Probability Theory Notes

Meow

September 29, 2025

Contents

Chapter 1

Probability and Statistics

1.3 Basics of Statistics

Definition 1.3.1. A **class** is one of the categories into which quantitative data can be classified.

Now, the **class frequency** is the number of observations in the data set falling in a particular class:

$$\text{class relative frequency} = \frac{\text{class frequency}}{n} \tag{1.1}$$

where n is the total number of observations in the data set.

Example 1.3.1. Consider the following data set of exam scores from a class of 20 students:

$$65, 72, 68, 85, 90, 62, 75, 78, 80, 82, 88, 92, 95, 70, 73, 77, 81, 84, 89, 93$$

If we create classes with intervals of 10 points (60-69, 70-79, 80-89, 90-99), the class frequencies would be:

- 60-69: 3 scores (65, 68, 62)
- 70-79: 6 scores (72, 75, 78, 70, 73, 77)
- 80-89: 7 scores (85, 80, 82, 88, 81, 84, 89)
- 90-99: 4 scores (90, 92, 95, 93)

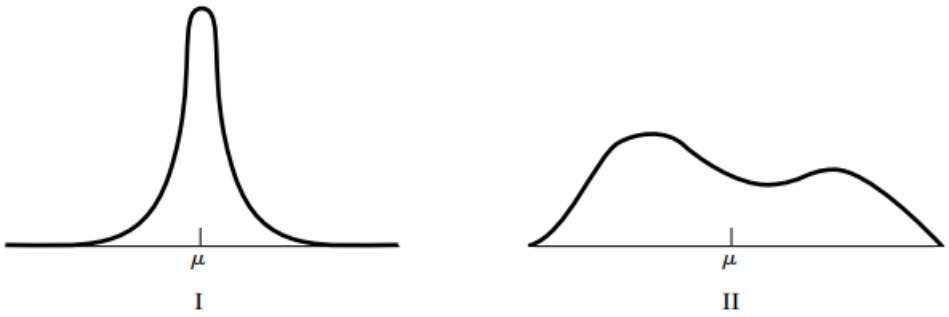
The relative frequencies would be:

- 60-69: $3/20 = 0.15$ or 15%
- 70-79: $6/20 = 0.30$ or 30%
- 80-89: $7/20 = 0.35$ or 35%
- 90-99: $4/20 = 0.20$ or 20%

The mean of a sample of n measured responses $y_1, y_2, \ldots, y_n$ is:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1.2}$$

The mean only shows the center of the distribution of the data. Different datasets may have equal means but different frequency distributions. In the figure below, what differs is the variation around the mean:



The **variance** of a sample $y_1, y_2, \ldots, y_n$ is the sum of squared differences between each measurement and the sample mean, divided by $n - 1$:

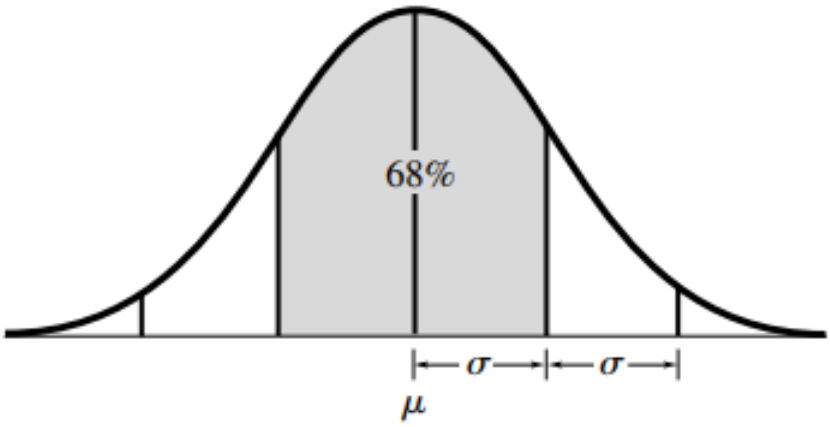
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1.3}$$

It shows how much the data values vary compared to each other, in a dataset.
The **standard deviation** is the square root of the variance:

$$s = \sqrt{s^2} \tag{1.4}$$

It gives an accurate picture of how far apart the values in a dataset are from the mean. For a distribution that is approximately bell-shaped, the following intervals apply:

- $\mu \pm \sigma$ contains approximately 68% of the measurements.
- $\mu \pm 2\sigma$ contains approximately 95%.
- $\mu \pm 3\sigma$ contains approximately all measurements.



