Statistics for Data Science

A random variable is neither random nor variable.

Leon Muscat

Contents

0 Course Description

0.1 Learning Objectives

- Understand fundamental concepts in statistics such as datasets, probability, probability distributions, and standard error.
- Utilize the R programming language to analyze datasets, including loading data, implementing sampling approaches, and performing statistical tests and Bayesian analysis.
- Evaluate sample calculations and their qualities using biases, standard error, and confidence intervals.

0.2 Course Content

- Comprehensive introduction to data analysis combining practical exercises with mathematical theory.
- Topics include statistical principles, probability theory, real-world examples, and code examples in R.
- Coverage of statistical inference using finite samples, modern statistical methods like Bayesian decision theory, equivalence testing, and statistical modelling.
- Emphasizes the importance of data analysis for computer science students, providing skills to identify patterns in datasets.

0.3 Event Structure and Teaching/Learning Design

- Course split into eight sections, each building on the previous with theoretical knowledge and practical application.
- Includes theory-based lectures and practice-oriented exercises consisting of workshops (practical implementation in R, problem-solving, discussion) and assignments.
- Encourages both individual and group study.

0.4 Literature

- Kaptein M. & van den Heuvel E. (2022). Statistics for Data Scientists – An Introduction to Probability Statistics and Data Analysis. Springer.

0.5 Examination

0.5.1 1st Examination Component (1/2)

- **Type**: Quiz
- **Form**: Written exam
- **Mode**: Digital
- **Time**: Term time
- **Location**: On Campus
- **Grading**: Individual work, individual grade
- **Weight**: 20%
- **Languages**: Questions and answers in English.
- **Aids**: Closed Book, Texas Instruments TI-30 series calculators, bilingual dictionaries without notes for non-language exams.

0.5.2 2nd Examination Component (2/2)

- **Type**: Analog written examination
- **Form**: Written exam
- **Mode**: Analog
- **Time**: Lecture-free period
- **Location**: On Campus
- **Grading**: Individual work, individual grade
- **Weight**: 80%
- **Duration**: 120 Min.
- **Languages**: Questions and answers in English.
- **Aids**: Information sheet provided by lecturers, no other aids allowed from home.

0.5.2.1 Examination Content

- Decentral Exam (20%): Covers sections 1-4 and assignment content.
- Central Exam (80%): Covers all sections (1-8) and assignment content.

0.5.2.2 Examination Literature

- Lecture materials and course literature by Kaptein & van den Heuvel (2022).

1 Introduction of Key Concepts

1.1 Sample space, events, partition of sample space

- **Random experiment:** A random experiment is a process that results in one of several possible outcomes. The key characteristic of a random experiment is that, although the individual outcome is unpredictable, the set of all possibe outcomes is well-defined and known in advance. An experiment can be repeated endlessly and is always performed under the same conditions and rules.
- **Sample space (Ω)**: The set of all possible outcomes of an experiment. It can be:
  - **Finite**: when the number of possible outcomes is limited (e.g., rolling a die has a sample space of six outcomes).
  - **Countable**: if there's a way to list all the outcomes in a sequence that can be matched one-to-one with the set of natural numbers (e.g., the outcomes of flipping a coin indefinitely).
  - A finite set is always countable, but a countable set isn't necessarily finite (e.g., the set of all integers).
- **Event** A: A collection of possible outcomes of an experiment ($A \subseteq \Omega$). Usually those outcomes share a specific property or characteristic, such as an even number when rolling a die. A single possible outcome $\omega \in \Omega$ is an elementary event (also called singleton).
- **Event Partition**: A collection of events that are mutually exclusive (no events have any outcomes in common) and collectively exhaustive (the events account for all possible outcomes). That means, for a collection of events $A_i, i \in I$ to be a partition, these rules must be fulfilled:
  - $A_i \cap A_j = \emptyset$ for any $i, j \in I$. That means all events are pairwise disjoint.
  - $\bigcup_{i \in I} A_i = \Omega$

1.1.1 Examples

Let's consider the experiment of drawing a card from a standard 52-card deck.

- **Sample space (Ω)**: The sample space for this experiment is all 52 cards in the deck. Each card is an elementary event, such as drawing the Ace of Spades.
- **Events**:
  - **Event** A: Drawing a heart. There are 13 possible outcomes in this event, from the Ace of Hearts to the King of Hearts.
  - **Event** B: Drawing a face card (Jack, Queen, King). There are 12 possible outcomes in this event.

  - **Partition of sample space**: The entire sample space can be partitioned into disjoint events, such as drawing a heart, drawing a club, drawing a diamond, and drawing a spade. These partitions are mutually exclusive and cover the entire sample space.

- **Sure event (S)**: A sure event is one that is certain to occur, such as drawing a card that is either red or black. Since all cards in the deck are either red or black, the probability of this event is 1.
- **Impossible event ($\emptyset$)**: An impossible event is an event that cannot possibly occur, such as drawing a card that is both a heart and a club at the same time. The probability of this event is 0.

This setup can be used to introduce basic concepts such as outcomes, events, and probabilities. It provides a concrete example that can be easily understood and visualized.

1.2 σ-Algebra, probability measure, probability space

- **Field**: A field is a set $\mathcal{K}$ with two operations: addition $+$ and multiplication $\cdot$, which obey certain axioms (for all $\alpha, \beta, \gamma \in \mathcal{K}$):. These axioms ensure that addition, subtraction, multiplication, and division (excluding division by zero) are possible within $\mathcal{K}$.

  - **A1**: $\alpha + \beta = \beta + \alpha$
  - **A2**: $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$
  - **A3**: $\alpha + 0 = 0 + \alpha = \alpha$
  - **A4**: $\forall \alpha \exists -\alpha \in K$ such that $\alpha + (-\alpha) = (-\alpha) + \alpha = 0$
  - **M1**: $\alpha \cdot \beta = \beta \cdot \alpha$
  - **M2**: $(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$
  - **M3**: $\alpha \cdot 1 = 1 \cdot \alpha = \alpha$
  - **M4**: $\forall \alpha \exists \alpha^{-1} \in K$ such that $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = 1$
  - **D**: $(\alpha + \beta) \cdot \gamma = \alpha \cdot \gamma + \beta \cdot \gamma$

- **σ-Algebra**: A σ-algebra $\mathcal{F}$ over a sample space Ω is a collection of subsets of Ω that includes the empty set $\emptyset$, is closed under complementation and countable unions. In other words, if A is in $\mathcal{F}$, then the complement A^c is also in $\mathcal{F}$, and if $A_1, A_2, \ldots$ are in $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$.
- **Algebra of Sets**: Any collection G of subsets of Ω that includes Ω itself, is closed under finite unions and intersections, and includes the complement of any set in G, is an algebra of sets.
- **Probability Measure**: A probability measure P is a function that assigns to each event A in a σ-algebra $\mathcal{F}$ a non-negative real number, representing the likelihood of the occurrence of A. It satisfies the properties that

  - $P(\Omega) = 1$,
  - $P(A) \geq 0$ for any $A \in \mathcal{F}$, and

  - for any countable sequence of mutually exclusive events A_i, $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

  The probability of an event A in a finite Ω is given by $P(A) = \sum_{\{i:\omega_i \in A\}} p_i$, where $p_i = P(\{\omega\})$.

- **Probability Space**: The triad $(\Omega, \mathcal{F}, P)$ forms a probability space, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra on Ω, and P is a probability measure on $\mathcal{F}$. This structure is the mathematical foundation for probability theory.

1.2.1 Examples

- **Real, Rational, and Complex Fields**: The sets $\mathbb{R}$, $\mathbb{Q}$, and $\mathbb{C}$ are examples of fields where all the axioms of addition and multiplication hold true, unlike the set of natural numbers $\mathbb{N}$ or integers $\mathbb{Z}$, where some axioms like those for multiplicative inverses do not hold.
- σ**-Algebras**: For a countable Ω, the easiest way to construct $\mathcal{F}$ is to take all possible subsets of Ω ($\mathcal{P}(\Omega)$). Here are some examples:
  - **Trivial** σ**-Algebra**: $\mathcal{F} = \{\emptyset, \Omega\}$
  - **Smallest** σ**-Algebra with** A: $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$
- **Borel** σ**-Algebra**: In the context of real numbers $\mathbb{R}$, the Borel σ-algebra is generated by the open intervals and contains all the intervals of various types ($[a, b), (a, b], (a, b), [a, b] \quad a, b \in \mathbb{R}$). It is the smallest σ-algebra containing all open sets of $\mathbb{R}$: $(-\infty, a]$, where $a \in \mathbb{Q}$
- **Probability Measure**: The easiest case is a finite Ω with uniform probability. That means all outcomes ω are equally likely. This simplifies the probability measure: $P(A) = \sum_{\{i:\omega_i \in A\}} p_i = \sum_{\{i:\omega_i \in A\}} \frac{1}{N} = \frac{|A|}{|\Omega|}$.
- **Probability Space**: Let's look at an example of a fair coin that is tossed once.
  - $\Omega = \{H, T\}$,
  - $\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$.
  - A candidate for probability measure $P : \mathcal{F} \to [0, 1]$ is given by:
    * $P(\emptyset) = 0$,
    * $P(\{H\}) = \frac{1}{2}$ with $p \in [0, 1]$,
    * $P(\{T\}) = 1 - \frac{1}{2} = \frac{1}{2}$ with $p \in [0, 1]$,
    * $P(\Omega) = 1$,
  - If $P(\{H\}) = P(\{T\})$ then $p = \frac{1}{2}$ and we say that the coin is fair or unbiased.

1.3 Counting Rules

Counting rules are fundamental in probability theory for determining the size of sample spaces and the likelihood of events. There are two main categories of counting problems:

1. **Counting with Replacement and Order Matters**: In this scenario, after an item is selected from the set, it is replaced, and the order in which items are selected is important. The number of possible outcomes is given by n^r, where n is the number of items to choose from, and r is the number of items to be chosen.
2. **Counting without Replacement and Order Matters**: This is used when items are not replaced once selected, and the order of selection is important. The number of possible outcomes is calculated using the permutation formula: $P(n, r) = \frac{n!}{(n-r)!}$, where n is the total number of items, r is the number of items to choose, and $n!$ is the product of all positive integers up to n.
3. **Counting with Replacement and Order Doesn't Matter**: Here, items are replaced after selection, and the order of selection does not matter. The formula to determine the number of outcomes is given by the combination with repetition formula: $C(n + r - 1, r) = \frac{(n+r-1)!}{r!(n-1)!}$.
4. **Counting without Replacement and Order Doesn't Matter**: In this case, once an item is selected, it is not replaced, and the order of selection does not matter. The number of outcomes is given by the combination formula: $C(n, r) = \frac{n!}{r!(n-r)!}$, also denoted as $\binom{n}{r}$.

