Probabilities and Distributions

This lecture offers a thorough introduction to probability theory and its use in data science. It is divided into key segments, addressing basic principles, various probability types, expectations and variances, probability principles, managing probabilities, order statistics, distributions, random variables, and their data science applications. Practical examples, using different datasets, will demonstrate these ideas.

A section focuses on probabilities and their various components. The initial subsection outlines the essential definitions in probability, such as events, sample space, and outcomes. We will cover the three principal probability axioms: non-negativity, normalization, and additivity. Understand the classical interpretation of probability, which relies on the ratio of favorable outcomes to the total number of possible outcomes. Investigate the empirical method of probability, which defines it as the limit of the relative frequency of an event happening over numerous trials. Comprehend subjective probability, rooted in personal belief or the degree of certainty regarding an event. Analyze conditional probability, the likelihood of an event occurring given that another event has taken place. Discuss joint probability, the likelihood of two or more events occurring simultaneously. Learn about marginal probability, the chance of an event happening regardless of another variable's outcome. Grasp Bayes' Theorem and its application in updating an event's probability based on new data. Distinguish between independent and dependent probability events. Talk about the expected value concept, which represents the long-term average outcome of the experiment repetitions. Learn about variance, which quantifies the spread of a set of values. Understand standard deviation, the square root of variance, providing a measure of the values' spread. Delve into the Law of Large Numbers, which asserts that as the trials number increases, the sample mean converges towards the expected value. Understand the Central Limit Theorem, which posits that the sample mean's distribution approaches a normal distribution as the sample size grows.

A brief segment on order statistics is included to explain first order statistics and provide the related Python code. Describe first order statistics, focusing on the distribution of the sample's minimum and maximum values. Understand how to compute first order statistics with Python.

A section concentrates on distributions and their various parts. Describe the mean, which is the average value of a set of observations. Review standard deviation and variance within the context of distribution moments. Grasp the concept of degrees of freedom, which indicate how many values in a computation have freedom to vary. Cover the distinction between biased and unbiased estimators. Discriminate between sample statistics and population statistics. Learn to compute skewness using either sample or population standard deviation. Understand Bessel's correction and its role in statistics. Study the Bernoulli distribution, representing the result of a single binary trial. Detail the binomial distribution, which counts the number of successes in a fixed number of binary trials. Investigate the Poisson distribution, illustrating the number of events occurring within a

fixed period or space. Comprehend the geometric distribution, which counts the number of trials to achieve the first success in a sequence of independent binary trials. Study the negative binomial distribution, which extends the geometric distribution to the number of trials needed to obtain a predetermined number of successes. Examine the hypergeometric distribution, representing the number of successes in a sample drawn without replacement from a finite population. Understand the uniform distribution, depicting a scenario where all outcomes are equally probable. Delve into the normal distribution, a paramount continuous probability distribution that is symmetric around the mean. Study the exponential distribution, indicating the interval between events in a Poisson process. Explore the gamma distribution, generalizing the exponential distribution to the sum of several exponential random variables. Understand the beta distribution, used to model random variables confined to finite-length intervals. Delve into the chi-square distribution, employed in hypothesis testing and variance confidence interval estimation. Learn about the t-distribution, utilized for estimating population parameters when the sample size is small and the population standard deviation is unknown. Compare two variances using the F-distribution. Understand the log-normal distribution, which models a variable whose logarithm follows a normal distribution. Investigate the Weibull distribution, applied in reliability and life data analysis. Learn about the multivariate normal distribution, extending the normal distribution to multiple variables.

A section is dedicated to Random Variables and their elements. It covers probability mass functions, which assign the probability to exact outcomes of a discrete random variable. It also explains probability density functions, which indicate the likelihood that a continuous random variable takes a specific value.

A section is allocated to discussing the applications of probability in data science and their components. Understand the Naive Bayes classifier, a probabilistic model that relies on Bayes' theorem and makes strong independence assumptions. Examine probabilistic graphical models, which employ graphs to depict and analyze the dependencies among random variables. Delve into Gaussian processes, utilized for defining distributions over functions and frequently employed in regression and optimization tasks. Grasp the concept of hypothesis testing, a method for making decisions based on data from scientific studies. Investigate p-values, which quantify the strength of evidence against the null hypothesis. Learn about confidence intervals that provide a span of values within which the true population parameter likely resides with a specific level of confidence. Recognize how to assess statistical significance in A/B testing, a technique for comparing two versions of a variable to ascertain which one performs better.

A section is dedicated that covers the types of algorithms and their characteristics. It addresses deterministic algorithms, which always yield the same result for a specific input. It also examines randomized algorithms that incorporate randomness in their logic and may generate different outcomes for the same input across multiple executions.

This document is an extension of the research and lecture notes completed at Johns Hopkins University, Whiting School of Engineering, Engineering for Professionals, Artificial Intelligence Master's Program, Computer Science Master's Program, and Data Science Master's Program.

Contents

1 Probabilities

Probability theory is a fundamental pillar of data science, providing the mathematical framework for making inferences from data, modeling uncertainty, and making predictions. When analyzing data an intuitive reaction is to graph the data in its raw format. This will allow you to see the patterns of the data in two or three dimensions at a time. This of course assumes our data consists of numerical values or a representation that can be visually represented. This has its limitations and in many cases is not a practical approach for large data sets with respect to the number of observations or features. When data is not represented in numerical form there is some preprocessing involved either prior to using an algorithm or within the algorithm itself. In this document the assumption is that data is purely represented by some real numerical values, it is complete and within allowable ranges, e.g., $\mathbf{X} = \mathbf{x}(n) = [\mathbf{x}(n)_1, \mathbf{x}(n)_2, \cdots, \mathbf{x}(n)_D]$ where $\mathbf{X} \in \Re^D$ and $d = 1, 2, \cdots, D$ represents the dimension of the data or the number of random variables (features) in the data for each of the N observations in $\mathbf{X}$, we let $n = 1, 2, \cdots, N$. In Data Science, when analyzing data using probability theory (a fundamental study) it is possible to use pattern recognition techniques as well as machine learning to in order to determine what the data represents. These foundational concepts are crucial for understanding more advanced topics and algorithms in data science.

2 Review

In this section, we will cover the basic concepts of probability, including events, sample spaces, and outcomes. We begin the discussion by reviewing some definitions and concepts from probability theory. These definitions and concepts are required to ensure we are using a common vocabulary and have the basic set of tools necessary to discuss algorithms probabilistically. We begin by considering a set of items $\mathbf{S}$.

2.1 Basic Definitions

Definition: A *sample space* is a set $\mathbf{S}$ whose elements are called elementary events. The sample space $\mathbf{S}$ is the set of all possible outcomes of a random experiment. For example, a single coin toss, the sample space is $\mathbf{S} = \{\text{Heads}, \text{Tails}\}$.

Definition: An *outcome* is a single result from the sample space of a random experiment. An elementary event can be treated as a possible outcome of an experiment. For example, for a single die roll, one possible outcome is 4

Definition: An *event* is a subset of the sample space. It is a set of outcomes to which a probability is assigned. An event is a subset of $\mathbf{S}$. For example, for a single die roll, the event $\mathbf{A}$ of rolling an even number can be expressed as $\mathbf{A} = 2, 4, 6$.

Definition: Two events, $\mathbf{A}$ and $\mathbf{B}$, are mutually exclusive if and only if $\mathbf{A} \cap \mathbf{B} = \emptyset$.

By definition, all elementary events are mutually exclusive. In this class we will treat the set **S** as a set of data observations in **X**.

2.2 Probability Axioms

Probability axioms are the foundational rules that govern the assignment of probabilities to events. They ensure that probability measures are consistent and logical. Understanding these axioms is crucial for anyone working with probabilistic models, algorithms, and data analysis in data science. The three fundamental axioms of probability are non-negativity, normalization, and additivity. These axioms provide a robust framework for reasoning about uncertainty and making informed decisions based on data.

Non-Negativity

The probability of any event A is a non-negative number:

$$P(A) \geq 0$$

For example, a fair six-sided die, the probability of rolling any particular number (e.g., 3) is:

$$P(\text{rolling a 3}) = \frac{1}{6} \geq 0$$

Normalization

The probability of the sample space S is equal to 1:

$$P(S) = 1$$

For example, a fair six-sided die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. The sum of the probabilities of all possible outcomes is:

$$P(S) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

Additivity

For any two mutually exclusive events A and B, the probability of their union is the sum of their individual probabilities:

$$P(A \cup B) = P(A) + P(B) \quad \text{if } A \cap B = \emptyset$$

For example, a fair six-sided die, let A be the event of rolling an even number and B be the event of rolling a number greater than 4. Assuming A and B are mutually exclusive events:

$$A = \{2\}, \quad B = \{5\}$$

Then,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

Examples Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Non-negativity Example

Let's define event A as selecting a flower with a sepal length greater than 5 cm. The probability $P(A)$ must be non-negative.

$$P(A) = \frac{\text{Number of flowers with sepal length greater than 5 cm}}{\text{Total number of flowers}} \geq 0$$

Given the dataset, suppose 120 flowers have a sepal length greater than 5 cm:

$$P(A) = \frac{120}{150} = 0.8 \geq 0$$

Normalization Example

The probability of selecting any flower from the dataset is 1 since the sample space S includes all 150 flowers.

$$P(S) = \frac{\text{Number of all flowers}}{\text{Total number of flowers}} = \frac{150}{150} = 1$$

Additivity Example

Define events A and B as follows:

- A: Selecting a Setosa flower.
- B: Selecting a Versicolor flower.

Since a flower cannot be both Setosa and Versicolor at the same time, A and B are mutually exclusive.

$$P(A) = \frac{50}{150} = \frac{1}{3}, \quad P(B) = \frac{50}{150} = \frac{1}{3}$$
$$P(A \cup B) = P(A) + P(B) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

2.3 Probability Law

2.3.1 Law of Large Numbers

The Law of Large Numbers (LLN) is a fundamental theorem in probability theory that describes the behavior of the average of a large number of independent, identically distributed random variables. It states that as the number of trials increases, the sample average converges to the expected value. This theorem provides the theoretical foundation for the reliability of statistical estimates and is essential for understanding why large samples provide more accurate estimates of population parameters. In data science, LLN is crucial for validating the use of large datasets and ensuring robust statistical inference.

The Law of Large Numbers asserts that as the size of a sample increases, the sample mean will converge to the population mean with high probability. There are two forms of the Law of Large Numbers: the Weak Law of Large Numbers (WLLN) and the Strong Law of Large Numbers (SLLN). For a sequence of independent, identically distributed random variables $X_1, X_2, \ldots, X_n$ with expected value μ and finite variance, the sample mean $\overline{X}_n$ is given by:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The Weak Law of Large Numbers (WLLN) states that for any $\epsilon > 0$:

$$P\left(\left|\overline{X}_n - \mu\right| \geq \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty$$

The Strong Law of Large Numbers (SLLN) states that:

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1$$

Examples

Example 1: Coin Tossing

Consider a fair coin with probability of heads $p = 0.5$. Let X_i be a random variable representing the outcome of the i-th toss, where $X_i = 1$ for heads and $X_i = 0$ for tails. The expected value μ is:

$$\mu = E(X_i) = 0.5$$

The sample mean after n tosses is:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

According to the Law of Large Numbers, as n increases, $\overline{X}_n$ will converge to $\mu = 0.5$.

Example 2: Rolling a Die

Consider rolling a fair six-sided die. Let X_i be a random variable representing the outcome of the i-th roll. The expected value μ is:

$$\mu = E(X_i) = \frac{1+2+3+4+5+6}{6} = 3.5$$

The sample mean after n rolls is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to the Law of Large Numbers, as n increases, $\overline{X}_n$ will converge to $\mu = 3.5$.

Example 3: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's calculate the sample average of sepal length for increasing sample sizes and observe its convergence to the population mean.

Step 1: Calculate Population Mean

The population mean μ of sepal length is given by:

$$\mu = \frac{1}{150}\sum_{i=1}^{150} \text{sepal length}_i \approx 5.84\,\text{cm}$$

Step 2: Calculate Sample Averages

Let X_i be the sepal length of the i-th flower. The sample average $\overline{X}_n$ for a sample size n is given by:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We will calculate $\overline{X}_n$ for $n = 10, 20, 50, 100, 150$.

Given the sepal length data:

$$\mathbf{X} = \{5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, \ldots\}$$

For $n = 10$:

$$\overline{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i \approx 4.86\,\text{cm}$$

For $n = 20$:

$$\overline{X}_{20} = \frac{1}{20}\sum_{i=1_5}^{20} X_i \approx 5.03\,\text{cm}$$

For $n = 50$:

$$\overline{X}_{50} = \frac{1}{50} \sum_{i=1}^{50} X_i \approx 5.26\,\text{cm}$$

For $n = 100$:

$$\overline{X}_{100} = \frac{1}{100} \sum_{i=1}^{100} X_i \approx 5.51\,\text{cm}$$

For $n = 150$:

$$\overline{X}_{150} = \frac{1}{150} \sum_{i=1}^{150} X_i \approx 5.84\,\text{cm}$$

As the sample size n increases, the sample average $\overline{X}_n$ converges to the population mean $\mu \approx 5.84\,\text{cm}$, demonstrating the Law of Large Numbers.

2.3.2 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental principle in probability theory and statistics that describes the behavior of the sum or average of a large number of independent, identically distributed random variables. The CLT states that, regardless of the original distribution of the variables, the distribution of their sum or average tends to follow a normal distribution as the number of variables increases. This theorem is crucial in data science for making inferences about population parameters, constructing confidence intervals, and conducting hypothesis tests.

The Central Limit Theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables. This is true as long as the number of variables is sufficiently large and the variables have a finite mean and variance. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, identically distributed random variables with mean μ and variance σ^2. The sample mean $\overline{X}_n$ is given by:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

According to the Central Limit Theorem, the standardized form of $\overline{X}_n$ approaches a standard normal distribution as n approaches infinity:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\mathcal{N}(0,1)$ denotes the standard normal distribution.

Examples

Example 1: Coin Tossing

Consider tossing a fair coin n times, where $X_i = 1$ for heads and $X_i = 0$ for tails. The mean and variance of X_i are:

$$\mu = E(X_i) = 0.5, \quad \sigma^2 = \text{Var}(X_i) = 0.25$$

The sample mean after n tosses is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to the Central Limit Theorem, for large n:

$$\frac{\overline{X}_n - 0.5}{\sqrt{0.25/n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Example 2: Rolling a Die

Consider rolling a fair six-sided die n times. Let X_i be the outcome of the i-th roll. The mean and variance of X_i are:

$$\mu = E(X_i) = 3.5, \quad \sigma^2 = \mathrm{Var}(X_i) = \frac{35}{12}$$

The sample mean after n rolls is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

According to the Central Limit Theorem, for large n:

$$\frac{\overline{X}_n - 3.5}{\sqrt{35/(12n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

Example 3: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's calculate the sample means of sepal length for multiple samples of increasing sizes and observe their distribution.

Step 1: Calculate Population Mean and Variance

The population mean μ and variance σ^2 of sepal length are given by:

$$\mu = \frac{1}{150}\sum_{i=1}^{150} \text{sepal length}_i \approx 5.84\,\text{cm}$$

$$\sigma^2 = \frac{1}{150}\sum_{i=1}^{150} (\text{sepal length}_i - \mu)^2 \approx 0.68\,\text{cm}^2$$

Step 2: Calculate Sample Means

Let X_i be the sepal length of the i-th flower. The sample mean $\overline{X}_n$ for a sample size n is given by:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

We will calculate $\overline{X}_n$ for $n = 10, 20, 50, 100, 150$.

Given the sepal length data:

$$\mathbf{X} = \{5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, \ldots\}$$

For $n = 10$:

$$\overline{X}_{10} = \frac{1}{10} \sum_{i=1}^{10} X_i \approx 4.86\,\text{cm}$$

For $n = 20$:

$$\overline{X}_{20} = \frac{1}{20} \sum_{i=1}^{20} X_i \approx 5.03\,\text{cm}$$

For $n = 50$:

$$\overline{X}_{50} = \frac{1}{50} \sum_{i=1}^{50} X_i \approx 5.26\,\text{cm}$$

For $n = 100$:

$$\overline{X}_{100} = \frac{1}{100} \sum_{i=1}^{100} X_i \approx 5.51\,\text{cm}$$

For $n = 150$:

$$\overline{X}_{150} = \frac{1}{150} \sum_{i=1}^{150} X_i \approx 5.84\,\text{cm}$$

Step 3: Illustrate the Distribution of Sample Means

To visualize the CLT, we can generate multiple samples of a fixed size (e.g., $n = 30$) and plot the distribution of their means.

For k samples of size $n = 30$:

$$\overline{X}_1 = \frac{1}{30} \sum_{i=1}^{30} X_i^{(1)}, \quad \overline{X}_2 = \frac{1}{30} \sum_{i=1}^{30} X_i^{(2)}, \ldots, \overline{X}_k = \frac{1}{30} \sum_{i=1}^{30} X_i^{(k)}$$

As k increases, the distribution of $\overline{X}_k$ will approach a normal distribution with mean $\mu \approx 5.84\,\text{cm}$ and variance $\frac{\sigma^2}{30}$.

2.4 Expectation and Variance

2.4.1 Expected Value

Expected value, often referred to as the mean, is a fundamental concept in probability and statistics that provides a measure of the central tendency of a random variable. It represents the average outcome one would expect if an experiment or a random process were repeated many times. In data science, the expected value is crucial for various applications, including decision making, risk assessment, and predictive modeling.

The expected value of a random variable is the long-run average value of repetitions of the experiment it represents. For a discrete random variable, the expected value is the sum of all possible values each multiplied by its probability of occurrence. For a continuous random variable, the expected value is the integral of the variable with its probability density function. For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $P(X = x_i)$, the expected value $E(X)$ is given by:

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$$

For a continuous random variable X with probability density function $f(x)$, the expected value $E(X)$ is given by:

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

Examples

Example 1: Discrete Random Variable

Consider a fair six-sided die. What is the expected value of the outcome when the die is rolled?

Solution: The possible values are $1, 2, 3, 4, 5, 6$, each with probability $\frac{1}{6}$. The expected value is:

$$E(X) = \sum_{i=1}^{6} x_i P(X = x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$E(X) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5$$

Example 2: Continuous Random Variable

Consider a continuous random variable X with probability density function $f(x) = 2x$ for $0 \leq x \leq 1$. What is the expected value of X?

Solution: The expected value is:

$$E(X) = \int_0^1 x f(x)\, dx = \int_0^1 x \cdot 2x\, dx = \int_0^1 2x^2\, dx$$

Evaluating the integral:

$$E(X) = 2\int_0^1 x^2\,dx = 2\left[\frac{x^3}{3}\right]_0^1 = 2\cdot\frac{1}{3} = \frac{2}{3}$$

Example 3: Expected Value of a Sum of Random Variables

Consider two independent random variables X and Y representing the outcome of rolling two fair six-sided dice. What is the expected value of the sum $X + Y$?

