PART I. PROBABILITY THEORY

Chapter 1. Probability Space

There are three approaches to the notion of *probability*:

- **classical**: intuitive, what most people are familiar with and think of when they hear the word "probability";
- **geometrical**: a natural extension of classical probability, for the case of infinite numbers of cases;
- **axiomatic**: rigorous, mathematical, enables proving probability formulas.

1 Experiments and Events

An **experiment** is any process or action whose outcome is not known (is random).
A **sample space**, denoted by S, is the set of all possible outcomes of an experiment.
Its elements are called **elementary events** (denoted by $e_i,\ i \in \mathbb{N}$).
An **event** is a collection of elementary events, i.e. it is a subset of S (events are denoted by capital letters, $A_i,\ i \in \mathbb{N}$).

Since events are defined as sets, we can employ set theory in describing them.

- two special events associated with every experiment:
  - the **impossible** event, denoted by $\emptyset$ ("never happens");
  - the **sure (certain)** event, denoted by S ("surely happens").
- for each event $A \subseteq S$, we define the event $\overline{A}$, the **complementary** event, to mean that $\overline{A}$ occurs if and only if A does not occur; $\overline{\overline{A}} = A$;

- we say that event A **implies (induces)** event B, $A \subseteq B$, if every element of A is also an element of B, or in other words, if the occurrence of A induces (implies) the occurrence of B; A and B are **equal**, $A = B$, if A implies B and B implies A;
- for any two events $A, B \subseteq S$, we define the following events:
  - **union** of A and B,

$$A \cup B = \{e \in S \,|\, e \in A \text{ or } e \in B\},$$

    the event that occurs if either A or B or both occur;
  - **intersection** of A and B,

$$A \cap B = \{e \in S \,|\, e \in A \text{ and } e \in B\},$$

    the event that occurs if both A and B occur;
  - **difference** of A and B,

$$A \setminus B = \{e \in S \,|\, e \in A \text{ and } e \notin B\} = A \cap \overline{B},$$

    the event that occurs if A occurs and B does not;
  - **symmetric difference** of A and B,

$$A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B),$$

    the event that occurs if A or B occur, but not both.

The operations of union, intersection and symmetric difference are

- **commutative**:

$$A \cup B = B \cup A, \quad A \cap B = B \cap A, \quad A \Delta B = B \Delta A;$$

– **associative**:

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

$$(A \Delta B) \Delta C = A \Delta (B \Delta C);$$

– **distributive**:

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

$$A \cap (B \Delta C) = (A \cap B) \Delta (A \cap C).$$

Definition 1.1.

- *Two events A and B are said to be **mutually exclusive (disjoint, incompatible)** if A and B cannot occur at the same time, i.e. $A \cap B = \emptyset$;*
- *Three or more events are mutually exclusive if any two of them are;*
- *A collection of events $\{A_i\}_{i \in I}$ from S is said to be **collectively exhaustive** if*

$$\bigcup_{i \in I} A_i = S;$$

- *A collection of events $\{A_i\}_{i \in I}$ from S is said to be a **partition** of S if the events are collectively exhaustive and mutually exclusive, i.e.*

$$\begin{aligned} \bigcup_{i \in I} A_i &= S \\ A_i \cap A_j &= \emptyset, \forall i, j \in I,\ i \neq j. \end{aligned}$$

Example 1.2. Consider the experiment of rolling a die. Then the sample space is

$$S = \{e_1, e_2, e_3, e_4, e_5, e_6\},$$

where the elementary events (outcomes) are $e_i,\ i = \overline{1,6}$, with e_i being the event that the face i shows.

Consider the following events:
A: face 1 shows,
B: face 2 shows,
C: an even number shows,
D: a prime number shows,
E: a composite number shows.
Then we have

$$A = \{e_1\},\ B = \{e_2\},\ C = \{e_2, e_4, e_6\},\ D = \{e_2, e_3, e_5\},\ E = \{e_4, e_6\}.$$

We also have

$$B \subseteq C,\ A \cap B = \emptyset,\ A \cap D = \emptyset,\ A \cap E = \emptyset,\ D \cap E = \emptyset,$$

$$C \cap D = B,\ A \cup D \cup E = S.$$

So, for example, events $\{A, B\}$ and $\{A, D, E\}$ are mutually exclusive. In fact, these last three are also collectively exhaustive. Thus, events $\{A, D, E\}$ form a partition of S.

Proposition 1.3.
For every collection of events $\{A_i\}_{i\in I}$, ***De Morgan's laws*** *hold:*

a) $\overline{\bigcup_{i\in I} A_i} = \bigcap_{i\in I} \overline{A}_i,$

b) $\overline{\bigcap_{i\in I} A_i} = \bigcup_{i\in I} \overline{A}_i.$

2 Sigma Fields, Probability and Rules of Probability

Definition 2.1. *A collection $\mathcal{K}$ of events from S is said to be a **σ-field** (**σ-algebra**) over S if it satisfies the following conditions:*

(i) $\mathcal{K} \neq \emptyset$;

(ii) *if $A \in \mathcal{K}$, then $\overline{A} \in \mathcal{K}$;*

(iii) *if $A_n \in \mathcal{K}$ for all $n \in \mathbb{N}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{K}$.*

*If $\mathcal{K}$ is a σ-field over the sample space S, then the pair $(S, \mathcal{K})$ is called a **measurable space**.*

Example 2.2. The power set $\mathcal{P}(S) = \{S' | S' \subseteq S\}$ is a σ-field over S.

