Imperial College London

Introduction to probability theory and statistical inference

Jesús Urtasun Elizari

Research Computing and Data Science

March 6, 2025

Contents

Index

Chapter 1

Introduction

In the following pages one will find an introductory text to one of the key subjects within mathematical sciences. The text is composed by four chapters, together with some appendix reviewing basic mathematical concepts, and a bibliographic note. The purpose of this lecture notes is to make both probability and statistical analysis an easy, interesting and engaging topic for anyone interested, without the need for prior experience with mathematical training.

First, we will introduce and explore the concept of probability itself, and we will discuss how to model information, surprise, and various random processes, also referred to as *stochastic*. Then we will introduce the idea of a function and how functions need to be adapted to implement uncertainty when discussing random events. In the second part we will address the difference between prediction and inference, and discuss a set of subjects normally grouped under the name of hypothesis testing. Here we will introduce how to quantify certainty and bias, and how to model significance and false positives, that is, to compute p-values.

Finally, we will discuss bayesian statistics (...).

Probability theory is the mathematical framework for quantifying uncertainty and randomness. It provides the foundation for understanding and modeling various real-world phenomena, ranging from gambling and statistical inference to machine learning and quantum mechanics. At its core, probability assigns numerical values to the likelihood of different outcomes in an experiment. The fundamental concepts include sample spaces, events, and probability measures that adhere to axioms established by Andrey Kolmogorov.

The idea of stochasticity and randomness has deep historical roots. Ancient civilizations, including the Babylonians, Egyptians, and Greeks, grappled with the concept of uncertainty in games of chance, commerce, and divination. The oldest known dice date back over 5,000 years, indicating an early human fascination with randomness. While these cultures did not develop formal mathematical probability, they recognized patterns in random events and attempted to predict outcomes based on empirical observations and superstitions. Greek philosophers such as Democritus and Aristotle debated the nature of chance and determinism. The Roman philosopher Cicero distinguished between chance events and those governed by fate, foreshadowing later discussions on probability. In medieval times, scholars like Gerolamo Cardano (1501–1576) made early contributions by analyzing gambling problems and laying the groundwork for probability theory. The formalization of probability as a mathematical discipline began in the 17th century with the correspondence between Blaise Pascal and Pierre de Fermat, who devised combinatorial methods to solve problems related to games of chance. Their work introduced fundamental ideas such as expected value and laid the foundation for later advances by Christiaan Huygens, Jacob Bernoulli, and Abraham de Moivre. Bernoulli's *Ars Conjectandi* (1713) introduced the Law of Large Numbers, establishing that observed frequencies converge to theoretical probabilities over many trials.

A stochastic process extends probability theory by modeling systems that evolve randomly over time. Unlike deterministic processes, where the outcome is fixed by initial conditions, stochastic processes incor-

porate uncertainty at each step. These models are essential in diverse fields such as finance (stock price fluctuations), physics (Brownian motion), and engineering (signal processing). A stochastic process is defined as a collection of random variables indexed by time or space, commonly represented as $\{X_t\}_{t \in T}$, where T is an index set.

One of the most fundamental stochastic processes is the Markov process, which exhibits the Markov property: the future state depends only on the present state, not on past history. This property makes Markov chains particularly useful in modeling systems with memoryless transitions, such as queueing networks and genetic sequences. Another significant class of stochastic processes is the Poisson process, which describes the occurrence of rare events over time, such as radioactive decay or network traffic.

Information theory, developed by Claude Shannon, is deeply intertwined with probability and stochastic processes. It quantifies the amount of uncertainty or surprise in a system and provides a mathematical foundation for data compression and communication. The fundamental measure in information theory is entropy, defined as:

$$H(X) = -\sum_i P(x_i) \log P(x_i), \tag{1.1}$$

where $P(x_i)$ is the probability of observing outcome x_i. Entropy captures the unpredictability of a source of information, with higher entropy indicating greater uncertainty.

Shannon also introduced the concept of mutual information, which measures the reduction in uncertainty about one random variable given knowledge of another. This concept is crucial in signal processing, machine learning, and cryptography, where efficient information transfer and encoding are essential.

Together, probability theory, stochastic processes, and information theory form a powerful toolkit for modelling and analyzing uncertainty. Their applications span numerous disciplines, from physics and biology to artificial intelligence and cybersecurity. Understanding these concepts allows for the development of efficient algorithms, accurate predictions, and optimized decision-making strategies in uncertain environments.