1.3.1 Counting Rules for Groups

When it comes to grouping items, the counting rules become a bit more complex, particularly when the groups are of different sizes or when there are restrictions on how items can be grouped.

1. **Partitioning into Groups of Equal Size**: If we want to divide n items into k groups of equal size, we use the multinomial coefficient: $\frac{n!}{(n/k)!^k}$, assuming n is divisible by k.
2. **Partitioning into Groups of Unequal Sizes**: If n items are to be divided into k groups of various sizes $n_1, n_2, \ldots, n_k$, such that $n_1 + n_2 + \ldots + n_k = n$, the number of ways to do this is given by the multinomial theorem: $\frac{n!}{n_1!n_2!\ldots n_k!}$.

1.4 Conditional Probability

If $A, B \in \mathcal{F}$ and $P(B) > 0$, then the conditional probability of event A given event B is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Conditional probability is a valid probability measure, as long as it satisfies the Kolmogorov axioms:

For any $B \in \mathcal{F}$ with $P(B) > 0$

1. $P(\cdot|B) \geq 0$
2. If $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P(\bigcup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$$

3. $P(\Omega|B) = 1$

For instance, consider the roll of a fair six-sided die. Let A be the event "the outcome is an odd number" and B be the event "the outcome is greater than two." Then $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$, and $P(A \cap B) = \frac{1}{3}$. Therefore, the conditional probability $P(A|B)$ is calculated as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}.$$

1.4.1 Partition equation

Let $\{A_i : i \in I\}$ be a finite or countable partition of Ω. Then for $\forall A \in \mathcal{F}$ it holds

$$P(A) = \sum_{i \in I} P(A|A_i)P(A_i).$$

1.4.2 Bayes Rule

Let $\{A_i : i \in I\}$ be a finite or countable partition of Ω. Let $B \in \mathcal{F}$ and $P(B) > 0$, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j \in I} P(B|A_j)P(A_j)}$$

Assuming we have two urns, Urn X and Urn Y. Urn X contains 5 red and 5 blue balls, and Urn Y contains 1 red and 9 blue balls. If a ball is drawn at random from one of the urns, and it is red, what is the probability it was drawn from Urn X? Let A_1 be the event of selecting from Urn X and A_2 from Urn Y, and B be the event of drawing a red ball. Then:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{10} \cdot \frac{1}{2}} = \frac{5}{6}.$$

The probability of drawing a red ball from Urn X given that a red ball was drawn is $\frac{5}{6}$.

1.4.3 Independence

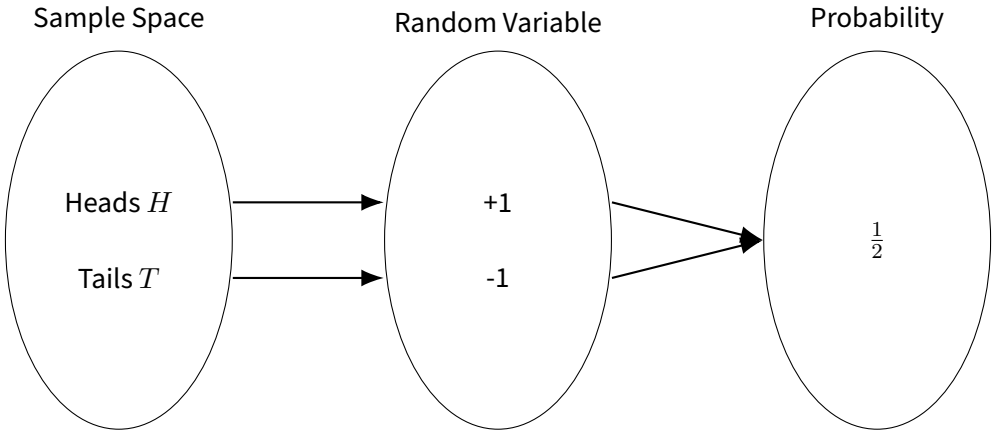
Two events A and B are said to be independent if the occurrence of one does not affect the probability of the occurrence of the other. In other words, events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$. This implies that the occurrence of B does not change the probability of A occurring, so we have $P(A|B) = P(A)$ and similarly $P(B|A) = P(B)$.

For example, if we have a fair coin and a fair six-sided die, the probability of getting a head on the coin toss (event A) is independent of the probability of rolling a six on the die (event B). Here, $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{6}$. The joint probability of both A and B occurring is $P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$, which is the product of the two individual probabilities, thus confirming their independence.

More generally, a family of events $\{A_i | i \in J\}$ is called independent if $P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ for every finite $I \subseteq J$.

2 Random Variables

A random variable is a mathematical formalization of a quantity or object which depends on random events. The term 'random variable' can be misleading as its mathematical definition is not actually random nor a variable, but rather it is a function from possible outcomes (e.g., the possible upper sides of a flipped coin such as heads H and tails T) in a sample space (e.g., the set $\{H, T\}$) to a measurable space (e.g., $\{-1, 1\}$ in which 1 is corresponding to H and -1 is corresponding to T, respectively), often to the real numbers. This can also be portrayed graphically:



The formal definition of a random variable is a function $X : \Omega \to T$, where T is a measurable space. If T is not countable (such as $\mathbb{R}$), then we use a subset $T' \subset T$, which is an image of Ω under X.

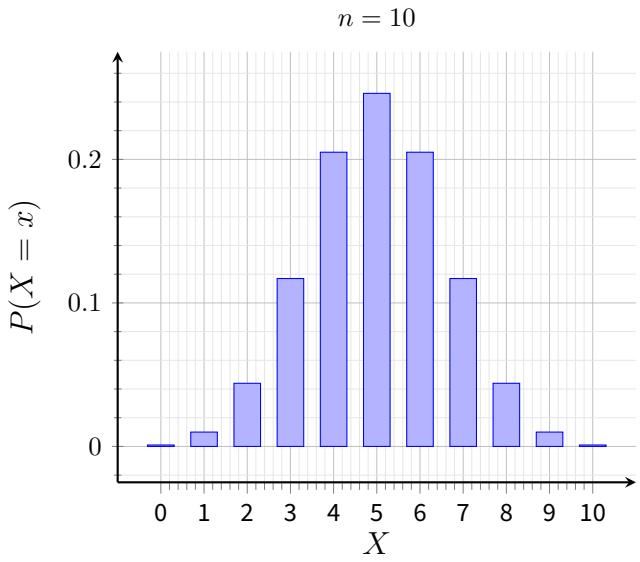
The probability that X takes a specific value x is calculated with $P(X = x) = P(\{\omega \in \Omega | X(\omega) = x\})$.

2.1 Probability Mass and Density Functions

When the image (= the range of) X is finite or infinitely countable ($T' \subset T$ is discrete), the random variable is called a **discrete random variable**. In that case, the distribution can be described by a **probability mass function (PMF)**, which assigns a probability to each value in the image of X. For a function $f(x)$ to be a valid PMF, it must satisfy these properties:

1. $f(x) = P(X = x) \geq 0$ for any $x \in T'$
2. $\sum_{x \in T'} f(x) = 1$

Let's look at an example. Let X be the number of heads flipped with 10 flips. We would expect a very low and very high number of heads to have a low probability. The middle section around 5 heads should have the heighest values.

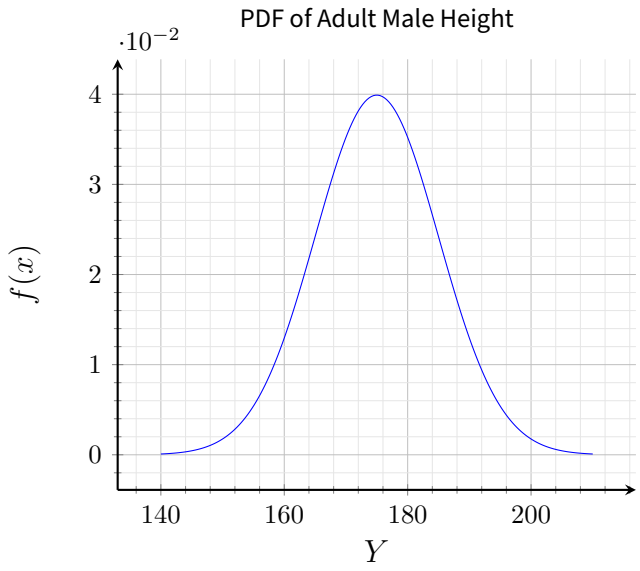


If the image is uncountably infinite then X is called a **countinuous random variable**. The distribution can be described by a **probability density function (PDF)**, which associates a probability with each range of realizations of X. For a function $f(x)$ to be a valid PDF, the following conditions must hold:

1. $f(x) \geq 0$ for any $x \in T'$
2. $\int_a^b f(x)dx = P(a < X < b) \geq 0$ for any $a, b \in T'$ satisfying $a < b$
3. $\int_{x \in T'} f(x)dx = 1$

Let's look at an example. Imagine we have a continuous random variable Y representing the height (in centimeters) of a randomly selected adults in St. Gallen. The height is a continuous variable because it can take any value within a certain range, say from 150 cm to 200 cm.

The PDF of Y might be normally distributed, centered around the average height of adults, say 175 cm, with a standard deviation (which measures the amount of variation or dispersion of a set of values) of 10 cm. This distribution is depicted as a bell-shaped curve, where the height of the curve at any given point represents the relative likelihood of that height occurring.