Theorem 2.3. *Let $\mathcal{K}$ be a σ-field over S. Then the following properties hold:*

a) $\emptyset, S \in \mathcal{K}$.

b) *for all $A, B \in \mathcal{K}$, $A \cap B$, $A \setminus B$, $A \Delta B \in \mathcal{K}$.*

c) *if $A_n \in \mathcal{K}$, for all $n \in \mathbb{N}$, then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{K}$.*

Definition 2.4. *Let $\mathcal{K}$ be a σ-field over S. A mapping $P : \mathcal{K} \to \mathbb{R}$ is called **probability** if it satisfies the following conditions:*

(i) $P(S) = 1$;

(ii) $P(A) \geq 0$, *for all $A \in \mathcal{K}$;*

(iii) *for any sequence $(A_n)_{n \in \mathbb{N}} \subseteq \mathcal{K}$ of mutually exclusive events,*

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n), \tag{2.1}$$

(P is σ-additive).

*The triplet $(S, \mathcal{K}, P)$ is called a **probability space**.*

Theorem 2.5. *Let $(S, \mathcal{K}, P)$ be a probability space, and let $A, B \in \mathcal{K}$. Then the following properties hold:*

a) $P(\overline{A}) = 1 - P(A)$ *and* $0 \leq P(A) \leq 1$.

b) $P(\emptyset) = 0$.

c) $P(A \setminus B) = P(A) - P(A \cap B)$.

d) *If $A \subseteq B$, then $P(A) \leq P(B)$, i.e. P is monotonically increasing.*

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

Proof.
a) We have $A, \overline{A} \in \mathcal{K}$, $A \cup \overline{A} = S$ and $A, \overline{A}$ are mutually exclusive. Then

$$1 = P(S) = P(A \cup \overline{A}) \overset{(2.1)}{=} P(A) + P(\overline{A}),$$

i.e. $P(\overline{A}) = 1 - P(A)$.

Since $P(\overline{A}) \geq 0$, it follows that $P(A) \leq 1$, so $0 \leq P(A) \leq 1$.

b) $P(\emptyset) = P(\overline{S}) = 1 - P(S) = 0$.

c) We have $A = (A \cap B) \cup (A \setminus B)$ and $A \cap B,\ A \setminus B$ are mutually exclusive. Thus

$$P(A) \overset{(2.1)}{=} P(A \cap B) + P(A \setminus B),$$

so $P(A \setminus B) = P(A) - P(A \cap B)$.

d) Since $A \subseteq B$, $A = A \cap B$. Then by c), we have

$$0 \leq P(B \setminus A) = P(B) - P(A),$$

which means $P(A) \leq P(B)$.

e) We have $A \cup B = A \cup (B \setminus (A \cap B))$ and $A, B \setminus (A \cap B)$ are mutually exclusive. Then using (3),

$$\begin{aligned} P(A \cup B) &\overset{(2.1)}{=} P(A) + P(B \setminus (A \cap B)) \\ &\overset{c)}{=} P(A) + P(B) - P(B \cap (A \cap B)) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

□

Part e) of Theorem 2.5 can be generalized to more than two events:

Theorem 2.6. *Let $(S, \mathcal{K}, P)$ be a probability space and $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{K}$ a sequence of events. Then Poincaré's formula (the inclusion-exclusion principle) holds*

$$\begin{aligned} P(\bigcup_{i=1}^{n} A_i) &= \sum_{i=1}^{n} P(A_i) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j) \\ &+ \sum_{1\leq i<j<k\leq n} P(A_i \cap A_j \cap A_k) \\ &+ \ \ldots \ + (-1)^{n-1} P(\bigcap_{i=1}^{n} A_i), \end{aligned} \tag{2.2}$$

for all $n \in \mathbb{N}$. As a consequence,

$$P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n), \tag{2.3}$$

i.e P is subadditive.

Example 2.7. Let us write formula (2.2) for three events $A, B, C \in \mathcal{K}$.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &- (P(A \cap B) + P(A \cap C) + P(B \cap C)) \\ &+ P(A \cap B \cap C). \end{aligned}$$

3 Classical Definition of Probability

Each event has an associated quantity which characterizes how likely its occurrence is; this is called the *probability* of the event. The classical definition of probability was given independently by B. Pascal and P. Fermat in the 17th century.