Solution: The expected value of the outcome of each die is 3.5 (as calculated in Example 1). Since the expected value of the sum of two independent random variables is the sum of their expected values:

$$E(X + Y) = E(X) + E(Y) = 3.5 + 3.5 = 7$$

2.4.2 Variance

Variance is a key concept in probability and statistics that measures the dispersion or spread of a random variable around its mean (expected value). It quantifies how much the values of a random variable differ from the expected value, providing insight into the variability and consistency of the data. In data science, variance is crucial for understanding the reliability and precision of models and predictions.

Variance is defined as the expected value of the squared deviation of a random variable from its mean. For a random variable X, the variance is denoted by $\mathrm{Var}(X)$ or σ^2. For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $P(X = x_i)$, the variance $\mathrm{Var}(X)$ is given by:

$$\mathrm{Var}(X) = E[(X - \mu)^2] = \sum_{i=1}^{n}(x_i - \mu)^2 P(X = x_i)$$

where $\mu = E(X)$ is the expected value of X.

For a continuous random variable X with probability density function $f(x)$, the variance $\mathrm{Var}(X)$ is given by:

$$\mathrm{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty}(x - \mu)^2 f(x)\,dx$$

Examples

Example 1: Discrete Random Variable

Consider a fair six-sided die. What is the variance of the outcome when the die is rolled?

Solution: The possible values are $1, 2, 3, 4, 5, 6$, each with probability $\frac{1}{6}$. The expected value is:

$$E(X) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

The variance is:

$$\text{Var}(X) = \sum_{i=1}^{6}(x_i - \mu)^2 P(X = x_i) = \sum_{i=1}^{6}(x_i - 3.5)^2 \cdot \frac{1}{6}$$

$$\text{Var}(X) = \frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6}$$

$$\text{Var}(X) = \frac{(2.5)^2 + (1.5)^2 + (0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2}{6}$$

$$\text{Var}(X) = \frac{6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25}{6} = \frac{17.5}{6} \approx 2.92$$

Example 2: Continuous Random Variable

Consider a continuous random variable X with probability density function $f(x) = 2x$ for $0 \leq x \leq 1$. What is the variance of X?

Solution: The expected value is:

$$E(X) = \int_0^1 x f(x)\, dx = \int_0^1 x \cdot 2x\, dx = \int_0^1 2x^2\, dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

The variance is:

$$\text{Var}(X) = \int_0^1 (x-\mu)^2 f(x)\, dx = \int_0^1 \left(x - \frac{2}{3}\right)^2 \cdot 2x\, dx$$

First, expand $(x - \frac{2}{3})^2$:

$$\left(x - \frac{2}{3}\right)^2 = x^2 - \frac{4}{3}x + \frac{4}{9}$$

Then, integrate each term:

$$\text{Var}(X) = \int_0^1 \left(x^2 - \frac{4}{3}x + \frac{4}{9}\right) 2x\, dx = 2\int_0^1 \left(x^3 - \frac{4}{3}x^2 + \frac{4}{9}x\right) dx$$

$$\text{Var}(X) = 2\left[\frac{x^4}{4} - \frac{4}{3}\cdot\frac{x^3}{3} + \frac{4}{9}\cdot\frac{x^2}{2}\right]_0^1 = 2\left[\frac{1}{4} - \frac{4}{9} + \frac{2}{9}\right]$$

$$\text{Var}(X) = 2\left(\frac{1}{4} - \frac{2}{9}\right) = 2\left(\frac{9-8}{36}\right) = 2\left(\frac{1}{36}\right) = \frac{2}{36} = \frac{1}{18} \approx 0.056$$

2.4.3 Standard Deviation

Standard deviation is a crucial statistical measure that quantifies the amount of variation or dispersion in a set of data values. It is derived from variance and provides an indication of how much the values in a dataset deviate from the mean. In data science, standard deviation is widely used to assess the variability and reliability of data, which is essential for model evaluation, risk assessment, and decision making.

Standard deviation is defined as the square root of the variance. It provides a measure of the spread of a set of values around the mean, expressed in the same units as the data. For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $P(X = x_i)$, the standard deviation σ is given by:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i)}$$

where $\mu = E(X)$ is the expected value of X.

For a continuous random variable X with probability density function $f(x)$, the standard deviation σ is given by:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx}$$

Examples

Example 1: Discrete Random Variable

Consider a fair six-sided die. What is the standard deviation of the outcome when the die is rolled?

Solution: The possible values are $1, 2, 3, 4, 5, 6$, each with probability $\frac{1}{6}$. The expected value is:

$$E(X) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

The variance is:

$$\text{Var}(X) = \frac{(1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2}{6}$$

$$\text{Var}(X) = \frac{6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25}{6} = \frac{17.5}{6} \approx 2.92$$

The standard deviation is:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{2.92} \approx 1.71$$

Example 2: Continuous Random Variable

Consider a continuous random variable X with probability density function $f(x) = 2x$ for $0 \leq x \leq 1$. What is the standard deviation of X?

Solution: The expected value is:

$$E(X) = \int_0^1 x f(x)\, dx = \int_0^1 x \cdot 2x\, dx = \int_0^1 2x^2\, dx = 2 \left[\frac{x^3}{3} \right]_0^1 = \frac{2}{3}$$

The variance is:

$$\text{Var}(X) = \int_0^1 (x - \mu)^2 f(x)\, dx = \int_0^1 \left(x - \frac{2}{3} \right)^2 \cdot 2x\, dx$$

First, expand $(x - \frac{2}{3})^2$:

$$\left(x - \frac{2}{3}\right)^2 = x^2 - \frac{4}{3}x + \frac{4}{9}$$

Then, integrate each term:

$$\text{Var}(X) = \int_0^1 2\left(x^3 - \frac{4}{3}x^2 + \frac{4}{9}x\right)\,dx$$

$$\text{Var}(X) = 2\left[\frac{x^4}{4} - \frac{4}{9}\cdot\frac{x^3}{3} + \frac{4}{9}\cdot\frac{x^2}{2}\right]_0^1 = 2\left[\frac{1}{4} - \frac{4}{27} + \frac{2}{9}\right]$$

$$\text{Var}(X) = 2\left(\frac{1}{4} - \frac{4}{27} + \frac{2}{9}\right) = 2\left(\frac{27}{108} - \frac{16}{108} + \frac{24}{108}\right) = 2\left(\frac{35}{108}\right) = \frac{70}{108} = \frac{35}{54}$$

The standard deviation is:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\frac{35}{54}} \approx 0.80$$

2.5 Handling Probabilities

Now that we have a basic understanding of the sample space and an initial data set, lets look at how we can handle probabilities. When considering a particular algorithm (or experiment), we represent some performance measure as a variable with some set of possible outcomes. When the outcomes are stochastic in nature (meaning we cannot determine a priori what the outcome will be), then we refer to these variables as random variables. For a particular random variable X, we will represent the probability that X takes on some value x as $P(X = x) = p$, where $0 \leq p \leq 1$. This defines a probability distribution over the values of the random variable X. For example, suppose X is a Boolean random variable (meaning that there are only two possible values for X). Suppose X can take on the values True and False. Then we can define the Boolean probability distribution over X to be $P(X = True) = p$ and $P(X = False) = (1 - p)$.

There are many different philosophical approaches to handling probabilities. Most of the probabilities handled in this class will tend to follow the "frequentist" approach in that probabilities will be derived from explicit experiments. However, an alternative approach that offers tremendous power in algorithm design is the"Bayesian" approach. In Bayesian probability, proposal distributions and belief estimates are used.

Both approaches make a distinction between "unconditional" and "conditional" probabilities. An unconditional probability is a probability that is determined based on no other information. In Bayesian probability, these probabilities are sometimes called "prior" or a priori probabilities. A conditional probability is a probability whose value is determined based on the existence of other information. In Bayesian probability, conditional probabilities tend to appear in one of two roles—as a likelihood estimate or as a "posterior" or a posteriori probability. Suppose we have two events a and b. The unconditional probability of a is denoted as $P(a)$. The conditional probability of a given that we know b, however, is denoted as $P(a|b)$. Conditional probabilities are defined as $P(a|b) = P(a, b)/P(b)$.

Notice the numerator to the definition of the conditional probability. The notation $P(a, b)$ denotes a joint probability over the two events a and b. One common task when manipulating probabilities is extracting the distribution of a subset of variables over a single variable. The process of extracting the distribution is called either marginalizing or conditioning depending on the form used. (In general, we will use the term marginalize even when mathematically we are conditioning. This is because, as we will see, the two approaches are mathematically equivalent.)

- Marginalizing: $P(Y) = \sum_z P(\mathbf{Y}, \mathbf{z})$
- Conditioning: $P(Y) = \sum_z P(\mathbf{Y}, \mathbf{z})P(\mathbf{z})$

Consider marginalizing. We calculate the probability distribution of event Y from the joint distribution $P(\mathbf{Y}, \mathbf{z})$ by summing the probabilities over all possible values of z. Recall the definition of conditional probability. Using this definition, we can rewrite $P(\mathbf{Y}, \mathbf{z})$ as $P(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$. Thus we see that conditioning and marginalizing are equivalent.

Another common use of probabilities is in estimating new probabilities given other probabilities.

One common rule used is Bayes' Rule (which forms the foundation of the Bayesian approach). Bayes' Rule can be derived directly from the definition of conditional probability. First, recall that $P(a,b) = P(a|b)(P(b)$. Observing that $P(a,b) = P(b,a)$, we can also show that $P(a,b) = P(b|a)P(a)$. Setting the right hand side of these two equations to be equal to each other, we derive Bayes' Rule:

$$P(a|b)P(b) = P(b|a)P(a) \tag{1}$$

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \tag{2}$$

As an alternative definition, we can "reverse" the process of conditioning by noting that $P(b) = \sum_i P(b|a_i)P(a_i)$. So we can now rewrite Bayes' Rule as

$$P(a|b) = \frac{P(b|a)P(a)}{\sum_i P(b|a_i)P(b)} \tag{3}$$

At times, we will also need to be concerned with the relationship between random variables. In particular, we will make use of the concept of independence. We say that two random variables X and Y and independent if any of the following hold:

$$P(X|Y) = P(X)$$
$$P(Y|X) = P(Y)$$
$$P(X,Y) = P(X)P(Y)$$

We can apply the same type of definitions with conditional probabilities. Specifically, given three random variables, X, Y , and Z, then we can say X and Y are conditionally independent given Z if any of the following hold:

$$P(X|Y,Z) = P(X|Z)$$
$$P(Y|X,Z) = P(Y|Z)$$
$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$

Given a probability distribution, other things we might want to know about that distribution are several summary statistics. Specifically, the mean or expected value of a distribution can be determined as $E[X] = \sum_x xP(X = x)$. Expectation has a nice property, called the linearity of expectation, in which we find that $E[X + Y] = E[X] + E[Y]$. It is interesting to note that this holds even when X and Y are not independent.

We do not want to end with expected value since the expected value says nothing about the shape of the distribution. There are several other statistics that can describe the shape (called "moments" of the distribution) such as variance, skew, and kurtosis. We will concern ourselves only with variance. (Actually, the mean is also a "moment" of the distribution—it is the first moment). Variance provides a measure of how "spread out" the distribution is. Variance is defined as an

expected value over the amount of variation in the distribution as follows:

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E^2[X]] \\ &= E[X^2] - 2E[XE[X]] + E^2[X] \\ &= E[X^2] - 2E[X]E[X] + E^2[X] \\ &= E[X^2] - E^2[X] \end{aligned} \tag{4}$$

Generally, it is not possible to determine the true mean and variance of a distribution based on a set of experiments that have been run. However, the sample mean and sample variance can be determined as approximations of the underlying mean and variance. These are defined for univariate data as follows:

- Sample Mean: $E[X] = \bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- Sample Variance: $Var[X] = s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{X})^2$

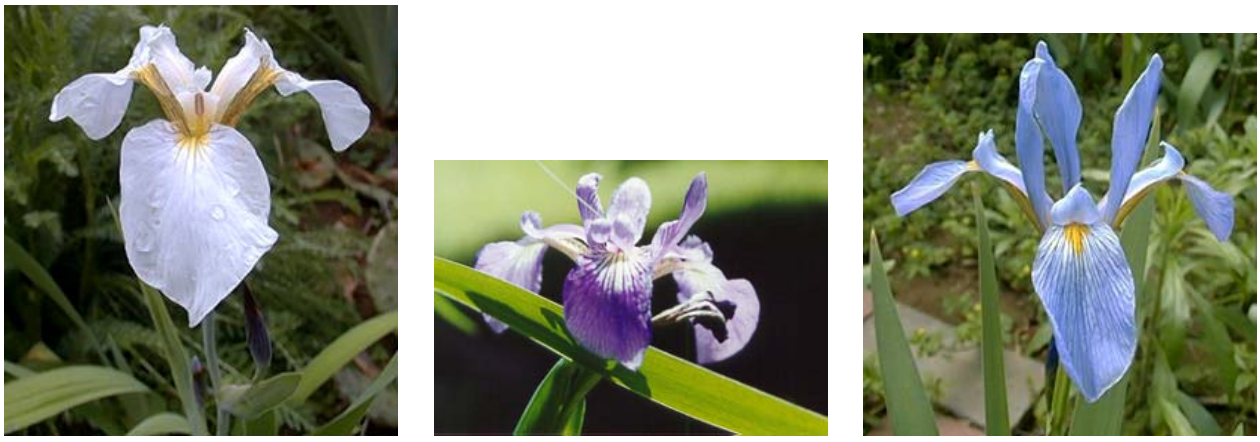
For multivariate data we define the mean and variance as follows:

- Sample Mean: $E[\mathbf{X}] = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{N} \mathbf{x}_n$
- Sample Variance: $Var[\mathbf{X}] = \mathbf{s}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{X}})^2$

Univariate data analysis involves the analysis of data with a single variable/feature while multivariate data analysis contains two or more variables/features. Most multivariate data analysis involves a dependent variable/featres and multiple independent variables/featres. Note that the square root of the variance is called the "standard deviation."

2.6 Example

Data of other forms and how to represent them will be discussed in the Data Processing module. In this example, let's consider the Iris data set containing 3 classes of 50 observations each, where each class refers to a type of Iris species. Four attributes/features were collected for each plant observation. To formalize the Iris data set mathematically, let it be represented by the $[150 \times 4]$ matrix $\mathbf{X}$, each observation is represented by a $[1 \times 4]$ vector $\mathbf{x}(n)$, the first 50 observations for $n = 1, \cdots, 50$) represent the Setosa species, the 50 observations from $n = 51, \cdots, 100$ represent the Versicolor species, and the final 50 observations $n = 101, \cdots, 150$ represent the Virginica species. The three species are the three classes of the data represented by $c = [1, 2, 3]$. The data in $\mathbf{X}$ contains four features ($D = 4$) that represent the sepal length, sepal width, petal length and petal width.



(a) Setosa (b) Versicolor (c) Virginica

Figure 1: Iris plant images courtesy of STPRTool

Let's look at the three different flowers and generate some probability values shown in Table 1.

From this complex table there is sufficient information to make determinations as to the data and how to best process the data. Now let's see what each of the test statistics means without evaluating the Plant Class for now. Let's consider the minimum and maximum values. In this case it can be determined that the data has the range of approximately 0.1 to 7.9 or mathematically represented as $\mathbf{x} \in [0.1, 7.9]$ in which case it will be easier to represent the range as $\mathbf{x} \in [0, 8]$. If we break this down even farther each of the features can have their own ranges as shown in in Table 2. In this case the ranges are in units of centimeters and are not that wide spread. In the data processing module data normalization will be discussed to ensure various statistical measures give the correct representation.

Now let's look at the mean and standard deviation of the data to get a representation of the data and what is meant to ask what is the "mean" and "standard deviation" of the data. Let's consider the data by features for all flower types, in this case the mean and corresponding standard deviation are shown in Table 3. Analyzing the results several conclusion can be drawn. First the largest mean is in the sepal length where the mean is 5.8433 with a corresponding standard deviation of

Table 1: Data Analysis Statistics

Test Statistics	**Flower Type**	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**	**Plant Class**
Minimum	Setosa	4.3000	2.3000	1.0000	0.1000	1
	Versicolor	4.9000	2.0000	3.0000	1.0000	2
	Virginica	4.9000	2.2000	4.5000	1.4000	3
Maximum	Setosa	5.8000	4.4000	1.9000	0.6000	1
	Versicolor	7.0000	3.4000	5.1000	1.8000	2
	Virginica	7.9000	3.8000	6.9000	2.5000	3
Mean	Setosa	5.0060	3.4180	1.4640	0.2440	1
	Versicolor	5.9360	2.7700	4.2600	1.3260	2
	Virginica	6.5880	2.9740	5.5520	2.0260	3
Standard Deviation +/-	Setosa	0.3525	0.3810	0.1735	0.1072	1
	Versicolor	0.5162	0.3138	0.4699	0.1978	2
	Virginica	0.6359	0.3225	0.5519	0.2747	3
Skewness	Setosa	0.1165	0.1038	0.0697	1.1610	1
	Versicolor	0.1022	-0.3519	-0.5882	-0.0302	2
	Virginica	0.1144	0.3549	0.5328	-0.1256	3
Kurtosis	Setosa	2.6542	3.6851	3.8137	4.2965	1
	Versicolor	2.4012	2.5517	2.9256	2.5122	2
	Virginica	2.9121	3.5198	2.7435	2.3387	3

Table 2: Data Minimum and Maximum Values by Feature

Test Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Minimun	4.3000	2.000	1.000	0.1000
Maximum	7.9000	4.4000	6.9000	2.5000

0.8281 and the smallest mean is the petal width with a mean of 1.1987 and corresponding standard deviation of 0.7632. Looking at the results of two features it may be concluded that the sepal length has a tighter distribution in the data. If the data is investigated further and the minimum and maximum feature values are included it is noted that the spread of data is $7.9 - 4.3 = 3.6$ for the sepal length and the spread for the petal width is $2.5 - 0.1 = 2.4$ which leads to an initial analysis that the larger mean with smaller standard deviation is not necessarily the best feature for evaluation.

This leads to the question as too which is the best feature(s) in the data for evaluation? It is not trivial to answer this question, meaning, what is it that is being evaluated in the data. Let's assume that currently we are interested in evaluating the four features and how their "test statistics" measure. So far only the minimum, maximum, mean and standard deviation of the data have been evaluated without regard to the plant class which the full set of test statistics are shown in

Table 3: Mean and Standard Deviation by Feature

Test Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.8433	3.0540	3.7587	1.1987
Standard Deviation	0.8281	0.4336	1.7644	0.7632

Table 4: Skewness and Kurtosis by Feature

Test Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Skewness	0.3118	0.3307	-0.2717	-0.1039
Kurtosis	2.4264	3.2414	1.6046	1.6648

Table 1. To complete the initial analysis let's consider the skewness and kurtosis by feature and not flower type.

