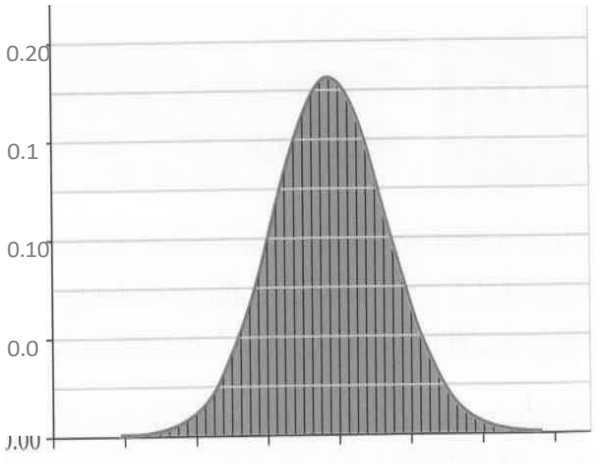


Figure 17: Smooth curve

For a continuous random variable X, the probability density function is a function $f(x)$ such that $f(x) > 0$ across its entire domain. The function may have different rules over various domains, and if only defined over a particular domain, it is assumed to be zero everywhere else. If the domain of the function is $a \leq x \leq b$, then $\int_a^b f(x)\,dx = 1$. In summary, a continuous random variable's probability density function is crucial for understanding probabilities associated with specific intervals, given that the probability of an exact value is zero.

For example, because for a continuous variable X, $P(X = x) = 0$ for all x, the probability that an apple will weigh exactly 72.9 g is zero.

If you were to weigh an apple on scales that weigh to the nearest 0.1 g, a reading of 72.9 g means the weight lies somewhere between 72.85 g and 72.95 g. No matter how accurate your scales are, you can only ever know the weight of an apple within a range.

So, for a continuous variable we can only talk about the probability that an event lies in an interval. A consequence of this is that $P(c \leq X \leq d) = P(c < X \leq d) = P(c \leq X < d) = P(c < X < d)$. This would not be true if X was discrete.

Expectation and variance calculations for continuous random variables involve integration:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)dx,$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)dx$$

$$Var(X) = E(X^2) - [E(X)]^2.$$

An example with a probability distribution function $f(x) = \begin{cases} \frac{3}{4}x(2-x); & 0 < x < 2, \\ 0 & ; \quad otherwise \end{cases}$

demonstrates these concepts, yielding an expected value

$$E(X) = \int_0^2 x\frac{3}{4}x(2-x)dx = \frac{3}{4}\int_0^2 x^2\,(2-x)dx = 1$$

and using $E(X^2) = \int_0^2 x^2\frac{3}{4}x(2-x)dx = \frac{3}{4}\int_0^2 x^3\,(2-x)dx = 1.2$ variance of X becomes $Var(X) = E(X^2) - [E(X)]^2 = 1.2 - 1^2 = 0.2$. Then the standard deviation approximately becomes $\sqrt{0.2} = 0.447$.

The identification of the median and mode is applicable to continuous distributions as well. The median is characterized by the property that half of the data points lie below this value and the other half above it. On the other hand, the mode represents the most probable value, and this interpretation can be framed in the context of probability.

The median m satisfies $\int_{-\infty}^{\infty} f(x)dx = \frac{1}{2}$.

The mode is the value of x at the maximum value of $f(x)$. The maximum value of $f(x)$ is not necessarily where $\frac{df}{dx} = 0$.

Continuous probability distributions can assume a variety of shapes. However, particularly the normal distribution, play a significant role in modeling real-world phenomena with a bell-shaped curve.

1.6.3.1. *The Normal distribution*

The preeminent form of a continuous random variable is the normal random variable. Many scenarios involve variables that tend to cluster around their average values, with values deviating further from the average becoming increasingly improbable. The normal distribution serves as a suitable model for such situations.

Here are examples of daily life situations illustrating Normal Distribution:

- Height:

The height of individuals follows a normal distribution. Most people in a specific population have an average height, with roughly equal numbers of individuals taller and shorter than the average. Only a small percentage of people fall into the extremes of being exceptionally tall or short due to a combination of genetic and environmental factors.

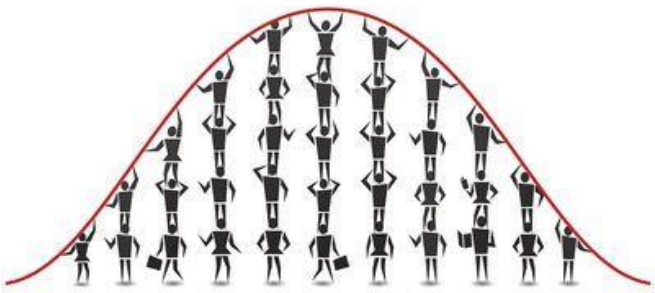


Figure 18: Normal distribution for height

- Rolling A Dice:

Rolling a fair dice exemplifies normal distribution. In experiments where a dice is rolled multiple times, the probability of getting specific outcomes, like ‘’1’’, remains consistent. With more rolls or multiple dice, the normal distribution graph becomes more elaborate. In an experiment, it was observed that when a die is rolled 100 times, the probability of getting a 1 is approximately 15-18%. Surprisingly, when the die is rolled 1000 times, the probability of getting 1 remains the same, averaging around 16.7% ($\frac{1}{6}$). If two dice are rolled simultaneously, yielding 36 possible combinations, the probability of rolling 1 (with six possible combinations) also averages around 16.7%, specifically ($\frac{6}{36}$). More the number of dice more elaborate will be the normal distribution graph.

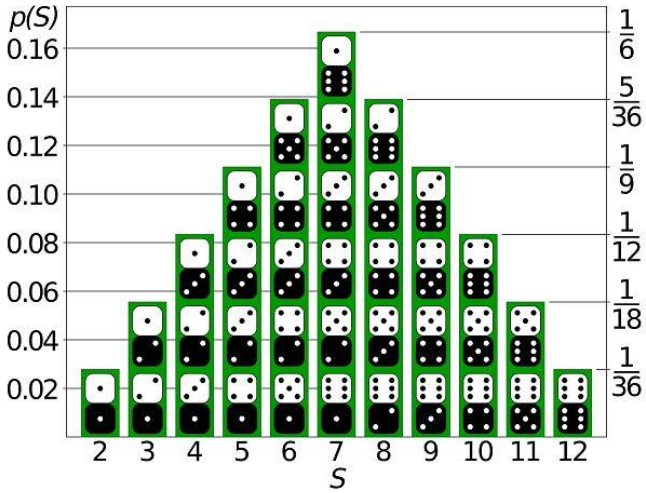


Figure 19: Normal distribution for dice

- Tossing A Coin:

Coin tossing, a common method for decision-making, is perceived as fair because it has equal chances for heads or tails. The probabilities of getting heads or tails in a single toss are both $\frac{1}{2}$, and their sum is always 1. Tossing coins multiple times maintains this cumulative probability.

- Intelligence Quotient (IQ):

IQ in a population forms a normal distribution curve. The majority of individuals fall within the normal IQ range, while a smaller percentage deviate from the average.

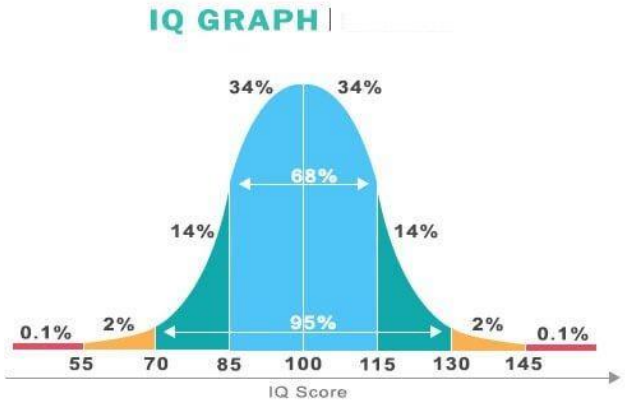


Figure 20: Normal IQ graph

- Technical Stock Market:

Many are familiar with the fluctuations in stock prices within the financial market, such as changes in the logarithmic values of Forex rates, price indices, and stock prices, which frequently exhibit a bell-shaped curve. In the context of stock returns, the standard deviation is commonly referred to as volatility. Assuming normal distribution of returns, over 99 percent of the expected returns are anticipated to fall within the deviations from the mean value. These properties of the bell-shaped normal distribution provide analysts and investors with the ability to draw statistical inferences about the anticipated return and risk associated with stocks.

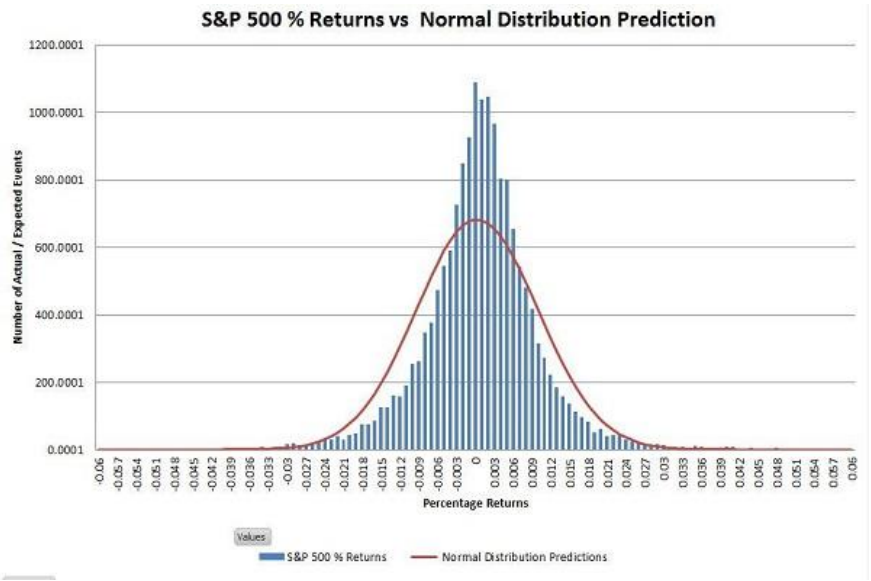


Figure 21: Normal distribution prediction for returns

- Blood Pressure:

Blood pressure in the general population tends to follow a Gaussian distribution (normal), making Gaussian mixture models suitable for modeling blood pressure behavior.

- Shoe Size:

Female shoe sales by size in Türkiye are normally distributed, reflecting the similar physical makeup of most women. This distribution is crucial for sizing considerations in the shoe industry. With this information one can calculate the probability of the glass slipper left by Cinderella at the prince's house fitted another woman's feet and he would have ended up marrying another woman.

Figure 22: Normal curve for shoe size

- Birth Weight:

The birth weight of newborns typically ranges from 2.5 to 3.5 kg and follows a normal distribution. The majority of newborns have a normal birth weight, with only a small percentage falling above or below the norm.

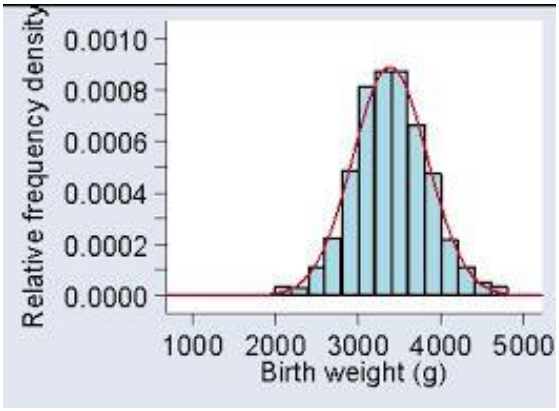


Figure 23: Normal curve for birth weight

The probability density function of a normal random variable x is determined by two parameters: the mean or expected value μ, and the standard deviation σ of the variable.

Despite all normal distributions sharing a general bell-shaped curve, the exact positioning and shape of the curve hinge on the mean μ (indicating the center of the distribution) and the standard deviation σ (indicating the distribution's spread). The normal probability density function forms a symmetric bell-shaped density curve about the mean μ, and its variability is measured by σ. Larger σ values result in increased variability, meaning a higher probability of encountering values further from the mean. Figure 24 illustrates different normal density functions with the same mean but varying standard deviations, showcasing how the curves flatten as σ increases while maintaining an area under the curve equal to 1.

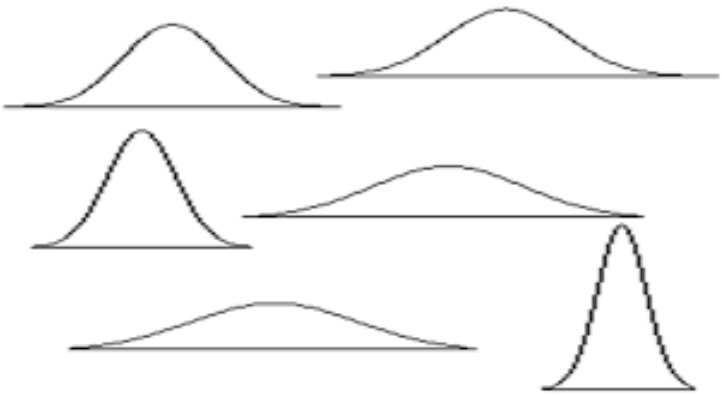


Figure 24: Some different normal density functions

For a perfect normal distribution:

- the curve is bell-shaped,
- the data is symmetrical about the mean (μ),
- The mean, mode, and median are identical.

If X is a normally distributed random variable with mean μ and standard deviation σ, then the probability density function of X is given by:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}, \quad -\infty < X < \infty.$$

This function, often referred to as the normal distribution curve or just normal curve, cannot be integrated in terms of other well-known functions and is crucial in statistics, serving as a suitable model for various naturally occurring random variables. Examples include the physical attributes of people, animals, and plants, crop yields, and specifications of mass-produced items. Additionally, the normal distribution can be applied as an approximation for scores on tests taken by large populations, student exam scores, time to complete work, reaction times, or IQ scores.

Consider the weight of oranges picked from a tree, which exhibits variation due to factors like genetics, fertilization times, sunlight exposure, and weather conditions. The resulting distribution of orange weights is bell-shaped, symmetrically centered around the mean, exemplifying a normal distribution.

The normal curve, also known as the Gaussian curve, is named after the German mathematician Carl Friedrich Gauss, who used it to analyze astronomical data. Early work on the normal curve involved French mathematicians Abraham De Moivre and Pierre Laplace, with De Moivre developing the normal curve as an approximation of the binomial theorem in 1733. Laplace used the normal curve to describe the distribution of errors in 1783 and, in 1810, to prove the Central Limit Theorem.