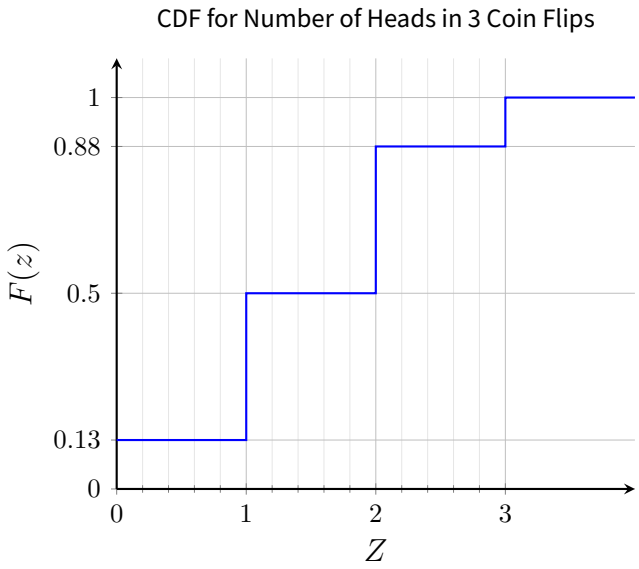


2.2 Cumulative Distribution Function

The **Cumulative Distribution Function (CDF)** is another important concept in the study of random variables. It describes the probability that a random variable X will take a value less than or equal to a specific value x. The CDF is defined for all real numbers and is denoted as $F(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\})$.

For a discrete random variable, the CDF is calculated as $F(x) = \sum_{x' \leq x} f(x')$, where $f(x')$ is the PMF. For a continuous random variable, the CDF is the integral of the PDF up to x, expressed as $F(x) = \int_{-\infty}^{x} f(t)dt$.

Let's look at an example: Consider a discrete random variable Z representing the number of heads obtained when flipping a coin 3 times. The CDF of Z would step up at each count of heads, reflecting the increasing cumulative probability of having 'up to that many' heads.



The CDF has several key properties:

1. **Right continuity**: A CDF is always continuous from the right: $\lim_{\Delta x \to 0} F(x + \Delta x) = F(x)$ at every point x.
2. **Non-decreasing**: $F_X(x) \leq F_X(y)$ if $x \leq y$.
3. **Limits**: A CDF approaches 0 as x approaches $-\infty$ and 1 as x approaches ∞: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

A CDF is continuous for a continuous random variable. For a discrete random variable, the CDF only increases at every $x \in T'$. This poses an issue when introducing boundaries, such as $P((a, b]) = F(b) - F(a)$, as we need to think about inclusion and exclusion of a and b:

- $P(a < X < b) = F(b) - F(a) - f(b)$
- $P(a \leq X \leq b) = F(b) - F(a) + f(a)$
- $P(a \leq X < b) = F(b) - F(a) + f(a) - f(b)$

2.3 Distributions of Discrete Random Variables

As alouded before a random variable can have different distributions. In the following section, we will take a look at the most common once.

2.3.1 Uniform Distribution

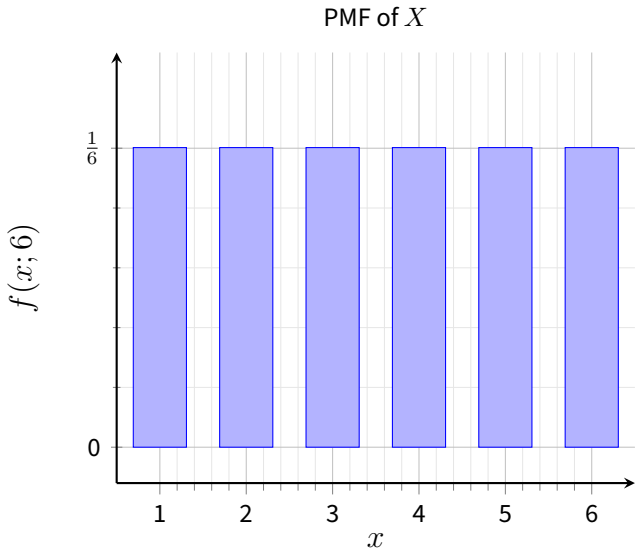
A random variable X is uniformly distributed if all x have the same probability. That means the PMF is given by

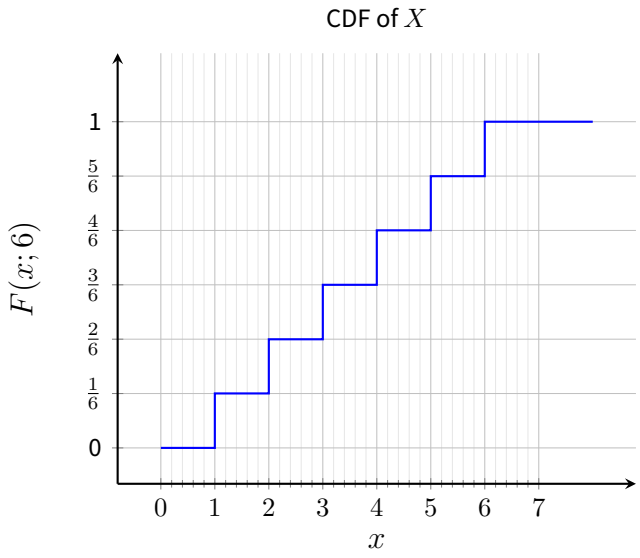
$$f(x;N) = P(X = x|N) = \begin{cases} \frac{1}{N}, & x = 1, \dots, N \\ 0, & \text{otherwise.} \end{cases}$$

An easy example is rolling a die once: The probability for $X =$ number rolled is given by

$$f(x;6) = P(X = x|6) = \begin{cases} \frac{1}{6}, & x = 1, \dots, 6 \\ 0, & \text{otherwise.} \end{cases}$$

In this example, the PMF and CDF for X look as follows:





2.3.2 Hypergeometric Distribution

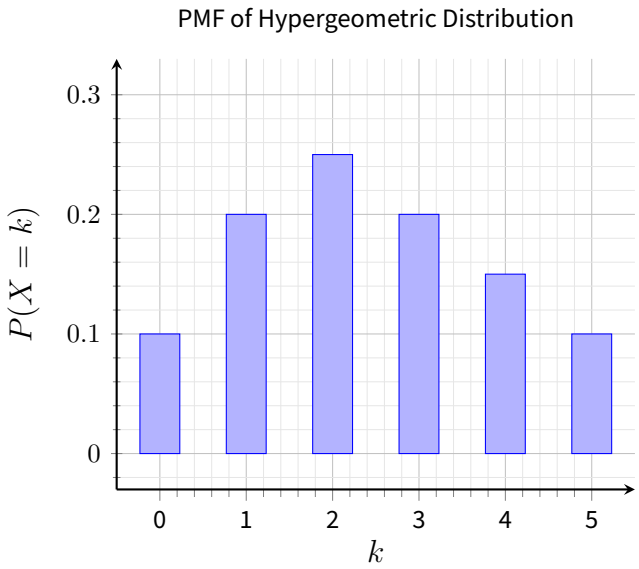
The hypergeometric distribution models the probability of a certain number of successes in a series of draws without replacement from a finite population. This distribution is characterized by three parameters: the population size (N), the number of successes in the population (K), and the number of draws (n). The probability mass function (PMF) for a hypergeometric distribution is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where X is the random variable representing the number of observed successes. From this we can derive that

$$K \geq x, \tag{1}$$
$$N - K \geq n - x, \tag{2}$$
$$x \geq n + K - N, \text{or} \tag{3}$$
$$n + K - N \leq x \leq K. \tag{4}$$

Assuming a population size of $N = 30$, a number of successes $K = 10$, and a number of draws $n = 5$, the PMF can be plotted as follows:



2.3.3 Bernoulli Distribution

The Bernoulli distribution is the simplest case of the binomial distribution, representing a single trial, often termed as a "success-failure" experiment. The probability mass function (PMF) of a Bernoulli random variable X is defined as:

$$f(x;p) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \\ 0, & otherwise, \end{cases}$$

where p is the probability of success. This distribution takes a value of 1 with probability p and a value of 0 with probability $1 - p$.

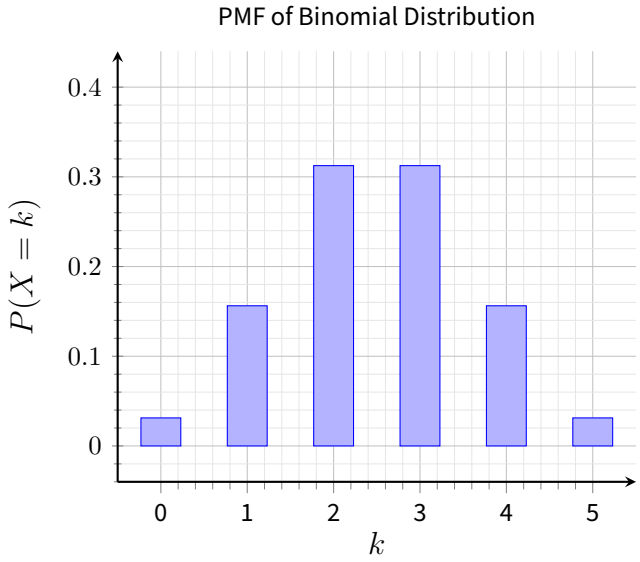
2.3.4 Binomial Distribution

The binomial distribution is a discrete probability distribution representing the number of successes in a fixed number of independent trials, each with the same probability of success. The probability mass function (PMF) for a binomial distribution with n trials and success probability p is:

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

where X is the random variable representing the number of successes.

For a binomial distribution with $n = 5$ trials and success probability $p = 0.5$, the PMF can be plotted as:



2.3.4.1 Binomial Distribution through Bernoulli Trials The binomial distribution can also be understood as the sum of outcomes from multiple Bernoulli trials. When we perform a fixed number n of independent Bernoulli trials, each with the same probability of success p, the resulting distribution of the number of successful outcomes is a binomial distribution.

Mathematically, if X_i represents the outcome of the i^{th} Bernoulli trial, then the total number of successes in n trials, represented by the random variable X, is given by:

$$X = \sum_{i=1}^{n} X_i,$$

where each X_i follows a Bernoulli distribution with probability p.