The results in Table 4 show the skewness and kurtosis for each feature in the data.

Let's look at the meaning of skewness and determine the feature(s) with the "best" skewness value. The measure of data asymmetry around the sample mean is know as skewness. In the event the skewness value is negative, the data has a spread to the left of the mean. When the value is positive the spread is to the right of the mean. The value returned is zero when perfect symmetry is about the mean. With this said the feature with the best skewness is the petal width of the Iris flower with a value of -0.1039.

Now let's look at the meaning of the kurtosis and determine the feature(s) with the "best" kurtosis value. So in this case the kurtosis provides a measure of how prone is the data to outliers. The introduction to outliers in data is provided in the data preprocessing module so for now outlier(s) is simply defined as a data point that differs significantly from other data points for a particular feature. A kurtosis value of 3 indicates the data has a normal distribution. Values that are above 3 are known to have outliers and values that are less than 3 do not have outliers. This can be normalized to have a mean value of 0 so anything above 0 will have outliers and below 0 do not have outliers. So in analyzing the results from Table 4, it can be seen that the sepal width has the value closest to 3.

This analysis provides an initial insight of the data. From the analysis of the test statistics used it is not clear which feature(s) are the best or have an ability to rank them in some form at this time. These statistics provide insight into what the features represent in terms of the range of the data, the mean and the spread of the data. In the upcoming data processing module additional preprocessing methods will be introduced for the purpose of pattern recognition and pattern classification. In addition it should be mentioned that the statistics in this section are best suited for data that is normally distributed.

3 First Order Statistics

The following table shows the mathematical equations of the mean, standard deviation, skewness, and kurtosis. The input is $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ where $\mathbf{X} \in \Re^D$. The maximum and minimum are also included.

Table 5: Data Analysis Statistics

Test Statistics	**Statistical Function** $F(\cdot)$
Minimum	$F_{\min}(\mathbf{X}) = \min(\mathbf{X}) = \mathbf{x}_{min}$
Maximum	$F_{\max}(\mathbf{X}) = \max(\mathbf{X}) = \mathbf{x}_{max}$
Mean	$F_{\mu}(\mathbf{X}) = \boldsymbol{\mu} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$
α Trimmed Mean	$F_{\mu_\alpha}(\mathbf{X}, \alpha) = \boldsymbol{\mu_\alpha} = \frac{1}{\lfloor N - 2N\alpha \rfloor} \sum_{\lfloor N\alpha \rfloor}^{\lfloor N - N\alpha \rfloor} \mathbf{x}_n$
p Trimmed Mean	$F_{\mu_p}(\mathbf{X}, p) = \boldsymbol{\mu_p} = \frac{1}{N-2p} \sum_{p}^{N-p} \mathbf{x}_n$
Standard Deviation	$F_{\sigma}(\mathbf{X}) = \boldsymbol{\sigma} = \left(\frac{1}{N-1} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu} \right)^2 \right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{X}) = \boldsymbol{\gamma} = \frac{\frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu} \right)^3}{\boldsymbol{\sigma}^3}$
Kurtosis	$F_{\kappa}(\mathbf{X}) = \boldsymbol{\kappa} = \frac{\frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu} \right)^4}{\boldsymbol{\sigma}^4}$

3.1 Calculating First Order Statistics in Python

When calculating first order statistics, patterns arise. Often one will have to sum the items in a list, whether it is for the mean or for the kurtosis, so we will create helper functions. We will also use the numpy and math libraries.

```
def makeList(data, column):
    """Makes a list of all values at index column

    Keyword arguments:
    data -- 2d np array
    column -- index to be put into list
    """

    dataList = []
    for l in data:
        dataList.append(l[column])

    return dataList

def sumList(l):
    """Sums items of list l

    Keyword arguments:
    l - list to sum
    """

    sum = 0
    for item in l:
        sum += item

    return sum
```

Figure 2: Helper Functions

To find the maximum and minimum, set a maximum or minimum variable to the first item in a list, and iterate through the list to see if any values are above or below the set maximum. Then, set the variable equal to the new value.

```
def maxCalc(data, column):
    """Calculate the max of a numpy array

    Keyword arguments:
    data -- the 2d numpy array
    column -- the index of data for the max to be calculated on
    """

    dataList = makeList(data, column)

    max = dataList[0]

    for item in dataList:
        if item > max:
            max = item

    return max
```

Figure 3: Maximum

```
def minCalc(data, column):
    """Calculate the min of a numpy array

    Keyword arguments:
    data -- the 2d numpy array
    column -- the index of data for the min to be calculated on
    """

    dataList = makeList(data, column)

    min = dataList[0]

    for item in dataList:
        if item < min:
            min = item

    return min
```

Figure 4: Minimum

Note that:

$$F_{\mu}(\mathbf{X}) = \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \frac{\sum_{n=1}^{N}\mathbf{x}_n}{N}. \tag{5}$$

To calculate the mean, sum all of the items in the data set, then divide by the number of items in the data set.

The mean represents the typical value in the data set.

```
def meanCalc(data, column):
    """Calculate the mean of a numpy array's column

    Keyword arguments:
    data -- the 2d numpy array
    column -- the index of data for the mean to be calculated on
    """

    dataList = makeList(data, column)

    # Calculate numerator
    sum = sumList(dataList)

    # Calculate denominator
    n = len(dataList)
    mean = sum/n

    return mean
```

Figure 5: Mean (μ)

The trimmed mean is a variation of the mean which is calculated by removing values from the beginning and end of a sorted set of data. The average is then taken using the remaining values. This allows any potential outliers to be removed when calculating the statistics of the data.

One variation of the trimmed mean is the alpha trimmed mean. In the alpha trimmed mean, a percentage from each end of the sorted data set is removed. If $\mathbf{x} = [x_1, ..., x_n]$ is sorted and α is the percentage to be trimmed, then $\mathbf{x}_\alpha = [x_{n\alpha}, ..., x_{n-n\alpha}]$Note that:

$$F_{\mu_\alpha}(\mathbf{X}, \alpha) = \boldsymbol{\mu_\alpha} = \frac{1}{\lfloor N - 2N\alpha \rfloor} \sum_{\lfloor N\alpha \rfloor}^{\lfloor N - N\alpha \rfloor} \mathbf{x}_n = \frac{\sum_{\lfloor N\alpha \rfloor}^{\lfloor N - N\alpha \rfloor} \mathbf{x}_n}{\lfloor N - 2N\alpha \rfloor}. \tag{6}$$

```
def alphaTrimmedMeanCalc(data, column, a):
    """Calculate the alpha trimmed mean of a numpy array's column

    Keyword arguments
    data -- the 2d numpy array
    column -- the index of data for the alpha trimmed mean to be calculated on
    a -- alpha, the percentage of items to trim
    """

    dataList = makeList(data, column)
    dataList.sort()

    n = len(dataList)
    k = int(n * a) # int() floors/truncates

    # Calculate numerator
    dataList = dataList[k: -k]
    numerator = sumList(dataList)

    # Calculate denominator
    r = len(dataList) # R = N - N * alpha, or the new length of the list

    aTrimmedMean = numerator/r

    return aTrimmedMean
```

Figure 6: α Trimmed Mean (μ_α)

Another variation of the trimmed mean is the p trimmed mean. In this variation, p values are removed from the beginning of the data set and the end of the data set, assuming the data set is sorted. The resulting data set is $\mathbf{x}_p = [x_p, ..., x_{n-p}]$. Note that:

$$F_{\mu_p}(\mathbf{X}, p) = \boldsymbol{\mu_p} = \frac{1}{N - 2p} \sum_{p}^{N-p} \mathbf{x}_n = \frac{\sum_{p}^{N-p} \mathbf{x}_n}{N - 2p}. \tag{7}$$

```
def pTrimmedMeanCalc(data, column, p):
    """Calculate the "p" trimmed mean of a numpy array's column

    Keyword arguments:
    data -- the 2d numpy array
    column -- the index of data for the "p" trimmed mean to be calculated on
    p -- the number of items to trim
    """

    dataList = makeList(data, column)
    dataList.sort()

    # Calculate numerator
    dataList = dataList[p: -p]
    numerator = sumList(dataList)

    # Calculate denominator
    r = len(dataList) # r = N - (2 * p), or the new length of the list

    # Calculate p Trimmed Mean
    pTrimmedMean = numerator/r

    return pTrimmedMean
```

Figure 7: p Trimmed Mean (μ_p)

The standard deviation can be calculated by subtracting each item in the data set by the mean, squaring it, and summing it with the rest of the values in the data set. Then, divide this sum by one less than the number of items in the data set and take the square root.

Standard deviation represents the typical distance a value is away from the mean.

Note that:

$$F_\sigma(\mathbf{X}) = \boldsymbol{\sigma} = \left(\frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^2\right)^{1/2} = \sqrt{\frac{\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^2}{N-1}}. \tag{8}$$

```
def standardDeviationCalc(data, column):
    """Calculate the standard deviation of a numpy array's column

    Keyword arguments:
    data -- the 2d numpy array
    column -- the index of data for the standard deviation to be calculated on
    """

    dataList = makeList(data, column)
    mean = meanCalc(data, column)


    # Calculate the numerator
    numeratorList = [] # Create a list for every individual value in the numerator

    for item in dataList:
        numeratorList.append((item - mean)**2)

    # Calculate denominator
    sum = sumList(numeratorList)
    n = len(numeratorList) - 1 # Set n equal to number of values in data set

    # Calculate variance
    variance = sum/n

    # Calculate standard deviation
    standardDeviation = math.sqrt(variance)

    return standardDeviation
```

Figure 8: Standard Deviation (σ)

Skewness can be calculated by subtracting each item from the mean, cubing it, and adding them will all of the other items. Then divide the sum by the number of items in the data set multiplied by the standard deviation cubed.

Skewness is used to determine if the distribution is skewed, or whether the distribution has a "tail".

Note that:

$$F_\gamma(\mathbf{X}) = \boldsymbol{\gamma} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^3}{\boldsymbol{\sigma}^3} = \frac{\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^3}{(N)(\boldsymbol{\sigma}^3)}. \tag{9}$$

```
def skewnessCalc(data, column):
    """Calculate the skewness of a numpy array's column

    data -- the 2d numpy array
    column -- the index of data for the skewness to be calculated on
    """

    dataList = makeList(data, column)
    mean = meanCalc(data, column)

    numeratorList = [] # Create a list for every individual value in the numerator

    # Calculate numerator
    for item in dataList:
        numeratorList.append(pow(item - mean, 3)) # Raise every value to power of 3
    numerator = sumList(numeratorList)

    # Calculate denominator
    n = len(numeratorList)
    denominator = n*(pow(standardDeviationCalc(data, column), 3))

    # Calculate skewness
    skewness = numerator/denominator

    return skewness
```

Figure 9: Skewness (γ)

Calculating the kurtosis is similar to calculating the skewness. However there are some differences. Each individual item subtracted by the mean is raised to the fourth power instead of the third. The standard deviation is also raised to the fourth power.

Kurtosis symbolizes how "sharp" a curve is compared to the rest of the distribution.

Note that:

$$F_{\kappa}(\mathbf{X}) = \boldsymbol{\kappa} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^4}{\boldsymbol{\sigma}^4} = \frac{\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^4}{(N)(\boldsymbol{\sigma}^4)}. \tag{10}$$

```
def kurtosisCalc(data, column):
    """Calculate the kurtosis of a numpy array's column

    data -- the 2d numpy array
    column -- the index of data for the kurtosis to be calculated on
    """

    dataList = makeList(data, column)
    mean = meanCalc(data, column)

    numeratorList = [] # Create a list for every individual value in the numerator

    # Calculate numerator
    for item in dataList:
        numeratorList.append(pow(item - mean, 4)) # Raise every value to power of 4
    numerator = sumList(numeratorList)

    # Calculate denominator
    n = len(numeratorList)
    denominator = (n) * (pow(standardDeviationCalc(data, column), 4))

    #Calculate kurtosis
    kurtosis = numerator/denominator

    return kurtosis
```

Figure 10: Kurtosis (κ)

4 Distributions

In the following, we will use $P(a)$ to denote the probability of event a. Then we say that a probability distribution $P()$ on $\mathbf{S}$ is a mapping from events in $\mathbf{S}$ to real numbers $\Re$ such that the following holds:

- $P(a) \geq 0$ for all $a \subseteq \mathbf{S}$,
- $P(\mathbf{S}) = 1$, and
- $P(a \cup b) = P(a) + P(b)$ for any two mutually exclusive events a and b.

Note that if the events a and b are not mutually exclusive, then we have $P(a \cup b) = P(a) + P(b) - P(a \cap b)$.

Given the axioms of probability and the notion of a probability distribution, we are now prepared to consider several types and examples of probability distributions. Generally, we are concerned with two types of distributions—discrete and continuous. A probability distribution is said to be discrete if it is defined over a finite (or countably infinite) sample space. A probability distribution is said to be continuous, on the other hand, if the sample space is neither finite nor countably infinite. When considering these distributions, we refer to the associated probability functions as probability mass functions and probability density functions respectively.

For a particular discrete probability distribution over a sample space $\mathbf{S}$ with event a, we can define the following:

$$P(a) = \sum_{s \in a \subseteq \mathbf{S}} P(s) \tag{11}$$

Thus, the probability distribution assigns a probability value to every event (elementary or otherwise) in the sample space $\mathbf{S}$. For example, suppose $\mathbf{S}$ is finite and every has a probability $P(s) = 1/|\mathbf{S}|$, where s is an elementary event. This particular distribution is called a uniform distribution. Two common examples of sample spaces with uniform probability distributions are coin flips and die tosses. If one has a fair coin with two sides, then $\mathbf{S} = \{Heads, Tails\}$, and $P(Heads) = P(Tails) = 1/2$. If one has a fair die with six sides, then $\mathbf{S} = \{1, 2, 3, 4, 5, 6\}$, and $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

For continuous probability distributions, it is common that these distributions are defined over ranges of values (typically closed intervals0 since the probability of any specific value is zero. Consider the uniform probability distribution for a continuous sample space. Then consider two closed intervals, $[a, b] \in \mathbf{S}$ and $[c, d] \in \mathbf{S}$, where $[c, d] \in [a, b]$, then $P([c, d]) = (d - c)/(b - a)$. An interesting corollary to working with intervals over a uniform distribution is that we can define events to be any subset of $[a, b]$ obtained by a finite or countable union of either open or closed intervals. Note specifically that for any $a, b, \in \mathbf{S}, P([a, a]) = P([b, b]) = 0$. Note also that the definition of a closed interval permits use to rewrite that interval as $[a, b] = [a, a] \cup (a, b) \cup [b, b]$. Therefore, we can substitute in the point probabilities to show that $P([a, b]) = P((a, b))$.

4.1 Moments of Distributions

When a set of data values has a strong central tendency, that is, a tendency to cluster around some particular value, then it may be useful to characterize the data set by a few values that have been calculated numerically, which are known as the data moments. This also assumes that the data has a random nature and is limited by a set range.

In this subsection, the moments described are the mean, standard deviation, skewness, and kurtosis. In addition the data minimum and maximum are introduced as they are important in the analysis of data. Table 6 shows the mathematical representation of the moments and the inputs variable $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ where $\mathbf{X} \in \Re^D$.

Table 6: Data Analysis Statistics

Test Statistics	**Statistical Function $\boldsymbol{F}(\cdot)$**
Minimum	$F_{\min}(\mathbf{X}) = \min(\mathbf{X}) = \mathbf{x}_{min}$
Maximum	$F_{\max}(\mathbf{X}) = \max(\mathbf{X}) = \mathbf{x}_{max}$
Mean	$F_{\mu}(\mathbf{X}) = \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$
Standard Deviation	$F_{\sigma}(\mathbf{X}) = \boldsymbol{\sigma} = \left(\frac{1}{N-1}\sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)^2\right)^{1/2}$
Skewness	$F_{\gamma}(\mathbf{X}) = \boldsymbol{\gamma} = \frac{\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)^3}{\boldsymbol{\sigma}^3}$
Kurtosis	$F_{\kappa}(\mathbf{X}) = \boldsymbol{\kappa} = \frac{\frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_n - \boldsymbol{\mu}\right)^4}{\boldsymbol{\sigma}^4}$

The difference between using $\frac{1}{N}$ for the mean and $\frac{1}{N-1}$ for the standard deviation or variance is derived from their purposes and from statistical theory.

4.1.1 Mean

The mean of a dataset is a measure of central tendency. It simply involves summing all the data points and dividing by the total number of data points N. Each data point contributes equally, and the objective is straightforward.

$$F_{\mu}(\mathbf{X}) = \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \tag{12}$$

4.1.2 Standard Deviation (STD) and Variance

The standard deviation is a measure of how spread out the values in a dataset are. It is determined by taking the square root of the variance. When calculating the sample variance, the term $\frac{1}{N-1}$ is used to make the estimator "unbiased."

$$F_{var}(\mathbf{X}) = \boldsymbol{var} = \frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^2 \tag{13}$$

$$F_{\sigma}(\mathbf{X}) = \boldsymbol{\sigma} = \left(\frac{1}{N-1}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^2\right)^{1/2} \tag{14}$$

The reason for using $N-1$ instead of N is related to the concept of degrees of freedom and the difference between a sample and a population.

4.1.3 Degrees of Freedom

In statistical theory, the concept of degrees of freedom (DoF) refers to the number of independent pieces of information that go into the calculation of a statistic, such as a sample mean or variance. The degrees of freedom are critical in determining various aspects of statistical inference, including the shape of different probability distributions that underlie hypothesis tests and confidence intervals.

For **Bias Correction**, the degrees of freedom help in making certain estimators unbiased. When calculating the sample variance, we first calculate the sample mean. This estimated mean is influenced by the sample data and may not be the actual population mean. For example, when estimating population variance using a sample, $\frac{1}{N-1}$ is used to correct for bias, where N is the sample size and $N-1$ is the degrees of freedom.

4.1.4 Biased vs. Unbiased Estimator

Using $\frac{1}{N}$ would give a "biased" estimate of the population variance, systematically underestimating the true population variance. Using $\frac{1}{N-1}$ makes it an "unbiased" estimator, providing a more accurate approximation of the true population variance when dealing with samples.

4.1.5 Sample vs. Population

If you're working with a complete population, $\frac{1}{N}$ is used because there's no sampling error. The discrepancy arises when you're trying to estimate population parameters (like variance) from a sample.

The **population** includes all individuals or items of interest for a particular study. For example, if you want to know the average height of all adult men in a country, the population would consist of the heights of all adult men in that country.