Carl Friedrich Gauss first characterized the normal distribution in 1809 to rationalize his method of least squares for linear regression. Although the normal distribution formula involves a Gaussian function, denoted as $f(x) = ae^{-\frac{1}{2}\left(\frac{X-b}{c}\right)^2}$, where a, b, and c are constants, it is a special case of the Gaussian function. The integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$ was first calculated by Laplace, leading to the correct normalization constant $\frac{1}{\sigma\sqrt{2\pi}}$ for the normal distribution. Lambert Quetelet, a Belgian scientist in the 19th century, applied the normal distribution to human characteristics, noting that traits like height, weight, and strength exhibited normal distributions.

Though the formula's direct use may not be necessary, understanding its properties aids in grasping the functioning of the normal distribution. Each normal curve is defined by its mean μ and standard deviation σ, referred to as the parameters of the distribution. The notation $X \sim N(\mu, \sigma^2)$ signifies that a random variable X is normally distributed with a mean of μ and a variance of σ^2. Regardless of μ and σ values, the total area under the curve remains 1, allowing partial areas to represent probabilities.

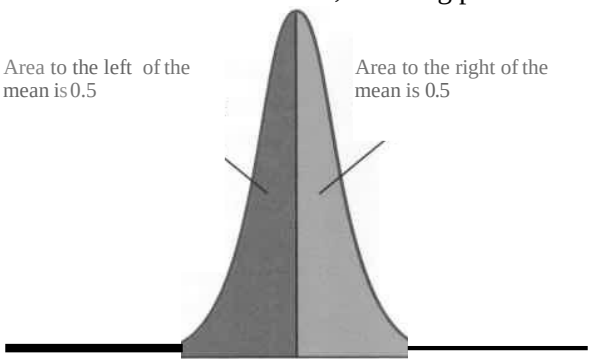
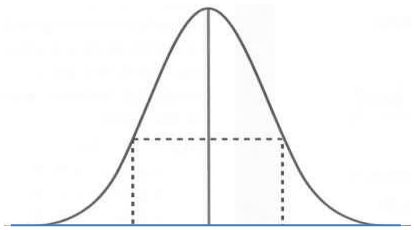


Figure 25: Normal probability distribution

Figure 25 illustrates the graph of a normal probability distribution. The mean or expected value serves as the center of the distribution, and the distribution is symmetric around this mean. Since the total area under the curve is 1, the symmetry implies that the areas to the right and left of the mean are both equal to 0.5. Larger values of σ decrease the curve's height and increase its spread, while smaller values of σ increase the height to compensate for the narrowness of the distribution.

The normal curve is symmetric about the vertical line $x = \mu$, with $f(x) > 0$ for all x. The x-axis functions as a horizontal asymptote, and the maximum occurs at $x = \mu$. Thus, the normal distribution is entirely determined by its mean μ and standard deviation σ. Changing μ without altering σ shifts the normal curve along the horizontal axis without affecting its spread. The standard deviation σ controls the curve's spread, and moving one σ to the right or left of the mean marks the point where the curvature changes, as indicated by the Empirical rule.

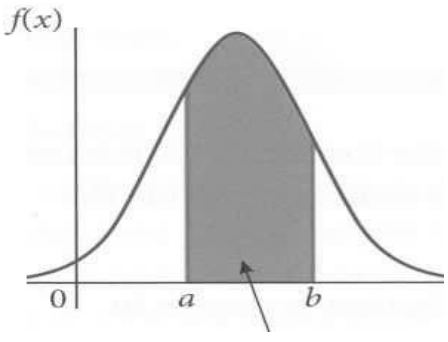
By differentiating the probability density function $f(x) = \frac{1}{\sigma^2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, we find $f'^{(x)} = \frac{-1}{\sigma^2\sqrt{2\pi}}\left(\frac{x-\mu}{\sigma}\right) e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. The first derivative $f'(x) = 0$ only when $x = \mu$, corresponding to the point where $f(x)$ is a maximum. Further differentiation yields the second derivative $f''(x) = \frac{-1}{\sigma^2\sqrt{2\pi}}\left[\frac{1}{\sigma} - \frac{(x-\mu)^2}{\sigma^3}\right] e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. Setting $f''(x) = 0$ results in $\frac{(x-\mu)^2}{\sigma^3} = \frac{1}{\sigma}$, leading to $\Rightarrow (x-\mu)^2 = \sigma^2 \Rightarrow x - \mu = \mp\sigma \Rightarrow x = \mu \mp \sigma$. Therefore, the points of inflection are at $x = \mu + \sigma$ and $x = \mu - \sigma$, indicating that the standard deviation is the horizontal distance from the line of symmetry $x = \mu$ to a point of inflection.



$x = \mu - \sigma$ μ $x = \mu + \sigma$

Figure 26: One σ the right or left of the mean μ marks the point where the curvature of the curve changes

In Figure 27, a graph of the density function is presented. For two values a and b of the random variable X, where $a < b$, the probability that x lies between a and b $[P(a < x < b)]$ is the area under the density function between these points, i.e., $P(a < x < b) = \int_a^b f(x)\,dx$.



$P(a < X < b) = P(a \leq X \leq b)$

Figure 27: Density function graph

Regrettably, the probability function (the equation defining the curve) for the normal distribution is intricate and challenging to manipulate. Utilizing integration to determine areas under this curve and consequently calculate probabilities would be overly complex. However, alternative methods exist for computing the area under the curve.

According to the empirical rule, all normal distributions are equivalent when measured in units of size σ around the mean μ as the center. The process of converting to these units is known as standardizing. To standardize a value, assess its distance from the mean and express that distance in terms of σ.

The quantity $x - \mu$ indicates the distance of a value from the mean; dividing by σ then indicates how many standard deviations that distance equals. Standardization transforms the normal curve.

For the sake of discussion, assume the mean μ to be positive. The transformation $x - \mu$ shifts the graph back μ units, making the new center 0. Dividing by σ scales the distances from the mean and expresses everything in terms of σ. Consequently, a point one standard deviation from the mean is positioned 1 unit above the new mean, represented by +1. Referring to the earlier empirical rule, points within one standard deviation from the mean will now be within a distance of 1 in the new distribution. Instead of being at $\mu + \sigma$ and $\mu - \sigma$, they will be at $0 + 1$ and $0 - 1$, respectively, denoted as -1 and +1.

This transformation gives rise to a new distribution called the standard normal distribution.

1.6.3.2. *The standard normal distribution*

The standard normal distribution is characterized by a mean (μ) of 0 and a standard deviation (σ) of 1. In mathematical notation, a random variable Z with a standard normal distribution is denoted as $Z \sim N(0, 1)$. This distribution is particularly useful for determining areas under any normal distribution through a process called standardization, which involves linear transformations.

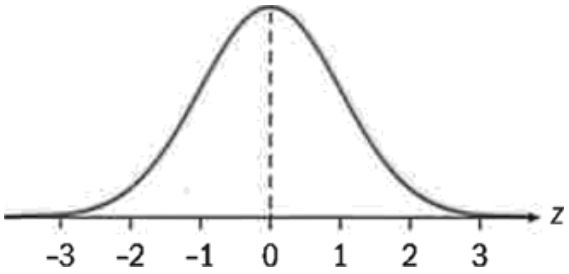


Figure 28: Standard normal distribution curve

Linear transformations can efficiently convert any normal distribution to a standard normal distribution. This is advantageous for calculating areas under normal distributions. Calculating the probability (P) that Z lies between two values, such as a and b ($P(a < Z < b)$), involves finding the area under the standard normal distribution curve. Notably, $P(Z = a)$ equals 0, reflecting the negligible width of a line.

The cumulative probability function, often called the cumulative distribution function, allows the calculation of $P(Z < z)$ from $Z = -\infty$ up to $Z = z$ for any value of z. Real-life random variables typically follow a normal distribution with a mean (μ) and standard deviation (σ) different from 0 and 1, respectively. However, through a process called standardization, any normally distributed variable $X \sim N(\mu, \sigma^2)$ can be transformed into a standard normally distributed variable $Z \sim N(0, 1)$ using the formula

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

This process, known as standardization, results in a standardized value called the z-score. This is possible because all normal distribution curves have the same basic bell-shape, and so can be obtained through translations and stretches.

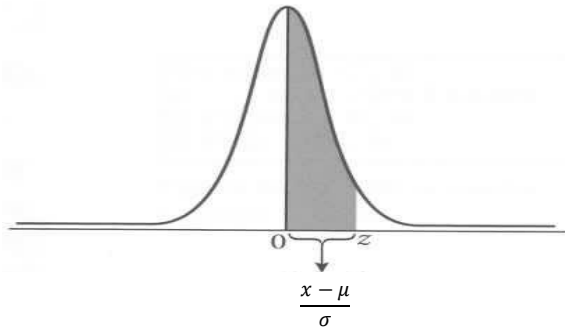


Figure 29: Transformation of the normal curve

No matter the parameters μ and σ of the original X-distribution, the standardization process yields the same standard normal distribution, $Z \sim N(0, 1)$.

To explain how this works, remember that a normal X-distribution with mean μ and standard deviation σ has probability density function

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

Consider first the normal distribution $N(0, 1)$ with $\mu = 0$, $\sigma = 1$, and probability density function becomes

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X^2}.$$

With $\mu = 0$ and $\sigma = 1$, the transformation $z = \frac{x-\mu}{\sigma}$ is simply $z = x$.

The probability density function for the standard normal distribution, denoted as $f(z)$, is $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$. It is symmetric about the mean $z = 0$, with points of inflection at $z = \pm 1$. The reverse transformation to go from $N(\mu, \sigma^2)$ to $N(0, 1)$ involves a horizontal stretch with a scale factor σ, a horizontal translation μ units to the right, and a vertical stretch with a scale factor $\frac{1}{\sigma}$ to maintain a total area under the curve of 1.

Z-scores, representing the standard deviation from the mean, are valuable for comparing populations with different μ and σ. Calculating probabilities for general normal distributions ($N(\mu, \sigma^2)$) often involved transforming data using the Z-transformation and consulting standard normal distribution tables.

The mean and variance of the standard normal variable Z are 0 and 1, respectively. This is evident from the expectation and variance calculations: $E(z) = 0$ and $Var(z) = 1$, reinforcing the standardization process's effectiveness.

Let $z = \frac{x-\mu}{\sigma}$ be the standard variable corresponding to a normal variable x.

$$E(z) = E\left(\frac{x-\mu}{\sigma}\right) = E\left(\frac{1}{\sigma}(x-\mu)\right) = \frac{1}{\sigma}E(x-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0$$

$$Var(z) = Var\left(\frac{x-\mu}{\sigma}\right) = Var\left(\frac{1}{\sigma}(x-\mu)\right) = \frac{1}{\sigma^2}Var(x-\mu) = \frac{1}{\sigma^2}Var(x) = \frac{1}{\sigma^2}\sigma^2 = 1$$

1.6.3.3. *The inverse normal distribution*

A different type of problem arises when we are provided with a cumulative probability and need to determine the corresponding value in our dataset that corresponds to this cumulative probability. In such cases, we are given a cumulative probability, $P(Z < z)$ for an unknown z, and our goal is to identify the z-value associated with this cumulative probability.

For instance, consider a scenario where the owner of a company, which fills cartons of juice with a mean volume of 150 ml, wants to find the volume below which 5% of the cartons fall, in order to reject them. In real-world applications, working backward from probabilities to estimate information about the data is often useful, necessitating the use of the inverse normal distribution.

When dealing with inverse normal calculations, particularly using the standard normal distribution $Z \sim N(0,1)$, it is important to note that in practical situations, the mean and standard deviation are unlikely to be 0 and 1, respectively. Therefore, variables need to be transformed to fit the standardized inverse normal distribution.

To illustrate, let's consider determining the age corresponding to the 95th percentile, meaning the age that is higher than or equal to 95% of the population. This involves finding the standard inverse normal number and then de-standardizing it to obtain the original data value corresponding to the given z-value.

For example, in a population of crabs where the shell length (X mm) is normally distributed with a mean of 70 mm and a standard deviation of 10 mm, a biologist aims to protect the population by allowing only the largest 5% of crabs to be harvested. The question is, ''95% of the crabs have lengths less than what?'' To answer this, we need to find the quantile, denoted as k, such that $P(X \leq k) = 0.95$. The obtained value of k is known as the 95% quantile.

Then k is found as,

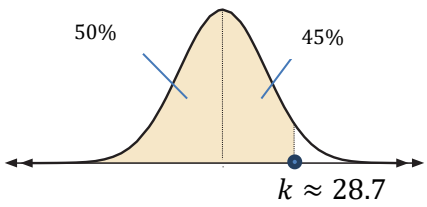


Figure 30: Inverse z-score calculation

In some instances, questions may require converting to z-scores to find the answer. Cumulative probabilities may be given, and one might be required to find either the mean (μ, if the standard deviation σ is known) or the standard deviation (σ, if the mean μ is known), or both μ and σ if both are unknown. Converting to z-scores is essential when seeking an unknown mean μ or standard deviation σ. The ability to standardize a normally distributed random variable is crucial for determining unknown values of μ and σ.

As an example, suppose an adult scallop population is known to be normally distributed with a standard deviation of 5.9 g. If 15% of scallops weigh less than 58.2 g, which means $X \sim N(\mu, 5.9^2)$, and we need to find the mean weight of the population (μ), we can use the inverse normal approach by converting to z-scores and leveraging the properties of $N(0, 1^2)$. Then,

$$P(X \leq 58.2) = 0.15 \Rightarrow P\left(Z \leq \frac{58.2 - \mu}{5.9}\right) = 0.15 \Rightarrow \frac{58.2 - \mu}{5.9} \approx -1.0364 \Rightarrow \mu \approx 64.3.$$

1.7. The Normal Approximation to the Binomial Distribution

Let's consider a scenario where X follows a binomial distribution, denoted as $X \sim B(n, p)$. Here, X represents the number of successes in n independent trials, each with a probability of success p. The probability of achieving x successes is given by the formula:

$$P(X = x) = \binom{n}{x} p^x q^{n-x},$$

where $\binom{n}{x}$ denotes the binomial coefficient, and x takes values $0, 1, 2, \cdots, n$.