2.3.5 Poisson Distribution

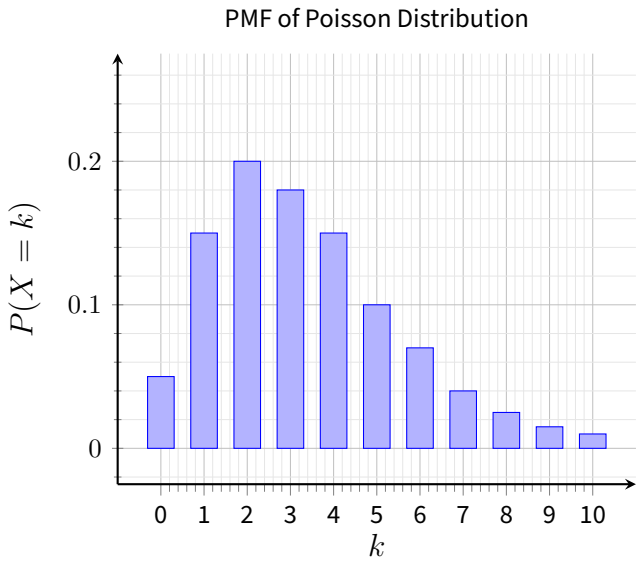
The Poisson distribution is used to model the number of events occurring within a fixed interval of time or space when these events occur with a known constant mean rate and independently of the time since the last event. The probability mass function (PMF) of a Poisson distributed random variable X is given by:

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

where λ is the average number of events in an interval and k is the actual number of events observed.

The Poisson distribution is approximation of the binomial distribution, but it is much faster to calculate, particularly for modeling rare events with a high number of trials.

For the Poisson distribution with a rate parameter $\lambda = 3$, the PMF can be plotted as follows:



2.3.5.1 Expected Value and Variance of a Poisson Distribution The Poisson distribution is a special case when it comes to the expected value and variance:

$$\mathbb{E}[X] = \mathbb{V}(x) = \lambda.$$

2.4 Distributions of Continuous Random Variables

2.4.1 Uniform Distribution

A random variable X is uniformly distributed if all x have the same probability. In case of a continuous random variable, the PDF must have the form

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & \text{if} x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

with parameters $-\infty < a < b < \infty$.

For instance, consider a random variable representing the landing position of a dart thrown at a 1-meter-long dartboard, where $a = 0$ and $b = 1$ meter, implying a uniform distribution over this range.

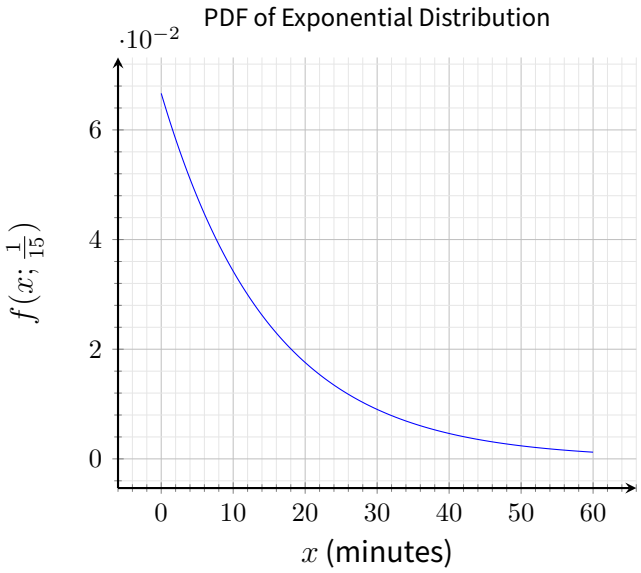
2.4.2 Exponential Distribution

The exponential distribution is commonly used to model the time between events in a Poisson process. It describes situations where events occur continuously and independently at a constant average rate. The probability density function (PDF) of an exponentially distributed random variable X is given by:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

where λ is the rate parameter, representing the average number of events in a given time interval.

An example could be the time until the next bus arrives at a station, assuming buses come on average every 15 minutes. In this case, $\lambda = 1/15$ per minute.



This means the likelihood of a bus arriving decreases as you wait longer, meaning it's more likely for the bus to arrive in the next 5 minutes than in 15 minutes.

2.4.2.1 Expected Value

The expected value of an exponential distribution is always equal to $\frac{1}{\lambda}$.

2.4.3 Gamma Distribution

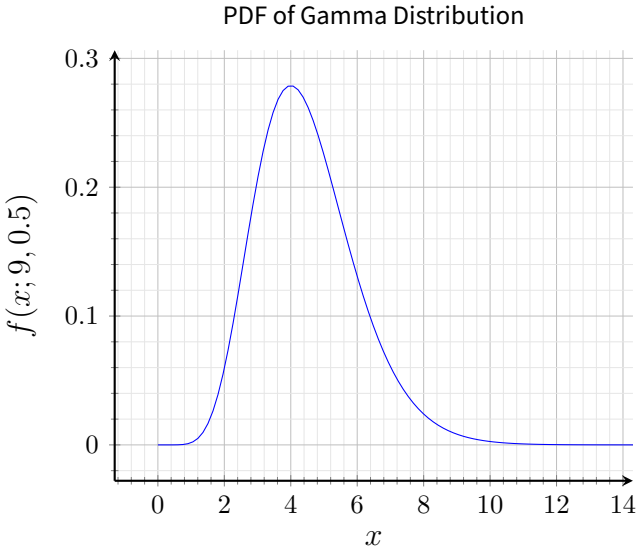
The gamma distribution is a two-parameter family of continuous probability distributions, which includes the exponential and Chi-squared distributions as special cases. The probability density

function (PDF) of a gamma distributed random variable X with shape parameter α and rate parameter β is:

$$f(x;\alpha,\beta) = \begin{cases} \frac{x^{\alpha-1}\cdot e^{-\frac{x}{\beta}}}{\beta^{\alpha}\cdot\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

where $\Gamma(\alpha)$ is the gamma function evaluated at α. The gamma distribution is often used to model waiting times for a sequence of events. The parameter β is calculated as $\beta = \frac{1}{\theta}$, where θ is the scale parameter, i.e. the rate for an event on average.

Assume a scenario where we are measuring the total time taken for a certain number of consecutive events, such as the total time taken for 9 buses to get from the start point to the finish line, assuming each bus takes an average of 2 minutes to arrive. Here, $\alpha = 9$ and $\beta = \frac{1}{2\text{min}}$.



The peak of the graph indicates the most probable total waiting time. This point is where the likelihood of the total waiting time for the nine buses is highest. As the curve moves away from this peak on either side, the probability of those waiting times decreases. Times significantly shorter than the peak are less likely because it's improbable all buses will arrive much earlier. Similarly, significantly longer times are also less likely, as it's rare for all buses to be much later than average. The shape of the curve reflects the range of most probable waiting times, tapering off as the times become more extreme in either direction.

2.4.4 Normal Distribution

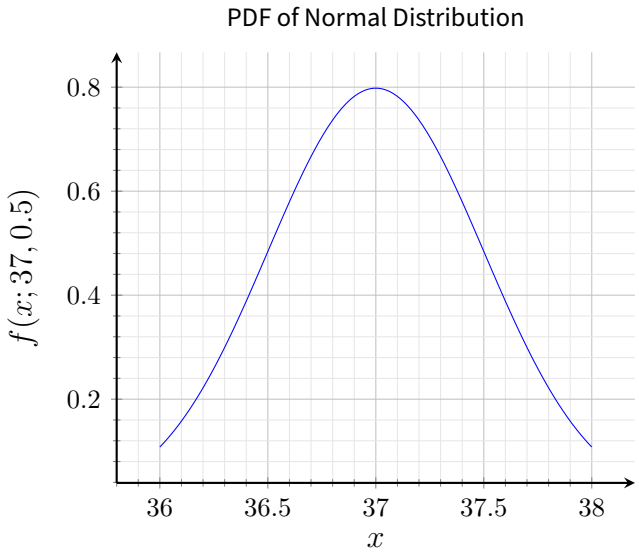
The normal (or Gaussian) distribution is one of the most commonly used probability distributions in statistics and natural sciences. It's defined by its mean (μ) and standard deviation (σ). The probability

density function (PDF) is given by:

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

The normal distribution is symmetric around its mean, and its shape is often referred to as a “bell curve”.

Let’s consider measuring adult human body temperatures, where the average body temperature is 37 °C with a standard deviation of 0.5 °C. Thus, $\mu = 37$ and $\sigma = 0.5$.



We see the peak at 37, which is the average body temperature in our example, and fewer occurences of temperatures much higher or lower.

Note for the exam: The equation of the normal distribution has to be learned by heart.

The integral of the normal distribution is called the Gaussian integral, and is not computable by hand. However, calculators can use numerical methods to “guess” the values.

2.4.5 Chi-Square Distribution

The Chi-Square distribution is a special case of the gamma distribution and is widely used in hypothesis testing and confidence interval estimation. It arises primarily in the context of estimating variances. The probability density function (PDF) of a Chi-Square distributed random variable X with k degrees of freedom is:

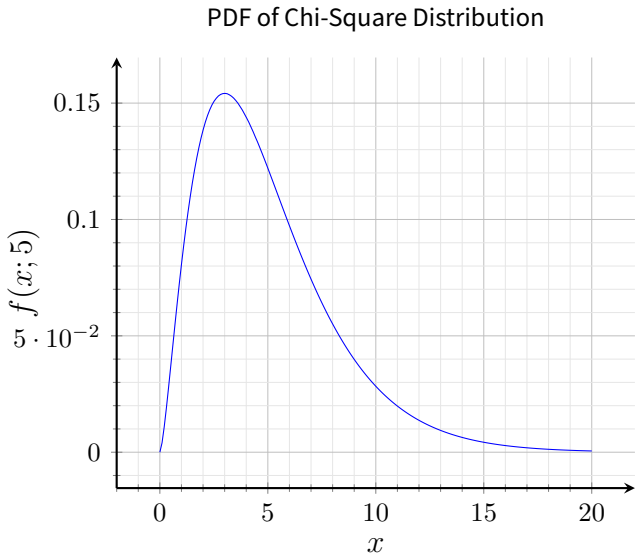
$$f(x;k) = \begin{cases} \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

where $\Gamma(k/2)$ is the gamma function evaluated at $k/2$. The Chi-Square distribution is used, for example, in the Chi-Square test for goodness of fit.

The parameter k relates to the number of independent factors in the data. The value of k influences the shape of the Chi-Square distribution. With an increase in k, the distribution becomes more symmetric and approaches a normal distribution. For smaller values of k, the distribution is skewed right.