A **sample** is a subset of individuals or items taken from the larger population. For instance, measuring the heights of 1,000 adult men randomly selected from the country constitutes a sample. Measures calculated from a sample are called sample statistics. The average height and variance

of height in your sample are sample statistics. Because a sample is a subset of the population, the sample statistics are subject to sampling error, which is the degree to which the sample differs from the population. For a sample to provide meaningful insights about a population, it should be representative. This is often achieved through random sampling techniques.

4.1.6 Skewness Calculation with Sample STD or Population STD

The choice between $\frac{1}{N}$ and $\frac{1}{N-1}$ when calculating the standard deviation (and consequently skewness) depends on whether you're dealing with a sample or the entire population. The use of $N-1$ Instead of N is known as Bessel's correction, and it's typically applied when estimating the population standard deviation based on a sample.

Sample Standard Deviation
When working with a sample and trying to generalize to a population, you typically use $1-N$ in the denominator. The formula for sample standard deviation σ is shown in Equation 14.

Population Standard Deviation
When working with an entire population or when you're not interested in generalizing to a broader population, you use N in the denominator. The formula for population standard deviation σ_p is shown as follows:

$$F_{\sigma_p}(\mathbf{X}) = \boldsymbol{\sigma_p} = \Big(\frac{1}{N-1}\sum_{n=1}^{N}\big(\mathbf{x}_n - \boldsymbol{\mu}\big)^2\Big)^{1/2} \tag{15}$$

Skewness Sample vs. Population

When calculating skewness, the same principle applies. If you're calculating sample skewness and want to generalize to a population, you would typically use the sample standard deviation with Bessel's correction (N-1 in the denominator). If you're calculating population skewness, then you would use the population standard deviation.

The formula for sample skewness is often given as:

$$F_{\gamma}(\mathbf{X}) = \boldsymbol{\gamma} = \frac{\frac{1}{N}\sum_{n=1}^{N}\big(\mathbf{x}_n - \boldsymbol{\mu}\big)^3}{\boldsymbol{\sigma}^3} \tag{16}$$

Here, σ is the sample standard deviation calculated with $N-1$ in the denominator.

Use Sample Standard Deviation when you have one of the following data sets.

- ◇ **You have a sample from a larger population**: If your data is a sample and you intend to make inferences about a larger population, you should use the sample standard deviation. The sample standard deviation is calculated with $\frac{1}{N-1}$ in the denominator, recall is known as Bessel's correction.
- ◇ **You are performing exploratory data analysis on a sample**: Even if you don't intend to generalize the results formally, using the sample standard deviation is more conservative when you're working with a subset of data.

The formula for population skewness is often given as:

$$F_\gamma(\mathbf{X}) = \boldsymbol{\gamma} = \frac{\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{x}_n - \boldsymbol{\mu}\right)^3}{\boldsymbol{\sigma_p^3}} \tag{17}$$

Here, σ_p is the population standard deviation calculated with N in the denominator.

Use Population Standard Deviation when you have one of the following data sets.

- ◇ **You have the entire population**: If you have data from every member of the population and are not making inferences to a larger group, then you should use the population standard deviation, calculated with $\frac{1}{N}$ in the denominator.
- ◇ **The data represents a "conceptual" population**: Sometimes, you might be working with sample data but are interested in describing only that specific dataset rather than generalizing to a larger population. In such cases, you might opt to use the population standard deviation.

4.1.7 Bessel's Correction

Bessel's correction is the use of $N-1$ instead of N in the formula to compute the sample variance and, consequently, the sample standard deviation. This correction is applied when estimating the population variance or standard deviation from a sample.

When working with a sample and want to make inferences about the population from which the sample is drawn, using N in the denominator tends to underestimate the population variance. This is because the sample mean $\bar{\mathbf{x}}$ is closer to the data points in the sample than the population mean μ would be. This makes the numerator (the squared differences) smaller than it would be if calculated with μ, leading to a biased estimate for the population variance.

Using $N-1$ instead of N corrects this bias, making the estimate unbiased. In other words, the expected value of the sample variance (calculated with $N-1$) is equal to the population variance.

You should use Bessel's correction when you are working with a sample and want to estimate the population variance or standard deviation. If you're calculating the variance of a whole population or don't intend to generalize the sample to a population, you would not apply Bessel's correction.

4.2 Covariance Matrix

The importance in understanding the covariance matrix is the the use of linear transformation using the covariance matrix. This is an important research are in understanding and using principal component analysis (PCA), singular value decomposition (SVD), Bays classifier and the Mahalanobis distance. Each of these topics are widely used in statistics and pattern recognition.

The covariance matrix is given by $\mathbf{C} \in \Re^{D \times D}$. This implies that the covariance matrix is a square matrix of D by D dimensions representing the features of the data matrix in $\mathbf{X}$. The covariance is calculated as follows:

$$\mathbf{C} = \frac{1}{N-1} \sum_{n=1}^{N} (\mathbf{X}_n - \bar{\mathbf{X}})(\mathbf{X}_n - \bar{\mathbf{X}})^T \tag{18}$$

The diagonal of the covariance matrix is the variance of the individual features in the data set.

The covariance matrix can be implemented in Python with the following code.

```
def covarianceCalc(data, row1, row2):
    """ Find the covariance matrix of a 2d numpy array

    Keyword arguments:
    data -- the 2d numpy array
    row1, row2 -- the features of the dataset to calculate covariance on
    """
    dataList1 = Eigen.makeList(data, row1)
    dataList2 = Eigen.makeList(data, row2)

    mean1 = meanCalc(data, row1)
    mean2 = meanCalc(data, row2)

    sum = 0
    for i, item in enumerate(dataList1):
        zeroMean1 = dataList1[i] - mean1
        zeroMean2 = dataList2[i] - mean2

    sum += zeroMean1 * zeroMean2

    # population covariance, N, not N-1
    n = len(dataList1)

    covariance = sum/n

    return covariance
```

Figure 11: Covariance Calculation

```
def covarianceMatrix(data):
    """ Find the covariance matrix of a 2d numpy array

    Keyword arguments:
    data -- the 2d numpy array
    """

    numRows = np.shape(data)[0]

    covarianceMatrix = np.empty(shape = [numRows, numRows])

    for i, row in enumerate(covarianceMatrix):
        for j, col in enumerate(row):
            covarianceMatrix[i][j] = covarianceCalc(data, i, j)

    return covarianceMatrix
```

Figure 12: Covariance Matrix ($\mathbf{C}$ or Σ)

4.2.1 Understanding Data Distributions

In the normal distribution, each standard deviation can be represent by a certain probability. Through the empirical rule, it is given that the probability of a data point being within $\mu - \sigma$ and $\mu + \sigma$ is 0.6827, the probability of a data point being within $\mu - 2\sigma$ and $\mu + 2\sigma$ is 0.9545, and the probability of a data point being within $\mu - 3\sigma$ and $\mu + 3\sigma$ is 0.9973.

The sigma ellipse plots is a different way of representing the standard deviations of different features of a data set. Each ellipse around the mean represents an additional standard deviation, which can then be mapped to the given distributions to check for normality in each feature.

For further interest, we can also calculate the confidence interval error ellipse with the following equation [2]

$$\left(\frac{x}{X\sigma_X}\right)^2 + \left(\frac{y}{X\sigma_Y}\right)^2 = S \tag{19}$$

where S is a critical value, such that $P(s < S) = 1 - X\sigma$ in the χ^2 distribution and X is the number of standard deviations the ellipse should be calculated for.

- describe what 1σ, 2σ, and 3σ mean for a distribution
- The confidence interval error ellipse are to be calculated with the following values: 68.27% 95.45% 99.73% 2.3 6.17 11.8

4.2.2 Common Probability Distributions

Probability distributions are fundamental in the field of data science, providing a mathematical framework for modeling and analyzing random phenomena. They describe how probabilities are

distributed over the values of a random variable and are classified into two main types: discrete and continuous distributions. Discrete distributions apply to scenarios where the random variable can take on a countable number of distinct values, while continuous distributions apply to scenarios where the random variable can take on any value within a given range. Understanding these distributions is essential for performing statistical analyses, making predictions, and developing probabilistic models in data science. This subsection will provide an overview of common discrete and continuous probability distributions, their characteristics, and applications.

5 Discrete Distributions

Discrete probability distributions are vital for modeling and analyzing countable outcomes in various data science applications. These distributions describe the probabilities of distinct, separate values and are often used when dealing with scenarios involving counting events, trials, or successes. Each discrete distribution has unique characteristics and is suitable for different types of data and problems. In this section, we will explore several common discrete distributions used in data science, including the Bernoulli, Binomial, Poisson, Geometric, Negative Binomial, and Hypergeometric distributions. Understanding these distributions is crucial for accurate data modeling, hypothesis testing, and decision-making based on discrete data.

5.1 Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution that models binary outcomes, such as success/failure or yes/no events. It is the simplest discrete distribution and is foundational for understanding more complex discrete distributions like the Binomial distribution. The Bernoulli distribution is characterized by a single parameter, p, which represents the probability of success. Understanding the Bernoulli distribution is crucial for data scientists as it forms the basis for binary classification problems and various statistical methods.

A discrete random variable X is said to follow a Bernoulli distribution with parameter $p \in [0, 1]$ if its probability mass function (PMF) is given by:

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

This distribution is denoted as $X \sim \text{Bernoulli}(p)$.

Properties

- **Mean (μ)**: Given by p.
- **Variance (σ^2)**: Given by $p(1 - p)$.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We create a binary variable X indicating whether the petal length is greater than 3 cm ($X = 1$) or not ($X = 0$).

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Let $X_i = 1$ if the petal length is greater than 3 cm and $X_i = 0$ otherwise. We can calculate the proportion p of petal lengths greater than 3 cm as follows:

$$p = \frac{\text{Number of petal lengths} > 3 \text{ cm}}{\text{Total number of samples}}$$

Suppose the number of petal lengths greater than 3 cm is 100 out of 150 samples. Then,

$$p = \frac{100}{150} = \frac{2}{3}$$

Therefore, $X \sim \text{Bernoulli}\left(\frac{2}{3}\right)$.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = x) = \begin{cases} \frac{2}{3} & \text{if } x = 1 \\ \frac{1}{3} & \text{if } x = 0 \end{cases}$$

Mean and Variance

The mean and variance of X are given by:

$$\mu = p = \frac{2}{3}$$

$$\sigma^2 = p(1-p) = \frac{2}{3}\left(1 - \frac{2}{3}\right) = \frac{2}{3} \times \frac{1}{3} = \frac{2}{9}$$

5.2 Binomial Distribution

The Binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. It is widely used in data science and statistics to model binary outcomes over multiple trials, such as the number of successes in a series of experiments or the number of positive responses in a survey. Understanding the Binomial distribution is crucial for performing various statistical analyses, including hypothesis testing and confidence interval estimation.

A discrete random variable X is said to follow a Binomial distribution with parameters n (number of trials) and $p \in [0, 1]$ (probability of success in each trial) if its probability mass function (PMF) is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

where $\binom{n}{k}$ is the binomial coefficient:

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

This distribution is denoted as $X \sim \text{Binomial}(n, p)$.

Properties

- **Mean (μ)**: Given by np.
- **Variance (σ^2)**: Given by $np(1-p)$.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We create a binary variable X_i indicating whether the petal length is greater than 3 cm ($X_i = 1$) or not ($X_i = 0$).

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Let $X_i = 1$ if the petal length is greater than 3 cm and $X_i = 0$ otherwise. We can calculate the proportion p of petal lengths greater than 3 cm as follows:

$$p = \frac{\text{Number of petal lengths} > 3 \text{ cm}}{\text{Total number of samples}}$$

Suppose the number of petal lengths greater than 3 cm is 100 out of 150 samples. Then,

$$p = \frac{100}{150} = \frac{2}{3}$$

Now, consider the number of samples with petal length greater than 3 cm in a subset of 10 samples. This can be modeled using a Binomial distribution with $n = 10$ and $p = \frac{2}{3}$. Therefore, $X \sim \text{Binomial}(10, \frac{2}{3})$.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = k) = \binom{10}{k} \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{10-k} \quad \text{for } k = 0, 1, 2, \ldots, 10$$

Mean and Variance

The mean and variance of X are given by:

$$\mu = np = 10 \times \frac{2}{3} = \frac{20}{3} \approx 6.67$$

$$\sigma^2 = np(1-p) = 10 \times \frac{2}{3} \times \left(1 - \frac{2}{3}\right) = 10 \times \frac{2}{3} \times \frac{1}{3} = \frac{20}{9} \approx 2.22$$

Example Calculation

To find the probability of exactly 7 samples having petal lengths greater than 3 cm in a subset of 10 samples:

$$P(X = 7) = \binom{10}{7} \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right)^3$$

Calculating the binomial coefficient and the probabilities:

$$\binom{10}{7} = \frac{10!}{7!3!} = 120$$

$$\left(\frac{2}{3}\right)^7 \approx 0.0577$$

$$\left(\frac{1}{3}\right)^3 \approx 0.0370$$

Therefore,

$$P(X = 7) = 120 \times 0.0577 \times 0.0370 \approx 0.256$$

Thus, the probability of exactly 7 out of 10 samples having petal lengths greater than 3 cm is approximately 0.256.

5.3 Poisson Distribution

The Poisson distribution is a discrete probability distribution that models the number of events occurring within a fixed interval of time or space, given the average number of times the event occurs over that interval. It is widely used in fields such as queueing theory, telecommunications, and reliability engineering. Understanding the Poisson distribution is essential for data scientists when analyzing count data and rare events.

A discrete random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ if its probability mass function (PMF) is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

This distribution is denoted as $X \sim \text{Poisson}(\lambda)$, where λ is the average rate (mean number of events) in the given interval.

Properties

- **Mean (μ)**: Given by λ.
- **Variance (σ^2)**: Given by λ.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We want to model the number of occurrences of petal lengths between 1 and 2 cm using a Poisson distribution.

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Count the number of occurrences of petal lengths between 1 and 2 cm. Suppose the count is 30 out of 150 samples. We can calculate the rate λ as follows:

$$\lambda = \frac{\text{Number of petal lengths between 1 and 2 cm}}{\text{Total number of samples}} = \frac{30}{150} = 0.2$$

Therefore, $X \sim \text{Poisson}(0.2)$.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = k) = \frac{0.2^k e^{-0.2}}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

Mean and Variance

The mean and variance of X are given by:

$$\mu = \lambda = 0.2$$

$$\sigma^2 = \lambda = 0.2$$

Example Calculation

To find the probability of exactly 2 occurrences of petal lengths between 1 and 2 cm:

$$P(X = 2) = \frac{0.2^2 e^{-0.2}}{2!} = \frac{0.04 \cdot e^{-0.2}}{2} \approx \frac{0.04 \cdot 0.8187}{2} \approx 0.0164$$

Thus, the probability of exactly 2 occurrences of petal lengths between 1 and 2 cm is approximately 0.0164.

5.4 Geometric Distribution

The Geometric distribution is a discrete probability distribution that models the number of trials needed to get the first success in a sequence of independent Bernoulli trials, each with the same probability of success. It is a fundamental distribution in probability theory and is widely used in various applications such as reliability testing, quality control, and survival analysis. Understanding the Geometric distribution is essential for data scientists when analyzing and modeling scenarios involving the waiting time until the first occurrence of an event.

A discrete random variable X is said to follow a Geometric distribution with parameter $p \in (0, 1]$ if its probability mass function (PMF) is given by:

$$P(X = k) = (1 - p)^{k-1}p \quad \text{for } k = 1, 2, 3, \ldots$$

This distribution is denoted as $X \sim \text{Geometric}(p)$.

Properties

- **Mean (μ)**: Given by $\frac{1}{p}$.
- **Variance (σ^2)**: Given by $\frac{1-p}{p^2}$.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We want to model the number of samples we need to check before finding a petal length greater than 3 cm using a Geometric distribution.

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Calculate the proportion p of petal lengths greater than 3 cm:

$$p = \frac{\text{Number of petal lengths} > 3 \text{ cm}}{\text{Total number of samples}} = \frac{100}{150} = \frac{2}{3}$$

Therefore, $X \sim \text{Geometric}\left(\frac{2}{3}\right)$.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = k) = \left(\frac{1}{3}\right)^{k-1}\left(\frac{2}{3}\right) \quad \text{for } k = 1, 2, 3, \ldots$$

Mean and Variance

The mean and variance of X are given by:

$$\mu = \frac{1}{p} = \frac{1}{2/3} = \frac{3}{2} = 1.5$$

$$\sigma^2 = \frac{1-p}{p^2} = \frac{1 - 2/3}{(2/3)^2} = \frac{1/3}{4/9} = \frac{1/3 \times 9/4}{=}\frac{3}{4} = 0.75$$

Example Calculation

To find the probability of finding the first petal length greater than 3 cm on the third sample:

$$P(X = 3) = \left(\frac{1}{3}\right)^{3-1} \left(\frac{2}{3}\right) = \left(\frac{1}{3}\right)^{2} \left(\frac{2}{3}\right) = \frac{1}{9} \times \frac{2}{3} = \frac{2}{27} \approx 0.0741$$

Thus, the probability of finding the first petal length greater than 3 cm on the third sample is approximately 0.0741.

5.5 Negative Binomial Distribution

The Negative Binomial distribution is a discrete probability distribution that models the number of trials needed to achieve a specified number of successes in a sequence of independent Bernoulli trials, each with the same probability of success. This distribution generalizes the Geometric distribution and is particularly useful for modeling count data with overdispersion, where the variance exceeds the mean. Understanding the Negative Binomial distribution is essential for data scientists when analyzing and modeling count data, especially in scenarios involving repeated trials until a set number of successes is reached.

A discrete random variable X is said to follow a Negative Binomial distribution with parameters r (number of successes) and $p \in (0, 1]$ (probability of success in each trial) if its probability mass function (PMF) is given by:

$$P(X = k) = \binom{k + r - 1}{k} (1 - p)^k p^r \quad \text{for } k = 0, 1, 2, \ldots$$

where $\binom{k+r-1}{k}$ is the binomial coefficient:

$$\binom{k + r - 1}{k} = \frac{(k + r - 1)!}{k!(r - 1)!}$$

This distribution is denoted as $X \sim \text{Negative Binomial}(r, p)$.

Properties

- **Mean (μ)**: Given by $\frac{r(1-p)}{p}$.
- **Variance (σ^2)**: Given by $\frac{r(1-p)}{p^2}$.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We want to model the number of samples we need to check to find 5 petal lengths greater than 3 cm using a Negative Binomial distribution.