This formula is typically employed for calculating probabilities associated with binomial random variables, particularly in cases where the number of trials (n) is relatively small. However, as the value of n increases, the calculation of probabilities becomes more challenging due to the substantial growth of the binomial coefficient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

Upon visualizing the binomial probability distribution of $X \sim B(n, p)$ and assigning true values for a large n and p, the shape of this distribution starts resembling that of the normal probability distribution, represented as $X \sim N(\mu, \sigma^2)$. Consequently, it becomes reasonable to approximate the binomial distribution with a normal distribution. This approximation is grounded in the understanding that, for a binomial distribution with parameters $\mu = np$ and $\sigma^2 = npq$, where $q = 1 - p$, when n is relatively large, $X \sim B(n, p)$ can be approximated as $X \sim N(np, npq)$. This process is termed the normal approximation, and it provides a means to estimate the probability of a binomial distribution when n is sufficiently large.

CHAPTER 2

STATISTICS

2. STATISTICS

Statistics theory play a role in our daily lives, and it's highly likely that you'll come across them in some form every day. For instance, the World Health Organization (WHO) regularly gathers and publishes data on the health of populations in all 192 UN member countries at 2019. One of the metrics they report is Health-Adjusted Life Expectancy (HALE). This metric is derived from life expectancy at birth but factors in time spent in poor health. Essentially, HALE represents the number of years a newborn can expect to live in full health, taking into account current rates of illness and mortality. According to WHO rankings, countries with lower income levels tend to experience more years lost due to disabilities. Several factors contribute to this disparity, including injuries, blindness, paralysis, and the debilitating effects of tropical diseases.

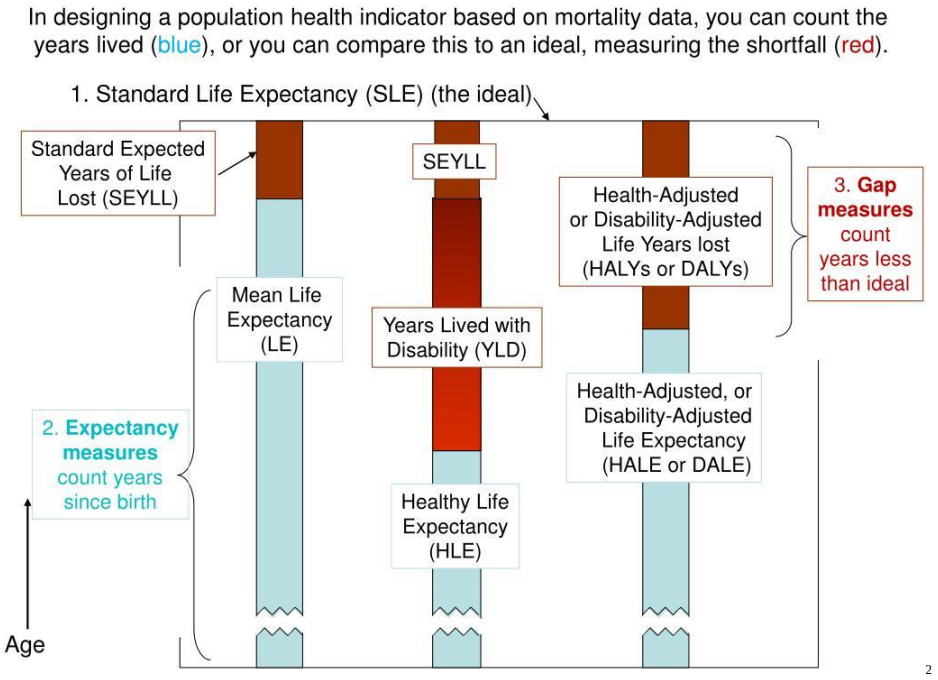


Figure 31: Standard life expectancy

Out of the 192 countries assessed by WHO, Japan boasts the highest healthy life expectancy at 75 years, while Sierra Leone has the lowest at 29 years.

Reports like these are commonplace in various publications, including business journals, newspapers, magazines, and online sources. Reading such reports often prompts questions: How was the data collected? Can we trust the results? What conclusions can be drawn? The increasing prevalence of statistical techniques across diverse fields, from business and agriculture to the social and natural sciences, underscores the importance of statistical literacy. A well-rounded education should include familiarity with the objectives and methods of these techniques.

[2]https://www.slideserve.com/murray/1-standard-life-expectancy-sle-the-ideal.

Given that statistical methods for summarizing and analyzing data provide powerful tools for interpreting the information we gather, this part begins by exploring data, evaluating randomness and then, we will introduce two fundamental components of most statistical inquiries: **populations** and **samples**. This will encompass essential techniques in descriptive statistics, which involves describing sets of measurements, both from samples and entire populations.

2.1. Exploring Data and Evaluating Randomness

Univariate data analysis focuses on processing data that encompasses a single variable ("uni" means "one"). For instance, the number of hours students spend studying each week is an example of univariate data. The primary objective of univariate analysis is to describe the data and identify patterns that enable generalizations about specific populations. Techniques such as charting, calculating central tendencies and other analytical methods can be employed to explore this data.

Figure 32: Representative sample of students, charts and average

The field of statistics enables mathematicians to gather a data set, such as this one, and then arrange, analyze, depict, and interpret it.

For example, 10 students participated in an exam with a maximum score of 72. Their individual scores are provided below.

63, 45, 10, 58, 72, 65, 22, 37, 44, 52

We may have the following questions regarding the data set:

- How should you handle this data?
- How can you arrange the data to present a clearer picture of the scores?

[3]https://www.agefotostock.com/age/en/details-photo/university-students-having-lunch-at-cafeteria/WR0596517.
[4]https://www.onlc.com/blog/10-types-tableau-charts-using/.
[5]https://www.google.com/search?q=average&rlz=1C1CHBD_trTR890TR890&tbm=isch&source=lnms&sa=X&ved=2ahUKEwjprKvNjcqAAxW6SPEDHY73AcIQ_AUoAXoECAIQAw&biw=1280&bih=571&dpr=1.5#imgrc=RqZm6r-YYGv-vM.

- What would be the most suitable way to present the scores?

2.2. Sampling

Ronald Fisher (1890-1962), a resident of both the United Kingdom and Australia, is renowned as a brilliant individual who played a nearly solitary role in establishing the fundamental principles of contemporary statistical science. His expertise in statistics was applied to examine issues across the realms of healthcare, agriculture, and the social sciences.

Let's try to think about sampling by trying to answer the questions given below:

A cohort of students aims to examine the mean average time spent on homework per night by each student in their school. They intend to survey all 1500 students in their school but realize that collecting and analyzing such a vast amount of data is time-consuming. As a result, they opt to survey a subset (sample) of the student population, hoping that the sample will provide a reliable approximation for the entire student body.

To determine the sample selection process:

- Ada proposes conducting interviews with only their close friends.
- Deniz suggests interviewing two individuals from each grade level.
- Özge recommends selecting 10 boys and 10 girls as the sample.
- Fatma suggests assigning a unique number to each student in the school and using a random number generator to choose the sample.
- Ali suggests obtaining an alphabetically organized list of all students in the school according to their surnames and selecting every 20th person on the list.

1. Engage in a group discussion to assess the merits and drawbacks of each of the five methods. Evaluate each suggestion based on the following criteria:

- Ease of obtaining the sample.
- Sample size generated by each method. Is the sample size sufficient?
- Whether the results obtained would be representative of all students in the school.

Document your findings in a table format similar to the one provided below.

Sampling Technique	Advantages	Disadvantages
Ada proposes conducting interviews with their close acquaintances.		
Deniz suggests interviewing two individuals from every grade level.		

Table 14: Sampling technique example

2. How can you integrate the suggestions of two or more students to achieve results that better represent the entire student population in the school?

3. Are there any other methods you can brainstorm to procure a sample in this particular scenario?

4. Why is it essential to take the scenario's context into account when selecting an appropriate sampling technique?

The data you gather can be categorized as either **qualitative** or **quantitative**.

Qualitative data (also known as categorical data) is expressed using words. When responding to questions such as, "How do you feel today?", the answers might include "happy," "sad," "low," or "excited".

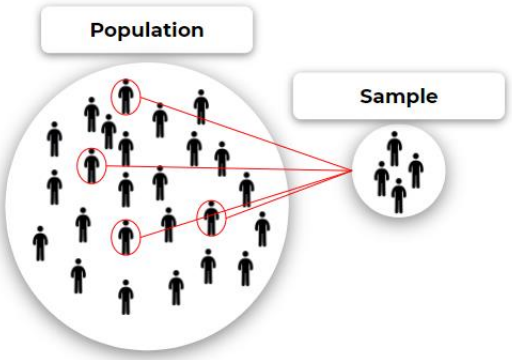
Quantitative data consists of numerical values. It involves measurable information that can be expressed in numbers. For instance, "How many individuals reside in your household?" (e.g., 2, 3, 4 or "What is the duration of your journey back home after school?" (e.g., 6 minutes, 20 minutes, 45 minutes).

Quantitative data may fall into either **discrete** or **continuous** categories.

Continuous data involves quantities that are measured and can take any real value within their range, such as height, weight, or age. On the other hand, **discrete data** pertains to quantities that are counted and can only have specific whole number values, like the number of T-shirts owned or the number of throws needed to hit a bulls eye in darts. Although these variables are typically reported with a certain level of precision, they can be measured to varying degrees of accuracy. For instance, a length measurement might be recorded as 5 cm, but in reality, it could be 5.1 cm, 5.08 cm, or 5.08236 cm. Prices are commonly considered **continuous**, despite the fact that in practice, most currencies are not subdivided below 0.01 units. Another example can be given as follows: Quantum theory, a branch of physics, proposes that at the atomic scale, all physical entities are fundamentally **discrete**. Interestingly, this concept finds support when closely examining the colors emitted by objects at various temperatures. If you're interested in delving deeper into this topic, consider researching the 'ultraviolet catastrophe.'

Could you explain the distinction between a population and a sample?

Imagine you wish to determine the count of individuals in your city who commute to work by car and have one or more passengers with them. In this case, the **population** comprises all the city residents who drive cars to work. Asking every single person in the population about their number of passengers would be highly impractical. Therefore, you would likely opt to inquire only a small group of drivers. This chosen group of drivers is referred to as a **sample**, representing a subset of the entire population. The **population** encompasses all individuals within the target group under investigation. On the other hand, a **sample** represents a smaller, representative subgroup of the population that provides insights into the entire population.



[6]

Figure 33: Population and sample

[6]https://sanket-m-kangle.medium.com/what-is-population-sample-and-sampling-error-6298d79c3771.

2.2.1. *Reasons for sampling*

Opting for sampling over conducting a complete census presents several advantages.

A census involves surveying the entire population, while a sample can save both time and money. For instance, if an eight-minute interview is planned, conducting these interviews with a sample of 100 people is clearly more cost-effective than surveying a population of 100,000. Aside from the cost benefits, the considerably smaller number of interviews generally reduces the overall time required.

Given limited resources, sampling can expand the breadth of a study. With fixed resources, a sample can yield more comprehensive information than gathering data from the entire population. Focusing on a smaller subset of individuals or items allows for a broader study scope that can accommodate more specialized inquiries.

Certain research procedures are detrimental to the product or item under investigation. For example, when testing light bulbs to determine their longevity or conducting taste tests on candy bars to assess their flavor, the products are often destroyed in the process.

In cases where access to the entire population is unattainable, sampling becomes the only viable option.

If sampling is deemed appropriate, the method for selecting a sample must be determined. Since the sample will be used to draw conclusions about the entire population, it is vital that the sample is representative of that population. It should closely mirror the pertinent parameter of the population being studied.

Selecting a sample that accurately represents the population

There are various methods available for drawing a sample from the population. It is essential to opt for a method that yields the best possible approximation of the entire population. In a cohort of students questions given above, you explored the process of selecting a sample that would provide a reliable representation of the whole population. Statisticians have defined the following key terms to clarify how a sample is taken:

- The **target population** refers to the population from which you intend to draw a sample.
- The **sampling frame** is a list of items or individuals (within the target population) from which you can select your sample.
- A **sampling unit** is a single member (e.g., an item or person) from the sampling frame that is chosen to be part of the sample.
- The **sampling variable** is the variable under investigation. This is the characteristic that you aim to measure for each sampling unit.
- The **sampling values** are the possible values that the sampling variable can take.

Let’s do the following example: To estimate the number of children in families residing in an apartment block, you need to conduct a survey. In this context, the following terms can be defined:

a) The target population consists of "all the apartments in the block."
b) The sampling unit is "each individual apartment."

c) The sampling frame is "a detailed list containing each individual apartment in the block."
d) The sampling variable is "the count of children residing in each apartment."
e) The sampling values are "0, 1, 2, 3, 4, 5, ... and so on, representing the possible number of children living in each apartment."

2.2.2. *Sampling methods*

As demonstrated in a cohort of students questions given above, various sampling techniques are employed in research. Let's delve into some of the most well-known methods:

Simple random sampling (SRS)
Each member of the population has an equal chance of being selected. A sample is chosen by drawing names from a hat or using a random number generator to assign numbers to the population. For example**,** to conduct a simple random sample and determine the mean time spent on homework, one might put the name of every student in the school into a hat and draw out 100 names to form the sample.

Systematic sampling
In this approach, the population members are listed, and a sample is selected based on a random starting point and a fixed interval. For example**,** if you wanted to create a systematic sample of 100 students from a school with an enrolled population of 1000, you would choose every tenth person from the list of all students.

Stratified sampling
The population is divided into non-overlapping smaller groups, known as strata, based on shared characteristics. A random sample is then chosen from each stratum, and these samples are combined to form the final sample. For example**,** in a high school with 1000 students, you could select 25 students from each of the four year groups to form a sample of 100 students.

Quota sampling
Similar to stratified sampling, but the sample size is taken from each stratum in proportion to the size of the population. For example**,** in a high school with 1000 students where 60% are female and 40% are male, your sample should also have 60% female and 40% male.

Convenience sampling
This is the easiest method for generating a sample. You select members of the population who are easily accessible or readily available. For example**,** to conduct a convenience sample and determine the mean time spent on homework, you might survey students who are in the same class as you.

Note: In common language, the term 'random' is often interpreted as something haphazard or without any particular pattern. However, in statistics, 'random' signifies that each member's selection probability is equal. The total of these probabilities amounts to 1.