Definition 1.3.2. The **range** of a sample of n measurements $y_1, y_2, \ldots, y_n$ is the difference between the largest and smallest measurements in the sample.

Example 1.3.2. If a sample consists of measurements 3, 1, 0, 4, 7, find the sample mean and range.

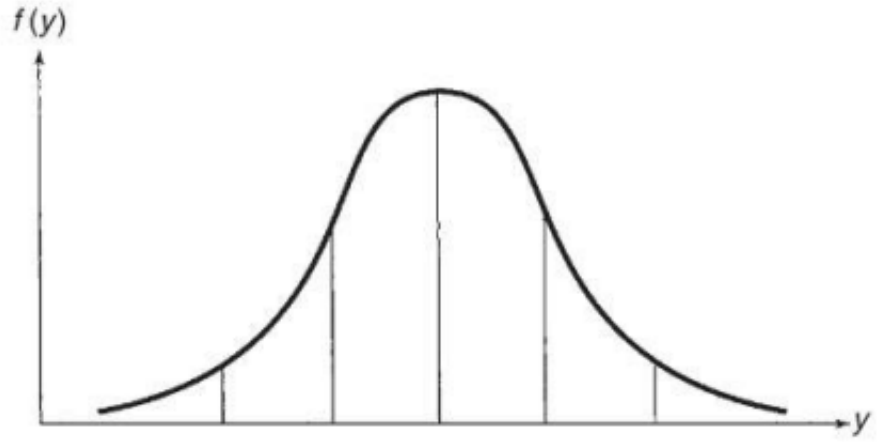
Solution:

$$\text{Sample mean} : \frac{1}{5}\sum_{i=1}^{5} y_i = 3$$

$$\text{Range} : 7 - 0 = 7$$

1.3.1 Normal Probability Distribution

This distribution is symmetric to its mean μ and the spread is determined by the value of the standard deviation σ.



The **z-score** measures the distance between the data points and the mean of the distribution, in number of standard deviations:

$$z = \frac{y - \mu}{\sigma} \tag{1.5}$$

But we can also use calculus to compute probabilities based on the z-score:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \tag{1.6}$$

Example 1.3.3. Let $X \sim \mathcal{N}(100, 15^2)$. Find the probability that X lies between 85 and 115.

Solution:

$$P(85 \leq X \leq 115) = \int_{85}^{115} \frac{1}{\sqrt{2\pi}\cdot 15}\, e^{-\frac{(x-100)^2}{2\cdot 15^2}}\, dx \tag{1.7}$$

which is equal to **0.6827** or 68.27%.

1.3.2 Sampling Distribution

It's the distribution of all possible sample statistic computed from all possible samples of size n from a population size of N.

Theorem 1.3.1. If $y_1, y_2, \ldots, y_n$ represent a random sample of n measurements from a large population with mean μ and standard deviation σ, the mean and standard error of estimate of the sampling distribution is:

$$\text{Mean} \ : E(\bar{y}) = \mu\bar{y} = \mu$$
$$\text{Standard error of estimate} : \sigma\bar{y} = \frac{\sigma}{\sqrt{n}}$$

The **standard error** shows how different is the sample mean between different samples and the population itself.

Theorem 1.3.2 (Central Limit Theorem)**.** For large sample sizes, the mean $\bar{y}$ of a sample from a population with mean μ and standard deviation σ has a sampling distribution approximately normal, regardless of the probability distribution of the sampled population. The larger the sample size, the better the approximation.

Now, the **covariance** demonstrates the direction of the linear relationship. It could be either positive or negative:

$$\text{Covariance}(Y, X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y}) \cdot (x_i - \bar{x})}{(n-1)} \tag{1.8}$$

Since covariance is affected by changes in the units of measurement, we standardize it:

$$z_i = \frac{y_1 - \bar{y}}{s_y} \tag{1.9}$$

where:

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}} \tag{1.10}$$

is the **standard deviation** of Y. We do the same for X.