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Calculate the proportion p of petal lengths greater than 3 cm:

$$p = \frac{\text{Number of petal lengths} > 3 \text{ cm}}{\text{Total number of samples}} = \frac{100}{150} = \frac{2}{3}$$

Therefore, $X \sim$ Negative Binomial$(5, \frac{2}{3})$.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = k) = \binom{k+4}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^5 \quad \text{for } k = 0, 1, 2, \ldots$$

Mean and Variance

The mean and variance of X are given by:

$$\mu = \frac{r(1-p)}{p} = \frac{5 \times \frac{1}{3}}{\frac{2}{3}} = \frac{5 \times 1}{2} = \frac{5}{2} = 2.5$$

$$\sigma^2 = \frac{r(1-p)}{p^2} = \frac{5 \times \frac{1}{3}}{\left(\frac{2}{3}\right)^2} = \frac{5 \times 1}{\frac{4}{9}} = \frac{5 \times 9}{4} = \frac{45}{4} = 11.25$$

Example Calculation

To find the probability of finding the 5th petal length greater than 3 cm on the 8th sample:

$$P(X = 3) = \binom{3+4}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^5$$

Calculating the binomial coefficient and the probabilities:

$$\binom{7}{3} = \frac{7!}{3!4!} = 35$$

$$\left(\frac{1}{3}\right)^3 = \frac{1}{27}$$

$$\left(\frac{2}{3}\right)^5 \approx 0.1317$$

Therefore,

$$P(X = 3) = 35 \times \frac{1}{27} \times 0.1317 \approx 0.1704$$

Thus, the probability of finding the 5th petal length greater than 3 cm on the 8th sample is approximately 0.1704.

5.6 Hypergeometric Distribution

The Hypergeometric distribution is a discrete probability distribution that models the number of successes in a fixed-size sample drawn without replacement from a finite population. Unlike the Binomial distribution, which deals with sampling with replacement, the Hypergeometric distribution is particularly useful in scenarios where the probability of success changes on each draw due to the finite population. This makes it relevant for applications such as quality control, ecological studies, and any situation where sampling without replacement occurs. Understanding the Hypergeometric distribution is essential for data scientists when analyzing such data.

A discrete random variable X is said to follow a Hypergeometric distribution with parameters N (population size), K (number of successes in the population), and n (sample size) if its probability mass function (PMF) is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \text{for } \max(0, n - (N - K)) \leq k \leq \min(n, K)$$

where $\binom{a}{b}$ is the binomial coefficient:

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

This distribution is denoted as $X \sim \text{Hypergeometric}(N, K, n)$.

Properties

- **Mean (μ)**: Given by $\frac{nK}{N}$.
- **Variance (σ^2)**: Given by $\frac{nK(N-K)(N-n)}{N^2(N-1)}$.

Example Using the Iris Dataset

Consider the petal length of iris flowers in the Iris dataset. We want to model the number of samples with petal lengths greater than 3 cm in a randomly drawn sample of 30 flowers from the dataset.

Given the petal length data:

$$\mathbf{X} = \{1.4, 1.4, 1.3, 1.5, 1.4, 4.7, 4.5, 4.9, 4.0, 4.6, \ldots\}$$

Suppose the number of petal lengths greater than 3 cm in the entire dataset is 100 out of 150 samples. Therefore, we have:

$$N = 150, \quad K = 100, \quad n = 30$$

We want to find the probability of getting exactly 20 samples with petal lengths greater than 3 cm in a sample of 30 flowers.

Probability Mass Function (PMF)

The PMF of X is:

$$P(X = 20) = \frac{\binom{100}{20}\binom{50}{10}}{\binom{150}{30}}$$

Calculating the binomial coefficients:

$$\binom{100}{20} = \frac{100!}{20!(100-20)!}, \quad \binom{50}{10} = \frac{50!}{10!(50-10)!}, \quad \binom{150}{30} = \frac{150!}{30!(150-30)!}$$

Using numerical software to calculate these values, we find:

$$P(X = 20) \approx \frac{2.696 \times 10^{23} \times 1.735 \times 10^{10}}{2.159 \times 10^{31}} \approx 0.0216$$

Thus, the probability of getting exactly 20 samples with petal lengths greater than 3 cm in a sample of 30 flowers is approximately 0.0216.

Mean and Variance

The mean and variance of X are given by:

$$\mu = \frac{nK}{N} = \frac{30 \times 100}{150} = 20$$

$$\sigma^2 = \frac{nK(N-K)(N-n)}{N^2(N-1)} = \frac{30 \times 100 \times 50 \times 120}{150^2 \times 149} \approx 5.37$$

6 Random Variables

Uniform deviates as defined by Press et al. (1986) [1] are just random number which lie within a specified range (typically 0 to 1), with any one number in the range just as likely as any other. They are, in other words, what you probably think "random numbers" are; however we want to distinguish uniform deviates from other sorts of "random numbers," for example, numbers drawn from a normal (Gaussian) distribution of specified mean and standard deviation. These other sorts of deviates are almost always generated by performing appropriate operations on one uniform deviates, as we will see in subsequent sections. So, a reliable source of random uniform deviates, the subject of this section, is an essential building block for any sort of stochastic modeling or Monte Carlo computer work.

6.1 Discrete Random Variables

Discrete random variables are an essential concept in probability theory and statistics, representing outcomes that take on a finite or countably infinite set of values. The probability mass function (PMF) is a fundamental tool for characterizing the probability distribution of discrete random variables. It provides the probability that a discrete random variable is exactly equal to each possible value. In data science, understanding PMFs is crucial for modeling and analyzing data that can be counted, such as the number of occurrences of an event.

6.1.1 Probability Mass Functions (PMF)

A probability mass function (PMF) of a discrete random variable X is a function that gives the probability that X is exactly equal to some value x. The PMF is denoted by $P(X = x)$ or $p(x)$. The PMF $p(x)$ must satisfy the following conditions:

- Non-negativity: $p(x) \geq 0$ for all x
- Normalization: $\sum_x p(x) = 1$

For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$, the PMF is given by:

$$p(x_i) = P(X = x_i)$$

Examples

Example 1: Fair Six-Sided Die

Consider a fair six-sided die. The possible values are $1, 2, 3, 4, 5, 6$. The PMF is:

$$p(x) = P(X = x) = \frac{1}{6} \quad \text{for } x = 1, 2, 3, 4, 5, 6$$

The PMF satisfies:

$$\sum_{x=1}^{6} p(x) = \sum_{x=1}^{6} \frac{1}{6} = 1$$

Example 2: Binomial Distribution

Consider a binomial random variable X representing the number of successes in n independent Bernoulli trials with success probability p. The possible values are $0, 1, 2, \ldots, n$. The PMF is:

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

Example 3: Poisson Distribution

Consider a Poisson random variable X representing the number of events occurring in a fixed interval of time or space with average rate λ. The possible values are $0, 1, 2, \ldots$. The PMF is:

$$p(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

Example 4: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Let X be the discrete random variable representing the species of an iris flower. X can take on three values: 1 (Setosa), 2 (Versicolor), and 3 (Virginica).

Step 1: Define the Random Variable

Define the random variable X as follows:

$$X = \begin{cases} 1 & \text{if the species is Setosa} \\ 2 & \text{if the species is Versicolor} \\ 3 & \text{if the species is Virginica} \end{cases}$$

Step 2: Calculate the PMF

Given the dataset, each species has 50 samples. The PMF $p_X(x)$ is given by:

$$p_X(x) = P(X = x) = \frac{\text{Number of flowers of species } x}{\text{Total number of flowers}}$$

$$p_X(1) = P(X = 1) = \frac{50}{150} = \frac{1}{3}$$

$$p_X(2) = P(X = 2) = \frac{50}{150} = \frac{1}{3}$$

$$p_X(3) = P(X = 3) = \frac{50}{150} = \frac{1}{3}$$

Step 3: Verify the PMF Properties

1. $0 \leq p_X(x) \leq 1$ for all x:

$$0 \leq \frac{1}{3} \leq 1$$

2. $\sum_x p_X(x) = 1$:

$$p_X(1) + p_X(2) + p_X(3) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$$

Thus, the PMF satisfies both properties.

6.2 Continuous Random Variables

Probability Density Functions (PDFs) are fundamental tools in probability theory and statistics for describing the distribution of continuous random variables. A PDF provides the relative likelihood of a random variable taking on a particular value and is crucial for understanding and modeling the behavior of continuous data. Unlike discrete random variables, where probabilities are assigned to specific values, PDFs describe the probability of the variable falling within a particular range. In data science, PDFs are essential for statistical inference, data analysis, and probabilistic modeling.

6.2.1 Probability Density Functions (PDF)

A probability density function (PDF) of a continuous random variable X is a function $f_X(x)$ that describes the relative likelihood for X to take on a given value. The PDF must satisfy the following conditions:

- Non-negativity: $f_X(x) \geq 0$ for all x
- Normalization: $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

For a continuous random variable X with PDF $f_X(x)$, the probability that X lies within the interval $[a, b]$ is given by:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

Examples

Example 1: Uniform Distribution

Consider a continuous random variable X that is uniformly distributed between 0 and 1. The PDF is:

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The probability that X lies between 0.2 and 0.5 is:

$$P(0.2 \leq X \leq 0.5) = \int_{0.2}^{0.5} f_X(x)\,dx = \int_{0.2}^{0.5} 1\,dx = 0.5 - 0.2 = 0.3$$

Example 2: Normal Distribution

Consider a continuous random variable X that is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. The PDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

The probability that X lies between -1 and 1 is:

$$P(-1 \leq X \leq 1) = \int_{-1}^{1} f_X(x)\, dx$$

This integral does not have a simple closed form, but it can be computed using numerical methods. The result is approximately:

$$P(-1 \leq X \leq 1) \approx 0.6827$$

Example 3: Exponential Distribution

Consider a continuous random variable X that is exponentially distributed with rate parameter $\lambda = 2$. The PDF is:

$$f_X(x) = \begin{cases} 2e^{-2x} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The probability that X lies between 1 and 2 is:

$$P(1 \leq X \leq 2) = \int_{1}^{2} 2e^{-2x}\, dx$$

Evaluating the integral:

$$P(1 \leq X \leq 2) = \left[-e^{-2x}\right]_1^2 = -e^{-4} + e^{-2} \approx 0.0902$$

Example 4: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's consider the sepal length of the iris flowers as a continuous random variable X.

Step 1: Calculate Population Mean and Variance

The population mean μ and variance σ^2 of sepal length are given by:

$$\mu = \frac{1}{150} \sum_{i=1}^{150} \text{sepal length}_i \approx 5.84\,\text{cm}$$

$$\sigma^2 = \frac{1}{150} \sum_{i=1}^{150} (\text{sepal length}_i - \mu)^2 \approx 0.68\,\text{cm}^2$$

Step 2: Define the PDF

Assume that the sepal length follows a normal distribution. The PDF of a normally distributed continuous random variable X with mean μ and variance σ^2 is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Substituting the values of μ and σ^2:

$$f_X(x) = \frac{1}{\sqrt{2\pi \cdot 0.68}} \exp\left(-\frac{(x-5.84)^2}{2 \cdot 0.68}\right)$$

Step 3: Calculate Probabilities

Let's calculate the probability that the sepal length is between 5 cm and 6 cm:

$$P(5 \leq X \leq 6) = \int_5^6 f_X(x)\, dx$$

$$\begin{aligned} P(5 \leq X \leq 6) &= \int_5^6 \frac{1}{\sqrt{2\pi \cdot 0.68}} \exp\left(-\frac{(x-5.84)^2}{2 \cdot 0.68}\right) dx \\ &\approx \frac{1}{\sqrt{4.28}} \int_5^6 \exp\left(-\frac{(x-5.84)^2}{1.36}\right) dx \\ &\approx \frac{1}{2.07} \int_5^6 \exp\left(-\frac{(x-5.84)^2}{1.36}\right) dx \end{aligned}$$

This integral can be evaluated using numerical methods or standard normal distribution tables. For simplicity, let's use a normal distribution calculator or software to find:

$$P(5 \leq X \leq 6) \approx 0.3829$$

Thus, the probability that the sepal length of an iris flower is between 5 cm and 6 cm is approximately 0.3829.

6.3 Cumulative Distribution Functions (CDF)

Cumulative Distribution Functions (CDFs) are fundamental tools in probability theory and statistics for describing the distribution of random variables. A CDF provides the probability that a random variable takes on a value less than or equal to a specific value. It is applicable to both discrete and continuous random variables and is essential for understanding the behavior and properties of probability distributions. In data science, CDFs are widely used for statistical inference, model evaluation, and hypothesis testing.

The cumulative distribution function (CDF) of a random variable X is a function that gives the probability that X will take a value less than or equal to x. It is denoted by $F_X(x)$. For a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$, the CDF $F_X(x)$ is given by:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

For a continuous random variable X with probability density function $f_X(x)$, the CDF $F_X(x)$ is given by:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$$

Examples

Example 1: Discrete Random Variable

Consider a fair six-sided die. The possible values are $1, 2, 3, 4, 5, 6$. The PMF is:

$$p(x) = P(X = x) = \frac{1}{6} \quad \text{for } x = 1, 2, 3, 4, 5, 6$$

The CDF is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{1}{6} & \text{for } 1 \leq x < 2 \\ \frac{2}{6} & \text{for } 2 \leq x < 3 \\ \frac{3}{6} & \text{for } 3 \leq x < 4 \\ \frac{4}{6} & \text{for } 4 \leq x < 5 \\ \frac{5}{6} & \text{for } 5 \leq x < 6 \\ 1 & \text{for } x \geq 6 \end{cases}$$

Example 2: Continuous Random Variable

Consider a continuous random variable X with probability density function $f_X(x) = 2x$ for $0 \leq x \leq 1$. The CDF is:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

For $0 \leq x \leq 1$:

$$F_X(x) = \int_{0}^{x} 2t\, dt = \left[t^2\right]_0^x = x^2$$

Thus, the CDF is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^2 & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

Example 3: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Discrete Random Variable: Species

Let X be the discrete random variable representing the species of an iris flower. X can take on three values: 1 (Setosa), 2 (Versicolor), and 3 (Virginica).

Step 1: Define the Random Variable

Define the random variable X as follows:

$$X = \begin{cases} 1 & \text{if the species is Setosa} \\ 2 & \text{if the species is Versicolor} \\ 3 & \text{if the species is Virginica} \end{cases}$$

Step 2: Calculate the PMF

Given the dataset, each species has 50 samples. The PMF $p_X(x)$ is given by:

$$p_X(1) = P(X = 1) = \frac{50}{150} = \frac{1}{3}$$

$$p_X(2) = P(X = 2) = \frac{50}{150} = \frac{1}{3}$$

$$p_X(3) = P(X = 3) = \frac{50}{150} = \frac{1}{3}$$

Step 3: Calculate the CDF

The CDF $F_X(x)$ is given by:

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} p_X(t)$$

$$F_X(1) = P(X \leq 1) = p_X(1) = \frac{1}{3}$$

$$F_X(2) = P(X \leq 2) = p_X(1) + p_X(2) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

$$F_X(3) = P(X \leq 3) = p_X(1) + p_X(2) + p_X(3) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$$

Continuous Random Variable: Sepal Length

Let Y be the continuous random variable representing the sepal length of the iris flowers.

Step 1: Calculate Population Mean and Variance

The population mean μ and variance σ^2 of sepal length are given by:

$$\mu = \frac{1}{150} \sum_{i=1}^{150} \text{sepal length}_i \approx 5.84\,\text{cm}$$

$$\sigma^2 = \frac{1}{150} \sum_{i=1}^{150} (\text{sepal length}_i - \mu)^2 \approx 0.68\,\text{cm}^2$$

Step 2: Define the PDF

Assume that the sepal length follows a normal distribution. The PDF of a normally distributed continuous random variable Y with mean μ and variance σ^2 is given by:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Substituting the values of μ and σ^2:

$$f_Y(y) = \frac{1}{\sqrt{2\pi \cdot 0.68}} \exp\left(-\frac{(y-5.84)^2}{2 \cdot 0.68}\right)$$

Step 3: Calculate the CDF

The CDF $F_Y(y)$ is given by:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,dt$$

Let's calculate the probability that the sepal length is less than or equal to 6 cm:

$$F_Y(6) = \int_{-\infty}^{6} \frac{1}{\sqrt{2\pi \cdot 0.68}} \exp\left(-\frac{(t-5.84)^2}{2 \cdot 0.68}\right)\,dt$$

This integral can be evaluated using numerical methods or standard normal distribution tables. For simplicity, let's use a normal distribution calculator or software to find:

$$F_Y(6) \approx 0.602$$

Thus, the probability that the sepal length of an iris flower is less than or equal to 6 cm is approximately 0.602.

7 Applications of Probability in Data Science

In this chapter, the application of probabilities in Data Science will include Machine Learning, data analysis, and A/B testing.

7.1 Machine Learning

Gaussian Processes

Gaussian Processes (GPs) are a powerful non-parametric approach in machine learning for regression and classification tasks. They provide a flexible way to define distributions over functions and can be used to make predictions with a measure of uncertainty. GPs are particularly useful in scenarios where the form of the underlying function is unknown, and they offer a principled way to incorporate prior knowledge and update beliefs as new data becomes available.

Gaussian Processes in Data Science

A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP is fully specified by its mean function and covariance function. For a set of inputs $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, the GP defines a distribution over possible functions $f(\mathbf{x})$ that fit the data.

Mean and Covariance Functions

The mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ are defined as:

$$m(\mathbf{x}) = E[f(\mathbf{x})]$$

$$k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

The GP is then written as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Gaussian Process Regression

In Gaussian Process regression, given training data $\{(x_i, y_i)\}_{i=1}^n$, where $y_i = f(x_i) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, we aim to predict the function values at new inputs $\mathbf{x}^*$.

The joint distribution of the training outputs $\mathbf{y}$ and the test outputs $\mathbf{f}^*$ under the prior is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix}\right)$$

The posterior distribution over function values at the test points is:

$$\mathbf{f}^* | \mathbf{X}, \mathbf{y}, \mathbf{X}^* \sim \mathcal{N}(\mu^*, \mathbf{\Sigma}^*)$$

where

$$\mu^* = K(\mathbf{X}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1}\mathbf{y}$$

$$\mathbf{\Sigma}^* = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1}K(\mathbf{X}, \mathbf{X}^*)$$

Examples

Example 1: Predicting a Noisy Sine Wave

Consider a dataset generated from a noisy sine wave: $y = \sin(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.1^2)$. We aim to predict the values of y at new points using Gaussian Process regression.

Given training data points (x_i, y_i), we define the mean function $m(x) = 0$ and use the squared exponential covariance function:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

Using the training data, we compute the posterior mean and covariance for the test points $\mathbf{x}^*$.

Example 2: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's perform Gaussian Process regression to predict the petal length based on the sepal length.

Step 1: Define the Data

Let x_i be the sepal length and y_i be the petal length. We have the training data $\{(x_i, y_i)\}_{i=1}^{150}$.

Step 2: Define the Covariance Function

We will use the squared exponential covariance function:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

where σ_f^2 is the signal variance and l is the length scale.