Let's try to think by trying to answer the questions given below:

A set of five unique digits can be created from the numbers 0-9, for example, (1, 2, 3, 4, 5), (2, 4, 6, 3, 5), (0, 3, 7, 8, 9), and so on. Each sample contains distinct digits without any repetition.

1. Create several samples yourself using a random number generator or any other method.
2. Zeynep utilizes a simple random sample to select five digits. Are all the digits equally likely to be chosen? Explain the reasoning behind your answer.
3. Define the population in Zeynep's sample and compute the mean of the population.
4. Calculate the mean of each of the three samples provided at the beginning of the exploration.
5. Generate your own samples randomly and calculate the mean of these samples.
6. When the mean of a sample is not a whole number, you might have observed that the first decimal digit (the "tenths" digit) is always even. Elaborate on the reason for this phenomenon.
7. While studying the binomial theorem, you learned about the number of ways to choose objects out of a total of 11 objects. Use this knowledge to determine the number of different five-digit samples you can select from the digits 0-9.
8. Below is a frequency table displaying the values of the means of all possible samples. Complete this table by adding the relative frequencies, which involve dividing the frequency of obtaining a sample mean in that interval by the total number of sample means.

Interval	1.8-2.4	2.6-3.2	3.4	4.0	4.2-4.8	5.0-6.4	6.6-7.2	Total
Frequency	4	24	59	78	59	24	4	
Relative Frequency								

Table 15: Frequency table example

9. What information does the relative frequency provide regarding the likelihood of obtaining a sample mean within that interval?
10. Compute the mean of the sample means. How does this relate to the population mean?
11. Based on your response to question 10, why can the mean of a serve as an estimation for the mean of a population?

Let’s do the following example: A researcher aims to comprehend the demographic of the potential market for a new model of Bluetooth headphones. It has been determined that the target market consists of individuals within the following age ranges.

Age	
16-25	58%
26-35	25%
36-45	12%
46-55	5%

Table 16: Bluetooth headphones example with percentages

a) Explain how quota sampling can be employed to obtain meaningful outcomes from a sample of 200 individuals. Establish the precise quotas for the 200 individuals.

Age	
16-25	116
26-35	50
36-45	24
46-55	10

Table 17: Bluetooth headphones example with numbers

b) Examine the pros and cons of this method.

Advantages:
The sampling process is rapid, uncomplicated, and convenient.

Disadvantages:
In practice, it might be challenging to locate the suitable number of participants, and inquiring about their age could be a sensitive matter.

The researcher must assess the distribution of subgroups within the population and maintain that proportion in the sample selected through this sampling method. For instance, if 58% of the potential buyers for Bluetooth headphones are between the ages of 16-25, the sample should also have the same percentage of individuals belonging to that specific age group. Quota sampling can effectively represent a population, enabling researchers to study specific subgroups within the population using predetermined quotas.

The purpose of sampling is to collect information that represents the broader population using an efficient and effective method. It is crucial to ensure that each sample is representative of the population and that bias is minimized or eliminated.

Sources of bias can include:

i) Exclusion of certain population members from the sampling frame
Conducting a survey by calling members of the population using a systematic sample may exclude individuals without a fixed phone line or those not listed in the directory.

ii) Non-response
A university sends out a mail survey, and only certain demographics, such as people over 50 or those with English as their first language, respond, potentially introducing bias to the survey results.

iii) Poor design
It is essential for the questionnaire to be clear, unambiguous, and free of leading questions to avoid eliciting misleading information from respondents.

iv) Bias by the respondent
This type of bias is challenging to eliminate as people may not be forthcoming with negative information. Personal questions about health, weight, or income are common areas where respondents may provide untruthful answers.

Let's try to think by trying to answer the questions given above:

In this investigation, you will compare the estimates of the mean length of a population of pencils using various sampling methods.

Required equipment: A selection of 50 colored pencils (preferably of five different colors) with different lengths.

1. Select four different samples of five pencils each, using the following sampling techniques: **A)** Simple random sample **B)** Systematic sample **C)** Stratified sample **D)** Quota sample

2. Determine the mean length for each of the four samples. Create frequency tables for the mean obtained from each sampling method.
3. Calculate the mean for each sample.
4. Compare the mean of each sample with the mean of the population lengths.
5. Which sampling technique resulted in the most accurate estimation of the population mean?
6. Which sampling technique would minimize bias and provide the best representation of the population?
7. How can you mitigate bias and ensure a more representative sample?

2.3. Reliability of Data

Data is considered reliable when the process of data collection can be replicated, resulting in consistent outcomes. For instance, if you were to conduct a survey multiple times, would you obtain similar findings each time? On the other hand, data is considered sufficient when there is an adequate amount of data available to support your conclusions. While determining the precise number of data items needed is not fixed, it is essential to gather a sufficient amount of data to ensure repeatability and representativeness, meaning that the results accurately reflect the entire population.

Two factors that can lead to unreliable data are as follows:

1. **Missing data** is a prevalent issue in various research studies. Missing data may arise due to

a) Non-responses to questionnaires or surveys.

b) The inability to record data under specific circumstances. For example, when surveying the number of cars on a road at different times of the day, data collection might not be feasible during nighttime as the researcher would likely be asleep.

Missing data compromises the validity of a sample and can distort inferences about the population. Missing values are automatically excluded from analysis. A small number of missing data points, like 10 in a sample of 1000, may not pose significant problems. However, if they constitute 20% or more of the sample, it becomes a serious concern.

In surveys:

• One way to minimize missing data is to avoid asking questions where respondents might choose "not available" or "don't know" as answers.

• Missing data might also result from certain questions being applicable only to specific individuals within the surveyed group. For example, if a question was relevant only to students who took a particular test, data might appear to be missing for other students in the school.

2. **Errors in handling data** can also lead to unreliable results. Data might be entered incorrectly, or columns within a table might become disorganized, affecting the final outcomes derived from the data.

To prevent these issues, it is crucial to closely monitor and verify the data collection process to minimize errors. Regular checks and validations should be conducted to ensure the accuracy and reliability of the data obtained.

2.4. Descriptive Statistics

The phrase "Lies, damned lies, and statistics" is commonly associated with Benjamin Disraeli (1804-81) and may have evolved from a courtroom saying that includes "liars, damned liars, and expert witnesses." Throughout time, statistics have been employed to advocate for beliefs and viewpoints by selectively crafting the analysis and portrayal of data.

2.4.1. *Measures of central tendency*

Measures of central tendency provide insights into the central values within a dataset. The three primary measures of central tendency are the mode, the mean, and the median.

Arithmetic mean

The arithmetic mean is the result of dividing the sum of all data values by the number of data values in the population. For calculating the arithmetic mean from grouped data, the midpoint of each class is used. Within a dataset, the mean is expressed as $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$, where $\sum_{i=1}^{k} f_i x_i$ represents the total of the data set and n represents the number of values in the data set. When considering the entire dataset, we express the mean as $\mu = \frac{\sum_{i=1}^{k} f_i x_i}{n}$.

Mode

The median was popularized by the 19th century German psychologist Gustav Fechner, although it had been previously employed by the French mathematician Pierre Laplace.

The mode represents the value (or values) that appear most frequently within a dataset. If there exists a singular value (or interval in the case of grouped data) that holds the highest frequency, that particular value becomes the mode, and its corresponding frequency is termed the modal value. In the context of a frequency table, the mode corresponds to the value or interval with the highest frequency.

Median

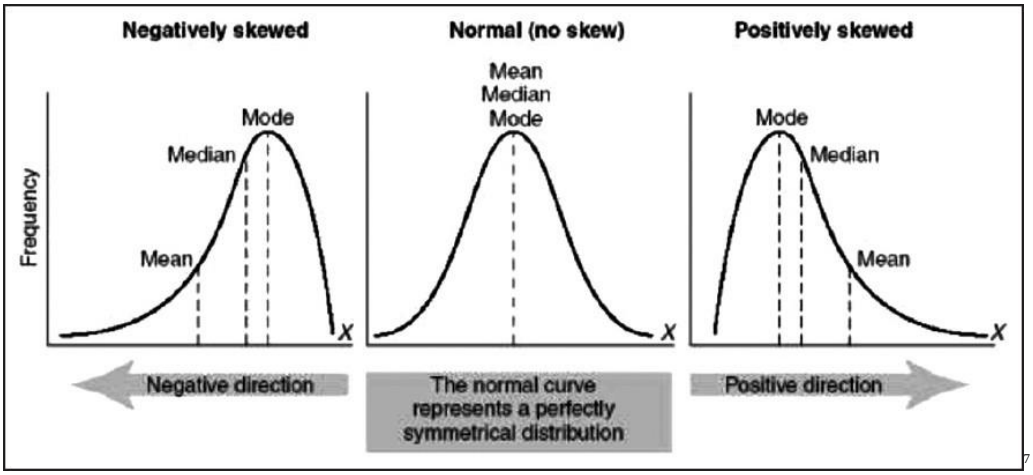
The median corresponds to the value located at the middle position when the data is organized in ascending order of magnitude. If the count of data values, denoted as n is odd, the median is the term situated at the position of $(n + 1/2)$. In situations where the count of data values is even, the median equates to the arithmetic mean of the terms at positions $(n/2)$th and $(n/2 + 1)$th.

2.4.2. *Measures of distribution*

The configuration of a distribution illustrates how the data is positioned in relation to the mean. Distributions can be either symmetrical or nonsymmetrical. When they are not symmetrical, we describe the distribution shape as either asymmetrical or skewed.

Symmetry: A distribution is deemed symmetrical when the data points are evenly dispersed around the mean. In a symmetrical distribution, the mean, median, and mode are all equal, as explained in the subsequent section.

Skewness: A distribution is considered skewed when the data points are not evenly distributed above and below the mean. A positively skewed (or right-skewed) distribution displays a tail extending towards positive values on the right side. Conversely, a negatively skewed (or left-skewed) distribution exhibits a tail extending towards negative values on the left side.



[7]

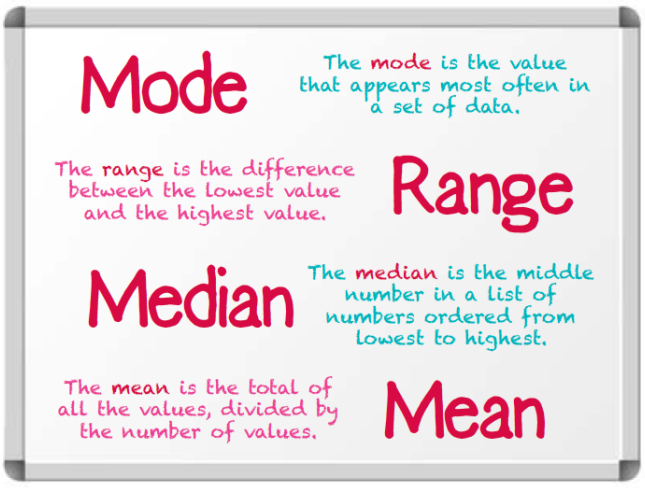
Figure 34: Skewed types

2.4.3. *Measures of variability*

Measures of variability provide insights into how the data set is spread out. The range, interquartile range, and standard deviation are the three most frequently used measures of variability.

Range

The range is calculated as the gap between the highest and lowest values within the dataset. The range serves as a straightforward indicator of the extent to which the data is dispersed. In the initial problem, you analyzed the examination scores of 10 students: 43, 25, 95, 72, 60, 52, 10, 100, 31, 67. Here, the minimum test score is 10 and the maximum test score is 100. So, the range value is calculated as 100-10=90. With this, gaining insight into whether the data is extensively dispersed or clustered near the mean enhances your comprehension of the overall data distribution. It's essential to take into account a broader range of values beyond solely the smallest and largest ones.



[8]

Figure 35: Mean, median, mode and range

[7] https://www.fromthegenesis.com/skewness/.

[8] https://www.theschoolrun.com/what-are-mode-mean-median-and-range.

Quartiles

Quartiles are values that partition your data into four equal parts.

• The first quartile (also known as $Q_1 = \left(\frac{n+1}{4}\right)$th or **lower quartile**) encompasses 25% of the data below it.

• The second quartile represents the median, which is the central value in the dataset. 50% of the data falls below the median, and the remaining 50% is above it.

• The third quartile (referred to as $Q_3 = 3\left(\frac{n+1}{4}\right)$th or **upper quartile**) includes 75% of the data below it.

Question:

"The term "interquartile range" is often referred to as the range that covers the central half of the data. Clarify why this description is accurate.

An important benefit of using the interquartile range is that it provides a more reliable measure of data spread when the presence of extreme values distorts the regular range

The interquartile range represents the spread measurement linked with the median: $IQR = Q_3 - Q_1$.

2.5. Histograms

When dealing with a large dataset, it is often convenient to organize it using a grouped frequency table. If the data is continuous, you can create a histogram.

A histogram bears resemblance to a bar chart, but a bar chart is appropriate only for discrete data. For continuous data, a histogram should be employed.

The key distinctions between a histogram and a bar chart are as follows:

- Histogram bars are contiguous, lacking gaps, since they depict continuous data.
- The horizontal axis of a histogram employs a continuous scale, whereas a bar chart's horizontal axis is based on discrete categories.
- In a bar chart, the width remains uniform and doesn't correlate with the x-scale. In a histogram, each rectangle's width should encompass the range of the represented data class.

To create a meaningful histogram, it's typically recommended to use between 5 and 15 classes. Begin by determining the minimum and maximum values, then establish appropriate class widths based on this information.

Steps to manually create a histogram with equal-sized classes:

• Record the class intervals on the horizontal axis.
• Plot the *frequency* on the vertical axis.
• Carefully determine the scale to ensure the histogram fits the page.
• Draw the upper boundary line of each class interval at the corresponding frequency.
• Utilize a ruler to draw rectangles representing the data. Ensure there are no gaps.

The class width is the range between the highest and lowest potential values within each class interval. For instance, in the class interval $18 < x \leq 20$, the class width is 2, derived from the calculation $20 - 18 = 2$. Although the weights of chicks are inherently continuous, they have been rounded to one decimal place for measurement purposes, resulting in discrete values. Despite this discretization, the data remains continuous in nature, thus warranting the use of a histogram.

Let's do the following example:

The recorded data in the table below pertains to the weights of 50 baby chicks that were hatched within a span of one week.