The **correlation coefficient** can be interpreted as the covariance of standardized variables:

$$Cor(Y, X) = \frac{\sum(y_i - \bar{y}(x_i - \bar{x}))}{\sqrt{\sum(y_1 - \bar{y})^2 \sum(x_i - \bar{x})^2}} \tag{1.11}$$

Correlation is useful to indicate both direction and strength of the linear relationship of your independent variables. The magnitude gives us the strength between Y and X. The closest the correlation is to -1 or 1, the strongest the relationship. The sign gives us the direction. Having a correlation of 0 does not mean there's no relationship between the variables; it only means their relationship isn't **linear**.

1.3.3 Confidence Interval

When estimating a population parameter using a sample statistic, we'll always have some error because we're making approximations. To express that error, we use interval estimate:

$$\text{Point estimate} \ \pm \text{Margin of error} \tag{1.12}$$

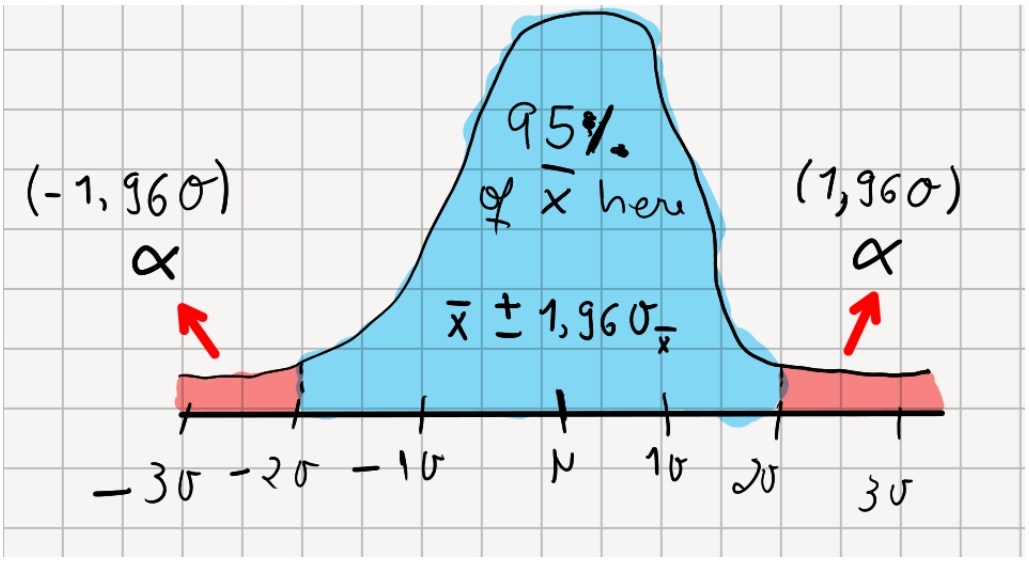
A **confidence interval** is the range of values that likely contains the true population parameter. A **point estimate** is a single value used to estimate a population parameter. The **confidence level** is the percentage of all possible samples that can be expected

to include the true population parameter. When the standard deviation is known the formula for the interval estimate is:

$$\bar{y} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{1.13}$$

If the population standard deviation is unknown, we can use :

$$\bar{y} \pm Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \tag{1.14}$$



Since our α is in 2 parts of our distribution, we divide it by 2, getting 0.25% on each side (our alpha is 5% because we have a 95% confidence interval in our case).

Remark 1.3.1. We can only use the standard normal curve when we **know** the population standard deviation. Otherwise, we gotta use the t-distribution curve.

Since the standard error depends on the sample size, our standard error of the mean will become larger and the distribution wider.

Example 1.3.4. A coffee shop owner wants to estimate the average daily coffee consumption per customer. She randomly selected 50 customers and recorded their daily coffee consumption. The average is 2.3 cups and the sample standard deviation is 0.5 cups. She wants a 95% confidence interval.

Solution:

$$\bar{y} \pm Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 2.3 \pm 1.96 \cdot \frac{0.5}{\sqrt{50}} \approx 2.3 \pm 0.14 \tag{1.15}$$

The answer is 2.16 - 2.44 cups per customer, approximately, with 95% confidence.