Definition 3.1. *Consider an experiment whose outcomes are finite and equally likely. Then the* ***probability of the occurrence of the event*** A *is given by*

$$P(A) = \frac{\text{number of favorable outcomes for the occurrence of } A}{\text{total number of possible outcomes of the experiment}} \overset{not}{=} \frac{N_f}{N_t}. \tag{3.1}$$

Remark 3.2. This approach can be used only when it is reasonable to assume that the possible outcomes of an experiment are equally likely (fair die, fair coin). Also, the two numbers have to be finite. When that is not the case, *geometrical probability* is used, when some continuous measure of a set is used (instead of the cardinality):

$$P(A) = \frac{\mu(A)}{\mu(S)}.$$

Remark 3.3. This notion is closely related to that of *relative frequency* of an event A: repeat an experiment a number of times N and count the number of times event A occurs, N_A. Then the relative frequency of the event A is

$$f_A = \frac{N_A}{N}.$$

Such a number is often used as an approximation to the probability of A. This is justified by the fact that

$$f_A \overset{N\to\infty}{\longrightarrow} P(A).$$

The relative frequency is used in computer simulations of random phenomena.

Example 3.4. Two dice are rolled. Find the probability of the events
A: a double appears;
B: the sum of the two numbers obtained is less than or equal to 5.

Solution We begin by computing the denominator in formula (3.1), because that number is common to both probabilities. The total number of possible outcomes is the number of elements of the sample space. The sample space is

$$S = \{e_{ij} \mid i, j = \overline{1,6}\},$$

where e_{ij} (identified by the pair (i, j), for simplicity) represents the event that number i showed on the first die and number j on the second. Hence $N_t = 36$.

For event A, $N_f = 6$ (there are six doubles out of 36 possible outcomes), so

$$P(A) = \frac{1}{6}.$$

For event B, we count the number of favorable outcomes (i.e. the number of pairs (i, j) for which $i + j \leq 5$). We have

$$\left.\begin{array}{llll} (1,1) & (1,2) & (1,3) & (1,4) \\ & (2,2) & (2,3) & \end{array}\right\} \ 6 \text{ outcomes}$$

By symmetry, we have $6 \times 2 = 12$, but two of the pairs were already symmetric, so

$N_f = 12 - 2 = 10$ cases. Thus

$$P(B) = \frac{5}{18}.$$

■

4 Conditional Probability and Independent Events

Many times, we have to compute the probability of an event that depends on another event to some extent, so the probability of that other event has to be considered, too.

Definition 4.1. *Let $(S, \mathcal{K}, P)$ be a probability space and let $B \in \mathcal{K}$ be an event with $P(B) > 0$. Then for every $A \in \mathcal{K}$, the* ***conditional probability of*** A ***given*** B *(or the* ***probability of*** A ***conditioned by*** B*) is defined by*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4.1}$$

Example 4.2. Ninety percent of flights depart on time. Eighty percent of flights arrive on time. Seventy-five percent of flights depart and arrive on time.
a) You are meeting a flight that departed on time. What is the probability that it will arrive on time?
b) You have met a flight, and it arrived on time. What is the probability that it departed on time?

Solution. Denote the events
A: a flight arrives on time ,
D: a flight departs on time.
Then we have:

$$P(A) = 0.8,\ P(D) = 0.9,\ P(A \cap D) = 0.75.$$

So,
a) $P(A|D) = \dfrac{P(A \cap D)}{P(D)} = \dfrac{0.75}{0.9} = 0.8333.$

b) $P(D|A) = \dfrac{P(A \cap D)}{P(A)} = \dfrac{0.75}{0.8} = 0.9375.$

■

An immediate consequence of Definition 4.1 is the following property:

Proposition 4.3. *Let $A, B \in \mathcal{K}$ with $P(A)P(B) \neq 0$. Then*

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B). \tag{4.2}$$

This rule can be generalized to any number of events.

Proposition 4.4. (The Multiplication Rule)
Let $\{A_i\}_{i=\overline{1,n}} \subseteq \mathcal{K}$, with $P(A_1 \cap A_2 \cap \ldots \cap A_n) \neq 0$. Then

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1 \cap \ldots \cap A_{n-1}). \tag{4.3}$$

Proof. We start with the right hand side (RHS) of (4.3) and get to the left hand side (LHS). By (4.1), we have

$$RHS \;=\; P(A_1)\cdot\frac{P(A_1\cap A_2)}{P(A_1)}\cdot\frac{P(A_1\cap A_2\cap A_3)}{P(A_1\cap A_2)}\cdot\ldots\cdot\frac{P\left(A_1\cap A_2\ldots\cap A_n\right)}{P\left(A_1\cap A_2\ldots\cap A_{n-1}\right)},$$

which, after cancellations, is $P(A_1\cap\ldots\cap A_n)$, the LHS of (4.3).

□

Proposition 4.5. *For every $A, B\in\mathcal{K}$ with $0<P(A)<1$, we have*

$$P(B) \;=\; P(A)P(B|A)+P\left(\overline{A}\right)P\left(B|\overline{A}\right). \tag{4.4}$$

Proof. Since $\{A,\overline{A}\}$ form a partition of S, we have

$$B=B\cap S=B\cap\left(A\cup\overline{A}\right)=\underbrace{(B\cap A)\cup\left(B\cap\overline{A}\right)}_{\text{m.e.}}.$$

Note that $B\cap A$ and $B\cap\overline{A}$ are mutually exclusive, since A and $\overline{A}$ are. Then

$$P(B)=P\left(B\cap A\right)+P\left(B\cap\overline{A}\right).$$

Using (4.2) for both terms on the right hand side, we obtain (4.4).