Here are three important textbooks in the field of probability and statistics:

- A simple, intuitive introduction to statistics with few mathematical concepts is provided in Spiegelhalter's *The Art of Statistics: How to Learn from Data* [1].
- A more foundational textbook, with more advanced mathematical approach, can be found at DeGroot and Schervish's *Probability and Statistics* [2].
- For a philosophical and historical perspective on probability and statistics, pleas find McFadden's *The Philosophy of Statistics* [3].

Chapter 2

Introduction to probability and random events

2.1 What is probability?

As already mentioned in the introduction, probability theory is one of the oldest subjects within mathematical studies. Ideas such as probability or chance, together with measurement, information, inference, can be traced back to ancient times. Paradoxically, almost every topic explained nowadays in modern courses of statistics is extremely new, ranging broadly a century. Concepts like distributions, gaussian behaviour, p-values, hypothesis testing or normalization, are introduced and formalized in the XXth century by mathematicians such as Pearson and Fisher. For the purpose of this course, we will assume a specific framework, where we will understand probability as a number representing information, or *surprise*. For a detailed discussion on foundations of these topics, please see [...]. For more mathematically advanced texts, check [...] and [...].

Broadly speaking, probability, and later statistical inference, are branches of mathematics dealing with chance, also referred to as *random* events, or *stochastic* processes. Indeed, the word stochastic comes from no other than the greek word στοχαστικός, which literally means *to guess*. Let's try to briefly introduce the idea of probability, as a quantity that allows as to describe such random events.

So let's first ask ourselves the question. What *is* probability in the first place? What do we mean by it and what does it describe? Probability is nothing more, and nothing less, that a number we make up, a quantity we come up with, to quantify certainty in a process whose outcome we ignore. A number we will use to describe the amount of information we have about a random, or stochastic, event. For simplicity, we can make it range from 0 to 1, in the following way.

- If I'm sure A will never happen, P(A) = 0.
- If I'm sure A will always happen, P(A) = 1.
- Anything in between, if exists degree of surprise, P(A) in [0, 1]

$$\sum_{i=0}^{n} p(x_i) = 1 \tag{2.1}$$

Once we have a definition for probability in the abstract case, we should have a way to compute for particular cases. A way of doing that, referred to as frequentist approach, is by dividing the number of favorable outcomes by the total number of outcomes.

$$P(\text{A happening}) = \frac{\text{Number of times A happens}}{\text{Total number of trials}} \tag{2.2}$$

Probabilities must follow a property we call unitarity. Unitarity ensures that, if we consider and add up the probabilities for all possible events in a given experiment, we get the total. That means, at least one of the scenarios will happen.

Indeed, the literal meaning of probability comes from latin probabilis. American logician and philosopher Richard Jeffrey, "Before the middle of the seventeenth century, the term "probable" (Latin probabilis) meant just approvable, and was applied in that sense, univocally, to opinion and to action. A probable action or opinion was one such as sensible people would undertake or hold, in the circumstances."[12] However, in legal contexts especially, "probable" could also apply to propositions for which there was good evidence.

The sixteenth-century Italian polymath Girolamo Cardano demonstrated the efficacy of defining odds as the ratio of favourable to unfavourable outcomes (which implies that the probability of an event is given by the ratio of favourable outcomes to the total number of possible outcomes[14]). Aside from the elementary work by Cardano, the doctrine of probabilities dates to the correspondence of Pierre de Fermat and Blaise Pascal (1654). Christiaan Huygens (1657) gave the earliest known scientific treatment of the subject.[15] Jakob Bernoulli's Ars Conjectandi (posthumous, 1713) and Abraham de Moivre's Doctrine of Chances (1718) treated the subject as a branch of mathematics.[16] See Ian Hacking's The Emergence of Probability[10] and James Franklin's The Science of Conjecture[17] for histories of the early development of the very concept of mathematical probability.

Like other theories, the theory of probability is a representation of its concepts in formal terms – that is, in terms that can be considered separately from their meaning. These formal terms are manipulated by the rules of mathematics and logic, and any results are interpreted or translated back into the problem domain.

There have been at least two successful attempts to formalize probability, namely the Kolmogorov formulation and the Cox formulation. In Kolmogorov's formulation (see also probability space), sets are interpreted as events and probability as a measure on a class of sets. In Cox's theorem, probability is taken as a primitive (i.e., not further analyzed), and the emphasis is on constructing a consistent assignment of probability values to propositions. In both cases, the laws of probability are the same, except for technical details.