Step 3: Compute the Covariance Matrices

Let X be the vector of training inputs (sepal lengths) and y be the vector of training outputs (petal lengths). Compute the covariance matrices $K(X, X)$, $K(X, X_*)$, $K(X_*, X)$, and $K(X_*, X_*)$.

Step 4: Make Predictions

Given a new sepal length x_*, compute the predictive mean and covariance:

$$\bar{f}_* = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y$$
$$\text{cov}(f_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

Let's compute these for $x_* = 5.5$ cm:

$$\bar{f}_* = K(5.5, X)[K(X, X) + \sigma_n^2 I]^{-1} y$$

Assume $\sigma_f^2 = 1$, $l = 1$, and $\sigma_n^2 = 0.1$.

For simplicity, we will show the computation for a subset of the data.

Step 5: Example Calculation

Given $X = [5.1, 4.9, 4.7, 4.6, 5.0]$ and $y = [1.4, 1.4, 1.3, 1.5, 1.4]$, compute:

$$K(X,X) = \begin{pmatrix} 1 & \exp(-0.02) & \exp(-0.08) & \exp(-0.125) & \exp(-0.005) \\ \exp(-0.02) & 1 & \exp(-0.02) & \exp(-0.08) & \exp(-0.005) \\ \exp(-0.08) & \exp(-0.02) & 1 & \exp(-0.02) & \exp(-0.08) \\ \exp(-0.125) & \exp(-0.08) & \exp(-0.02) & 1 & \exp(-0.125) \\ \exp(-0.005) & \exp(-0.005) & \exp(-0.08) & \exp(-0.125) & 1 \end{pmatrix}$$

$$K(5.5,X) = \begin{pmatrix} \exp(-0.2) & \exp(-0.36) & \exp(-0.64) & \exp(-0.81) & \exp(-0.25) \end{pmatrix}$$

$$\bar{f}_* = K(5.5,X)[K(X,X) + 0.1I]^{-1} y$$

Perform the matrix inversion and multiplication to obtain the predictive mean.

7.2 Naive Bayes Classifier

The Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' Theorem, with the "naive" assumption that features are independent given the class label. Despite its simplicity and often unrealistic assumption of feature independence, the Naive Bayes classifier has proven to be effective for a variety of classification tasks, especially with large datasets and text classification problems.

Naive Bayes Classifier in Data Science

The Naive Bayes classifier applies Bayes' Theorem to compute the posterior probability of each class given the observed features. The class with the highest posterior probability is assigned to the instance. The "naive" assumption is that all features are conditionally independent given the class label. Bayes' Theorem for a given class C_k and features $x_1, x_2, \ldots, x_n$ is given by:

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|C_k)P(C_k)}{P(x_1, x_2, \ldots, x_n)}$$

Under the independence assumption, the joint likelihood $P(x_1, x_2, \ldots, x_n|C_k)$ simplifies to the product of individual likelihoods:

$$P(x_1, x_2, \ldots, x_n|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

Thus, the posterior probability becomes:

$$P(C_k|x_1, x_2, \ldots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(x_1, x_2, \ldots, x_n)}$$

Since $P(x_1, x_2, \ldots, x_n)$ is constant for all classes, it can be omitted for classification purposes:

$$P(C_k|x_1, x_2, \ldots, x_n) \propto P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

The predicted class $\hat{C}$ is the one that maximizes the posterior probability:

$$\hat{C} = \arg\max_{C_k} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

Examples

Example 1: Spam Email Classification

Consider a simple spam email classifier with two features: the presence of the word "win" (x_1) and the presence of the word "prize" (x_2). The possible classes are Spam (C_1) and Not Spam (C_2).

Given the probabilities:

$$P(C_1) = 0.4, \quad P(C_2) = 0.6$$
$$P(x_1 = \text{yes}|C_1) = 0.8, \quad P(x_1 = \text{yes}|C_2) = 0.1$$
$$P(x_2 = \text{yes}|C_1) = 0.7, \quad P(x_2 = \text{yes}|C_2) = 0.2$$

For an email with "win" and "prize" present ($x_1 = \text{yes}$, $x_2 = \text{yes}$), we compute:

$$P(C_1|x_1 = \text{yes}, x_2 = \text{yes}) \propto P(C_1)P(x_1 = \text{yes}|C_1)P(x_2 = \text{yes}|C_1)$$
$$P(C_1|x_1 = \text{yes}, x_2 = \text{yes}) \propto 0.4 \times 0.8 \times 0.7 = 0.224$$

$$P(C_2|x_1 = \text{yes}, x_2 = \text{yes}) \propto P(C_2)P(x_1 = \text{yes}|C_2)P(x_2 = \text{yes}|C_2)$$
$$P(C_2|x_1 = \text{yes}, x_2 = \text{yes}) \propto 0.6 \times 0.1 \times 0.2 = 0.012$$

The predicted class is Spam (C_1) because:

$$0.224 > 0.012$$

Example 2: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's classify an iris flower with the following features:

$$X = \{\text{sepal length} = 5.1, \text{sepal width} = 3.5, \text{petal length} = 1.4, \text{petal width} = 0.2\}$$

Step 1: Calculate Priors

The prior probability $P(C)$ for each class is given by the proportion of each class in the dataset:

$$P(\text{Setosa}) = P(C = \text{Setosa}) = \frac{50}{150} = \frac{1}{3}$$

$$P(\text{Versicolor}) = P(C = \text{Versicolor}) = \frac{50}{150} = \frac{1}{3}$$

$$P(\text{Virginica}) = P(C = \text{Virginica}) = \frac{50}{150} = \frac{1}{3}$$

Step 2: Calculate Likelihoods

Assume the likelihoods $P(x_i \mid C)$ are normally distributed. For each feature x_i and class C, the likelihood is given by:

$$P(x_i \mid C) = \frac{1}{\sqrt{2\pi\sigma_{C,i}^2}} \exp\left(-\frac{(x_i - \mu_{C,i})^2}{2\sigma_{C,i}^2}\right)$$

Given the feature values for each class, calculate $\mu_{C,i}$ and $\sigma_{C,i}^2$.

For Setosa:

$$P(\text{sepal length} = 5.1 \mid \text{Setosa}) = \frac{1}{\sqrt{2\pi \cdot 0.12}} \exp\left(-\frac{(5.1 - 5.0)^2}{2 \cdot 0.12}\right)$$

$$P(\text{sepal width} = 3.5 \mid \text{Setosa}) = \frac{1}{\sqrt{2\pi \cdot 0.14}} \exp\left(-\frac{(3.5 - 3.4)^2}{2 \cdot 0.14}\right)$$

$$P(\text{petal length} = 1.4 \mid \text{Setosa}) = \frac{1}{\sqrt{2\pi \cdot 0.03}} \exp\left(-\frac{(1.4 - 1.5)^2}{2 \cdot 0.03}\right)$$

$$P(\text{petal width} = 0.2 \mid \text{Setosa}) = \frac{1}{\sqrt{2\pi \cdot 0.01}} \exp\left(-\frac{(0.2 - 0.2)^2}{2 \cdot 0.01}\right)$$

For Versicolor:

$$P(\text{sepal length} = 5.1 \mid \text{Versicolor}) = \frac{1}{\sqrt{2\pi \cdot 0.26}} \exp\left(-\frac{(5.1 - 5.9)^2}{2 \cdot 0.26}\right)$$

$$P(\text{sepal width} = 3.5 \mid \text{Versicolor}) = \frac{1}{\sqrt{2\pi \cdot 0.10}} \exp\left(-\frac{(3.5 - 2.8)^2}{2 \cdot 0.10}\right)$$

$$P(\text{petal length} = 1.4 \mid \text{Versicolor}) = \frac{1}{\sqrt{2\pi \cdot 0.22}} \exp\left(-\frac{(1.4 - 4.2)^2}{2 \cdot 0.22}\right)$$

$$P(\text{petal width} = 0.2 \mid \text{Versicolor}) = \frac{1}{\sqrt{2\pi \cdot 0.02}} \exp\left(-\frac{(0.2 - 1.3)^2}{2 \cdot 0.02}\right)$$

For Virginica:

$$P(\text{sepal length} = 5.1 \mid \text{Virginica}) = \frac{1}{\sqrt{2\pi \cdot 0.40}} \exp\left(-\frac{(5.1 - 6.6)^2}{2 \cdot 0.40}\right)$$

$$P(\text{sepal width} = 3.5 \mid \text{Virginica}) = \frac{1}{\sqrt{2\pi \cdot 0.10}} \exp\left(-\frac{(3.5-3.0)^2}{2 \cdot 0.10}\right)$$
$$P(\text{petal length} = 1.4 \mid \text{Virginica}) = \frac{1}{\sqrt{2\pi \cdot 0.22}} \exp\left(-\frac{(1.4-5.6)^2}{2 \cdot 0.22}\right)$$
$$P(\text{petal width} = 0.2 \mid \text{Virginica}) = \frac{1}{\sqrt{2\pi \cdot 0.05}} \exp\left(-\frac{(0.2-2.0)^2}{2 \cdot 0.05}\right)$$

Step 3: Calculate Posterior Probabilities

Combine the priors and likelihoods to calculate the posterior probability for each class:

$$P(\text{Setosa} \mid X) \propto P(\text{Setosa}) \cdot P(\text{sepal length} = 5.1 \mid \text{Setosa}) \cdot P(\text{sepal width} = 3.5 \mid \text{Setosa}) \cdot P(\text{petal length} = 1.4 \mid \text{Setosa}) \cdot P(\text{petal width} = 0.2 \mid \text{Setosa})$$

$$P(\text{Versicolor} \mid X) \propto P(\text{Versicolor}) \cdot P(\text{sepal length} = 5.1 \mid \text{Versicolor}) \cdot P(\text{sepal width} = 3.5 \mid \text{Versicolor}) \cdot P(\text{petal length} = 1.4 \mid \text{Versicolor}) \cdot P(\text{petal width} = 0.2 \mid \text{Versicolor})$$

$$P(\text{Virginica} \mid X) \propto P(\text{Virginica}) \cdot P(\text{sepal length} = 5.1 \mid \text{Virginica}) \cdot P(\text{sepal width} = 3.5 \mid \text{Virginica}) \cdot P(\text{petal length} = 1.4 \mid \text{Virginica}) \cdot P(\text{petal width} = 0.2 \mid \text{Virginica})$$

Step 4: Classify

Select the class with the highest posterior probability as the predicted class $\hat{C}$:

$$\hat{C} = \arg\max_{C} P(C \mid X)$$

By evaluating the posterior probabilities, we determine the class with the highest probability, thus classifying the iris flower based on its features.

7.3 Probabilistic Graphical Models

Probabilistic Graphical Models (PGMs) are a powerful framework for representing complex probability distributions through graphs. They combine principles from probability theory and graph theory to model the conditional dependencies between random variables. PGMs are widely used in data science and machine learning for tasks such as inference, learning, and prediction. They provide a compact and intuitive way to encode the relationships among variables and facilitate efficient computation of probabilities and conditional independencies.

Probabilistic Graphical Models in Data Science

Probabilistic Graphical Models are graphical representations where nodes represent random variables, and edges represent probabilistic dependencies between these variables. There are two main types of PGMs: Bayesian Networks (directed graphs) and Markov Random Fields (undirected graphs).

Bayesian Networks

A Bayesian Network (BN) is a directed acyclic graph (DAG) where each node represents a random variable, and each edge represents a conditional dependency. The joint probability distribution of a set of random variables $X_1, X_2, \ldots, X_n$ in a BN is given by:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i))$$

Markov Random Fields

A Markov Random Field (MRF) is an undirected graph where each node represents a random variable, and edges represent the dependencies between variables. The joint probability distribution of a set of random variables $X_1, X_2, \ldots, X_n$ in an MRF is given by:

$$P(X_1, X_2, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C)$$

where $\mathcal{C}$ denotes the set of cliques in the graph, ψ_C denotes the potential function for clique C, and Z is the partition function ensuring normalization.

Examples

Example 1: Bayesian Network

Consider a simple Bayesian Network with three random variables: A, B, and C, where A influences B, and B influences C. The network structure can be represented as:

$$A \rightarrow B \rightarrow C$$

The joint probability distribution is:

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

Given the conditional probabilities:

$$\begin{aligned} P(A = a) = 0.6, \quad & P(A = \neg a) = 0.4 \\ P(B = b|A = a) = 0.7, \quad & P(B = b|A = \neg a) = 0.2 \\ P(C = c|B = b) = 0.9, \quad & P(C = c|B = \neg b) = 0.3 \end{aligned}$$

We can compute:

$$P(A = a, B = b, C = c) = P(A = a)P(B = b|A = a)P(C = c|B = b)$$

$$P(A = a, B = b, C = c) = 0.6 \times 0.7 \times 0.9 = 0.378$$

Example 2: Markov Random Field

Consider a simple Markov Random Field with three random variables: X, Y, and Z, where X and Y are connected, and Y and Z are connected. The graph structure can be represented as:

$$X - Y - Z$$

The joint probability distribution is:

$$P(X, Y, Z) = \frac{1}{Z}\psi_{XY}(X, Y)\psi_{YZ}(Y, Z)$$

Given the potential functions:

$$\psi_{XY}(X, Y) = \exp(\alpha XY), \quad \psi_{YZ}(Y, Z) = \exp(\beta YZ)$$

The partition function Z ensures that the distribution sums to 1:

$$Z = \sum_{X,Y,Z} \psi_{XY}(X, Y)\psi_{YZ}(Y, Z)$$

The joint probability can be computed accordingly.

Example 3: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's construct a Bayesian Network for the Iris dataset to model the relationships among the features and the species.

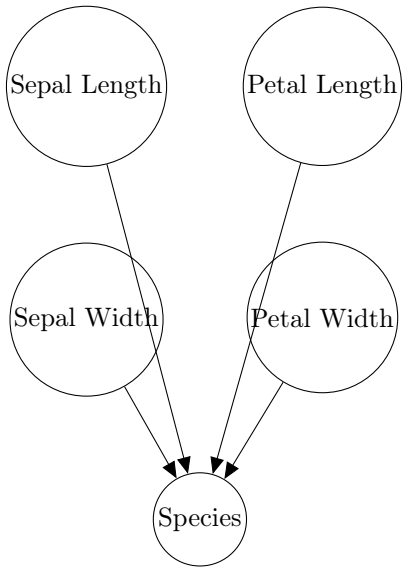


Figure 13: Bayesian Network for the Iris Dataset

Step 1: Define Conditional Probability Distributions

Assume the conditional probability distributions are Gaussian for the features given the species. For example:

$$P(\text{Sepal Length} \mid \text{Species} = \text{Setosa}) \sim \mathcal{N}(\mu_{\text{Setosa, Sepal Length}}, \sigma^2_{\text{Setosa, Sepal Length}})$$

$$P(\text{Sepal Width} \mid \text{Species} = \text{Setosa}) \sim \mathcal{N}(\mu_{\text{Setosa, Sepal Width}}, \sigma^2_{\text{Setosa, Sepal Width}})$$

Similar distributions are assumed for Versicolor and Virginica.

Step 2: Specify Parameters

Using the Iris dataset, we estimate the parameters (mean and variance) for each feature given the species.

For Setosa:

$$\mu_{\text{Setosa, Sepal Length}} \approx 5.006, \quad \sigma_{\text{Setosa, Sepal Length}} \approx 0.352$$
$$\mu_{\text{Setosa, Sepal Width}} \approx 3.428, \quad \sigma_{\text{Setosa, Sepal Width}} \approx 0.379$$

For Versicolor:

$$\mu_{\text{Versicolor, Sepal Length}} \approx 5.936, \quad \sigma_{\text{Versicolor, Sepal Length}} \approx 0.516$$
$$\mu_{\text{Versicolor, Sepal Width}} \approx 2.770, \quad \sigma_{\text{Versicolor, Sepal Width}} \approx 0.313$$

For Virginica:

$$\mu_{\text{Virginica, Sepal Length}} \approx 6.588, \quad \sigma_{\text{Virginica, Sepal Length}} \approx 0.636$$
$$\mu_{\text{Virginica, Sepal Width}} \approx 2.974, \quad \sigma_{\text{Virginica, Sepal Width}} \approx 0.322$$

Step 3: Perform Inference

Given a new sample with the following features:

$$X = \{\text{sepal length} = 5.5, \text{sepal width} = 3.5, \text{petal length} = 1.3, \text{petal width} = 0.2\}$$

We want to infer the probability of each species given this sample.

Using Bayes' Theorem, we compute the posterior probabilities for each species:

$$P(\text{Setosa} \mid X) \propto P(\text{Setosa}) \cdot P(X \mid \text{Setosa})$$
$$P(\text{Versicolor} \mid X) \propto P(\text{Versicolor}) \cdot P(X \mid \text{Versicolor})$$
$$P(\text{Virginica} \mid X) \propto P(\text{Virginica}) \cdot P(X \mid \text{Virginica})$$

Assuming equal priors:

$$P(\text{Setosa}) = P(\text{Versicolor}) = P(\text{Virginica}) = \frac{1}{3}$$

We then calculate the likelihoods $P(X \mid \text{Species})$ using the Gaussian distributions defined earlier.

7.4 Data Analysis

7.4.1 Hypothesis Testing

Now that the basis of statistical data analysis has been conducted it should be desired to determine what tests can be done on the data or variable of interest. Statistical hypothesis testing allows a null hypothesis H_0 to represent a position that is desired in the variable of interest. When using the hypothesis test it is assumed that the null hypothesis is true. This allows the model or in our case the algorithm to be trained so that the outcome of the model is reasonable. For example, assume the data being collected is to determine if a flower types are the same. The model can be trained to determine if the flower type is a Setosa, the null hypothesis would determine that the new observation is the hypothesis of interest or not. On the other hand if an observation is determined not to be in the null hypothesis then it is said to be some alternative hypothesis H_1. Examining this approach using an observation as x_0, the null hypothesis H_0 and the alternative hypothesis H_1 would yield the following:

- $x_0 = H_0$ or true if the input observation is within a threshold σ_{Setosa} of the Setosa data mean μ_{Setosa}.
- $x_0 = H_1$ or false if the input observation is outside the threshold σ_{Setosa} of the Setosa data mean.

In the formal sense of the problem, the threshold is known as the significance level (α). The significance level is a probability threshold, not necessarily the standard deviation as indicated above, that is used to determine if the null hypothesis is rejected in favor of the alternative hypothesis. It should be noted that the exact null hypothesis and alternative hypothesis will depend on the specific test model.

When testing, we may conduct a one-sided test or two-sided test, depending on H_1. In a two-sided test, we take the probability of getting a sample statistic more extreme than the one we observed from both sides of the distribution, then compare that value (known as a p-value) to α. This is usually used when H_1 contains $\neq$. In a one-sided test, the probability is only taken from one side of the distribution, this is used if H_1 contains $>$ or $<$.