□

This result can also be generalized, for any partition of S.

Proposition 4.6. (The Total Probability Rule)
Let $\{A_i\}_{i\in I}$ be a partition of S and let $A\in\mathcal{K}$. Then

$$P(A) \;=\; \sum_{i\in I}P\left(A_i\right)P\left(A|A_i\right). \tag{4.5}$$

Proof. Just as before, we have

$$A=A\cap S=A\cap\left(\bigcup_{i\in I}A_i\right)=\bigcup_{i\in I}(\underbrace{A\cap A_i}_{\text{m.e.}}),$$

with $\{(A \cap A_i)\}_{i \in I}$ mutually exclusive and then

$$P(A) = \sum_{i \in I} P\left(A \cap A_i\right) \overset{(4.2)}{=} \sum_{i \in I} P(A_i) P(A|A_i).$$

□

Example 4.7. A test for a certain viral infection is 95% reliable for infected patients (i.e. it gives a correct positive result) and 99% reliable for the healthy ones (i.e. gives a correct negative result). It is known that 4% of the population is infected with that virus.
a) How reliable is the test in general (i.e. what is the probability that it shows a correct result)?
b) If a patient got a positive result, how likely is it that she truly is infected?

Solution. Denote the events
C: the test gives a correct result,
PR: the test gives a positive result,
V: a person has the virus (is infected).
What is given:

$$P(C|V) = P(PR|V) = 0.95,\ P(C|\overline{V}) = P(\overline{PR}|\overline{V}) = 0.99 \text{ and } P(V) = 0.04.$$

a) What we want is $P(C)$ (without any condition).
Notice that $\{V, \overline{V}\}$ form a partition of the sample space. By the Total Probability Rule (4.5), we have

$$\begin{aligned} P(C) &= P(C|V)P(V) + P(C|\overline{V})P(\overline{V}) \\ &= 0.95 \times 0.04 + 0.99 \times 0.96 = 0.9884. \end{aligned}$$

b) Here, we want $P(V|PR)$, which is given by

$$P(V|PR) = \frac{P(V \cap PR)}{P(PR)}.$$

The numerator is

$$P(V \cap PR) \overset{(4.2)}{=} P(V)P(PR|V) = 0.04 \times 0.95 = 0.038.$$

For the denominator, we use (4.5) again, with the same partition $\{V, \overline{V}\}$:

$$\begin{aligned} P(PR) &= P(PR|V)P(V) + P(PR|\overline{V})P(\overline{V}) \\ &= P(PR|V)P(V) + \Big[1 - P(\overline{PR}|\overline{V})\Big]P(\overline{V}) \\ &= 0.95 \times 0.04 + 0.01 \times 0.96 \;=\; 0.0476. \end{aligned}$$

Thus, the probability that the patient is indeed infected, is

$$P(V|PR) = \frac{0.038}{0.0476} = 0.7983.$$

■

Closely related to conditional probability is the notion of *independence* of events.

Definition 4.8. *Two events $A, B \in \mathcal{K}$ are said to be **independent** if*

$$P(A \cap B) = P(A)P(B). \tag{4.6}$$

*The events $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{K}$ are said to be **(mutually) independent** if*

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \ldots P(A_{i_k}),$$

for any finite subset $\{i_1, \ldots, i_k\} \subset \mathbb{N}$.

Remark 4.9. If the events $A, B \in \mathcal{K}$ are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$. The converse is also true.

Example 4.10. Refer again to Example 4.2. Are the events, departing on time and arriving on time, independent?

Solution. No, because

$$0.75 = P(A \cap D) \neq P(A)P(D) = 0.8 \times 0.9 = 0.72.$$

Also notice that

$$P(A|D) = 0.8333 \neq 0.8 = P(A) \;\text{ and }\; P(D|A) = 0.9375 \neq 0.9 = P(D).$$

Further, we see that $P(A|D) > P(A)$ and $P(D|A) > P(D)$. In other words, departing on time

increases the probability of arriving on time, and having arrived on time, it is more likely (probable) that the flight departed on time. ■

Proposition 4.11. *If $A = \emptyset$ (the impossible event, $P(A) = 0$) or $A = S$ (the certain event, P(A) = 1) and $B \in \mathcal{K}$ is any event, then A and B are independent.*

Proof. For the impossible event, we have

$$P(A \cap B) \;=\; P(\emptyset \cap B) \;=\; P(\emptyset) \;=\; 0 \;=\; P(A)P(B).$$

If A is the certain event, then

$$P(A \cap B) \;=\; P(S \cap B) \;=\; P(B) \;=\; P(A)P(B).$$

□

Proposition 4.12. *Let $A, B \in \mathcal{K}$ be independent events. Then A and $\overline{B}$ are also independent.*

Proof. We simply check the condition for independence:

$$\begin{aligned} P\left(A \cap \overline{B}\right) &= P(A \setminus B) \;=\; P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \;=\; P(A)(1 - P(B)) \\ &= P(A)P\left(\overline{B}\right). \end{aligned}$$

□

Remark 4.13.