22.4	20.0	23.3	25.2	24.1	23.6	19.7	21.3	24.0	24.7
23.6	22.3	21.4	22.4	23.9	25.0	22.2	23.5	21.9	23.0
24.3	21.6	19.9	21.2	23.5	24.6	23.5	19.9	20.2	20.3
23.0	24.6	23.5	22.5	23.0	24.5	24.9	21.3	20.5	20.4
21.4	22.4	23.6	24.5	23.4	22.9	23.8	24.6	21.0	21.3

Table 18: The weighted of 50 baby chicks data

a) Arrange the information into a tabular form by classifying the data into intervals of the same width.

Class interval	Frequency
$19.5 < x \leq 20$	3
$20 < x \leq 20.5$	5
$20.5 < x \leq 21$	1
$21 < x \leq 21.5$	6
$21.5 < x \leq 22$	2
$22 < x \leq 22.5$	6
$22.5 < x \leq 23$	4
$23 < x \leq 23.5$	6
$23.5 < x \leq 24$	5
$24 < x \leq 24.5$	5
$24.5 < x \leq 25$	6
$25 < x \leq 25.5$	1

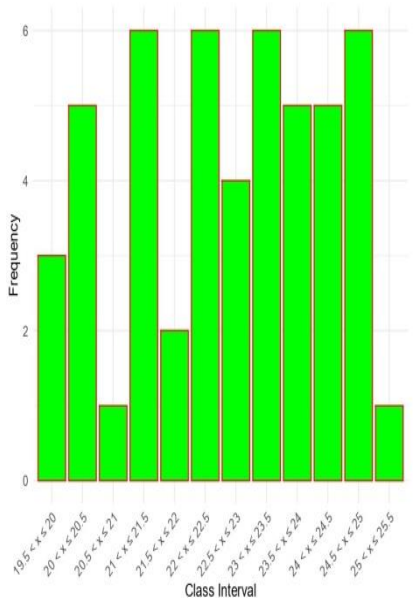
Table 19: Classified 50 baby chicks data

Start by finding the minimum, maximum and range:

- Minimum= 19.7
- Maximum= 25.2

• Range=5.5. Given that the range equals 5.5 and our objective is to have a distribution of 5 to 15 categories, opting for a class interval of 0.5 seems appropriate, as it results in $5.5/0.5 = 11$ intervals. For the sake of tidiness and simplicity, commencing at 19.5 instead of 19.7 as the initial value would be more convenient.

b) Generate a frequency histogram to portray the data, depicting the distribution.



R 4.2.1 code

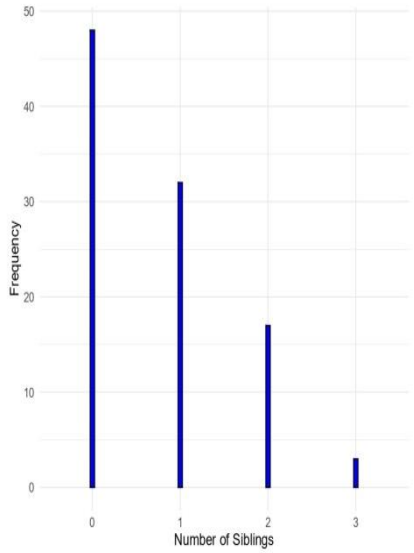
```
class_intervals <- c("19.5 < x ≤ 20", "20 < x ≤ 20.5", "20.5 < x ≤ 21", "21 < x ≤ 21.5", "21.5 < x ≤ 22", "22 < x ≤ 22.5", "22.5 < x ≤ 23", "23 < x ≤ 23.5", "23.5 < x ≤ 24", "24 < x ≤ 24.5", "24.5 < x ≤ 25", "25 < x ≤ 25.5")

frequency <- c(3, 5, 1, 6, 2, 6, 4, 6, 5, 5, 6, 1)

data <- data.frame(Class_Interval = class_intervals, Frequency = frequency)

library(ggplot2)
ggplot(data, aes(x = Class_Interval, y = Frequency)) +
  geom_bar(stat = "identity", fill = "green", color = "red") +
  labs(x = "Class Interval", y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggsave(filename = "histogram.jpg", plot = plot)
```

Figure 36: Plot of class interval versus frequency

Histograms of frequencies for discrete data

The count of siblings among 100 students in a class	
Number of siblings	**Frequency**
0	48
1	32
2	17
3	3

Table 20: Siblings data



R 4.2.1 code

```
number_of_siblings <- c(0, 1, 2, 3)
frequency <- c(48, 32, 17, 3)

# install.packages("ggplot2")
library(ggplot2)

data <- data.frame(Number_of_Siblings = factor(number_of_siblings), Frequency = frequency)

ggplot(data, aes(x = Number_of_Siblings, y = Frequency)) +
  geom_col(fill = "blue", color = "black", width = 0.05) +
  labs(x = "Number of Siblings", y = "Frequency") +
  theme_minimal()
```

Figure 37: Plot of number of siblings versus frequency

When dealing with discrete data, as shown in the table above, it's common practice to use a vertical bar graph, as demonstrated on the left side. However, this approach poses challenges when dealing with grouped data. For instance, if we want to represent the category '1-3 eggs,' simply drawing a single vertical bar would be misleading. Likewise, drawing separate bars for 1, 2, and 3 could lead to confusion. Hence, to address this issue, we can opt to depict the '1-3 eggs' category as a rectangular shape that starts at 0.5 and ends at 3.5.

When inspecting a histogram, you can begin to visualize specific characteristics of the dataset from which the histogram was generated. When providing comments on **the data distribution**, you are essentially summarizing some of the fundamental aspects of how the data is structured. There are four essential elements that you should always aim to cover, and you can use the acronym CSOS to help remember them:

Center

- Seek the mean, median, and mode.

Spread

- Observe the range, maximum, and minimum.

Outliers

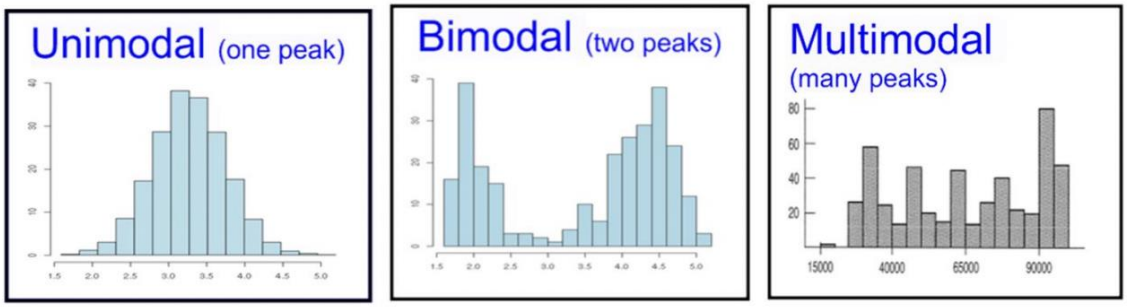
- In the histogram, there may be data points that appear unusual. An outlier is defined as a data point that falls more than 1.5 times the Interquartile Range (IQR) above the upper quartile or less than 1.5 times the IQR below the lower quartile.

Shape

- Analyze whether it is unimodal, bimodal, or multimodal, and determine whether it exhibits symmetric or skewed.

When comparing two distributions, it is crucial to be concise and clear in your description of the distributions. Use terms like "wider," "narrower," "more varied," or "less varied."

If direct calculations for the mean, median, or mode cannot be made, you can provide estimates. To estimate the mean for grouped data, you would need to organize the data into a frequency table."



[9]

Figure 38: Frequency tables

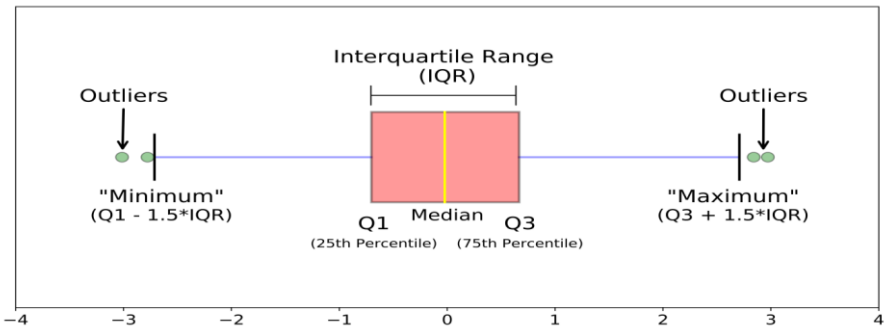
2.6. Cognizance for Statistical Approaches

2.6.1. *Box-and-Whisker diagrams*

Statistical methods can be justified through the use of summary statistics, which encompass:

1. the minimum value
2. the first quartile
3. the median, also known as the second quartile
4. the third quartile
5. the maximum value.

These five key data points can be visually represented as a **box-and-whisker diagram**, alternatively referred to as a box plot.



[10]

Figure 39: Box-and-whisker diagram

[9]https://datasci.soniaspindt.com/StatisticalFoundations/DescriptiveStats/CentralTendency/mode.html.
[10]https://www.kdnuggets.com/2019/11/understanding-boxplots.html.

The first and third quartiles are positioned at the extremities of the rectangular shape (the box), while the median is represented by a vertical line within the box. The minimum and maximum data points are situated at the extremities of the horizontal lines (the whiskers). The diagram should be accurately drawn along a sole horizontal axis. This box-and-whisker diagram displays the test data for class scores presented earlier in this chapter.

2.6.2. *Outliers*

Outliers may appear within a dataset. Such occurrences can result from data collection errors or inherent irregularities within the population. One approach to detecting outliers involves applying the concept of IQR. According to this principle:

An outlier is characterized as a data point that falls either more than 1.5 times below the lower quartile or exceeds 1.5 times the upper quartile.

2.6.3. *Variance and standard deviation*

While the range and quartiles provide valuable insights into the dispersion of data around the median, they do not utilize the entire data set to convey a comprehensive understanding of its spread. In contrast, variance and standard deviation incorporate all data values within a dataset to offer a measure of the data's dispersion relative to its mean. Specifically, the standard deviation quantifies the average extent to which each data point deviates from the mean.

Let's try to think by trying to answer the questions given below:

The temperature (°C) has been documented for the initial day of the opening five months in the calendar year within two distinct municipalities, Whitby and Mullion Cove.

	Whitby	**Mullion Cove**
January	10	18
February	16	10
March	14	13.5
April	12	14
May	18	14.5

Table 21: Whitby and Mullion data

1. Compute the three central tendency measures (mean, mode, and median) along with the range.

2. Compare these results. Now, you can explore if there exists another statistical measure that can effectively differentiate between the two sets of data.

For the data from Whitby:

3. Determine the deviation of each individual value from the mean.

4. Observe the sign (positive or negative) of these deviations.

5. How can you ensure that all deviations share the same sign?

6. What is the unit of measurement at this stage?

7. Calculate the mean of these deviation values.

8. To revert the unit back to its original quantity, what mathematical operation is necessary?

9. Compute this final value as the mean of deviations from the mean.

10. How does this method identify and describe data dispersion?

11. Can you formulate this algorithm into a mathematical formula?

The standard deviation is the square root of the variance.

Standard deviation provides a measure of the average distance between each data point and the mean.

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

Variance calculates the mean average of the squared differences between each data point and the mean.

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

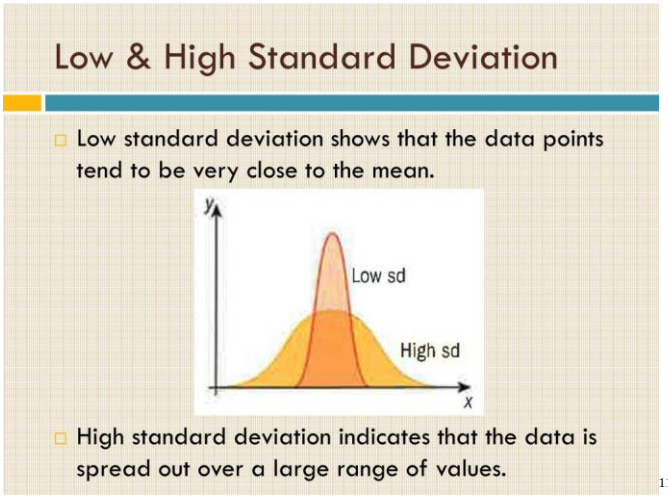
Characteristics of the standard deviation:

• The standard deviation is exclusively utilized to gauge the extent or dispersion of data points around the dataset's mean.

• Variance (and consequently, the standard deviation) is always non-negative, as it represents an average of squared distances $(x - \bar{x})^2$.

• The standard deviation is highly responsive to outliers. A single outlier can significantly increase the standard deviation, thereby distorting the perception of data spread.

• For data sets with approximately similar means, a larger spread results in a correspondingly greater standard deviation.



[11]

Figure 40: Low and high standard deviation

[11]https:// www.slideserve.com/tammy/variance-standard-deviation .

When the data set exhibits minimal dispersion, the standard deviation will be small. Conversely, if the data is widely distributed, the standard deviation will be substantial. The standard deviation shares units with the original data set.

Let's do the following examples:

An ensemble of 40 students was surveyed about their annual dentist visits. They provided the following responses: 3, 0, 2, 5, 7, 6, 8, 0, 4, 1, 6, 3, 0, 5, 6, 5, 3, 6, 2, 7, 6, 0, 4, 4, 6, 6, 5, 7, 0, 1, 2, 5, 8, 0, 4, 3, 4, 6, 7, 5. Calculate the average and standard deviation for this data set.

$$\bar{x} = \frac{3 + 0 + \cdots \ldots \ldots \ldots . + 7 + 5}{40} = \frac{160}{40} = 4$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - 4)^2}{40}} = \sqrt{\frac{(3-4)^2 + (0-4)^2 + \cdots \ldots \ldots \ldots + (7-4)^2 + (5-4)^2}{40}}$$
$$= 2.39$$

a) Calculate the average and median of the number of text messages sent by a group of 15 students during a one-week residential trip. The data points are as follows: 36, 40.12, 0, 15, 25.25, 78, 45, 28, 18, 3, 15, 19, 20.

$$\bar{x} = \frac{36 + 40 + \cdots \ldots \ldots \ldots . + 19 + 20}{15} = 25.3$$

If the count of data values, denoted as $n = 15$ is odd, the median is the term situated at the position of $(n + 1/2)$ =8. When we list the numbers from smallest to largest, we can see that the 8th number, the median, is 20.

b) Determine the range between the lower quartile (Q_1) and the upper quartile (Q_3) to find the interquartile range.