In hypothesis testing, especially in z-Tests and t-Tests, we use the sampling distribution of sample means to find the probability of getting a certain result. To do this we need to know the mean and standard deviation of the distribution of sample means. This is given by the central limit theorem (CLT), which says that the sampling distribution of sample means of a continuous random variable X will approach $\mathcal{N}(\mu_X, \sigma_X^2)$. Note that, however, in the rest of the section when calculating a test statistic, any probabilities found from it will be based on $\mathcal{N}(0,1)$ or t_ν as the statistic is standardized. The general test statistic for z-Tests and t-Tests is given by

$$\textbf{test statistic} = \frac{\textbf{observed statistic} - \textbf{expected statistic}}{\textbf{standard error of statistic}}. \tag{20}$$

There are a handful of conditions that are needed for the central limit theorem to be applicable, and thus we need to check them before conducting hypothesis tests. There are also test-specific

conditions. All conditions will be omitted in this section.

It is also worth noting that getting results from z-Tests and t-Tests are largely similar. There are small differences that require the use of different tests, but the procedure is essentially identical. The general process is stating the hypotheses, obtaining a test statistic (t or z) and finding a p-value, or the probability of observing results this or more extreme. After getting the p-value, we can compare it to α. If our p-value is less than α, we reject H_0 and if it is greater than α, we fail to reject H_0 (not accept H_0, because we need to take more samples).

7.4.2 z-Test

The z-Test is a hypothesis test used to determine the probability of an outcome given a sample of a certain population. It differs from the t-Test because when using a z-Test, we know σ. In the following section, let the continuous random variable $Z \sim \mathcal{N}(0, 1)$.

One Sample z-Test The one sample z-Test is the test used to determine if an estimated mean (μ_0) differs from the true mean (μ).

We first identify the hypotheses. These are usually

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

because we are testing for equality. Note that the $\neq$ can be changed to $>$ or $<$, but then the test will be one-sided.

Then we calculate the test statistic z.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \tag{21}$$

Since our test statistic is now standardized, we can switch to the normal distribution $\mathcal{N}(0, 1)$, accounting for if the test is one-sided or two-sided.

Usually a one sample z-Test is used to compare a sample to a known distribution, for example, IQ scores. This is important because we already know μ and σ for IQ, 100 and 15, respectively. If we know the distribution of sepal length for setosa flowers, we could see if the iris data set has a sample that represents the whole species.

Two Sample z-Test The two sample z-Test is used to compare two populations and check for a sample mean, again when the means and standard deviations of the population are known. Usually the hypotheses for a two sample z-Test are

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the means of their respective distributions.

Since H_0 can also be stated as $\mu_1 - \mu_2 = 0$ and H_a can be stated as $\mu_1 - \mu_2 \neq 0$, in reality we are not testing for $\mu_1 = \mu_2$, we are testing for the difference in sample means $\mu_1 - \mu_2$. This means our distribution now follows $\mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Thus, our test statistic can be found from

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{22}$$

Note that since because of H_0, $\mu_1 - \mu_2$ is usually assumed to be 0.

We then continue as we did with a one sample z-Test by finding the p-value using $\mathcal{N}(0, 1)$, comparing to α and seeing if we reject or fail to reject H_0.

Usually a two sample z-Test is to compare two population with known standard deviations. It cannot be used, however, to compare a population with a known standard deviation before and after a treatment. We then need to perform a paired test, which is explained later.

7.4.3 t-Test

One of the most used tests with the statistical hypothesis testing is the t-Test. In this subsection the one-sample t-test, two-sample t-test and paired t-test will be explained. The t-Test is used when the population standard deviation, σ is unknown. This will make it impossible to know the standard deviation of sampling distribution of sample means, $\sigma_{\bar{x}}$. This is resolved through estimating $\sigma_{\bar{x}}$ using the sample standard deviation, s and finding $s_{\bar{x}}$, which is known as the standard error. The specific calculation of $s_{\bar{x}}$ depends on the type of t-Test being conducted.

An extra parameter, known as the degrees of freedom, ν, is needed for the t-Test. The calculation of the degrees of freedom depends on the type of t-Test being conducting. The degrees of freedom affects the shape of the t-distribution. The tails of the t-distribution are more prominent than the normal distribution, but as ν goes to infinity, the t-distributions begins to resemble a normal distribution.

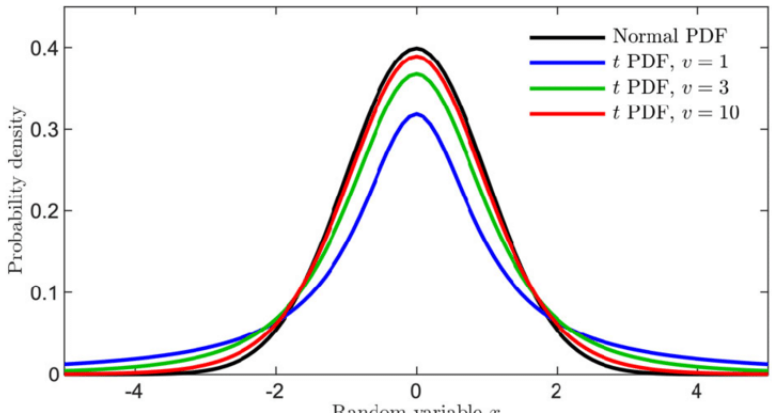


Figure 14: T distribution vs Z Distribution

In the following section, the t distribution will be denoted as t_ν and $T \sim t_\nu$. It is worth noting that t-Tests can be applied in many different places, including testing the slope of a least squares regression line.

One-Sample t-Test The one-sample t-test checks if there is a sample mean difference from the data population mean.

The following formula can be used to calculate the degrees of freedom for a one sample t-Test.

$$\nu = n - 1. \tag{23}$$

The hypothesis for the one sample t-Test are usually the same as the hypotheses for the one sample z-Test

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0.$$

And the standard error of the sampling distribution of sample means is given by the following formula

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{24}$$

where s is the standard deviation of the sample.

When putting this together, out test statistic t can be found from

$$t = \frac{\bar{x} - \mu}{65\frac{s}{\sqrt{n}}} \tag{25}$$

.

Finally, we find the probability of getting a t greater than or equal to the t we received from the test to obtain our p-value, then we compare the p-value to α and decide whether to reject H_0.

Suppose we want to check if the mean sepal length of the setosa flower could be $\mu = 5.1500$ or if it less than 5.1500, based on the value from the iris data set. We can first set up our hypotheses.

$$\begin{aligned} H_0 &= \mu = 5.1500 \\ H_1 &= \mu < 5.1500. \end{aligned}$$

Note that since H_1 contains $<$, we will be using a one-sided test.

Then we should calculate the parameters of our sampling distribution of sample means

$$\begin{aligned} \mu_{\bar{x}} &= \mu = 5.1500 \\ s_{\bar{x}} = \frac{s}{\sqrt{n}} &= \frac{0.3525}{\sqrt{50}} = 0.0499 \\ \nu &= n - 1 = 50 - 1 = 49. \end{aligned}$$

Then, we can calculate our statistic

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{5.0060 - 5.1500}{0.0499} = -2.8858.$$

We then check the probability of getting a t this extreme

$$P(t < -2.8858) = 0.0029 < \alpha = 0.05.$$

The probability of the true mean of setosa sepal lengths being 5.1500 is 0.0029. Since this value is less than a significance level of 0.05, we can reject H_0.

Two-Sample t-Test The two-sample t-test determines if there is a difference from two independent data samples. In this test the null hypothesis is the the means of the two samples are the same.

The degrees of freedom for a two sample t-Test is given by

$$\nu = n_1 + n_2 - 2. \tag{26}$$

The two-sample t-Test is combination of the two-sample z-Test and the one-sample t-Test. Our hypotheses are identical to the two-sample z-Test

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2.$$

The we calculate the standard error of the sampling distribution of sample means. Note that since this is t-Test we use s instead of σ and since we are dealing with two samples, we must use the sample standard deviations and sample sizes from both samples. When combined, we get

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{27}$$

Four our test statistic, we can reference 20

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}. \tag{28}$$

From here, we follow the one-sample t-Test, and compute the p-value. This is the same as the one-sample t-Test, as t is a standardized statistic, we only have to change p-value computation if the test is one-sided or two-sided. After finding our p-value, we compare to α and decide whether or not to reject H_0.

We can conduct a two-sample t-Test to see if the sepal length for the setosa and versicolor species differ in the iris data set. First, identify the hypotheses

$$H_0 : \mu_{\mathbf{set}} = \mu_{\mathbf{ver}}$$
$$H_1 : \mu_{\mathbf{set}} \neq \mu_{\mathbf{ver}}.$$

Let us also identify the data we already know

$$\bar{x}_{\mathbf{set}} = 5.0060$$
$$\bar{x}_{\mathbf{ver}} = 5.9360$$
$$s_{\mathbf{set}} = 0.3525$$
$$s_{\mathbf{ver}} = 0.5162$$
$$n_{\mathbf{set}} = 50$$
$$n_{\mathbf{ver}} = 50.$$

We can determine $\sigma_{\mathbf{set}-\mathbf{ver}}$

$$s_{\mathbf{set}-\mathbf{ver}} = \sqrt{\frac{(0.3525)^2}{50} + \frac{(0.5162)^2}{50}} = 0.08798.$$

And the test statistic

$$t = \frac{5.0060 - 5.9360}{0.08798} = -10.57058.$$

with $50 + 50 - 2 = 98$ degrees of freedom.

Since we assume H_0, we assume that $\mu_{\mathbf{set}} - \mu_{\mathbf{set}} = 0$.

Finally, we can calculate $P(t < -10.57058) + P(t > 10.57058)$, since the test is two-tailed (H_1 has $\neq$, not $>$ or $<$). This probability turns out to be extremely small ($7.0179642754 \times 10^{-18}$ exactly) which we would expect, as we know that the two species do not have the same mean sepal length.

Paired t-Test In this test the mean difference between samples of the same group at different points in time are determined.

Other statistical significance tests of importance not covered in this section are the ANOVA, and the Chi-Square Test.

7.5 Hypothesis Testing Algorithm

In Algorithm 1 outlines a generic procedure for hypothesis testing:

The objective of this algorithm design is to perform hypothesis tests in the fields of data science and AI. It computes a test statistic using the given data and assesses the statistical significance of the results with respect to the null hypothesis H_0.

The steps in analysing this algorithm are as follows:

1. Initialization: The process begins by establishing the null and alternative hypotheses and selecting a significance level (α), typically 0.05.
2. Test Statistic Calculation: The CalculateTestStatistic function is utilized to determine a numerical value that measures the extent to which the sample data differs from the null hypothesis.
3. P-Value Calculation: A different operation, called CalculatePValue, determines the likelihood of observing the obtained test statistic, or a value even more extreme, assuming the null hypothesis is true.

Algorithm 1 Hypothesis Testing with P-Value

```
function HypothesisTest(data, α)
    H_0 ← "Null hypothesis"
    H_a ← "Alternative hypothesis"
    α ← Significance level (e.g., 0.05)

    testStatistic ← function CalculateTestStatistic(data)
    pValue ← function CalculatePValue(testStatistic, data)
    print "Test Statistic: ", testStatistic
    print "P-Value: ", pValue

if pValue < α then
    print "Reject H_0: Evidence supports H_a"
else
    print "Fail to reject H_0: Not enough evidence against H_0"

function CalculateTestStatistic(data)
    ...computation based on the nature of data... return testStatistic

function CalculatePValue(testStatistic, data)
    ...computation based on test statistic and data... return pValue
```

4. Decision Making: The algorithm makes a decision on whether to reject H_0 based on the p-value and the significance level.

CalculateTestStatistic and CalculatePValue represent generic statistical calculations that will differ depending on the specific test being performed (e.g., t-test, chi-square test). These calculations should be implemented based on the hypothesis being tested and the characteristics of the data. The algorithm outputs the computed test statistic and p-value, and then provides a conclusion on whether to reject the null hypothesis based on a given significance level. If the p-value is less than the significance level (α), it suggests that the observed data is unlikely under the null hypothesis (H_0), and therefore the null hypothesis is rejected. On the other hand, a higher p-value indicates insufficient evidence to reject the null hypothesis (H_0).

7.5.1 p-Values

P-values have a crucial role in hypothesis testing, which is a commonly used technique in statistics and data science to draw conclusions about a population using sample data. The p-value represents the likelihood of obtaining results that are as extreme as, or more extreme than, the observed data, assuming that H_0 is true.The role of the p-value has a statistical significance, interpretation and misinterpretation as follows:

Statistical Significance - If the p-value is less than the selected significance level (α), the findings are deemed statistically significant. This implies that the observed data is unlikely to occur under the null hypothesis, resulting in the rejection of H_0. Conversely, if the p-value is greater than or equal to α, the findings are not statistically significant, indicating a lack of sufficient evidence to reject H_0.

Interpreting P-Values - A low p-value (usually ≤ 0.05) indicates strong evidence against the null hypothesis, leading to its rejection. Conversely, a high p-value suggests weak evidence

against H_0, resulting in its failure to be rejected. It is important to note that p-values do not quantify the probability of the hypothesis being true or false. Rather, they indicate the likelihood of obtaining the observed data assuming the null hypothesis is true.

Misinterpretations - The p-value does not provide information about the likelihood of either the null or alternative hypothesis being true. The practical significance of an effect cannot be determined solely based on a small p-value. Instead, a small p-value may suggest that a small effect has been observed with a high level of precision. However, repeatedly conducting experiments or altering hypotheses in order to obtain a small p-value can result in false positive results, which is commonly referred to as p-hacking.

P-values play a crucial role in hypothesis testing by helping determine if the observed data significantly differ from what would be anticipated under the null hypothesis. They serve as a statistical inference tool rather than providing absolute proof. Interpreting p-values requires taking into account various factors such as the magnitude of the effect, the reliability of the data, and the practical significance of the findings.

7.6 Confidence Intervals

Confidence intervals are a crucial concept in statistical analysis, providing a range of values within which we can expect a population parameter to lie with a certain level of confidence. They offer a way to quantify the uncertainty associated with sample estimates and are widely used in data science for hypothesis testing, parameter estimation, and predictive modeling. Confidence intervals allow analysts to make informed decisions and draw conclusions about population parameters based on sample data, incorporating the inherent variability of the data.

Confidence Intervals in Data Science

A confidence interval (CI) is an interval estimate, computed from the sample data, that is likely to cover the true population parameter with a specified confidence level. The confidence level (e.g., 95%, 99%) represents the proportion of intervals that would contain the parameter if the estimation process were repeated multiple times. For a population mean μ with a known standard deviation σ, the confidence interval is given by:

$$\bar{x} \pm z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

where:

- $\bar{x}$ is the sample mean
- z^* is the critical value from the standard normal distribution corresponding to the desired confidence level
- n is the sample size

For a population mean μ with an unknown standard deviation, the confidence interval is given by:

$$\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}} \right)$$

where:

- s is the sample standard deviation
- t^* is the critical value from the t-distribution with $n-1$ degrees of freedom

Examples

Example 1: Known Standard Deviation

Suppose we have a sample of 100 measurements with a sample mean $\bar{x} = 50$ and a known population standard deviation $\sigma = 10$. We want to compute a 95% confidence interval for the population mean.

The critical value z^* for a 95% confidence level is 1.96. The confidence interval is:

$$50 \pm 1.96 \left(\frac{10}{\sqrt{100}} \right) = 50 \pm 1.96 \times 1 = 50 \pm 1.96 = [48.04, 51.96]$$

Example 2: Unknown Standard Deviation

Suppose we have a sample of 25 measurements with a sample mean $\bar{x} = 100$ and a sample standard deviation $s = 15$. We want to compute a 95% confidence interval for the population mean.

The critical value t^* for a 95% confidence level with 24 degrees of freedom is approximately 2.064. The confidence interval is:

$$100 \pm 2.064 \left(\frac{15}{\sqrt{25}} \right) = 100 \pm 2.064 \times 3 = 100 \pm 6.192 = [93.808, 106.192]$$

Example 3: Using the Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Let's calculate the 95% confidence interval for the mean sepal length of the Setosa species.

Step 1: Extract the Data

The sepal length values for Setosa species are:

$$\begin{aligned} X = \{&5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, \\ &4.8, 4.8, 4.3, 5.8, 5.7, 5.4, 5.1, 5.7, 5.1, 4.6, 5.2, 5.0, \\ &5.1, 4.9, 5.2, 5.4, 5.2, 5.1, 5.3, 5.5, 5.0, 4.9, 5.2, 5.0, \\ &5.1, 5.0, 5.0, 5.1, 5.1, 5.0, 5.0, 5.4, 5.2, 5.3, 5.0, 4.9, 4.9, 5.0, 5.1\} \end{aligned}$$

Step 2: Calculate the Sample Mean and Standard Deviation

The sample mean $\bar{x}$ and sample standard deviation s are given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \approx 5.006$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \approx 0.352$$

Step 3: Determine the Critical Value

For a 95% confidence level, the critical value z^* is approximately 1.96.

Step 4: Calculate the Confidence Interval

The 95% confidence interval for the mean sepal length is given by:

$$\bar{x} \pm z^* \left(\frac{s}{\sqrt{n}} \right)$$

$$5.006 \pm 1.96 \left(\frac{0.352}{\sqrt{50}} \right)$$

$$5.006 \pm 1.96 \left(\frac{0.352}{7.071} \right)$$

$$5.006 \pm 1.96 \times 0.0498$$

$$5.006 \pm 0.0976$$

$$(4.9084, 5.1036)$$

Thus, the 95% confidence interval for the mean sepal length of the Setosa species is approximately (4.91, 5.10) cm.