2.2 Discrete and continuous

Discrete and continuous

2.3 Probability distributions

Once we are comfortable with the idea of random events, and we have been introduced to probability as a number to quantify surprise, we can agree that not all random phenomena are equal. A basic way to classify and separate random events is according to how their probabilities are distributed.

2.3.1 Binomial distribution

The binomial distribution models the number of successes in a fixed number of independent trials, each with the same probability of success. It was developed by Jacob Bernoulli in the 17th century while studying the probability of repeated Bernoulli trials. His work laid the foundation for the Law of Large Numbers.

Intuitively, this distribution is useful when considering repeated experiments with two possible outcomes (success or failure). For example, flipping a fair coin multiple times follows a binomial pattern.

The probability mass function is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2.3}$$

where n is the number of trials and p is the probability of success.

Example: Suppose we flip a fair coin 5 times ($n = 5$) and want to find the probability of getting exactly 3 heads ($p = 0.5$):

$$P(X = 3) = \binom{5}{3} (0.5)^3 (0.5)^2 = 10 \times 0.125 \times 0.25 = 0.3125. \tag{2.4}$$

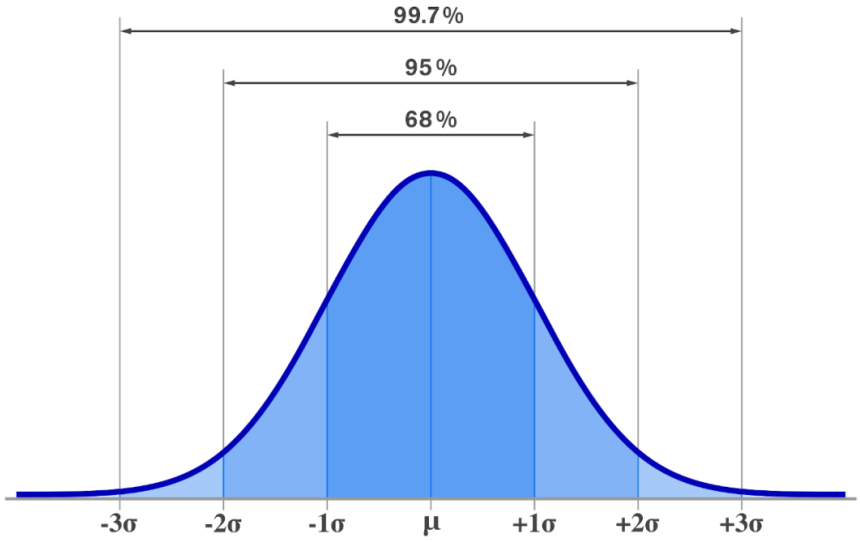

Figure 2.1: Binomial distribution.

2.3.2 Poisson distribution

Named after Siméon Denis Poisson, this distribution was introduced in the early 19th century to model rare events. It is particularly useful for counting occurrences over a fixed interval of time or space.

Intuitively, the Poisson distribution applies to events happening randomly but with a known average rate, such as the number of emails received per hour.

Named after Siméon Denis Poisson, this distribution was introduced in the early 19th century to model rare events. It is particularly useful for counting occurrences over a fixed interval of time or space.

Intuitively, the Poisson distribution applies to events happening randomly but with a known average rate, such as the number of emails received per hour.

The probability mass function is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{2.5}$$

where λ is the expected number of occurrences.

Example: If a bookstore sells an average of 3 books per hour ($\lambda = 3$), the probability of selling exactly 5 books in an hour is:

$$P(X = 5) = \frac{3^5 e^{-3}}{5!} = \frac{243 e^{-3}}{120} \approx 0.1008. \tag{2.6}$$