To illustrate with a concrete example, consider a genetic study analyzing the distribution of eye color in a population. In this scenario, the Chi-Square distribution could be applied to compare observed frequencies of eye color with expected frequencies based on genetic theories. Assuming the study involves five categories of eye color (e.g., brown, blue, green, hazel, and other), we would use $k = 5$ degrees of freedom. This parameter choice reflects the count of independent eye color categories minus one (since one category can be determined if the others are known).

A graphical representation of the Chi-Square distribution for this example, with $k = 5$ degrees of freedom, would look like this:



Interpreting the graph within our genetics study context, the PDF of the Chi-Square distribution with 5 degrees of freedom illustrates how likely different Chi-Square statistic values align with expected genetic patterns. The peak of the curve reflects the most probable value of the statistic, while the rightward skew indicates that higher values are possible but less likely. If our calculated Chi-Square statistic from the eye color data falls near the peak, this would suggest little deviation from expected frequencies. However, if our statistic falls in the tail to the right, it may indicate a significant deviation, potentially questioning the expected genetic inheritance patterns.

2.5 The Expectation of a Random Variable

Let X be a random variable and f be its probability or density function. The expected value of X is defined as the weighted average of all possible values that the random variable can take.

The expectation can be calculated with

$$\mathbb{E}[X] = \sum_x xf(x) = \sum_x xp_x, \quad \text{if} X \text{is discrete or}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx, \quad \text{if} X \text{is continuous.}$$

Let's look at a concrete example:

- Suppose we have a six-sided fair die. The random variable X here represents the outcome of a die roll, so X can take any value from 1 to 6 with equal probability. Therefore, the probability function $f(x)$ or p_x is $\frac{1}{6}$ for each outcome. The expected value of X can be calculated as:

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

Instead of $\mathbb{E}[X]$, we can also write μ_X.

The central property of the expected value μ_X is the difference between the random variable and the expected value equaling zero, that is

$$\mathbb{E}[X - \mu_X] = 0.$$

Instead of x, we can also use a real-valued function $g(x)$. This means that the expected value of a function of a random variable, $g(X)$, can be found similarly. For a discrete random variable, this is expressed as

$$\mathbb{E}[g(X)] = \sum_x g(x)f(x) = \sum_x g(x)p_x,$$

and for a continuous random variable, it is given by

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Let's continue with our example:

- Returning to the die example, let's define a function $g(X) = X^2$. The expected value of $g(X)$ is:

$$\mathbb{E}[g(X)] = \sum_{x=1}^{6} x^2 \cdot \frac{1}{6} = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6}.$$

For linear functions $g(x)$, we can also take a shortcut. Let's say $Y = g(X) = aX + b$, then the expected value of Y equals

$$\mathbb{E}[Y] = \mathbb{E}[aX + b] = a \cdot \mathbb{E}[X] + b.$$

In general, the operation $\mathbb{E}$ is linear, meaning we can use linear operations. For instance,

$$\mathbb{E}[g(x)] = \mathbb{E}[3X + 2X^2] = \mathbb{E}[3X] + \mathbb{E}[2X^2] = 3\mathbb{E}[X] + 2\mathbb{E}[X^2].$$

2.5.1 Density Transformation Theorem

Assume we have a random variable X with a known PDF $f_X(x)$, and we want to transform it using a function g, to get a new random variable $Y = g(X)$. To find the PDF of Y, denoted as $f_Y(y)$, we can use the density transformation theorem:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |\frac{d}{dy} g^{-1}(y)|$$

Here's what each component means:

- $g^{-1}(y)$ is the inverse function of g, which we use to express x in terms of y.
- $\frac{d}{dy} g^{-1}(y)$ is the derivative of the inverse function with respect to y, which accounts for the rate of change of X with respect to Y.
- The absolute value of the derivative, $|\frac{d}{dy} g^{-1}(y)|$, is used to ensure the resulting PDF is non-negative, as probabilities cannot be negative.

It is important to note that this theorem assumes certain regularity conditions on the function g, such as g being a differentiable and monotonic function, so that the inverse function g^{-1} exists and is unique.

2.5.2 Variance

An important concept related to the expectation is the variance, denoted as $\mathbb{V}(X)$ or σ_X^2, which measures the spread of a distribution. It is defined as the expected value of the squared deviation of X from its mean μ_X, that is,

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X] - (\mathbb{E}[X])^2.$$

This assumes the sum or the integral of the expected value exists. If the mean of X does not exist, then $\mathbb{V}(X)$ also does not exist.

For a discrete random variable, the variance can be calculated as

$$\mathbb{V}(X) = \sum_x (x - \mu_X)^2 f(x) = \sum_x (x - \mu_X)^2 p_x,$$

and for a continuous one, it is

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx.$$

The standard deviation, σ_X, is simply the nonnegative square root of the variance ($\sigma_X = +\sqrt{\mathbb{V}(X)}$) and gives a measure of the average distance of the values of X from the mean.

The variance of a sum of two variables can be calculated with $\mathbb{V}(X+Y) = \mathbb{V}(X)+\mathbb{V}(Y)+2\text{Cov}(X,Y)$, where Cov is the coveriance of the 2 random variables.

Similarly to the shortcut for calculting the expected value of a linear function, there are simple rules for calculating the variance of a linear function. Let's say X is a random variable with an existing variance $\mathbb{V}(X)$ and $Y = aX + b$, then

$$\mathbb{V}(Y) = a^2 \cdot \mathbb{V}(X)$$

and

$$\sigma_Y = |a| \cdot \sigma_X$$

To illustrate the concept of variance, let's consider a few concrete examples:

1. Discrete random variable example: Suppose we have a six-sided fair die. The probability of each face (1 through 6) is equal, i.e., $\frac{1}{6}$. Here, the random variable X represents the outcome of a die roll. The mean μ_X can be calculated as $\frac{1+2+3+4+5+6}{6} = 3.5$. The variance $\mathbb{V}(X)$ is then calculated using the formula for discrete random variables:

$$\mathbb{V}(X) = \sum_{x=1}^{6} (x - 3.5)^2 \cdot \frac{1}{6}$$

2. Linear transformation of a random variable: If we have a random variable X with a known variance, say $\mathbb{V}(X) = 4$, and we define a new random variable $Y = 3X + 5$, then the variance of Y can be calculated using the transformation rule:

$$\mathbb{V}(Y) = 3^2 \cdot \mathbb{V}(X) = 3^2 \cdot 4 = 36$$

   And the standard deviation of Y, σ_Y, would be $|3| \cdot \sqrt{4} = 6$.

3 Multiple Random Variables

We have looked at how a random variable maps from the sample space to real numbers. What about when we are interested in the outcome of an event that is characterized by multiple random variables? We can generalize probability mass and density over multiple random variables, which we will cover in this chapter.

3.1 Joint Distributions

The combination of random variables is denoted as a random vector. For example, X and Y is denoted as the bivariate random vector (X, Y).

If all random variables are descrite (a discrete random vector), then they are governed by a joint probability mass function; if all the random variables are continuous (a continuous random vector), then they are governed by a joint probability density function. For instance, let (X, Y) be a discrete bivariate random vector, then the function $f(x, y)$ from $\mathbb{R}^2$ to $\mathbb{R}$ defined by $f(x, y) = P(X = x, Y = y)$ is called the joint PMF, also denoted as $f_{X,Y}(x, y)$.

Let's look at a practical example:

Consider a scenario where we roll two fair dice. Let X be the random variable representing the sum of the numbers on the two dice, and Y be the random variable representing the absolute difference between the numbers on the two dice. Here, X can take values from 2 to 12 (inclusive), and Y can take values from 0 to 5 (inclusive).

The joint PMF $f_{X,Y}(x, y)$ in this case gives the probability of any particular combination of X and Y. For instance, the probability that the sum of the dice is 7 and their absolute difference is 1, is calculated as $f_{X,Y}(7, 1)$. To find this, we identify all dice roll outcomes that result in a sum of 7 and an absolute difference of 1, and then divide by the total number of possible outcomes (36, since each die has 6 faces). There are two outcomes that meet the criteria: $(3, 4)$ and $(4, 3)$, so $f_{X,Y}(7, 1) = \frac{2}{36} = \frac{1}{18}$.

More generally, let $A \subset \mathbb{R}^2$, then

$$P((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y).$$

As with univariate random variables, we can define a real valued function $g(x, y)$ over all possible values (x, y) of the discrete random vector (X, Y). Then $g(X, Y)$ represents a new random variable. Its expected value is given by

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x, y)f(x, y),$$

where $f(x, y)$ is the multivariate PMF.

For example, let's consider $g(X, Y) = XY$:

$$\mathbb{E}[XY] = 2 \times 0 \times \frac{1}{36} + \cdots = \frac{245}{18}.$$

More generally, any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for any $(x,y) in $\mathbb{R}^2$ and

$$\sum_{(x,y)\in\mathbb{R}^2} f(x, y) = P((X, Y) \in \mathbb{R}^2) = 1$$

is the joint PMF of some discrete bivariate random vector (X, Y).

The excpected value continues to have the same properties as univariate random variables:

$$\mathbb{E}[ag_1(X, Y) + bg_2(X, Y) + c] = a\mathbb{E}[g_1(X, Y)] + b\mathbb{E}[g_2(X, Y)] + c.$$

The same goes for joint PDFs. A function $f(x, y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is called joint probability density function of the continuous random vector (X, Y) if, for every $A \subset \mathbb{R}^2$,

$$P((X, Y) \in A) = \int_A \int f(x, y)dxdy.$$

The notation $\int_A \int$ is $\int_{A_x} \int_{A_y}$, meaning that the limits of integration are set so that the function is integrated over all $(x, y) \in A$.

Any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

is the joint PDF of some continuous bivariate random vector (X, Y).

The expected value for any $g(X, Y)$ is

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$$

The joint comulative distribution function or joint CDF is the function $F(x, y)$ defined by $F(x, y) = P(X \leq x, Y \leq y)$ for all $(x, y) \in \mathbb{R}^2$.