7.7 Generating Sigma Ellipse Plots

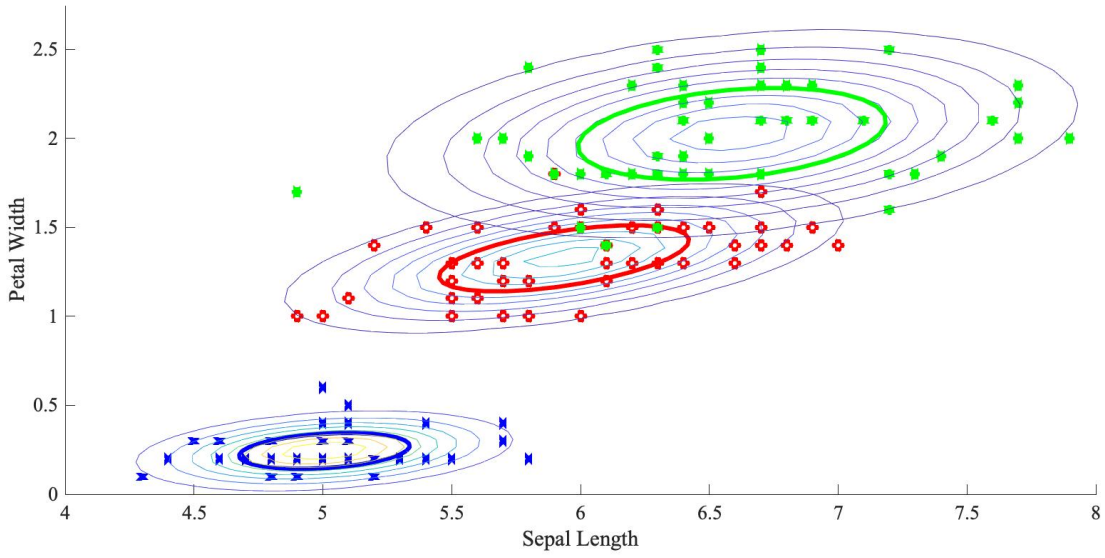


Figure 15: Gaussian Example using the iris data set

In data analysis you may also intend on using statistics to understand where outliers are in a set of data. Mahalanobis distance which is built on the covariance of a set of data can tell us where an outlier is more likely to occur. These can then be plotted in 2-Dimensions in a sigma ellipse

plot. The sigma ellipse plot (a.k.a error ellipse) can be generated by first outlining the number of standard deviations needed. For now, we will use three to emulate Figure 15.

We will first get the area under the normal distribution for each standard deviation. Since we are just working with one, two, and three standard deviations, these will be 0.6827, 0.9545, and 0.9973, respectively, as shown in Section 4.2.1. However, a generalized equation for this can be given by

$$P(-\sigma < X < \sigma) = CDF(\sigma) - CDF(-\sigma) \tag{29}$$

where CDF represents the cumulative distribution function for area to the left of the argument and $X \sim \mathcal{N}(0, 1)$.

We can then use the probability point function (PPF), to calculate the χ^2 critical values of the outputs of the cumulative distribution function. When these χ^2 values are multiplied by $1/2$, we can get values that are crucial in drawing the sigma ellipse plot.

$$\chi^2 = PPF(P(-\sigma < X < \sigma), df = 2). \tag{30}$$

If we do this for every standard deviation (in this case 1, 2, and 3), we can form the sigma ellipses for each standard deviation. Also, the Mahalanobis Distance (D_i) from the mean to each position on the ellipse will be the same.

$$D_i = \left((x_i - \bar{x})\Sigma^{-1}(y_i - \bar{y})^T\right)^{1/2} \tag{31}$$

where Σ represent the covariance matrix of the data.

The χ^2 distribution is the distribution that D_i^2 follows, and there are two degrees of freedom because the data is bivariate. The Mahalanobis Distance should always satisfy

$$D_i = \sqrt{\chi^2}. \tag{32}$$

Ultimately, the sigma ellipse plot will account for the angle of rotation (which is calculated by the two eigenvectors) and the spread along both axes. **[Briggs2007]**

This process can be implemented in Python with the following sigma_ellipse_plot class, notice the get_chisquare_vals() method.

Python Code

```
class sigma_ellipse_plot:

    def __init__(self, df=None, target='setosa', target_header='species',
        feature1='sepal_length', feature2='petal_width', std_devs=[1, 2, 3]):

        self.data = df
        self.target = target
        self.feature1 = feature1
        self.feature2 = feature2
        self.target_header = target_header
        self.std_devs=std_devs
        self.largest_eigenvalue = None
        self.largest_eigenvector = None
        self.smallest_eigenvalue = None
        self.smallest_eigenvector = None
        self.angle = None
        self.mean = None
        self.r_ellipses = None
        self.mu_X = None
        self.mu_Y = None
        self.chisquare_val = None

    def get_data(self):

        self.data = self.data[self.data[self.target_header] == self.target].drop(
            self.target_header, axis =1)[[self.feature1, self.feature2]]
        #self.data = self.data[self.data['species'] == self.target].drop(
        # 'species', axis =1)[[self.feature1, self.feature2]]

        return
    def get_eigens(self):

        covariance_matrix = self.data.cov()
        eigenvalues, eigenvectors = eigh(covariance_matrix)

        self.largest_eigenvector = eigenvectors[np.argmax(eigenvalues)]
        self.largest_eigenvalue = np.max(eigenvalues)
        self.smallest_eigenvector = eigenvectors[np.argmin(eigenvalues)]
        self.smallest_eigenvalue = np.min(eigenvalues)

        return
```

Figure 16: Sigma Ellipse Plot part 1

```
42
43      def get_angle(self):
44
45          self.angle = math.atan2(self.largest_eigenvector[1], self.largest_eigenvector[0])
46
47          return
48
49      def shift_angle(self):
50
51          if self.angle < 0:
52              self.angle = self.angle + 2*math.pi
53
54          return
55
56      def get_mean(self):
57
58          self.mean = self.data.mean()
59
60          return
61
62      def get_chisquare_vals(self):
63
64          self.chisquare_val = []
65          for i in range(0, len(self.std_devs)):
66              #percent_covered = stats.norm.cdf(i+1) - stats.norm.cdf((i+1) * -1)
67              percent_covered = stats.norm.cdf(self.std_devs[i]) - stats.norm.cdf(self.std_devs[i] * -1)
68              self.chisquare_val.append((chi2.ppf(percent_covered, df=2))**0.5)
69
70          return self.chisquare_val
71
72      def get_ellipses(self):
73
74          chisquare_val = self.get_chisquare_vals()
75
76          self.r_ellipses = []
77          for i in range(0, len(self.std_devs)):
78              theta_grid = np.linspace(0,2*math.pi, 100)
79              phi = self.angle
80              self.mu_X = self.mean[0]
81              self.mu_Y = self.mean[1]
82              a = chisquare_val[i] * math.sqrt(self.largest_eigenvalue)
83              b = chisquare_val[i] * math.sqrt(self.smallest_eigenvalue)
```

Figure 17: Sigma Ellipse Plot part 2

```
84
85              ellipse_x_r = a * np.cos(theta_grid)
86              ellipse_y_r = b * np.sin(theta_grid)
87
88              R = [[math.cos(phi), math.sin(phi)], [-math.sin(phi), math.cos(phi)]]
89
90              ellipses = np.array([ellipse_x_r, ellipse_y_r])
91
92              r_ellipse = ellipses.T.dot(R).T
93
94              self.r_ellipses.append(r_ellipse)
95
96          return
97
98      def get_labels(self, special_phrase=None):
99
100         labels = []
101         for i in range(0, len(self.std_devs)):
102
103             if special_phrase is None:
104                 label = str(self.std_devs[i]) + " std. dev. from mean"
105                 labels.append(label)
106             else:
107                 label = special_phrase + str(self.std_devs[i]) + " std. dev. from mean"
108                 labels.append(label)
109
110         return labels
111
112     def pipeline(self):
113
114         self.get_data()
115         self.get_eigens()
116         self.get_angle()
117         self.shift_angle()
118         self.get_mean()
119         self.get_ellipses()
120
121         return self.data, self.r_ellipses, self.mu_X, self.mu_Y
```

Figure 18: Sigma Ellipse Plot part 3

8 A/B Testing

A/B testing, also known as split testing, is a fundamental technique in data science and statistics used to compare two versions of a variable to determine which one performs better. This method is widely used in various fields, including marketing, web development, and product management, to make data-driven decisions. By randomly assigning subjects to two groups—A (control) and B (treatment)—and analyzing the outcomes, A/B testing helps determine the effectiveness of changes or interventions.

8.1 A/B Testing in Data Science

A/B testing is an experimental method used to compare two variants, A and B, to identify which one performs better based on a specific metric. The null hypothesis (H_0) typically states that there is no difference between the two variants, while the alternative hypothesis (H_1) suggests that a difference exists.

To determine if there is a statistically significant difference between the two groups, we use a two-sample t-test for the means of two independent samples. The test statistic for the two-sample t-test is calculated as follows:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

where:

- $\bar{X}_A$ and $\bar{X}_B$ are the sample means of groups A and B, respectively
- s_A^2 and s_B^2 are the sample variances of groups A and B, respectively
- n_A and n_B are the sample sizes of groups A and B, respectively

The degrees of freedom for the t-test can be approximated using the following formula:

$$df = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1}}$$

Examples

Example 1: Website Conversion Rate

Suppose we want to compare the conversion rates of two versions of a website: Version A (control) and Version B (treatment). We collect the following data:

$$\text{Version A: } n_A = 1000, \bar{X}_A = 0.12, s_A = 0.03$$
$$\text{Version B: } n_B = 1000, \bar{X}_B = 0.15, s_B = 0.04$$

The test statistic is:

$$t = \frac{0.12 - 0.15}{\sqrt{\frac{0.03^2}{1000} + \frac{0.04^2}{1000}}} = \frac{-0.03}{\sqrt{\frac{0.0009}{1000} + \frac{0.0016}{1000}}} = \frac{-0.03}{\sqrt{0.0000025}} = \frac{-0.03}{0.00158} \approx -18.99$$

The degrees of freedom can be approximated as:

$$df = \frac{\left(\frac{0.03^2}{1000} + \frac{0.04^2}{1000}\right)^2}{\frac{\left(\frac{0.03^2}{1000}\right)^2}{999} + \frac{\left(\frac{0.04^2}{1000}\right)^2}{999}} = \frac{(0.0000009 + 0.0000016)^2}{\frac{0.0000009^2}{999} + \frac{0.0000016^2}{999}} = \frac{0.0000025^2}{\frac{0.00000000081}{999} + \frac{0.00000256}{999}} \approx 1997.99 \approx 1998$$

We compare the test statistic t to the critical value from the t-distribution with 1998 degrees of freedom to determine if the difference is statistically significant.

Example 2: Email Campaign Click-through Rate

Suppose we want to compare the click-through rates of two email campaigns: Campaign A (control) and Campaign B (treatment). We collect the following data:

$$\text{Campaign A: } n_A = 800, \bar{X}_A = 0.05, s_A = 0.01$$
$$\text{Campaign B: } n_B = 800, \bar{X}_B = 0.06, s_B = 0.015$$

The test statistic is:

$$t = \frac{0.05 - 0.06}{\sqrt{\frac{0.01^2}{800} + \frac{0.015^2}{800}}} = \frac{-0.01}{\sqrt{\frac{0.0001}{800} + \frac{0.000225}{800}}} = \frac{-0.01}{\sqrt{0.000000125 + 0.00000028125}} = \frac{-0.01}{\sqrt{0.00000040625}} = \frac{-0.01}{0.0006374} \approx$$

The degrees of freedom can be approximated as:

$$df = \frac{\left(\frac{0.01^2}{800} + \frac{0.015^2}{800}\right)^2}{\frac{\left(\frac{0.01^2}{800}\right)^2}{799} + \frac{\left(\frac{0.015^2}{800}\right)^2}{799}} = \frac{(0.000000125 + 0.00000028125)^2}{\frac{0.000000000125}{799} + \frac{0.0000004225}{799}} \approx 1597.99 \approx 1598$$

We compare the test statistic t to the critical value from the t-distribution with 1598 degrees of freedom to determine if the difference is statistically significant.

8.2 Statistical Significance in A/B Testing

Statistical significance is a crucial concept in A/B testing, which is widely used in data analysis to compare two versions of a variable and determine which one performs better. Statistical significance helps quantify whether the observed difference between two groups (e.g., control and treatment) is likely to be due to chance or represents a true effect. Understanding and correctly applying statistical significance in A/B testing ensures that data-driven decisions are based on reliable and valid results.

Statistical Significance in A/B Testing in Data Science

Statistical significance is a measure of whether the observed difference between two groups in an A/B test is likely to be due to chance. It is determined using a p-value, which represents the probability of observing the data, or something more extreme, if the null hypothesis is true. A result is considered statistically significant if the p-value is less than a predefined significance level (α), commonly set at 0.05.

To determine statistical significance in A/B testing, we use a two-sample t-test to compare the means of two independent samples. The test statistic for the two-sample t-test is calculated as follows:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

where:

- $\bar{X}_A$ and $\bar{X}_B$ are the sample means of groups A and B, respectively

- s_A^2 and s_B^2 are the sample variances of groups A and B, respectively
- n_A and n_B are the sample sizes of groups A and B, respectively

The degrees of freedom for the t-test can be approximated using the following formula:

$$df = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B - 1}}$$

The p-value is then calculated based on the t-distribution with the calculated degrees of freedom. If the p-value is less than the significance level (α), the result is considered statistically significant.

Examples

Example 1: Website Conversion Rate

Suppose we want to compare the conversion rates of two versions of a website: Version A (control) and Version B (treatment). We collect the following data:

$$\text{Version A: } n_A = 1000, \bar{X}_A = 0.12, s_A = 0.03$$
$$\text{Version B: } n_B = 1000, \bar{X}_B = 0.15, s_B = 0.04$$

The test statistic is:

$$t = \frac{0.12 - 0.15}{\sqrt{\frac{0.03^2}{1000} + \frac{0.04^2}{1000}}} = \frac{-0.03}{\sqrt{\frac{0.0009}{1000} + \frac{0.0016}{1000}}} = \frac{-0.03}{\sqrt{0.0000025}} = \frac{-0.03}{0.00158} \approx -18.99$$

The degrees of freedom can be approximated as:

$$df = \frac{\left(\frac{0.03^2}{1000} + \frac{0.04^2}{1000}\right)^2}{\frac{\left(\frac{0.03^2}{1000}\right)^2}{999} + \frac{\left(\frac{0.04^2}{1000}\right)^2}{999}} = \frac{(0.0000009 + 0.0000016)^2}{\frac{0.00000000081}{999} + \frac{0.00000256}{999}} \approx 1997.99 \approx 1998$$

Using a t-table or statistical software, we find the p-value corresponding to $t = -18.99$ and $df = 1998$. If the p-value is less than 0.05, the result is statistically significant.

Example 2: Email Campaign Click-through Rate

Suppose we want to compare the click-through rates of two email campaigns: Campaign A (control) and Campaign B (treatment). We collect the following data:

$$\text{Campaign A: } n_A = 800, \bar{X}_A = 0.05, s_A = 0.01$$
$$\text{Campaign B: } n_B = 800, \bar{X}_B = 0.06, s_B = 0.015$$

The test statistic is:

$$t = \frac{0.05 - 0.06}{\sqrt{\frac{0.01^2}{800} + \frac{0.015^2}{800}}} = \frac{-0.01}{\sqrt{\frac{0.0001}{800} + \frac{0.000225}{800}}} = \frac{-0.01}{\sqrt{0.000000125 + 0.00000028125}} = \frac{-0.01}{\sqrt{0.00000040625}} = \frac{-0.01}{0.0006374}$$

The degrees of freedom can be approximated as:

$$df = \frac{\left(\frac{0.01^2}{800} + \frac{0.015^2}{800}\right)^2}{\frac{\left(\frac{0.01^2}{800}\right)^2}{799} + \frac{\left(\frac{0.015^2}{800}\right)^2}{799}} = \frac{(0.000000125 + 0.00000028125)^2}{\frac{0.000000000125}{799} + \frac{0.0000004225}{799}} \approx 1597.99 \approx 1598$$

Using a t-table or statistical software, we find the p-value corresponding to $t = -15.69$ and $df = 1598$. If the p-value is less than 0.05, the result is statistically significant.

8.3 Hypothesis Testing in A/B Testing

Statistical significance is a measure of whether an observed effect or difference is unlikely to have occurred due to chance alone. It is determined using a hypothesis test, where the null hypothesis (H_0) represents no effect or difference, and the alternative hypothesis (H_1) represents an effect or difference. The significance level (α) is the threshold for rejecting the null hypothesis, commonly set at 0.05. A result is statistically significant if the p-value, the probability of obtaining a result at least as extreme as the observed one under the null hypothesis, is less than α.

- **Null Hypothesis (H_0)**: There is no difference between the two groups.
- **Alternative Hypothesis (H_1)**: There is a difference between the two groups.
- **Significance Level (α)**: Typically 0.05.
- **Test Statistic**: Calculated based on the test type (e.g., t-test, chi-square test).
- **p-value**: The probability of obtaining the observed result, or more extreme, assuming H_0 is true.
- **Decision Rule**: If $p \leq \alpha$, reject H_0; otherwise, do not reject H_0.

Example 3: Using the Iris Dataset

We compare the mean sepal lengths of Setosa and Versicolor species using a two-sample t-test.

Given the sepal length data:

$$\text{Setosa: } n_1 = 50, \bar{X}_1 = 5.01, s_1 = 0.35$$
$$\text{Versicolor: } n_2 = 50, \bar{X}_2 = 5.94, s_2 = 0.52$$

Formulate Hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Select Significance Level

$$\alpha = 0.05$$

Calculate Test Statistic

The test statistic for a two-sample t-test is given by:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Substituting the given values:

$$t = \frac{5.01 - 5.94}{\sqrt{\frac{0.35^2}{50} + \frac{0.52^2}{50}}} = \frac{-0.93}{\sqrt{\frac{0.1225}{50} + \frac{0.2704}{50}}} = \frac{-0.93}{\sqrt{0.00245 + 0.00541}} = \frac{-0.93}{0.0886} \approx -10.50$$

Determine p-value

Using a t-distribution table or statistical software, find the p-value for $t = -10.50$ with $df \approx 98$.

Compare p-value with α

If $p \leq 0.05$, reject H_0; otherwise, do not reject H_0.

Assume the p-value is very small (e.g., $p < 0.001$):

Since $p < 0.05$, we reject H_0 and conclude that there is a significant difference in mean sepal lengths betwee

9 Module Questions

Probabilities and Random Variables

1. Q: What is the difference between discrete and continuous random variables? Give examples.
2. Q: Define Probability Mass Function (PMF) and Probability Density Function (PDF).
3. Q: What is the purpose of the Cumulative Distribution Function (CDF)?

Probability Distributions

1. Q: Why is the Normal distribution significant in statistics and data science?
2. Q: When would you use Binomial versus Poisson distributions?
3. Q: Explain the Exponential and Gamma distributions in time-to-event data context.

Order Statistics

1. Q: Define the k-th order statistic in a sample.

Hypothesis Testing

1. Q: What are the key steps in conducting a hypothesis test?
2. Q: Explain Type I and Type II errors.
3. Q: How do you interpret a p-value?

Algorithm Types

1. Q: Distinguish between deterministic and randomized algorithms with examples.
2. Q: Discuss the trade-offs between deterministic and randomized algorithms.

References

[1] William Press et al. *Numerical Recipes in Fortran: The Art of Scientific Computing*. 1st. Cambridge University Press, 1986. ISBN: 0-521-30811-9.

[2] Vincent Spruyt. *How to draw an error ellipse representing the covariance matrix?* https://www.visiondummy.com/2014/04/draw-error-ellipse-representing-covariance-matrix/. 2014.