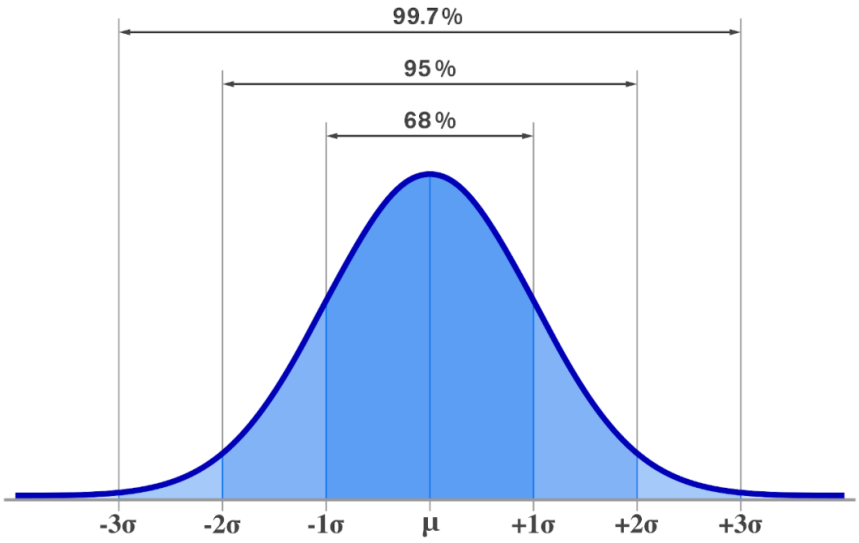

Figure 2.2: Poisson distribution.

2.3.3 Uniform distribution

The uniform distribution represents a scenario where all outcomes in an interval $[a, b]$ are equally likely. This distribution has been used since antiquity, especially in early probability and gambling studies.

Intuitively, it is useful when every possible outcome is equally probable, such as rolling a fair die.

The uniform distribution represents a scenario where all outcomes in an interval $[a, b]$ are equally likely. This distribution has been used since antiquity, especially in early probability and gambling studies.

Intuitively, it is useful when every possible outcome is equally probable, such as rolling a fair die.

The probability density function is:

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b. \tag{2.7}$$

Example: If a random number is chosen from the interval [2, 10], the probability density is:

$$f(x) = \frac{1}{10-2} = 0.125. \tag{2.8}$$

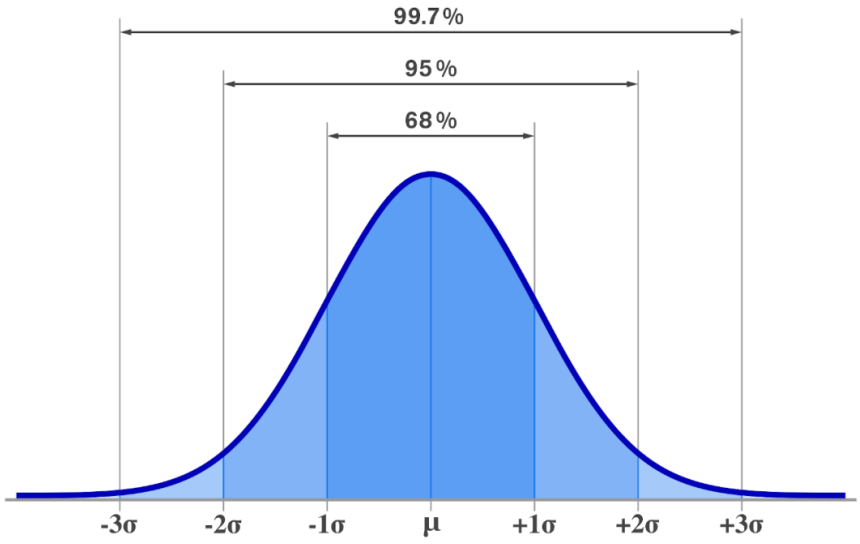

Figure 2.3: Uniform distribution.

2.3.4 Gaussian distribution

Introduced by Carl Friedrich Gauss, the normal distribution became central to statistics due to the Central Limit Theorem (CLT). It describes how averages of large samples tend to form a bell-shaped curve.

Intuitively, many natural and social phenomena follow a normal distribution, such as human heights and test scores.

Introduced by Carl Friedrich Gauss, the normal distribution became central to statistics due to the Central Limit Theorem (CLT). It describes how averages of large samples tend to form a bell-shaped curve.

Intuitively, many natural and social phenomena follow a normal distribution, such as human heights and test scores.

The probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.9}$$

Example: If human heights are normally distributed with mean $\mu = 170$ cm and standard deviation $\sigma = 10$ cm, the probability density of someone being exactly 180 cm is:

$$f(180) = \frac{1}{10\sqrt{2\pi}} e^{-\frac{(180-170)^2}{2(10)^2}} \approx 0.0242. \tag{2.10}$$