1. A direct consequence of proposition 4.12 is that if $A, B \in \mathcal{K}$ are independent, then so are $\overline{A}, B$ and $\overline{A}, \overline{B}$.

2. More generally, if $A_1, A_2, ..., A_n \in \mathcal{K},\ n \in \mathbb{N}$ are independent, then so are $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$ and any combination of events and contrary events.

Chapter 2. Classical Probabilistic Models

In probability theory, one can notice that some experiments follow the same "patterns", so they are said to be in the same "class of experiments". Therefore, for each such class, we design a **probabilistic model**, which depends on certain parameters. For each model, we then find the corresponding general computational formulas, which then are applied to each experiment from that class, giving specific values to each parameter.

Sometimes, the easiest setup for describing a probabilistic model is to consider one (or more) box(es) containing a number (known or unknown) of balls, having a certain color distribution. The experiment consists of extracting one (or more) ball(s) from the box(es) (with or without putting it back) and noting its (their) color.

There is one important distinction that must be made! For an experiment, we can have

$-$ **sampling with replacement**, meaning that once an object (a ball) is selected (extracted), it is replaced (returned to the box), so it can be selected again,

or

$-$ **sampling without replacement**, which means that once an object is selected, it is NOT replaced, so it cannot be selected again.

If nothing else is specified, then the sampling is considered to be done *with* replacement.

1 Binomial Model

This model is used when the trials of an experiment satisfy three conditions, namely

(i) they are independent,

(ii) each trial has only two possible outcomes, which we refer to as "success" (A) and "failure" ($\overline{A}$) (i.e. the sample space for each trial is $S = A \cup \overline{A}$),

(iii) the probability of success $p = P(A)$ is the same for each trial (we denote by $q = 1-p = P(\overline{A})$ the probability of failure).

Trials of an experiment satisfying (i) $-$ (iii) are known as **Bernoulli trials**.

Model: Given n Bernoulli trials with probability of success p, find the probability $P(n; k)$ of exactly k ($0 \leq k \leq n$) successes occurring.

Remark 1.1. For the Binomial model, the parameters are n (number of trials) and p (probability of success). These are the numbers that describe the model. The number k is **not** a parameter of the model. It varies from 0 to n (all possible numbers of successes in n trials), depending on which probability we are interested in computing.

Proposition 1.2. *The probability $P(n;k)$ in a Binomial model is given by*

$$P(n;k) \;=\; C_n^k p^k (1-p)^{n-k} \;=\; C_n^k p^k q^{n-k}, \;\; k = 0, 1, \ldots, n. \tag{1.7}$$

Remark 1.3.
1. The probability $P(n;k)$ is the coefficient of x^k in the Binomial expansion

$$(px+q)^n \;=\; \sum_{k=0}^{n} C_n^k p^k q^{n-k} x^k \;=\; \sum_{k=0}^{n} P(n;k) x^k,$$

hence the name of this model.
2. As a consequence (let $x = 1$ above),

$$\sum_{k=0}^{n} P(n;k) \;=\; 1.$$

This also follows from the fact that the events $\{k \text{ successes occur}\}_{k=0}^{n}$ form a partition of S.

Example 1.4. A die is rolled 5 times. Find the probability of the events
a) A : getting three 6's,
b) B : getting at least two even numbers.

Solution. Here a trial is a roll of the die. Therefore, $n = 5$.
a) For the first part, "success" means rolling a 6. Hence, $p = \dfrac{1}{6}$. This is a Binomial model with parameters $n = 5, p = 1/6$ and we have

$$P(A) = P(5;3) = C_5^3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^2 \approx 0.0322.$$

b) For part two, "success" means getting an even number, so $p = \dfrac{1}{2}$. This a Binomial model with $n = 5$ and $p = 1/2$. To obtain at least 2 successes (out of 5 trials), means to obtain 2, 3, 4 or 5 successes. These events are mutually exclusive (only one at a time can happen), thus,

$$P(B) = P(5;2) + P(5;3) + P(5;4) + P(5;5).$$

However, in this case it is easier to compute the probability of the contrary event, which is "at most 1 success", since there are fewer cases (0 or 1). Thus,

$$\begin{aligned} P(B) &= 1 - P(\overline{B}) = 1 - \Big(P(5;0) + P(5;1)\Big) \\ &= 1 - \left(C_5^0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 + C_5^1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4\right) \approx 0.8125. \end{aligned}$$

■

2 Hypergeometric Model

This is the version of the Binomial model, *without* replacement. That will make a great difference, not only in the computational formulas, but in the parameters of the model.

Model: There are N ($N \in \mathbb{N}$) objects, n_1 ($n_1 \leq N$) of which have a certain trait (we could call that "success"). A number of n ($n \leq N$) objects are selected, one at a time, **without** replacement. Find the probability $P(n;k)$ of exactly k ($0 \leq k \leq n$) of the n objects selected, having that trait (i.e. k successes).