This large-sample method works when the sample size ≥ 30, so that the sample standard deviation μ can be a good approximation of σ. When it's smaller, it's required that the sampled population have a normal probability distribution. We can use the **t-distribution**:

$$\bar{y} \pm t_{\frac{\alpha}{2}} s_{\bar{y}} \tag{1.16}$$

where $s_{\bar{y}} = \frac{s}{\sqrt{n}}$ is the estimated standard error of $\bar{y}$.

1.3.4 Hypothesis Testing

Hypothesis testing is a statistical method used to evaluate claims about population parameters based on sample data.

Key Definitions

- **Null Hypothesis ($\mathbf{H_0}$)**: The default assumption, postulated to be true unless evidence suggests otherwise.
  Example: $H_0 : \mu = \mu_0$ (no effect or no difference).
- **Alternative Hypothesis ($\mathbf{H_a}$)**: Counters the null hypothesis, representing the effect or difference we want to test.
  Examples:
  - $H_a : \mu > \mu_0$ (one-tailed, right-tailed)
  - $H_a : \mu < \mu_0$ (one-tailed, left-tailed)
  - $H_a : \mu \neq \mu_0$ (two-tailed)

Test Statistic

A standardized value used to assess the null hypothesis:

$$z = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}}, \quad \text{where} \quad \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

- $\bar{y}$: Sample mean
- μ_0: Hypothesized population mean under H_0
- $\sigma_{\bar{y}}$: Standard error of the mean (measures sampling variability)

Decision Rules

1. **Rejection Region Approach:**
   Compare the test statistic (z) to a critical value (z_α) based on the chosen significance level (α).
   - For $H_a : \mu > \mu_0$, reject H_0 if $z > z_\alpha$.
   - For $H_a : \mu < \mu_0$, reject H_0 if $z < -z_\alpha$.
   - For $H_a : \mu \neq \mu_0$, reject H_0 if $|z| > z_{\alpha/2}$.
2. **p-Value Approach:**
   p-value: Probability of observing a test statistic as extreme as (or more extreme than) the sample result, assuming H_0 is true.
   Rule: Reject H_0 if p-value $< \alpha$.

Errors in Hypothesis Testing

- **Type I Error (α)**: Rejecting H_0 when it is true (false positive).
- **Type II Error (β)**: Failing to reject H_0 when it is false (false negative).
- **Power ($1 - \beta$)**: Probability of correctly rejecting H_0 when H_a is true.

1.3.5 Confidence Interval for Two Populations

To construct a confidence interval for two populations, we use the difference of their means. For large samples ($n \geq 30$), the sampling distribution of the sample means $\bar{x}_1$ and $\bar{x}_2$ is approximately normal and we use the *z-score* as our test statistic. The variance of the populations are rarely known so we use their variances to estimate the population mean:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sigma_{\bar{x}_1 - \bar{x}_2} = (\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{1.17}$$

We're assuming here that the samples were randomly and independently selected and their sizes are large enough so that their sampling distributions are approximately normal.

Example 1.3.5. A technology company is evaluating the feasibility of expanding its operations in either São Paulo or Rio de Janeiro. To compare the average annual salaries of software engineers in these two cities, the company collects two independent random samples:

- **São Paulo (Population 1):**
  - Sample size: $n_1 = 100$
  - Sample mean salary: $\bar{x}_1 = \text{R\$120,000}$
  - Sample standard deviation: $s_1 = \text{R\$15,000}$
- **Rio de Janeiro (Population 2):**
  - Sample size: $n_2 = 120$
  - Sample mean salary: $\bar{x}_2 = \text{R\$112,000}$
  - Sample standard deviation: $s_2 = \text{R\$18,000}$

Assume both samples are independent, random, and sufficiently large such that the Central Limit Theorem applies.

Question: Construct a 95% confidence interval for the difference in mean salaries, $\mu_1 - \mu_2$, between São Paulo and Rio de Janeiro.

1.3.6 Chi-square Test

The chi-square (χ^2) test determines whether there is a statistically significant association between two categorical variables by comparing observed frequencies to expected frequencies under the null hypothesis.