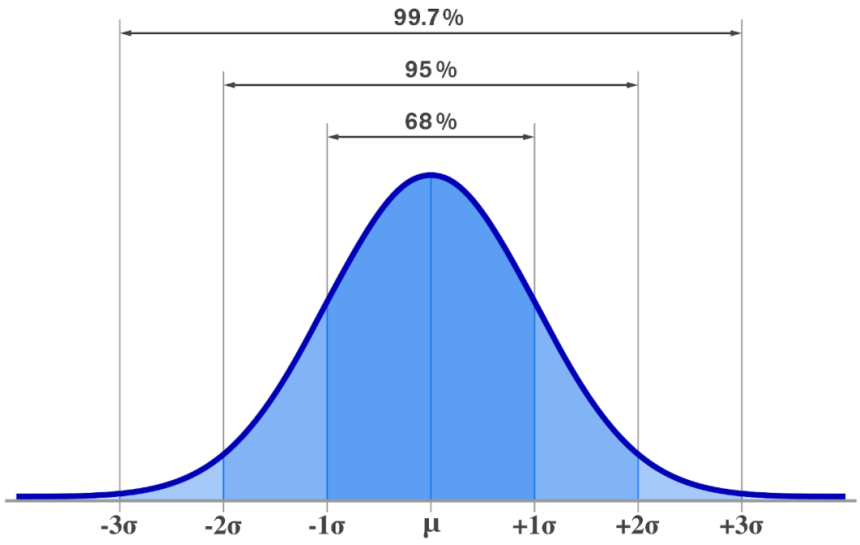

Figure 2.4: Gaussian distribution.

2.3.5 Exponential distribution

The exponential distribution models waiting times between Poisson process events. It has been widely applied in reliability analysis and survival studies.

Intuitively, it describes situations where the probability of waiting a certain time between events remains constant, such as time between bus arrivals.

The exponential distribution models waiting times between Poisson process events. It has been widely applied in reliability analysis and survival studies.

Intuitively, it describes situations where the probability of waiting a certain time between events remains constant, such as time between bus arrivals.

The probability density function is:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0. \tag{2.11}$$

Example: If a call center receives calls at an average rate of $\lambda = 2$ per minute, the probability that the next call arrives after more than 2 minutes is:

$$P(X > 2) = e^{-2(2)} = e^{-4} \approx 0.0183. \tag{2.12}$$

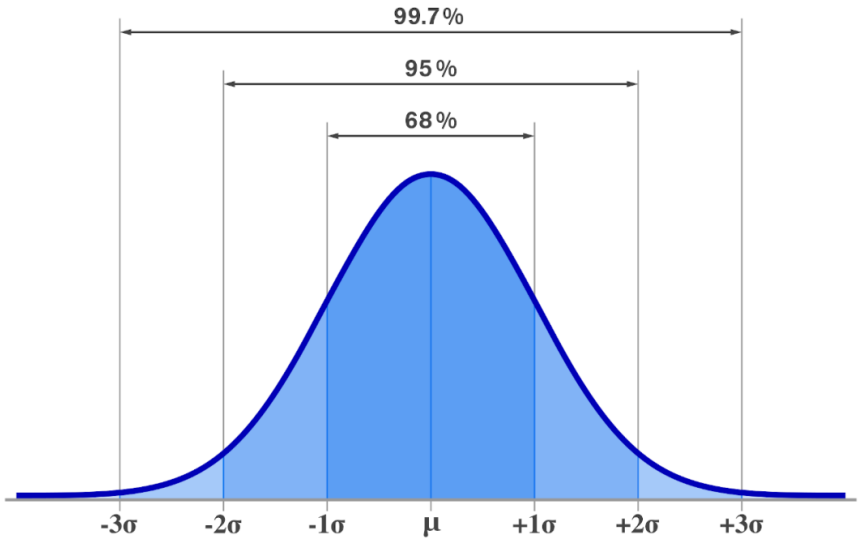

Figure 2.5: Exponential distribution.

Chapter 3

Parameter estimation

3.1 Prediction vs inference

Prediction vs inference

3.2 Parameter estimation

Parameter estimation

3.3 Law of Large Numbers (LLN)

3.3.1 Definition

The Law of Large Numbers states that as the sample size increases, the sample mean approaches the expected value. Formally, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X] = \mu$, then:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \quad \text{as } n \to \infty. \tag{3.1}$$

3.3.2 Intuition

Consider flipping a fair coin multiple times. The proportion of heads observed converges to 0.5 as the number of flips increases. This illustrates that the observed average stabilizes around the theoretical probability.

3.4 Law of Large Numbers (LLN)

3.4.1 Introduction

The Law of Large Numbers (LLN) is one of the fundamental theorems of probability theory. It was first formulated by Jacob Bernoulli in the late 17th century and later refined by other mathematicians, including Pafnuty Chebyshev. Bernoulli's work aimed to formalize how relative frequencies of events stabilize as the number of trials increases, providing the foundation for statistical inference. LLN plays a crucial role in statistics, finance, and machine learning, ensuring that averages computed from large samples are reliable estimates of expected values.