Based on a continuous bivariate random vector, we have

$$F(x, y) = \int_{-\infty}^{x} \int_{\infty}^{y} f(s, t)dtds.$$

3.2 Marginal Distributions

Often we have direct access to a joint density function, but we are more interested in the probability of an outcome of a subset of the random variables in the joint density. Let (X, Y) be a discrete bivariate random vector with joint PMF $f_{X,Y}(x,y), then the marginal PMFs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x, y).$$

The joint PMF, $f_{X,Y}(x, y)$, cannot be determined from knowledge of only the marginal PMFs, $f_X(x)$ and $f_Y(y)$, since there are many different joint distributions that have the same marginal distributions.

The marginal probability density functions of X and Y are also defined as in the discrete case with integrals replacing sums:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy \quad -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \quad -\infty < y < \infty$$

3.3 Conditional Distributions

Let (X, Y) be a discrete (/ continuous) bivariate random vector with joint PMF (/ PDF) $f(x, y)$ and marginal PMFs (/ PDFs) $f_X(x)$ and $f_Y(y)$.

For any x such that $P(X = x)$ $f_X(x) > 0$, the conditional PMF (/ PDF) of Y given that $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x, y)}{f_X(x)}.$$

A couple of things apply to the conditional PMF / PDF:

- $f(y|x) \geq 0$ since $f(x, y) \geq 0$ and $f_X(x) > 0$
- $\sum_y f(y|x) = \frac{\sum_y f(x,y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$

Thus, the conditional PMF / PDF is indeed valid and can be used in the usual way to compute probabilities involving Y given the knowledge that $X = x$ occurred.

We can use conditional PMFs / PDFs to calculate conditional expected values. Given $g(Y)$ and $X = x$,

the excpected value is

$$\mathbb{E}[g(Y)|x] = \sum_y g(y)f(y|x) \quad \text{and} \quad \mathbb{E}[g(Y)|x] = \int_{-\infty}^{\infty} g(y)f(y|x)dy$$

in the discrete and continuous case, respectively.

Furthermore, $\mathbb{E}[Y|X]$ represents the best guess or prediction of Y when X is known. This is because $\mathbb{E}[Y|X]$ minimizes the expected squared difference between Y and any other possible function $g(X)$:

$$\text{min}\mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2].$$

3.3.1 Example of Calculating Conditional PDFs

This example demonstrates the calculation of conditional probability density functions (PDFs) for a continuous random vector.

3.3.1.1 Joint PDF of Random Vector (X, Y) Consider a continuous random vector (X, Y) with a joint probability density function (PDF) defined as:

$$f(x, y) = e^{-y}, \text{ for } 0 < x < y < 1.$$

This function describes the likelihood of the vector (X, Y) falling within a specific range.

3.3.1.2 Marginal PDF of X The marginal PDF of X, denoted as $f_X(x)$, is derived by integrating the joint PDF over all possible values of Y. This gives us the probability distribution of X regardless of Y.

- **Case 1:** When $x \leq 0$, the joint PDF $f(x, y) = 0$ for all values of y, leading to $f_X(x) = 0$.
- **Case 2:** When $x > 0$, we consider the range where $f(x, y) > 0$, which occurs only if $y > x$. Hence, we calculate:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_x^{\infty} \exp(-y)\, dy = \exp(-x)$$

3.3.1.3 Conditional PDF of Y Given X The conditional PDF of Y given X, denoted as $f(y|x)$, is the ratio of the joint PDF to the marginal PDF of X. It represents the distribution of Y when X is known.

- **For $y > x$:**

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{\exp(-y)}{\exp(-x)} = \exp(-(y - x))$$

- **For $y < x$:**

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{0}{\exp(-x)} = 0$$

Given $X = x$, Y follows an exponential distribution with location parameter x and scale parameter $\beta = 1$.

3.3.1.4 Conditional Mean and Variance of Y Given $X = x$

- **Conditional Mean of Y:** The conditional expectation $\mathbb{E}[Y|X = x]$ is calculated as:

$$\mathbb{E}[Y|X = x] = \int_x^\infty y \exp(-(y - x))\, dy = 1 + x$$

- **Conditional Variance of Y Given $X = x$:** The conditional variance $V(Y|X = x)$ is the difference between the expectation of Y^2 and the square of the expectation of Y:

$$V(Y|X = x) = \mathbb{E}[Y^2|X = x] - (\mathbb{E}[Y|X = x])^2 = \int_x^\infty y^2 e^{-(y-x)}\, dy - \left(\int_x^\infty y e^{-(y-x)}\, dy\right)^2 = 1$$

Note: The marginal distribution of Y is $\Gamma(2, 1)$, which implies $V(Y) = 2$. However, with the knowledge that $X = x$, the variability in Y is significantly reduced.

3.4 Independence

Let (X, Y) be a bivariate random vector with joint PDF or PMF $f(x, y)$ and marginal PDFs or PMFs $f_X(x)$ and $f_Y(y)$, then X and Y are called **independent random variables** if, for every $x, y \in \mathbb{R}$,

$$f(x, y) = f_X(x) f_Y(y).$$

If X and Y are independent, the conditional PDF of Y igven $x = x$ is

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y).$$

Using that we can simplify

$$P(Y \in A|x) = \int_A f(y|x) dy = \int_A f_Y(y) dy = P(Y \in A)$$

for any $A \subset \mathbb{R}$ and $x \in \mathbb{R}$. This means the knowledge that $X = x$ gives us no additional information about Y and is not needed for calculating probabilities of Y.

Using this current definition of independence requires f_X and f_Y. However, there is an easier check: If there exist functions $g(x)$ and $h(y)$ such that, for every $x, y \in \mathbb{R}$,

$$f(x, y) = g(x)h(y),$$

then the X and Y are independent.

Independent random variables make certain calculations easier:

- For any $A, B \subset \mathbb{R}$: $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.
- Let $g(x)$ be a function only of x and $h(y)$ be a function only of y, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.
- Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\gamma, \tau^2)$ be independent normal random variables. Then the random variable (RV) $Z = X + Y$ has a $\mathcal{N}(\mu + \gamma, \sigma^2 + \tau^2)$ distribution.

3.5 Covariance

Covariance is a measure of how much two random variables vary together. It's a concept that helps in understanding the relationship between two variables. For two random variables X and Y, their covariance is defined as the expected value of the product of their deviations from their respective means. Mathematically, it is expressed as:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Where $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are the expected values (means) of X and Y respectively.

Depending on the result, we can understand the variables' covariance:

- If $\text{Cov}(X, Y) > 0$, then X and Y tend to move in the same direction.
- If $\text{Cov}(X, Y) < 0$, then X and Y tend to move in opposite directions.
- If $\text{Cov}(X, Y) = 0$, there's no linear relationship between X and Y.

3.5.1 Properties of Covariance

1. **Symmetry:** $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. **Linearity:** $\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y)$ for constants a, b, c, and d.
3. **Zero Covariance of Independent Variables:** If X and Y are independent, then $\text{Cov}(X, Y) = 0$. Note that the converse is not necessarily true; zero covariance does not imply independence.

3.6 Correlation

Correlation, denoted as $\rho(X, Y)$, measures the strength and direction of a linear relationship between two variables. It's a normalized form of covariance that provides a unitless measure of linear relationship.

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where σ_X and σ_Y are the standard deviations of X and Y respectively.

We say that X and Y are positively correlated if $\rho_{X,Y} > 0$, that X and Y are negatively correlated if $\rho_{X,Y} < 0$, and that X and Y are uncorrelated if $\rho_{X,Y} = 0$.

3.6.1 Properties of Correlation

1. **Range:** The value of $\rho(X, Y)$ lies between -1 and 1.
2. **Symmetry:** $\rho(X, Y) = \rho(Y, X)$.
3. **Independence:** If X and Y are independent, then $\rho(X, Y) = 0$. Again, the converse is not necessarily true.

3.6.2 Limitations

- Correlation only assesses linear relationships.
- Correlation does not imply causation.
- It can be affected by outliers in the data.

3.7 Central Limit Theorem

A sequence of n random variables $X_1, \ldots, X_n$ is called a random sample of size n if they are independent and identically distributed (i.i.d.) and $\mathbb{E}[X_i] = \mu < \infty$ and $\mathbb{V}(X_i) = \sigma^2 < \infty$.

The central limit theorem states that the distribution of a sample approximates a normal distribution as the sample size become larger.

Mathematically, we first define $S_n = X_1 + \cdots + X_n$ and $\overline{X}_n = \frac{S_n}{n}$. Then, the random variable

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

converges in distribution to the standard normal random variable for $n \to \infty$. That is:

$$lim_{n\to\infty} P(Z_n \leq x) = \mathcal{N}(0, \sigma^2).$$

3.7.1 Limit Theorem of De Moivre and Laplace

The theorem of De Moivre and Laplace is an extension of the Central Limit Theorem, applied to binomially distributed random variables. It focuses on a binomially distributed random variable S_n with parameters n (number of trials) and p (probability of success in each trial). As the number of trials n increases, the distribution of S_n approaches a normal distribution.

Mathematically, the distribution function of S_n, denoted as $F_{Bi}(s_n; n, p)$, converges to a normal distribution $F_N(s_n; np, np(1-p))$ as n approaches infinity. This means that for large values of n, the binomial distribution can be approximated by a normal distribution with mean np and variance $np(1-p)$.

3.7.2 Examples Demonstrating the Theorems

3.7.2.1 Example 1: Tossing a Coin Consider a fair coin (where the probability of heads, p, is 0.5) tossed 100 times ($n = 100$). We want to find the probability of getting between 45 and 55 heads.

For a binomial distribution with $n = 100$ and $p = 0.5$, the mean (np) is 50 and the standard deviation ($\sqrt{np(1-p)}$) is approximately 5. Using the normal approximation, the probability of getting between 45 and 55 heads is approximately 68.27%. The actual probability is 72.87%

3.7.2.2 Example 2: Rolling a Die Consider a six-sided die rolled 60 times ($n = 60$). We want to find the probability of getting exactly 15 sixes, where the probability of rolling a six, p, is $\frac{1}{6}$.

For this binomial distribution, the mean (np) is 10 and the standard deviation ($\sqrt{np(1-p)}$) is approximately 2.89. Applying the normal approximation to find the probability of getting exactly 15 sixes (using continuity correction), we get approximately 3.11%. The actual probability is 3.09%

4 Descriptive Statistics

The task of descriptive statistics is to introduce a number of descriptive, in most cases graphical methods to summarize all the information about the variables under investigation and illustrate the main features without distorting the picture.