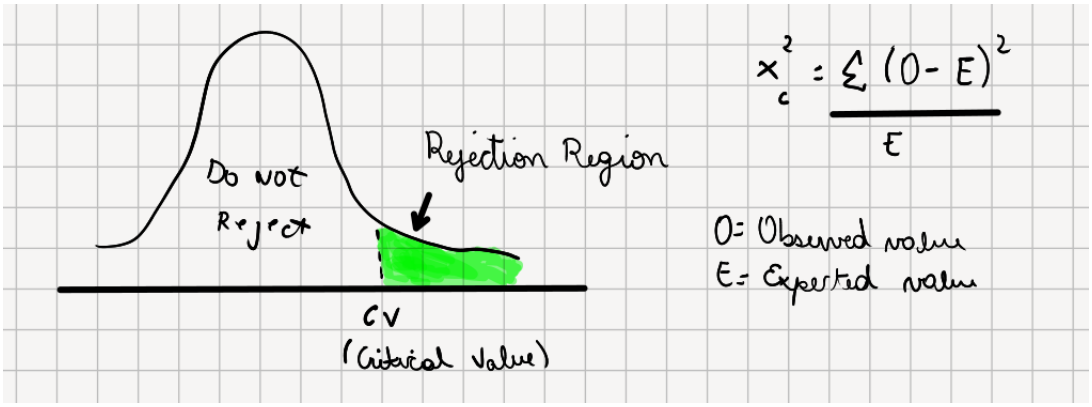
A school principal would like to know in which days of the week the students are more likely to be absent. The principal expects that students will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which day of the week they had the highest number of student absences. The observed and expected results are shown below. Based on these days, do the days for the highest number of absences occur with equal frequencies? (Use a 5% significance level).

The Chi-squared distribution is an uneven distribution.

Table 1.1: Observed vs. Expected Student Absences by Weekday

Day of Week	Observed Absences	Expected Absences
Monday	25	20
Tuesday	15	20
Wednesday	20	20
Thursday	18	20
Friday	22	20
Total	**100**	**100**

H_0: Equal frequencies of absences across weekdays
H_a: Unequal frequencies of absences across weekdays



1.3.7 Simple Linear Regression Model

The linear regression model expresses the relationship between Y and X as shown below:

$$Y = B_0 + B_1 X_i + \epsilon \tag{1.18}$$

where B_0 is the y-intercept, B_1 is the slope and ϵ accounts for the error in the approximation (the random disturbance). The word **linear** could mean that the relationship between Y and X is linear, but it could also mean that the coefficients enter in a linear fashion in the model.

To estimate the parameters, we find the best fitting line of the points in our scatter plot. For that, we can use the *least squares method*, which gives us the line that minimizes the sum of squares of the difference between the observed values and the estimations:

$$\epsilon_i = y_i - \beta_0 - \beta_1 X_i \tag{1.19}$$

The sum is:

$$S(B_0, B_1) = \sum \epsilon_i^2 = \sum (y_i - \beta_0 - \beta_1 X_i)^2 \tag{1.20}$$

To minimize it we can use:

$$B_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \tag{1.21}$$

and:

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1.22}$$

1.4 Basics of set theory

Let Ω denote the sample space, and let ω denote an element of that sample space. We say that ω is an element of Ω, written as $\omega \in \Omega$. The negation, indicating that ω is not an element of Ω, is written as $\omega \notin \Omega$.

A set A is a *subset* of Ω if every element of A is also an element of Ω; that is,

$$A \subseteq \Omega \iff \forall \omega \in A,\ \omega \in \Omega.$$

A set A is a *proper subset* of Ω, denoted by $A \subset \Omega$, if it is a subset of Ω but not equal to Ω. Formally,

$$A \subset \Omega \iff A \subseteq \Omega \wedge A \neq \Omega.$$

The *complement* of A, denoted by A^c represents all elements that do not belong in A. The *union* represents all elements that are in at least one of the sets:

$$A_1 \cup A_2 \cdots \cup A_n = \bigcup_{i=1}^{n} A_i \quad \text{with } i = 1, 2, \ldots, n$$

The **intersection** of two sets A_1 and A_2, denoted by:

$$A_1 \cap A_2$$

represents the elements belonging to both sets. We can further extend this notion for any number of sets:

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n}, \text{ with } i = 1, 2, \ldots, n$$