The lower quartile (Q_1) value lies in the $\frac{n+1}{4}$th position when arranged in order. Therefore, 4th position=15. The upper quartile (Q_3) value lies in the $\frac{3(n+1)}{4}$th position when arranged in order. Therefore, 12th position=36.

$$IQR = Q_3 - Q_1 = 36 - 15 = 21.$$

c) Identify if any of the data points can be classified as outliers.

An outlier is defined as a data point that falls below 1.5 times the interquartile range (1.5xIQR) below Q_1 or above 1.5 times the interquartile range (1.5xIQR) above Q_3.

In this case:

$$1.5xIQR \; = \; 31.5$$

$$15 \; - \; 31.5 \; = -16.5 \; (which\ is\ not\ possible)$$

$$36 \; + \; 31.5 \; = \; 67.5$$

Therefore, we can conclude that the data point 78 can be considered an outlier because it is greater than 67.5, which is 1.5 times the interquartile range above Q_3.

When computing the standard deviation, it is essential to determine whether your sample accurately represents the entire population. To make predictions about the standard deviation of the entire population, it's crucial to acknowledge that the population size is significantly larger than the sample, which increases the likelihood of a wider dispersion. As an illustration, consider a sample consisting of the weights of 25 lemons (in grams) as presented below.

132	122	132	125	134
129	130	131	133	129
126	132	133	133	131
133	138	135	135	134
142	140	136	132	135

Table 22: The weights of 25 lemons data

You've received the unprocessed data, so the mean and standard deviation are computed using the following formulas:

$$\bar{x} = \frac{\sum x}{n} = 132.48 \text{ g and } s_n = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = 4.28 \text{ g.}$$

You do not possess access to all available data, which necessitates the estimation of the population's mean and standard deviation from this sample. To *estimate* the population mean, you can employ the sample mean, which is an unbiased estimator of the population mean. We denote this estimator as 'μ.'

There exists a single distinction in the calculation process for estimating the population standard deviation. When determining the population standard deviation, you divide the sum of squared deviations from the mean by the total number of items in the population (in this case, 25).

However, when computing an estimate of the population standard deviation from a sample, you divide the sum of squared deviations from the mean by the number of items in the sample minus one. In this instance, you would divide by 24 (equivalent to 25-1).

Consequently, the computed sample standard deviation will be slightly higher compared to if you had used the population standard deviation formula. This adjustment aims to provide a more accurate and unbiased estimate of the population's standard deviation, considering that the sample may not precisely reflect the entire population's full range.

The unbiased estimate for the population mean equals the sample mean.

$$\bar{x} = \mu$$

Similarly, the unbiased estimate for the population standard deviation is

$$s_{n-1} = \sqrt{\frac{\sum(x-\mu)^2}{n-1}}$$

When dealing with a sample and seeking to make predictions about the entire population, employ these values as unbiased estimates.

It's always essential to use $'n-1'$ as the denominator in the formula. When computing the population standard deviation, we rely on the sample mean and predict each deviation from this sample mean. In cases where one of these measurements (either the sample mean or a deviation) is missing, it can be

calculated using the available information. So, if you have 'n' data points, only $'n-1'$ of them possess independent variation. Consequently, $'n-1'$ is referred to as our 'degrees of freedom,' and we substitute $'n'$ with $'n-1'$. This adjustment is known as **Bessel's correction**.

Let's do the following example:

An apparatus manufactures 1000 marbles daily, each having an average diameter of 1 cm. A sample of eight marbles was selected from the manufacturing process, and their diameters were measured. The recorded measurements in centimeters were as follows: 1.0, 1.1, 1.0, 0.8, 1.4, 1.3, 0.9, 1.1. Calculate the standard deviation of the machine's production for the entire day.

$$\bar{x} = \frac{1.0 + 1.1 + \cdots \ldots \ldots \ldots . + 0.9 + 1.1}{8} = 1.075$$

$$s_{n-1} = \sqrt{\frac{\sum(x-\mu)^2}{n-1}}$$

$$= \sqrt{\frac{(1.0-1.075)^2 + (1.1-1.075)^2 + \cdots \ldots \ldots \ldots . + (0.9-1.075)^2 + (1.1-1.075)^2}{8-1}} = \sqrt{\frac{0.275}{7}}$$

$$= 0.1982\ cm$$

2.6.4. *Cumulative frequency*

When dealing with unprocessed data (such as a list of numbers), you can use the formula median = $(n+1)/2$th value to determine the median and quartiles when the data is sorted in ascending order. However, in the case of grouped data, accurately identifying the median or quartile becomes challenging when the $(n+1)/2$th value falls within a group.

Cumulative frequency curves offer a solution for determining the median and quartiles from a dataset that has been grouped. To find the median, you draw a horizontal line extending from the frequency axis to the curve at the $(n+1)/2$th value, and then extend it downward to intersect with the x-axis, which gives you the median data point. The process for finding quartiles is quite similar.

Cumulative frequency distributions

It is beneficial to include a cumulative frequency column when calculating the class interval that contains the median.

The representation of the cumulative frequency curve can also be utilized for computing the five-point summary and other percentiles when dealing with grouped data.

You can take the five-point summary derived from the cumulative frequency curve and depict it on the same graph either above or below the x-axis, forming a box-and-whisker plot.

The actual cumulative frequency curve is commonly known as an **ogive**, and different frequency distributions lead to distinct ogive shapes.

Important guidelines for constructing a cumulative frequency graph include:

- Cumulative frequency values are depicted on the y-axis, while the variable is shown on the x-axis.

- Start by plotting 0 (cumulative frequency) against the lower boundary of the lowest class, as all values must be above this point.
- Each cumulative frequency value is plotted against the upper boundary of its corresponding class.
- The cumulative frequency value associated with the highest class must equal the total frequency.
- Connect all points with a smooth curve.

Note: When confronted with grouped data, the process of approximating percentiles and quartiles is made more straightforward through the utilization of a cumulative frequency curve.

Note: Outliers in the box-and-whisker diagram should be marked with an X.

Note: The term '0give' has its roots in the French language, signifying a pointed or gothic arch. Galton was the pioneer statistician to employ it in the context of denoting a cumulative frequency curve.

2.7. Covariance, Correlation, Causation and Linear Regression

Bivariate data

In 1956, the Australian statistician Oliver Lancaster presented the initial compelling argument for a connection between sunlight exposure and the development of skin cancer, employing statistical techniques such as correlation and regression.

Several techniques presented in this part are utilized when analyzing single-variable data. In certain scenarios, multiple distinct attributes can be assessed for each data point, and an examination can be conducted to determine if these variables are correlated. This is referred to as analyzing **bivariate data**.

Scatterplots

A scatterplot illustrates the connection between two quantitative variables for each individual within the data set. One variable's values are plotted along the x-axis, while the other variable's values are represented on the y-axis.

Each data point corresponds to an individual in the data set.

Note: A **response variable (or dependent variable)** quantifies the result or outcome of a research endeavor, while an **explanatory variable (or independent variable)** elucidates the alterations in the response variable. For instance, in a study examining the relationship between body weight and blood pressure, body weight serves as the explanatory variable, and blood pressure serves as the response variable. In a study aimed at investigating the connection between fertilizer dosage and crop yield during a farming season, the fertilizer dosage constitutes the explanatory variable, while the crop yield represents the response variable. Ensure that the x-axis represents the independent variable, while the y-axis represents the dependent variable in every plot.

In contrast to univariate graphs, bivariate graphs typically display a spread or "scatter" of points that exhibit considerable variability. Even when a strong underlying linear relationship exists between the variables, the actual data points may not align precisely with the specific linear equation.

When describing a scatterplot, it's essential to seek out patterns and categorize your observations in terms of:

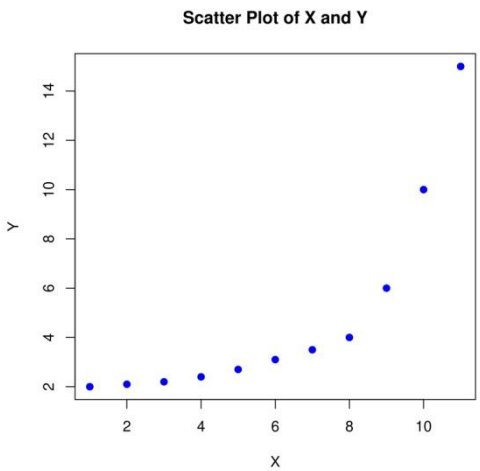
1. **Association:** If the pattern inclines upward from left to right, it indicates a **positive** association; if it slopes downward from left to right, it signifies a **negative** association.
2. **Form:** This refers to the overall shape of the pattern, often resembling a **linear** shape.
3. **Strength:** Determined by how closely the points adhere to the specified association and form.

Pearson product moment correlation coefficient

The Pearson product moment correlation coefficient (PMCC) assesses the degree of linear association between two variables. What occurs when there seems to be an association, but it doesn't seem to follow a linear pattern? You can illustrate these data on a scatter plot as depicted:

X	1	2	3	4	5	6	7	8	9	10	11
Y	2	2.1	2.2	2.4	2.7	3.1	3.5	4	6	10	15

Table 23: X and Y values



R 4.2.1. code

```
# Define the data

x<-c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)

y<-c(2, 2.1, 2.2, 2.4, 2.7, 3.1, 3.5, 4, 6, 10, 15)

# Create a scatter plot

plot(x, y, main="Scatter Plot of x and y", xlab="x", ylab="y", pch=19, col="blue")
```

Figure 41: Plot of X versus Y

- ✓ Do you believe there exists a connection between the variables x and y?
- ✓ Is the association robust? Describe the connection.
- ✓ Calculate the Pearson's correlation coefficient, denoted as 'r,' for this data set.
- ✓ What insights does it provide regarding the correlation in this data set?
- ✓ Elaborate on why the connection is highly significant, yet the correlation coefficient fails to represent this accurately.

Sperman rank correlation

You can also utilize Spearman's rank correlation to analyze the association:

Spearman's correlation assesses the connection between variables, which need not be linear. The Spearman correlation coefficient relies on the **ranked values** of each variable rather than the **original data**.

What are the primary distinctions between the Pearson correlation and the Spearman correlation? The formula for Spearman's correlation, denoted as 'r_s' is as follows:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

where D represents the difference between pairs of scores, and 'n' is the count of rank pairs. Investigate how to interpret Spearman's rank correlation. What is the Spearman's rank value for the provided data set? What insights does this 'r_s' value provide about these data?

Covariance

Intuitively, when we consider the relationship between two variables, X and Y, we typically expect that one variable, such as Y, will either rise or fall as X undergoes changes. Our focus will be on two indicators of this relationship: the **covariance** between two random variables and their **correlation coefficient**.

If X and Y are random variables with means μ_X and μ_Y, the covariance of X and Y is calculated as

$$cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

A larger absolute value of the covariance between X and Y signifies a stronger linear relationship between the two variables. Positive values indicate that Y tends to increase as X increases, while negative values suggest that Y tends to decrease as X increases. A covariance value of zero indicates that the variables are linearly uncorrelated, meaning there is no linear connection between X and Y.

Here are some key points to understand about covariance:

1. If you ever need to calculate covariance without relying on built-in functions in your calculator or software, you can use this shortcut formula:

$$\begin{aligned} cov(X,Y) = E\big((X - \mu_X)(Y - \mu_Y)\big) &= E(XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y) \\ &= E(XY) - E(X\mu_Y) - E(\mu_X Y) + E(\mu_X\mu_Y) \\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y \end{aligned}$$

2. This result also leads to:

$$cov(X,X) = E(XX) - \mu_X\mu_X = E(X^2) - \mu_X^2 = V(X)$$

3. If X and Y are **not independent**, then:

$$V(X + Y) = V(X) + 2cov(X,Y) + V(Y)$$

4. If X and Y are **independent**, then:

$$cov(X,Y) = E(XY) - \mu_X\mu_Y = E(X)E(Y) - \mu_X\mu_Y = 0$$

   Consequently:

$$V(X + Y) = V(X) + V(Y)$$

It's important to note that the converse of this theorem is not necessarily true: if $cov(X,Y) = 0$, X and Y may still not be independent. This issue with covariance can be resolved by standardizing its value using the **correlation coefficient**, 'ρ' (or 'r'):

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

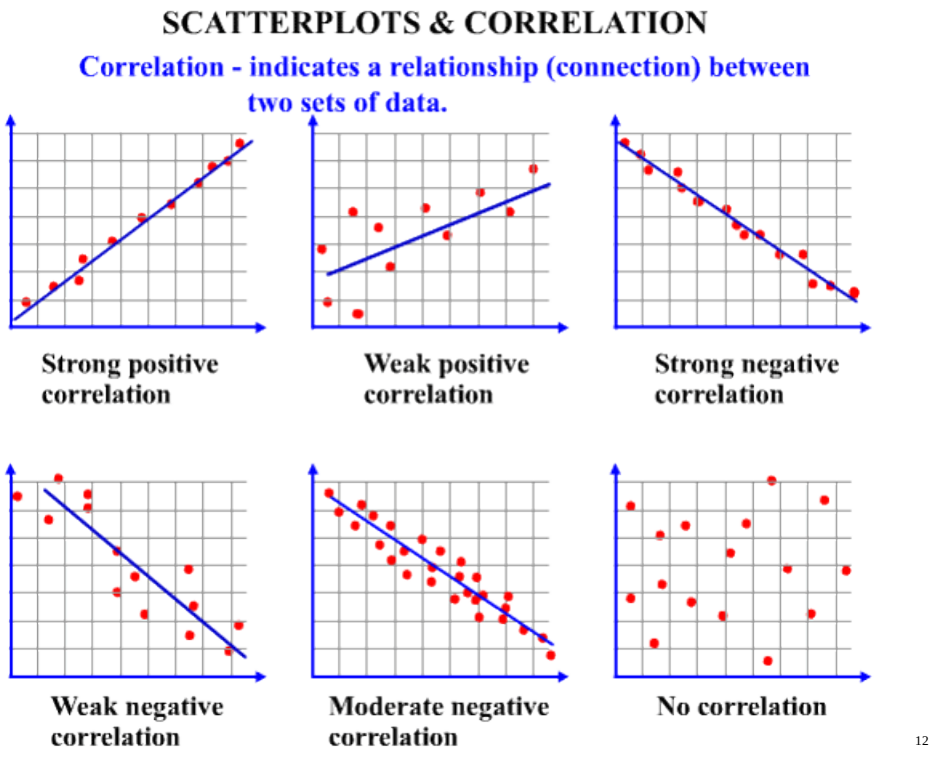
Since both σ_X and σ_Y are positive, the sign of the correlation coefficient matches that of the covariance. It's important to note that all the models discussed regarding correlation and regression assume that the data are samples from normal populations. Correlation can also be characterized as strong, moderate, or weak.