Remark 2.1. The parameters in a Hypergeometric model are N (total number of objects), n_1 (number of objects with a certain property) and n (number of trials). Again, k is not a parameter of the model.

Proposition 2.2. *The probability $P(n;k)$ in a Hypergeometric model is given by*

$$P(n;k) = \frac{C_{n_1}^k C_{N-n_1}^{n-k}}{C_N^n}, \quad k = 0, 1, \ldots, n. \tag{2.1}$$

Remark 2.3.

1. Intuitively, the probability $P(n;k)$ in (2.1) can be computed using the classical definition of probability. The total number of possible outcomes for the experiment is C_N^n. There are $C_{n_1}^k$ ways of choosing the k objects from the first category and $C_{N-n_1}^{n-k}$ ways of choosing the remaining $n-k$ objects from the rest (without replacement), and the two actions are independent of each other, so the number of favorable outcomes is $C_{n_1}^k C_{N-n_1}^{n-k}$.

2. As before,

$$\sum_{k=0}^{n} P(n;k) = 1, \text{ i.e. } \sum_{k=0}^{n} C_{n_1}^k C_{N-n_1}^{n-k} = C_N^n.$$

Example 2.4. There are 15 boys and 20 girls in a probability class. Ten people are selected for a certain project. Find the probability that the group contains

a) an equal number of boys and girls (event A),

b) at least one girl (event B).

Solution.

This is a Hypergeometric model with $N = 35, n_1 = 15$ (or $n_1 = 20$) and $n = 10$.

a) For event A, an equal number of boys and girls out of 10 people, means 5 boys and 5 girls.

Therefore,

$$P(A) = P(10;5) = \frac{C_{15}^5 C_{20}^5}{C_{35}^{10}} \approx 0.2536\,.$$

b) For event B, it is easier to compute the probability of the complementary event, which would be "no girls at all", or "10 boys". Thus,

$$P(B) = 1 - P\left(\overline{B}\right) = 1 - P(10;10) = 1 - \frac{C_{15}^{10} C_{20}^0}{C_{35}^{10}} = 1 - \frac{C_{15}^{10}}{C_{35}^{10}} \approx 0.9999.$$

■

3 Poisson Model

This model is a generalization of the Binomial model, in the sense that it allows the probability of success to vary at each trial. Everything else is the same. So, instead of one probability of success p, we will have probabilities of success $p_1, p_2, \ldots, p_n$, one for each of the n trials.

Model: Consider an experiment where in each trial there are two possible outcomes, "success", A, and "failure", $\overline{A}$. The probability of success in the ith trial is p_i (and, accordingly, the probability of failure is $q_i = 1 - p_i$). Find the probability $P(n;k)$ that in n independent such trials, exactly k ($0 \leq k \leq n$) successes occur.

The parameters of a Poisson model are n and $p_1, p_2, \ldots, p_n$.

Proposition 3.1. *The probability $P(n;k)$ in a Poisson model is given by*

$$P(n;k) \;=\; \sum_{1 \leq i_1 < \ldots < i_k \leq n} p_{i_1} \ldots p_{i_k} q_{i_{k+1}} \ldots q_{i_n}, \quad k = 0, 1, \ldots, n, \tag{3.1}$$

where $i_{k+1}, \ldots, i_n \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$.

Remark 3.2.

1. The number $P(n;k)$ is the coefficient of x^k in the polynomial expansion

$$(p_1 x + q_1) \ldots (p_n x + q_n) \;=\; \sum_{k=0}^{n} P(n;k) x^k$$

and, for the Poisson model, *this* is the computational formula that we will use.

2. Again, as a consequence (let $x = 1$ above),

$$\sum_{k=0}^{n} P(n;k) \;\; = \;\; 1.$$

3. If $p_i = p$ (and consequently, $q_i = q$), $\forall i = \overline{1,n}$, then this becomes the Binomial model and (3.1) is reduced to (1.7) in Lecture 2.

Example 3.3. (The Three Shooters Problem) Three shooters aim at a target and they hit it (independently of each other) with probabilities 0.4, 0.5 and 0.7, respectively. Each of them shoots once. Find the probability p that the target is hit once.

Solution. Define "success" as "the target is hit". Then we have a Poisson model with $n = 3$ independent trials and $p_1 = 0.4$, $p_2 = 0.5$, $p_3 = 0.7$. We want the probability of 1 success occurring. Hence $p = P(3;1)$ and by Remark 3.2 above, it is equal to the coefficient of x in the polynomial

$$(0.4x + 0.6)(0.5x + 0.5)(0.7x + 0.3) = 0.14x^3 + 0.41x^2 + 0.36x + 0.09,$$

i.e. $p = 0.36$. ■

4 Pascal (Negative Binomial) Model

This model is a little different from the previous ones, in the sense that, we are not only interested in number of successes and failures, but also in the order that they occur, in the rank of a success. Another novelty is that in this model we have (theoretically) an infinite number of trials.