Theorem 1.4.1. For any three events A, B and C, defined on a sample space Ω:

1. Commutativity
   - $A \cup B = B \cup A$
   - $A \cap B = B \cap A$
2. Associativity
   - $A \cup (B \cup C) = (A \cup B) \cup C$
   - $A \cap (B \cap C) = (A \cap B) \cap C$
3. Distributive
   - $A \cup (B \cap C) = (A \cap B) \cup (A \cap C)$
   - $A \cap (B \cup C) = (A \cup B) \cap (A \cup C)$
4. DeMorgan's Laws
   - $(A \cup B)^c = A^c \cap B^c$
   - $(A \cap B)^c = A^c \cup B^c$

To prove the distributive law $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, we establish set equality by proving containment in both directions. First, suppose $x \in A \cap (B \cup C)$. Then $x \in A$ and $x \in B \cup C$, meaning $x \in B$ or $x \in C$. If $x \in B$, then $x \in A \cap B$; if $x \in C$, then $x \in A \cap C$. In either case, $x \in (A \cap B) \cup (A \cap C)$, proving $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

Conversely, let $x \in (A \cap B) \cup (A \cap C)$. Then $x \in A \cap B$ or $x \in A \cap C$. If $x \in A \cap B$, then $x \in A$ and $x \in B$, so $x \in B \cup C$, implying $x \in A \cap (B \cup C)$. Similarly, if $x \in A \cap C$, then $x \in A$ and $x \in C$, so $x \in B \cup C$, again implying $x \in A \cap (B \cup C)$. Thus, $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. Since both inclusions hold, the equality follows.

Definition 1.4.1. Two events A and B are *disjoint* if their intersection is empty. Also, two events $A_1, A_2, \ldots$ are *pairwise disjoint* if $A_i \cap A_j = \emptyset, \forall i \neq j$.

Definition 1.4.2. If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a **partition** of the sample space.

1.4.1 Sequences and Their Limits

Monotone sequences can be classified as either **nondecreasing** or **nonincreasing**.

Nondecreasing Sequences

A sequence of sets $\{A_n\}$ is **nondecreasing** if:

$$A_n \subseteq A_{n+1} \quad \text{for all } n \in \mathbb{N}$$

The limit of such a sequence is given by:

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$$

For a numerical sequence $\{a_n\}$, nondecreasing means:

$$a_{n+1} \geq a_n \quad \forall n \in \mathbb{N}$$

Example:

$$a_n = n \quad \text{yields} \quad 1, 2, 3, 4, \ldots$$

Nonincreasing Sequences

A sequence of sets $\{A_n\}$ is **nonincreasing** if:

$$A_n \supseteq A_{n+1} \quad \text{for all } n \in \mathbb{N}$$

The limit in this case is:

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$$

Numerical Example:

$$b_n = \frac{1}{n} \quad \text{yields} \quad 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$$

Key Properties:

- Nondecreasing: $\limsup A_n = \liminf A_n = \bigcup_{n=1}^{\infty} A_n$
- Nonincreasing: $\limsup A_n = \liminf A_n = \bigcap_{n=1}^{\infty} A_n$

1.4.2 Probability of Unions, complements and intersections

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive:

$$P(A \cup B) = P(A) + P(B)$$

for complements:

$$P(A^c) = 1 - P(A)$$

Two events are **independent** if and only if the probability of occurrence of event B has no effect on the probability of occurrence of A, and vice-versa.

1.4.3 Probability Measure

A **probability measure** or **probability distribution** is a function that assigns for each event A a real number and satisfies three axioms:

1. $\mathcal{P}(A) \geq 0$
2. $\mathcal{P}(\Omega) = 1$
3. If $A_1, A_2, \ldots$ are disjoint, then:

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

1.4.4 Conditional Probability

We can reason about the outcome of an experiment, based on partial information:

Example 1.4.1. In a word guessing game, the first letter of the word is h. What's the probability that the second letter is an h?

The formula is:

$$P(A|B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of B}} \tag{1.23}$$

which corrresponds to:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.24}$$

where we assume that $P(B) \geq 0$.