3.4.2 Definition

The Law of Large Numbers states that as the sample size increases, the sample mean approaches the expected value. Formally, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X] = \mu$, then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{as } n \to \infty. \tag{3.2}$$

3.4.3 Intuition

Consider flipping a fair coin multiple times. The proportion of heads observed converges to 0.5 as the number of flips increases. This illustrates that the observed average stabilizes around the theoretical probability.

3.4.4 Types of LLN

- **Weak Law of Large Numbers (WLLN)**: Convergence in probability, i.e., for any $\epsilon > 0$,

$$P(|\bar{X}_n - \mu| \geq \epsilon) \to 0 \quad \text{as } n \to \infty. \tag{3.3}$$

- **Strong Law of Large Numbers (SLLN)**: Almost sure convergence, i.e.,

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{3.4}$$

3.4.5 Example

Suppose we roll a fair six-sided die multiple times. The expected value of a roll is:

$$\mathbb{E}[X] = \frac{1+2+3+4+5+6}{6} = 3.5. \tag{3.5}$$

As we roll more dice, the sample mean of observed values gets closer to 3.5.

3.5 Central Limit Theorem (CLT)

3.5.1 Introduction

The Central Limit Theorem (CLT) was first discovered in the 18th century by Abraham de Moivre and later developed by Pierre-Simon Laplace and Carl Friedrich Gauss. It formalizes the idea that the distribution of sample means tends toward a normal distribution, regardless of the shape of the original population distribution. The CLT is fundamental in inferential statistics, allowing researchers to make predictions and construct confidence intervals for population parameters based on sample data.

3.5.2 Definition

The Central Limit Theorem states that for a large enough sample size, the sampling distribution of the sample mean follows a normal distribution, regardless of the original population distribution. Formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then the standardized sample mean:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{3.6}$$

converges in distribution to a standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$.

3.5.3 Intuition

No matter the shape of the original distribution, when we take many samples and compute their means, the histogram of these sample means will resemble a normal curve as the sample size grows.

3.5.4 Example

Consider rolling a fair six-sided die multiple times and computing the average outcome for groups of n rolls. As n increases, the distribution of these sample means approaches a normal distribution, centered at $\mu = 3.5$.

3.5.5 Applications

- Used in inferential statistics to approximate sampling distributions.
- Forms the basis for hypothesis testing and confidence intervals.
- Justifies the normality assumption in many statistical models.

Chapter 4

Introduction to statistical inference

4.1 Prediction vs inference

4.2 Hypothesis testing

Probability theory is one of the oldest subjects within mathematical studies. Paradoxically, almost every topic explained nowadays in modern courses of statistics is extremely new, ranging broadly a century. (...)

So let's first ask ourselves the question. What is probability in the first place? What do we mean by it and what does it describe? Probability is nothing more, and nothing less, that a number we make up, a quantity we come up with, to quantify certainty. A number we will use to describe the amount of information we have about a random, or stochastic, event. For simplicity, we can make it range from 0 to 1, in the following way.

4.3 Statistic tests

Statistic tests

4.4 P-values and significance

P-values and significance

Compare sample mean with hypothesized value - One sample t-test

The one-sample t-test is used to determine whether the mean of a single sample differs significantly from a known or hypothesized population mean. It assumes that the sample data are drawn from a normally distributed population. The test statistic is given by:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.

Compare sample means of two groups - Two sample t-test

Compare sample variances of two groups - Fisher test

Compare more than two groups - ANOVA

Compare distributions - χ^2 test

4.5 Parametric and non-parametric

4.6 Comparing data and normalization

Comparing data and normalization

Chapter 5

Introduction to bayesian statistics

5.1 The Bayes' theorem

The Bayes' theorem.

5.2 Bayesian vs frequentist

Bayesian vs frequentist.

5.3 Bayesian statistics

Bayesian statistics.

Chapter 6

Stochasticity and Markov processes

6.1 Stochasticity and Markov processes

6.2 Markov chains and hidden Markov models

6.3 Modern applications

Bibliography

[1] David Spiegelhalter. *The Art of Statistics: How to Learn from Data*. Basic Books, 2019.

[2] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics* (4th ed.). Pearson, 2012.

[3] J. A. F. McFadden. *The Philosophy of Statistics*. Wiley-Blackwell, 2011.