Model: Consider an infinite sequence of Bernoulli trials with probability of success p (and probability of failure $q = 1 - p$) in each trial. Find the probability $P(n,k)$ of the nth success occurring after k failures ($n \in \mathbb{N},\ k \in \mathbb{N} \cup \{0\}$).

Remark 4.1. For the Pascal model, again the parameters are n (rank of the success we want) and p (probability of success), but n has a *very different* meaning than the one in the Binomial model. Again k is not a parameter of the model, it varies from 0 to ∞.

Proposition 4.2. *The probability $P(n,k)$ in a Negative Binomial model is given by*

$$P(n,k) \;\; = \;\; C^k_{n+k-1} p^n q^k, \;\; k = 0, 1, \ldots. \tag{4.1}$$

Remark 4.3.

1. The probability $P(n;k)$ is the coefficient of x^k in the expansion

$$\left(\frac{p}{1-qx}\right)^n = \sum_{k=0}^{\infty} P(n,k)x^k,\ |qx|<1,$$

hence the name.

2. As before,

$$\sum_{k=0}^{\infty} P(n,k) = 1.$$

5 Geometric Model

Although a particular case for the Pascal Model (case $n=1$), the Geometric model comes up in many applications and deserves a place of its own.

Model: Consider an infinite sequence of Bernoulli trials with probability of success p (and probability of failure $q=1-p$) in each trial. Find the probability p_k that the first success occurs after k failures ($k \in \mathbb{N} \cup \{0\}$).

There is only one parameter for this model, p.

Proposition 5.1. *The probability p_k in a Geometric model is given by*

$$p_k = pq^k,\ \ k=0,1,\ldots. \tag{5.1}$$

Remark 5.2.

1. The number p_k is the coefficient of x^k in the Geometric expansion (series)

$$\frac{p}{1-qx} = \sum_{k=0}^{\infty} p_k x^k,\ |qx|<1,$$

hence the name.

2. Again,

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} pq^k = 1$$

(the Geometric series).

Example 5.3. When a die is rolled, find the probability of the following events:
a) A: the first 6 appears after 5 throws;
b) B: the 3^{rd} even appears after 5 throws.

Solution.
a) For event A, success means that face 6 appears, hence $p = 1/6$. We want the first success to occur after 5 failures, so this is a Geometric model. By (5.1), we have

$$P(A) = p_5 = \frac{1}{6}\left(\frac{5}{6}\right)^5 \approx 0.067.$$

b) For event B, success means that an even number shows, so $p = 1/2$. This fits the Pascal model with $n = 3$ and $p = 1/2$. The 3^{rd} even appears after 5 throws (on the 6^{th} throw), which means after 3 odds, i.e. after 3 failures. Thus, using (4.1), we have

$$P(B) = P(3,3) = C_5^3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^3 \approx 0.1562.$$

■

Chapter 3. Random Variables and Random Vectors

In order to do a more rigorous study of random phenomena, we need to give them a more general quantitative description. That materializes in *random variables*, variables whose observed values are determined by chance. Random variables are the fundamentals of modern Statistics. They fall into one of two categories: *discrete* or *continuous*.

1 Discrete Random Variables and Probability Distribution Function

Let $(S, \mathcal{K}, P)$ be a probability space.

Definition 1.1. *A **random variable** is a function $X : S \to \mathbb{R}$ satisfying the property that for every*

$x \in \mathbb{R}$, *the event*

$$(X \leq x) \quad := \quad \{e \in S \mid X(e) \leq x\} \in \mathcal{K}. \tag{1.2}$$

Definition 1.2. *A random variable $X : S \to \mathbb{R}$ is a **discrete random variable** if the set of values that it takes, $X(S)$, is at most countable in $\mathbb{R}$.*

Example 1.3. Consider the experiment of rolling a die. Then the sample space is $S = \{e_1, \ldots, e_6\}$, where e_i represents the event that face i shows on the die, $i = \overline{1,6}$. Let $\mathcal{K} = \mathcal{P}(S)$ (all subsets of S) and P be given by classical probability. Define $X : S \to \mathbb{R}$ by

$$X(e_i) \quad = \quad i, \ i = 1, \ldots, 6.$$

Let us check that this is a discrete random variable.
For any $x \in \mathbb{R}$, the event (set) $(X \leq x) \subseteq S$, so it obviously belongs to $\mathcal{K}$. Thus X is a well-defined random variable (it satisfies (1.2)).
Since the set of values that it takes $X(S) = \{1, \ldots, 6\}$ is finite, X is also a discrete random variable.