Note: The correlation coefficient 'ρ' (or 'r') gauges both the direction and the magnitude of the linear connection between two variables. It can be expressed either in words or as a numerical value.

Note: At this point, we utilize the table provided below to assess the degree of correlation. The 'ρ' values span from -1 to 1.

Strong negative	**Moderate negative**	**Weak negative**	**Very weak negative**	**Zero**	**Very weak positive**	**Weak positive**	**Moderate positive**	**Strong positive**
-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1

Table 24: The degrees of correlation



[12]

Figure 42: Scatterplots and correlation

This part also offers several methods to determine the correlation from a scatter plot. **Method 1** involves deriving the correlation from the best-fit line, while **Method 2** utilizes PMCC. Additionally, it could be said that the correlation coefficient 'ρ' or 'r' equals to PMCC value. In the upcoming exploration, you will employ the first method to visually identify the best-fit line.

Let's try to think by trying to answer the questions given below:

[12]https://www.aplustopper.com/scatter-plots-correlation/.

A dozen students participated in a ski race where rankings were determined based on the cumulative times of two separate runs. One of the students aimed to investigate if there existed a connection between the durations of these two runs and diligently documented the time data. The outcomes are displayed in the table.

Pupil	A	B	C	D	E	F	G	H	I	J	K	L	Total	Mean
Run 1	53.3	56.7	53.8	54	61.3	62.5	56.7	58.9	61.0	58.7	70.1	56.8		
Run 2	54.3	57.6	53.9	55.6	67.5	63.4	55.1	57.8	68.9	62.5	66.6	57.7		

Table 25: A ski race data

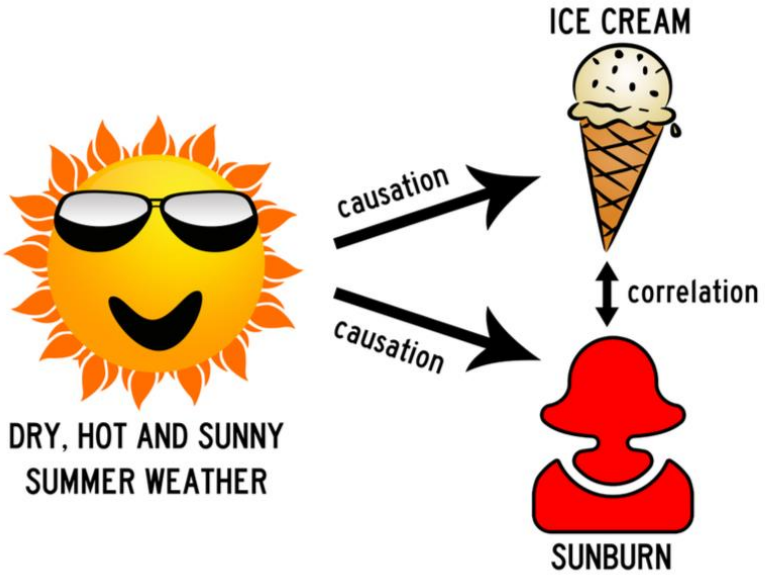
1. Utilize either 1mm or 2mm grid paper to mark the coordinates (Run 1, Run 2).
2. Calculate the mean value for each run and mark this point on the graph.
3. Draw a line that intersects the point determined in step 2 and aligns as closely as feasible with as many data points as possible.
4. Determine the equation for your drawn line.
5. Assess the degree of correlation between the two variables.

Causation

The correlation between two variables doesn't necessarily imply a cause-and-effect relationship between them. For instance:

1. When we measured the arm length and running speed of a group of young children, we found a strong, positive correlation. However, this doesn't mean that having shorter arms makes you run slower or that running faster causes your arms to grow longer. Instead, the strong, positive correlation arises because both arm length and running speed are closely tied to a third variable, age. Up to a certain age, both arm length and running speed tend to increase.
2. We recorded the number of television sets sold in London and the number of stray dogs collected in Boston over several years, and we discovered a strong, positive correlation between these variables. Clearly, the sale of television sets in London isn't influencing the number of stray dogs collected in Boston. This correlation is coincidental and stems from both variables increasing over the same time frame.

If a change in one variable directly leads to a change in another, then we can speak of a **causal relationship** between them. In such cases, we can say that the independent variable explains the dependent variable. Alternatively, we may refer to them as **explanatory variable** and **response variable**. However, when a causal relationship isn't evident, we cannot conclude that one exists solely based on a high correlation.



[13]

Figure 43: Causation example

For example, during the summertime, many regions experience dry, hot, and sunny weather. This type of weather often prompts people to seek ways to cool down and enjoy the outdoors. One popular choice is to indulge in ice cream, a delightful treat known for its refreshing and sweet qualities.

However, the connection between hot weather and ice cream consumption goes beyond mere preference. The causal link lies in the desire for relief from the heat. People naturally gravitate toward cold and refreshing foods, such as ice cream, during hot and sunny days as a means of staying cool and refreshed. This cause-and-effect relationship is driven by the human instinct to seek comfort and relief from uncomfortable temperatures.

While enjoying ice cream on a scorching summer day can provide a temporary respite from the heat, it's important to note that it doesn't provide any protection against the harmful effects of prolonged sun exposure. In fact, excessive sun exposure during hot and sunny weather can increase the risk of sunburn.

Sunburn occurs when the skin is exposed to the sun's ultraviolet (UV) radiation for an extended period without adequate protection, such as sunscreen or protective clothing. The UV radiation can damage the skin cells and lead to painful and uncomfortable burns. Ice cream consumption, though enjoyable, does not offer any sunburn protection and should not be seen as a preventative measure.

To avoid sunburn and its potential health risks, it's essential to take proper sun safety precautions, such as wearing sunscreen, sunglasses, and protective clothing, seeking shade during peak sunlight hours, and staying hydrated. While ice cream can provide relief from the heat, it should not replace these important sun safety practices during hot and sunny summer weather.

[13]https://www.lukeworthington.com/causality-correlation-or-just-coincidence/.

The coefficient of determination

The quantity

$$\rho^2 = r^2 = R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

is called the **coefficient of determination**. SS_T measures the variability in *y* without considering the influence of the explanatory variable *x*, while SS_{Res} represents the remaining variability in *y* after accounting for *x*. Consequently, R^2 is often referred to as the proportion of variance explained by the regressor x. Since $0 \leq SS_{Res} \leq SS_T$, it follows that $0 \leq R^2 \leq 1$. High R^2 values, close to 1, indicate that the regression model explains most of the variance in y.

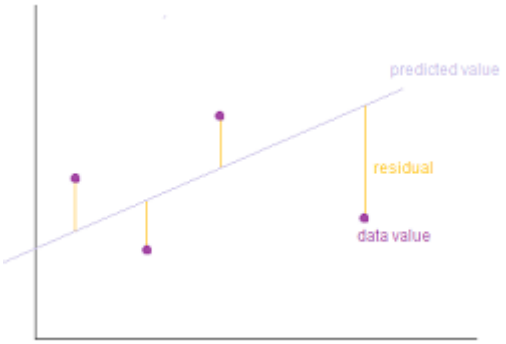
Note: It's essential to exercise caution when using the R^2 statistic because it's always possible to inflate R^2 by adding enough terms to the model. For example, with no repeated data points (more than one y value for the same x value), a polynomial of degree *n*–1 will yield a perfect fit ($R^2 = 1$) for *n* data points. However, when there are repeated data points, R^2 can never reach 1 because the model cannot account for the variability related to pure error. Furthermore, adding an explanatory variable to the model won't decrease R^2, but it doesn't necessarily mean the new model is superior. Unless the error sum of squares in the new model is reduced by an amount equivalent to the original error mean square, the new model will have a larger error mean square than the old one due to the loss of one degree of freedom for error. Consequently, the new model may actually perform worse than the old one.

Linear regression

Karl Pearson (1857-1936) was an English attorney and mathematician renowned for his significant contributions to the field of statistics. Among his notable achievements are the development of the product-moment correlation coefficient and the establishment of the chi-squared test. In 1911, he also founded the inaugural university statistics department in the world at the University College of London.

The term "regression" in statistics is employed in a manner distinct from its usage in other contexts. Its initial application was to explore the connection between the heights of fathers and sons. It was observed that, on average, tall fathers tended to have sons shorter than them, while short fathers tended to have sons taller than them. The heights of sons tended to move closer to the mean height. Nowadays, the term "**regression**" is employed for various types of curve fitting.

In this part, you've created scatter plots of data, assessed the strength and direction of the relationship, and utilized a best-fit line to represent the data. You may have noticed that, visually, there can be more than one line that appears to be a good fit for the data. A best-fit line is often an imprecise measure. Consequently, a technique known as **linear regression** is employed to ascertain the most suitable line equation for the data set. The accuracy of the line can be enhanced by considering **residuals**. A residual represents the vertical gap between a data point and the plotted regression line.



[14]

Figure 44: Plotted regression line example

Correlation evaluates both the direction and strength of the connection between two quantitative variables. When a scatterplot exhibits a linear correlation, we emphasize this connection by adding a line to the scatterplot. This regression line should only be introduced when one of the variables contributes to explaining or predicting the values of the other variable.

Regression necessitates an independent variable and a dependent variable.

To accurately draw the best-fit line, it is essential to employ a technique known as **least squares regression**. This method guarantees that the line is drawn in a manner where the sum of the squares of the deviations from the data points to the line is minimized. When focusing on vertical deviations, this line becomes the regression line for 'y on x' with 'x' as the independent variable and 'y' as the dependent variable. This is the most commonly used regression line. When minimizing the horizontal deviations, it results in 'x on y' for the line. You've observed that graphical representations of two variables can provide a general sense of the information within a data set, but for precision, a numerical measure is necessary.

A calculator can identify each data point, calculate 'x' and 'y' position the line accordingly, and then compute the vertical distance from each point to the line, creating corresponding squares. The line is then adjusted until the total area of the squares reaches the minimum possible value. This process yields the best-fit line. Once the line is determined, its equation can be calculated for use as a model.

Let y represent the response variable, and x the explanatory variable. Since, for a given value of the explanatory variable x, we can expect multiple values of the response variable y, our linear model allows us to predict, on average, the y value when given a specific x value. Thus, we express the equation of the linear regression line as: $E(y) = \alpha + \beta x$. In other words, for a particular x value, the expected y value is $\alpha + \beta x$. Here, α corresponds to the value when $x = 0$, and β represents the slope, indicating how the response variable changes for each one-unit change in the explanatory variable (the gradient).

It's worth noting that the regression model can also be formally stated as $E(y|X = x) = \alpha + \beta x$. In cases like this, our data are merely samples from a larger population, so we can only estimate the regression equation.

From our sample data, we estimate the regression equation, which we express as: $\hat{y} = \hat{\alpha} + \hat{\beta} x$. In this equation, $\hat{\beta}$ is the slope of the line, serving as an estimate of β. It reflects how the response variable y

[14] http:// wiki.engageeducation.org.au/further-maths/data-analysis/residuals/.

changes with changes in the explanatory variable x. $\hat{\alpha}$ is an estimate of α and represents the y value corresponding to an x value of zero. In the example of height and lean mass, the equation is:

$$Lean\ mass\ (kg) = 56.1 + 0.0966\ Height\ (cm)$$

So, $\hat{\beta} = 0.0966$ and $\hat{\alpha} = 56.1$. This means that, on average, for every 1 cm increase (or decrease) in height, we predict a 0.0966 kg increase (or decrease) in weight.

The interpretation of $\hat{\alpha}$ is unique. As you may know from algebra, $\hat{\alpha}$ represents the y value (which is Lean mass in this case) corresponding to an x value of zero (which is height in this case). However, for this problem, the interpretation is not straightforward because it implies a height of zero. The general guideline is that if zero is not a part of the explanatory variable's domain, then attempting to interpret the intercept becomes irrelevant.

Let's do the following example:

The tables below display the percentages achieved by 10 students in their German and Turkish exams.

German	92	18	88	20	30	80	60	54	46	40
Turkish	100	42	80	54	60	66	80	68	62	54

Table 26: German and Turkish exams data

a) Calculate the regression equation.

b) Create a scatterplot and incorporate the equation line derived in part a. Assess the degree of correlation between English exam scores and Humanities exam scores.

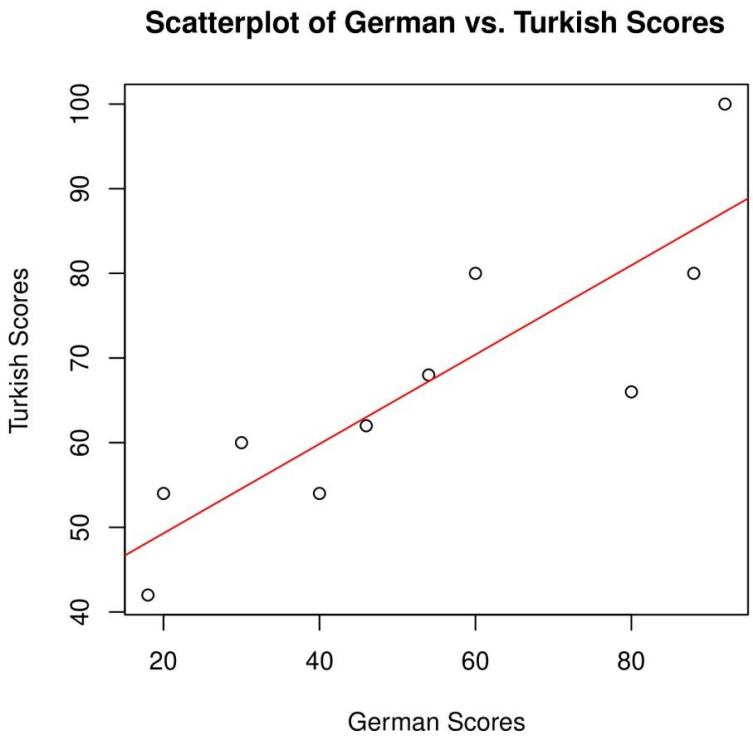


Figure 45: Plot of German scores versus Turkish scores

R 4.2.1 code

```
# Enter the data
German <- c(92, 18, 88, 20, 30, 80, 60, 54, 46, 40)
Turkish <- c(100, 42, 80, 54, 60, 66, 80, 68, 62, 54)

# Calculate the regression equation
regression_model <- lm(Turkish ~ German)

# Create a scatterplot
plot(German, Turkish, main="Scatterplot of German vs. Turkish Scores",
    xlab="German Scores", ylab="Turkish Scores")

# Add the regression line
abline(regression_model, col="red")

# Assess the degree of correlation
correlation <- cor(German, Turkish)
cat("Correlation between German and Turkish scores:", correlation)
```

R Output

Correlation between German and Turkish scores: 0.8626339

```
> regression_model
Call:
lm(formula = Turkish ~ German)
Coefficients:
(Intercept)      German
   38.7376       0.5277
```

that is; $\hat{y} = 38.7376 + 0.5277x$.