4.1 Scale of a Random Variable

A very important characteristic of statistical variables is the scale in which they are measured. We will look at three basic scales. [1]

4.1.1 Nominal Scale

A variable measured on a nominal scale is a variable that does not really have any evaluative distinction. One value is really not any greater than another. A good example of a nominal variable is sex (or gender). Information in a data set on sex is usually coded as 0 or 1, 1 indicating male and 0 indicating female (or the other way around: 0 for male, 1 for female). 1 in this case is an arbitrary value, and it is not any greater or better than 0. There is only a nominal difference between 0 and 1. With nominal variables, there is a qualitative difference between values, not a quantitative one.

4.1.2 Ordinal

Something measured on an ordinal scale does have an evaluative connotation. One value is greater or larger or better than the other. Product A is preferred over product B, and therefore A receives a value of 1 and B receives a value of 2. Another example might be rating your job satisfaction on a scale from 1 to 10, with 10 representing complete satisfaction. With ordinal scales, we only know that 2 is better than 1 or 10 is better than 9; we do not know by how much. It may vary. The distance between 1 and 2 maybe shorter than between 9 and 10.

4.1.3 Ratio

A variable measured on a ratio scale gives exact information on betterness compared to ordinal scales. The distance between each value is the same and there is an absolute zero point. Temperature measured in Kelvin is an example. There is no value possible below 0 degrees Kelvin, it is absolute zero. Weight is another example, 0 lbs. is a meaningful absence of weight. Your bank account balance is another. Although you can have a negative or positive account balance, there is a definite and nonarbitrary meaning of an account balance of 0.

4.2 Numerical Techniques

Using numberical techniques, we can derive key information from raw data points, which allows us to (sort of) summarize the data.

[1] https://web.pdx.edu/~newsomj/pa551/lecture1.htm

4.2.1 Measures of Location

The most familiar measueres of location are [2]

- The arithmatic mean (requires a ratio scale):

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

- The median (requires ordinal or ratio scale):

$$x_{med} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ is even} \end{cases}$$

- The mode, a measure of frequency (requires nominal, ordinal, or ratio scale)

4.2.1.1 Empirical Quantiles A more nuanced analysis to the arithmatic mean are empirical α-quantiles. While the mean describes only the central tendency of a distribution or random sample, quantiles are able to describe the whole distribution. [3]

Taking an ordered list of observations x_1, …, x_n, we can compute the quantile at level $\alpha \in (0, 1)$ as follows:

1. $K = \lfloor \alpha n \rfloor + 1$
2. $$\begin{cases} x_K & \text{if } \alpha \times n \text{ is not an integer number} \\ \frac{1}{2}(x_K + x_{K-1}) & \text{if} \alpha \times n \text{ is an integer number} \end{cases}$$

The resulting number is the α-percent of observations that lie below the empirical α-quantile.

4.2.2 Measures of Distribution

Location measures only give some information about the central tendency of a distribution (empircal quantiles are similar to moving the center to α). Two distributions might have the same location parameter while being different. A measure of dispersion can help distinguish among distributions with the same location measure.

Common measures of dispersion are

[2] https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/mldap
[3] https://lorentzen.ch/index.php/2023/02/11/quantiles-and-their-estimation/

- The range, which is the difference between the smallest and largest observations:

$$r = x_{\text{max}} - x_{\text{min}}.$$

- The mean quartile range, which is the average quartile size:

$$MQA = \frac{Q_3 - Q_2 + Q_2 - Q_1}{2} = \frac{IQA}{2},$$

  where $IQA = Q_3 - Q_1$ is called the interquartile range.
- The variance

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x - i - \overline{x})^2$$

  and the standard deviation

$$s_x = +\sqrt{s_x^2}.$$

However, these are all absolute measures of dispersion and are dependent on the unit in which the variable is measured. Relative measures of dispersion are calculated as ratios or percentages. They are always dimensionless, and they are particularly useful for making comparisons between separate data sets or different experiments that might use different units. [4]

To compore degrees of dispersion of two variables, we can use a relative measure called coefficient of variation:

$$VK_x = \frac{s_x}{|\overline{x}|}.$$

4.3 Visualizations

We can visualize the results of numerical analysis to better convay the information. There are many different plots, but we will focus on three: Boxplots, QQ-plots, and scatterplots.

4.3.1 Boxplots

Boxplots, also known as box-and-whisker plots, offer a five-number summary of a data set: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. These plots provide a visual representation of the distribution of the data, highlighting its central tendency, variability, and skewness.

4.3.1.1 Construction of a Boxplot

[4] https://www.statisticshowto.com/relative-dispersion-absolute-dispersion/

1. **Central Box**: The central box of the boxplot represents the interquartile range (IQR), which is the distance between the first and third quartiles (Q3 - Q1). This box shows the middle 50% of the data.
2. **Median Line**: Within this box, a line is drawn at the median (Q2) of the dataset, offering a clear view of the data's central tendency.
3. **Whiskers**: Lines, or 'whiskers', extend from either end of the box to the maximum and minimum values within 1.5 times the IQR from the quartiles. These whiskers represent the range of the bulk of the data.
4. **Outliers**: Points that lie beyond the whiskers are often plotted individually as dots and considered outliers, highlighting potential anomalies in the data.

4.3.1.2 Interpretation

- **Spread**: The length of the box and whiskers indicates the spread of the data. A longer boxplot suggests greater variability.
- **Symmetry**: If the median is equidistant from the quartiles, the distribution is symmetrical. Otherwise, it indicates skewness.
- **Outliers**: Outliers can indicate either data variability or errors in data collection.

4.3.1.3 Example Let's illustrate the concept of a boxplot with an example. We are given this dataset:

i	x[i]	Median	Quartile
1	7		
2	7		
3	31		Q1 = 31
4	31		
5	47		
6	75		
7	87	Q2 = 87	
8	115		
9	116		
10	119		Q3 = 119
11	119		

i	x[i]	Median	Quartile
12	155		
13	177		

For this dataset, the interquartile range (IQR) is calculated as IQR = Q3 - Q1 = 119 - 31 = 88.

Now, let's visualize this data set in a boxplot. In the boxplot, the central box will represent the interquartile range (IQR) between Q1 and Q3, the median (Q2) will be marked within this box, and the whiskers will extend to the minimum and maximum values that are within 1.5 times the IQR from the quartiles.

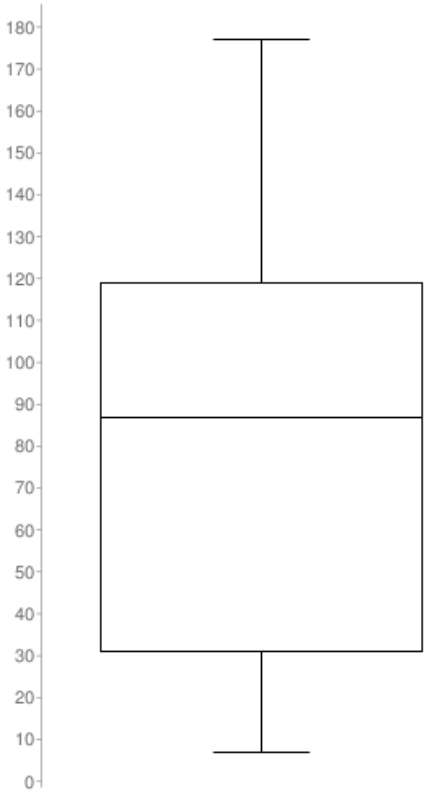


Figure 1: Box plot

This boxplot visually summarizes the distribution of the data, allowing for easy comparison and understanding of the spread, central tendency, and outliers.

4.3.2 Quantile-Quantile-plot (QQ-plot)

While QQ-plots exist, nobody really understands what they do. Something for checking your theoretical distribution against your actual results. Yada yada yada, if you get a line close to $y = x$ you're good or smth.

4.3.3 Scatter plot

In a scatter plot, each value is plotted. This gives the viewer an idea on how to group the data or find relations between variables.

There are two often used types of scatter plots:

- One dimensional: All points are plotted on a single line
- Two dimensional: Points are plotted on an x-y plane

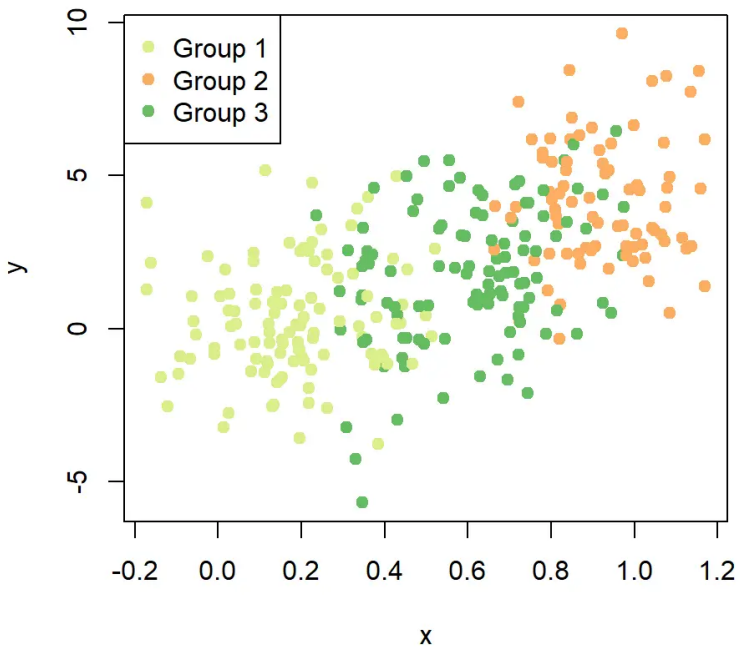


Figure 2: Scatter Plot

5 Estimation of Parameters

Estimators are fundamental concepts in statistics, primarily used for making inferences about population parameters based on sample data.

An estimator is a rule or a formula that tells you how to calculate an estimate of a population parameter based on sample data. For instance, if you want to estimate the average height of all students in a school, you might take a sample of students, calculate the average height of just those in the sample, and use that as your estimate for the entire school's average height.

For creating an estimate of the unknown parameters, we have two main approaches:

- Point estimation: We calculate a single value based on the sample.
- Confidence intervals: We calculate an interval in which the true value of the unknown parameter is likely in.