Example 1.4.2. We toss a fair coin three successive times. We want to find the conditional probability $P(A|B)$, where A and B are the events:

$$A = \{\text{more heads than tails come up}\}, \quad B = \{\text{1st toss is a head}\}$$

The sample space is:

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Event B consists of 4 events:

$$\{HHH, HHT, HTH, HTT\}$$

If we assume they're equally likely:

$$P(B) = \frac{4}{8}$$

Now, $P(A \cap B)$ consists of three elements:

$$P(A \cap B) = \{HHH, HHT, HTH\} = \frac{3}{8}$$

So, the conditional probability of A given B is:

$$P(A|B) = \frac{\frac{3}{8}}{\frac{4}{8}} = \frac{3}{4}$$

We can extend this notion to n events:

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot \ldots \frac{P(\bigcap_{i=1}^{n} A_1)}{P(\bigcap_{i=1}^{n-1} A_i)} \tag{1.25}$$

Events A and B are **conditionally independent** given event C if:

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

1.4.5 Random Variables

If X represents some unknown value of interest or if this value could change, we have a **random variable**. The set $\mathcal{X}$ represents all the possible values for X. When the sample space $\mathcal{X}$ is finite or countably infinite, X is a **discrete random variable**. Then, the probability of X having value x is denoted by:

$$P(X = x)$$

Probability mass function

It describes the probabilities of possible discrete outcomes for our random variable. Some conditions must be met:

- $0 \leq P(X = x) \leq 1$
- $\sum_{\text{all } x} P(X = x) = 1$

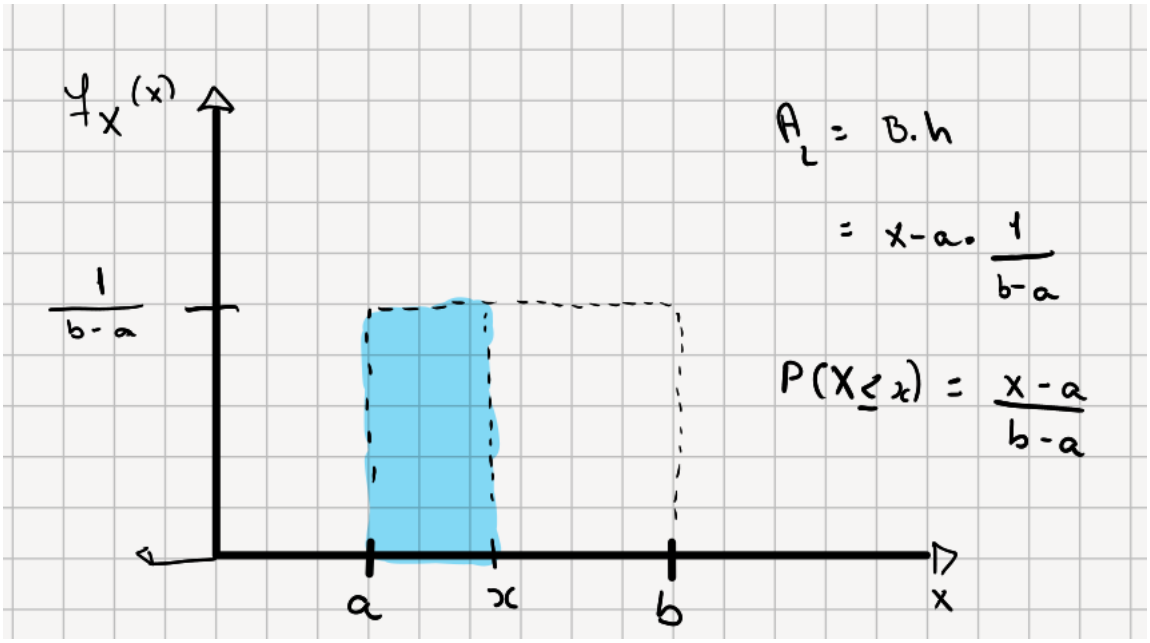
If $X \in \mathbb{R}$ we use **probability density function** instead.

Cumulative probability distributions

They're used to calculate the area under the curve to the left of a point of interest. It's used to evaluate the accumulated probability.

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

Let us consider a rectangular distribution, also called a **uniform distribution**.



1.4.6 Expectation

The **expected value** is the mean of the random variable:

$$E(X) = \mu = \sum xP(x)$$