This illustrates a moderate positive correlation between German exam outcomes and Turkish exam results.

Let's do the following example:

a) Using the data and the regression line equation from German and Turkish exams example, provide a rationale for predicting the grade a student who scored 75 on an German exam would receive on the Turkish exam, and calculate their overall exam grade. Given that the score of 75 falls within the specified range of English exam scores, which spans from 18 to 92, and considering the moderate positive correlation observed in the data, it is appropriate to employ the regression equation for predicting the student's Humanities exam grade.

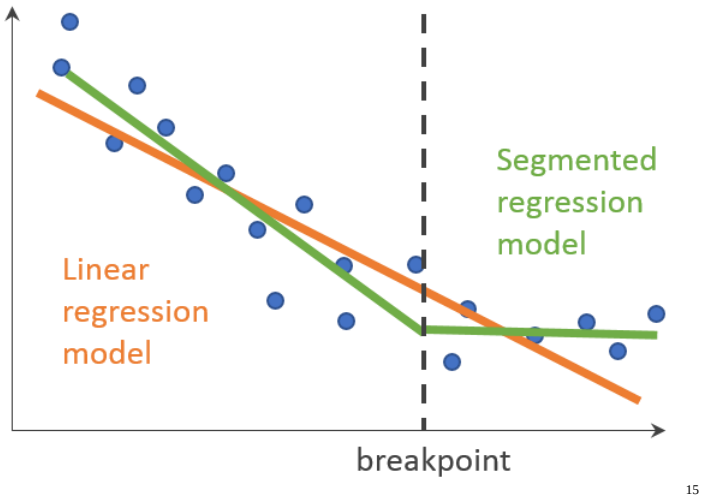
$$\hat{y} = 38.7376 + 0.5277x = 38.7376 + 0.5277(75) = 78.3$$

The projected grade for the student on the Humanities exam is 78.

b) In addition, if a different student missed the German exam and scored 20 on it, explain whether it is reasonable to predict their German grade using the regression equation, providing your reasoning. As 20 falls beyond the scope of the Turkish exam data range, it is not advisable to predict the German grade based on the Turkish grade. When making predictions, if the data used falls outside the provided range, extrapolation is typically considered unreliable for making accurate predictions.

Note: Segmented regression, also referred to as **piecewise regression** or **broken-stick regression**, is a technique within regression analysis where the independent variable is divided into segments, and a distinct line segment is fitted to each of these segments. For instance, when examining data related to

traffic accidents before and after the implementation of a particular law, we gather accident data for the period preceding the law and the period following its enactment. At this level, our approach involves fitting a separate regression line to each segment of the data. If we observe from a scatterplot that the data suggests a change in the pattern at a certain point, known as a break point, we then partition our data into two or more segments, as illustrated.

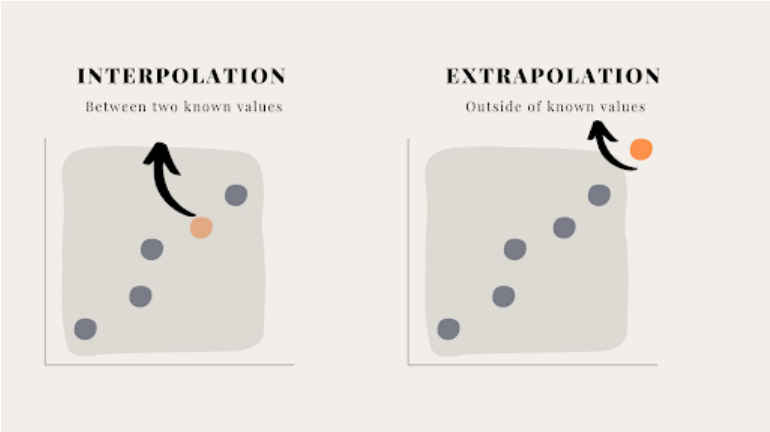


[15]

Figure 46: Segmented regression model example

Interpolation and extrapolation

Interpolation refers to the estimation of a value within the range of the smallest and largest observed values. **Extrapolation**, on the other hand, involves predicting a value beyond the scope of the smallest and largest data points. The **accuracy of interpolation** relies on the linearity of the data, which can be assessed using the correlation coefficient. In contrast, the **accuracy of extrapolation** not only relies on the original data set but also assumes that the underlying relationship extends beyond the collected data. This aspect is highly contingent on the specific context being investigated and is often **not a dependable method**.



[16]

Figure 47: Interpolation and extrapolation

[15]https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/changingtrendsinmortalityinenglandandwales1990to2017/experimentalstatistics
[16]https://builtin.com/data-science/extrapolation.

Correlation fallacies

Here are the provided definitions of six common correlation fallacies and an analysis of which ones are predominant in the following examples (there may be more than one for each):

1. **Correlation vs. causation fallacy**: This fallacy is evident in the example of the ski race. It is not a valid explanation to claim that the time in run 2 was slow because the time in run 1 was slow just because there is a correlation between the two variables. This highlights the distinction between correlation and causation.

2. **Correlation is only linear fallacy**: The text acknowledges that the PMCC measures linear correlation. However, it also mentions that technology allows for the consideration of non-linear relationships. This suggests an awareness of the limitation but doesn't directly demonstrate this fallacy in the examples.

3. **Theory of the third variable fallacy**: This fallacy is prevalent in the example of detox juice bars and knife crime in London. The high correlation observed between these two variables is likely influenced by a third variable, "time," which plays a significant role in the correlation study. This example illustrates the importance of considering third variables in correlation analysis.

4. **Spurious mathematical relationships fallacy**: This fallacy is emphasized in the example involving a student's Maths HL grade and total Diploma points. When a mathematical relationship exists between variables, such as a component of one variable being part of another, it can lead to an invalid correlation if not carefully examined. This example highlights the need for caution in such cases.

5. **Combinations from separate populations fallacy**: The text doesn't directly demonstrate this fallacy in the examples, but it stresses the importance of using data from a single, homogeneous population to avoid this issue.

6. **Incorrect sampling techniques fallacy**: The text mentions that data collection methods are crucial in correlation analysis. While it doesn't provide a specific example of incorrect sampling techniques, it underscores the significance of proper data collection practices in correlation studies.

In summary, the examples primarily illustrate the **Correlation vs. causation Fallacy** and the **Theory of the third variable fallacy**, with some awareness of the **Correlation is only linear fallacy** and the **Spurious mathematical relationships fallacy**. The other fallacies are mentioned as important considerations in correlation analysis but are not directly demonstrated in the given examples.

POSTSCRIPT

As you reach the final pages of this book, we hope that the journey through the realms of probability, statistics and the indispensable role of R in data analysis has been as enlightening for you as it has been for our in crafting these pages. The world around us is a tapestry of uncertainty, variability, and chance, and through the lens of probability and statistics, we gain a clearer understanding of the patterns that emerge from this seemingly chaotic fabric.

In the pursuit of knowledge, we have explored real-life examples that illustrate the power and applicability of these mathematical tools. From predicting the outcome of a political election to understanding the fluctuations in financial markets, from analyzing health data to making informed decisions in everyday life, the principles of probability and statistics serve as indispensable guides.

The collaborative spirit of the R community and the language's robust capabilities in statistical modeling and data visualization stand as a testament to the dynamic evolution of tools that empower individuals across various domains. Just as probability provides a systematic framework for dealing with uncertainty, and statistics transforms raw data into meaningful insights, R emerges as a reliable companion in the quest for extracting knowledge from the vast sea of information.

As we navigate through the complexities of our existence, we are often faced with decisions that involve an element of risk. This book aimed to equip you with the skills to navigate these uncertainties with confidence, making informed choices based on evidence and data. Remember, probability is not just about numbers; it's about understanding the likelihood of events and making sense of the inherent randomness in the world.

We encourage you to continue exploring and applying the concepts you've encountered here. Embrace the challenges that come with uncertainty, for it is within these challenges that opportunities for growth and discovery lie. Whether you are a student, a professional, or simply a curious mind, may the knowledge gained from this exploration serve you well in your endeavors.

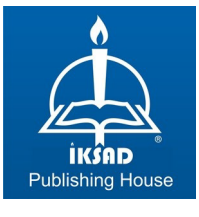
In closing, we extend our gratitude to those who contributed to this journey—teachers, researchers, and thinkers who have dedicated their lives to unraveling the mysteries of probability and statistics. As we conclude this chapter, let us remember that the pursuit of knowledge is an ongoing adventure, and the world of probability and statistics is a rich landscape waiting to be explored further.

We wish you a future filled with understanding, informed choices, and a profound appreciation for the fascinating interplay of chance and order in our lives.

REFERENCES

1. C.H. Saraswathi, S.K. Durga Moulali, A. Nagamani, 2017, The Real Life Applications of Probability in Mathematics, International Journal of Management and Applied Science, 3(4), 62.
2. D.C., Montgomery, E.A., Peck, and G.G., Vining, 2012, *Introduction to Linear Regression Analysis*. John Wiley and Sons Inc, Hoboken.
3. H., Tijms, 2007, *Understanding Probability, Chance Rules in Everyday Life*. Cambridge University Press, Newyork.
4. I., Wazir and T., Garry, 2019, *Mathematics: Analysis and Approaches for IB Diploma: SL.* Pearson Education Limited, London.
5. I., Wazir and T., Garry, 2019, *Mathematics: Analysis and Approaches for IB Diploma: HL.* Pearson Education Limited, London.
6. J., Owen, R., Haese, S., Haese and M., Bruce, 2004, *Mathematics for the International Student: Mathematics SL (First edition)*. Raksar Nominees Pty Ltd, Australia.
7. J.C., Wathall, J., Harcet, R., Harrison, L., Heinrichs and M., Torres-Skoumal, 2019, *Mathematics: Analysis and Approaches HL*. Oxford University Press, U.K.
8. M., Haese, M., Humphries, C., Sangwin and N., Vo, 2019, *Mathematics: Core Topics SL 1*. Haese Mathematics, Australia
9. M., Haese, M., Humphries, C., Sangwin and N., Vo, 2019, *Mathematics: Analysis and Approaches SL 2*. Haese Mathematics, Australia
10. M., Haese, M., Humphries, C., Sangwin and N., Vo, 2019, *Mathematics: Core Topics HL 1*. Haese Mathematics, Australia
11. M., Haese, M., Humphries, C., Sangwin and N., Vo, 2019, *Mathematics: Analysis and Approaches HL 2*. Haese Mathematics, Australia
12. N., Awada, L., Buchanan, J.C., Wathall, E., Kemp, P., La Rondie, J., Stevens and E., Thompson, 2019, *Mathematics: Analysis and Approaches SL*. Oxford University Press, U.K.
13. P., Fannon, V., Kadelburg, B., Woolley and S., Ward, 2012, *Mathematics Higher Level for the IB Diploma.* Cambridge University Press, U.K.
14. R., Haese, S., Haese, M., Haese, M., Maenpaa and M., Humphries, 2012, *Mathematics for the International Student: Mathematics SL (Third edition)*. Haese Mathematics, Australia.
15. https://www.investopedia.com/terms/c/conditional_probability.asp
16. https://flexbooks.ck12.org/cbook/ck-12-interactive-geometry-for-ccss/section/11.5/primary/lesson/everyday-examples-of-independence-and-probability-geo-ccss/
17. https://flexbooks.ck12.org/cbook/ck-12-middle-school-math-concepts-grade-8/section/11.7/primary/lesson/theoretical-probability-msm8/
18. https://www.cuemath.com/data/theoretical-probability/
19. https://byjus.com/maths/theoretical-probability/
20. https://study.com/learn/lesson/geometric-probability-formula-examples.html
21. https://studiousguy.com/poisson-distribution-examples/
22. https://studiousguy.com/binomial-distribution-examples/
23. https://studiousguy.com/real-life-examples-normal-distribution/
24. https://blog.drustvo-evo.hr/en/2020/09/13/simulating-the-bean-machine/
25. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/ISSS_SC/Module004/isss_m4_normal/node4.html
26. https://www.slideserve.com/murray/1-standard-life-expectancy-sle-the-ideal.
27. https://www.agefotostock.com/age/en/details-photo/university-students-having-lunch-at-cafeteria/WR0596517.
28. https://www.onlc.com/blog/10-types-tableau-charts-using/.
29. https://www.google.com/search?q=average&rlz=1C1CHBD_trTR890TR890&tbm=isch&source=lnms&sa=X&ved=2ahUKEwjprKvNjcqAAxW6SPEDHY73AcIQ_AUoAXoECAIQAw&biw=1280&bih=571&dpr=1.5#imgrc=RqZm6r-YYGv-vM.
30. https://sanket-m-kangle.medium.com/what-is-population-sample-and-sampling-error-6298d79c3771.
31. https://www.fromthegenesis.com/skewness/.

32. https://www.theschoolrun.com/what-are-mode-mean-median-and-range.
33. https://datasci.soniaspindt.com/StatisticalFoundations/DescriptiveStats/CentralTendency/mode.html.
34. https://www.kdnuggets.com/2019/11/understanding-boxplots.html.
35. https:// www.slideserve.com/tammy/variance-standard-deviation .
36. https://www.aplustopper.com/scatter-plots-correlation/.
37. https://www.lukeworthington.com/causality-correlation-or-just-coincidence/.
38. http:// wiki.engageeducation.org.au/further-maths/data-analysis/residuals/.
39. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/changingtrendsinmortalityinenglandandwales1990to2017/experimentalstatistics
40. https://builtin.com/data-science/extrapolation.



ISBN: 978-625-367-517-2