Given random variables of interest $X_1, \ldots, X_n$, any arbitrary real-valued function $\hat{\theta_n}(X_1, \ldots, X_n)$ is called an estimator. The hat means that this is an estimator (I like to call it cap). Ideally, that estimate is close to $\mu = \mathbb{E}[X]$. The estimated mean is denoted $\mu_{\hat{\theta_n}}$.

An easy example is calculating the mean of the sample, which is a point estimation:

$$\hat{\mu} = \hat{\theta_n}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

However, this gives us no information about how precise the estimate is. Since $\hat{\theta_n}$ is a random variable itself, we can calculate its variance:

$$\mathbb{V}(\hat{\theta_n}) = \mathbb{E}[(\hat{\theta_n} - \mu_{\hat{\theta_n}})^2] = \sigma^2_{\hat{\theta_n}}.$$

5.1 Bias

An estimator $\hat{\theta_n}$ is unbiased for θ if $\mathbb{E}[T] = \mu_{\hat{\theta_n}} = \theta$ for all values of θ. The difference between the expected value of $\hat{\theta_n}$ and the parameter θ is called the Bias. That is

$$\text{Bias}(\hat{\theta_n}) = \mathbb{E}[\hat{\theta_n}] - \theta.$$

1. **Example 1:** An unbiased estimator for $\mu = \mathbb{E}[X]$ (where $\mu = \mathbb{E}[X_i], i = 1, \ldots, n$) can be calculated as follows:
$$T = T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We can show that $\mathbb{E}[T] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu$.

2. **Example 2:** We aim to prove that the estimator $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - X_n)^2$ for the population variance $\sigma^2 = \mathsf{Var}(X)$ is biased. S^2 is the sample variance.

First, we calculate the expected value of S^2:

$$\mathbb{E}[S^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - X_n)^2]$$

Expanding and simplifying:

$$\mathbb{E}[S^2] = \frac{1}{n}\sum_{i=1}^{n}\Big[\mathsf{Var}(X_i) + (\mu - X_n)^2\Big]$$

Since $\mathsf{Var}(X_i) = \sigma^2$ and simplifying further:

$$\mathbb{E}[S^2] = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mu - X_n)^2$$

As $\mathbb{E}[S^2]$ is not equal to σ^2, S^2 is a biased estimator. To obtain an unbiased estimator, we use:

$$T = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_n)^2$$

This adjusted estimator is unbiased for σ^2.

5.1.1 Goodness

If the bias of an estimator becomes monotonically smaller when the sample size increases, and vanishes as $n \to \infty$, then the estimator is asymptotically unbiased:

$$\lim_{n\to\infty}\mathbb{E}[\hat{\theta_n}] = \theta$$

This means that an unbiased estimator is good.

As the sample size n grows for an estimator $\hat{\theta_n} = T(X_1, \ldots, X_n)$, it may converge to the unknown parameter θ. In this case, we the estimator is consistent. Mathematically, this means that

$$P(|\hat{\theta_n} - \theta| > \epsilon) \to 0 \text{ as } n \to \infty$$

An easier check for consistency is analyzing if the following two conditions are satisfied:

- It is unbiased (or asymptotically unbiased)
- As $n \to \infty$, its variance vanishes

5.1.2 Effiency

If the variance of an unbiased estimator $\hat{\theta_n}$ is smaller (or equal) to other unbiased estimators, then $\hat{\theta_n}$ is called efficient.

5.1.3 Mean Squared Error

The mean squared error (MSE) compelemnts the criteria introduced so far to judge the goodness of an estimator.

$$MSE(\theta) = \mathbb{E}[(\hat{\theta_n} - \theta)^2]$$

The MSE can also be transformed to

$$MSE(\theta) = \mathbb{V}(\hat{\theta_n}) + \text{bias}^2$$

If $MSE(\theta) = \mathbb{V}(\hat{\theta_n})$, then we know there is no bias. This means that the estimator is good.

5.2 Point Estimators

5.2.1 Method of Moments

Now that we have discussed how to characterize an estimator, let's look at how to get an estimator. The easiest is the method of moments. I will leave out some details here because they are out of scope.

A moment is a specific characteristic of the whole population (the whole data). Mathematically, the k-th moment is given by

$$\mu_k = \mathbb{E}[X^k]$$

We need n moments to estimate n unknown parameters of a distribution.

We can use a sample to estimate those moments μ_k. This is called a sample moment and is defined as

$$\hat{\mu_k} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Each X_i represents one data point of our sample.

What is obvious is that the first moment μ_1 is simply the mean of the population: $\mu_1 = \mathbb{E}[X] = \mu$.

Now we know how to calculate an estimate for the moments. What we really want is to calculate the parameters of a distribution.

Let's say, we believe our population to follow a normal distribution, which has 2 parameters: μ and σ. We now need to express the moments through these parameters and then use our sample moments to estimate the moments. At the end this equation will leave us with the value of the parameters. How do we do that?

Before we start replacing things, let's write down the equation we are starting out with:

$$\begin{aligned} \mu_1 &= \mathbb{E}[X] \\ \mu_2 &= \mathbb{E}[X^2] \end{aligned} \tag{5}$$

We already know that the first moment, μ_1, is one of our parameters, the mean. Now we need to think of how to express $\mathbb{E}[X^2]$ with the parameters. For that, we can use the definition of variance: $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, which equates to $\mathbb{E}[X^2] = \mathbb{V}(X) + \mathbb{E}[X]^2$. Let's plug these into the equation:

$$\begin{aligned} \mu_1 &= \mu \\ \mu_2 &= \mathbb{V}(X) + \mathbb{E}[X]^2 = \sigma^2 + \mu^2 \end{aligned} \tag{6}$$

Now we use our sample moments to estimate the moments μ_k. For that we use the afformentioned definition of sample moments:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \mu \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \sigma^2 + \mu^2 \end{aligned} \tag{7}$$

Let's rearrange this a bit:

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2 \end{aligned} \tag{8}$$

And that's it! Now we have all the parameters of the distribution in terms of our sample data. In an actual calculation we would be given this data, for example $X =$ `[ 5.91026825, 5.6001254 , 6.82458987, 1.22446064, 14.93518372, 3.20409614, 0.16061157, 1.83804317, 4.18228092, 7.62702544]`. Each number is one X_i. So, at the end, we

only need to sum these number and their squares to get an estimate of μ and σ (given these numbers you should get $\hat{\mu} = 5.15, \hat{\sigma} = 4.24$).

5.2.2 Least-Squares Method

This method was barely covered in the lecture and is therefore not relevant.

The basic idea is to minimize the following sum:

$$\text{argmin} \sum_{i=1}^{n} (X_i - \mu)^2$$

As the minimum is found, we will have calculated an estimate $\hat{\mu}_{LS}$ of μ.

5.2.3 Maximumum Likelihood Method

The maximumum likelihood method is often used in linear regression. It allows us to estimate the unknown mean parameter μ.

The core idea of maximum likelihood is to find the set of parameters that make the observed data most probable. In other words, we're looking for the parameter values that maximize the likelihood that our sample was generated by a given distribution.

Let's take a known distribution f_x with a parameter θ that we want to estimate. The likelihood function for a given sample $X_1, \ldots, X_n$ returns the probablity of observing the sample, given the parameter θ. Mathematically, it's denoted as

$$L(\theta) = f_x(X_1, \ldots, X_n|\theta).$$

For independent and identically distributed (i.i.d.) samples this becomes:

$$L(\theta) = \prod_{i=1}^{n} f_x(X_i|\theta).$$

Our goal is to maximize that function to get our maximum likelihood estimate $\hat{\theta}$:

$$\hat{\theta} = \text{argmax} L(\theta).$$

This is where the "maximum likelihood" name comes from. I practice, it's often easier to work the log of the likelihood function, as it turns the product into a sum:

$$\hat{\theta} = \log L(\theta) = \sum_{i=1}^{n} f_x(X_i|\theta).$$

The resulting maximum likelihood estimator (MLE) is consistent and asymptotically unbiased.

Let's consider a practical example using the maximum likelihood method. Assume we have a dataset of heights (in centimeters) and we want to find the most likely mean height, assuming that the data follows a normal distribution. Let's also assume we know the standard deviation (σ) of the population, which is 10 cm. Our dataset is as follows:

$X = [160, 165, 170, 175, 180, 185]$

Since we're assuming a normal distribution, the probability density function (PDF) f for a normal distribution is:

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

The likelihood function for our dataset is the product of the pdfs for each data point:

$$L(\mu) = \prod_{i=1}^{6} \frac{1}{10\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{200}}$$

It's easier to maximize the log of the likelihood function:

$$\log L(\mu) = \sum_{i=1}^{6} \log\left(\frac{1}{10\sqrt{2\pi}}\right) - \frac{(X_i-\mu)^2}{200}$$

To find the value of μ that maximizes $\log L(\mu)$, we take its derivative with respect to μ and set it to zero.

The derivative of the log-likelihood function:

$$\frac{d}{d\mu} \log L(\mu) = \sum_{i=1}^{6} \frac{X_i - \mu}{100}$$

Set the derivative equal to zero:

$$\sum_{i=1}^{6} \frac{X_i - \mu}{100} = 0$$

Solve for μ:

$$\mu = \frac{\sum_{i=1}^{6} X_i}{6}$$

Now, we need to plug in our sample values X, and we get an estimated mean height $\hat{\mu} = 172.5cm$. This $\hat{\mu}$ maximizes the probability of getting our sample given a normal distribution with $\sigma = 10$.

5.3 Confidence Intervals

A confidence interval (or CI) is a range of estimates for an unknown parameter. The interval theoretically contains the true value of the parameter with x confidence. Usually, the interval is computed with 90 confidence level.

CIs are often used in hypothesis testing. Usually, there are 2 hypotheses:

1. Null hypothesis (H0): This is a statement that suggests there is no effect or no difference. Usually, this is what we are testing against.
2. Alternative hypothesis (H1): This states that there is an effect or a difference.

If the calculated CI lies within the null hypothesis, then there is not enough evidence to reject the null hypothesis. If H0 lies outside the CI, there is a high probability the H0 can be rejected. The probability that H0 is falsely rejected is the significance level $\alpha = 1 -$ confidence. Rejecting H0 when it's actually true is called a type I error.

I am not going to go into detail about the calculation as it likely not relevant.