Example 1.4. (The **indicator** of an event) Consider a probability space $(S, \mathcal{K}, P)$ over the sample space S of some experiment. For any event $A \in \mathcal{K}$, define $X_A : S \to \mathbb{R}$ by

$$X_A(e) \quad = \quad \begin{cases} 0, & e \notin A \ \ (e \in \overline{A}) \\ 1, & e \in A \end{cases} \tag{1.3}$$

First off, $X_A(S) = \{0, 1\}$, which is obviously countable. Let us check condition (1.2).
Let $x < 0$. Since all the values that X_A takes are nonnegative, there is no way that $X_A(e)$ could be $\leq x$, i.e.

$$(X_A \leq x) = \{e \in S \mid X_A(e) \leq x\} = \emptyset \in \mathcal{K},$$

since any σ-field contains the impossible event (empty set).
If $0 \leq x < 1$, the event from (1.2) is

$$\begin{aligned} (X_A \leq x) &= \{e \in S \mid X_A(e) \leq x\} \\ &= \{e \in S \mid X_A(e) = 0\} \\ &= \overline{A} \in \mathcal{K}, \end{aligned}$$

because $A \in \mathcal{K}$.

Finally for $x \geq 1$,

$$(X_A \leq x) = \{e \in S \mid X_A(e) \leq x\} = A \cup \overline{A} = S \in \mathcal{K},$$

again, by the properties of a σ-field.

Remark 1.5. A discrete random variable that takes only a finite set of values is called a **simple discrete random variable**. All of the examples above are simple discrete random variables.

The previous example can easily be generalized to any countable partition of S.

Example 1.6. Let I be a countable set of indexes, $\{A_i\}_{i\in I} \subseteq \mathcal{K}$ a partition of S and $\{x_i\}_{i\in I} \subseteq \mathbb{R}$ a sequence of distinct real numbers. Define $X : S \to R$ by

$$X(e) \;=\; \sum_{i\in I} x_i X_{A_i}(e), \tag{1.4}$$

where X_{A_i} is the indicator of $A_i, i \in I$. Then X is a discrete random variable satisfying

$$X(e) = x_i \;\; <=> \;\; e \in A_i, \tag{1.5}$$

for all $i \in I$.

This is more than just a general example, relation (1.4) gives the general expression of a discrete random variable. *Any* discrete random variable can be put in the form (1.4). Having the set of values that X takes, $\{x_i\}_{i\in I}$, X can be written as in (1.4), with $A_i = (X = x_i)$. This justifies the next definition. Instead of defining a discrete random variable as a function $X : S \to \mathbb{R}$, we emphasize directly the values $\{x_i\}_{i\in I}$ that it takes and the probabilities of taking each value, $p_i = P(A_i) = P(X = x_i)$.

Definition 1.7. *Let $X : S \to \mathbb{R}$ be a discrete random variable. The **probability distribution function (pdf)** of X is an array of the form*

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}, \tag{1.6}$$

where $x_i \in \mathbb{R},\ i \in I$, are the values that X takes and $p_i = P(X = x_i)$ are the probabilities that X takes each value x_i.

Remark 1.8.

1. All values $x_i, i \in I$, in (1.6) are distinct. If some are equal, they only appear once, with the added corresponding probability.

2. All probabilities $p_i \neq 0, i \in I$. If for some $i \in I,\ p_i = 0$, then the corresponding value x_i is not included in the pdf (1.6).

3. If X is a discrete random variable with pdf (1.6), then

$$\sum_{i \in I} p_i = 1$$

(a necessary and sufficient condition for such an array to represent a pdf of a discrete random variable). Indeed, since the events $\{(X = x_i)\}_{i \in I}$ form a partition of S, we have

$$\sum_{i \in I} p_i = \sum_{i \in I} P(X = x_i) = P(S) = 1.$$

4. Henceforth, we will identify a discrete random variable with its pdf and use (1.6) to describe it.

Example 1.9. The pdf of the random variable in Example 1.3 (rolling a die) is

$$X \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Example 1.10. The pdf of the random variable in Example 1.4 (the indicator of an event) is

$$X_A \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},\ p = P(A).$$

2 Cumulative Distribution Function

Definition 2.1. *Let X be a random variable (of any type, discrete or continuous). The function $F = F_X : \mathbb{R} \to \mathbb{R}$, defined by*

$$F_X(x) \;=\; P(X \leq x), \tag{2.1}$$

*is called the **(cumulative) distribution function (cdf)** of X.*

Example 2.2. Let us go back to Example (1.4) (or (1.10)) in Lecture 3 (the indicator of an event). Its pdf is

$$X_A \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},\ p = P(A).$$

From the analysis we did, let us recall:
For $x < 0$,

$$P(X_A \leq x) \;=\; P(\emptyset) \;=\; 0.$$

If $0 \leq x < 1$,

$$P(X_A \leq x) \;=\; P(X_A = 0) \;=\; 1 - p.$$

Finally for $x \geq 1$,

$$P(X_A \leq x) \;=\; P(\{X_A = 0\} \cup \{X_A = 1\}) \;=\; 1 - p + p \;=\; 1.$$

So, we find now the cdf of X_A to be

$$F_A(x) \;=\; \begin{cases} 0, & \text{if} \quad x < 0 \\ 1-p, & \text{if} \quad 0 \leq x < 1 \\ 1, & \text{if} \quad x \geq 1. \end{cases}$$

The graphic representation of F_A is given in Figure 1.

Remark 2.3. It easily follows from the previous example that for a discrete random variable with pdf

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I},$$

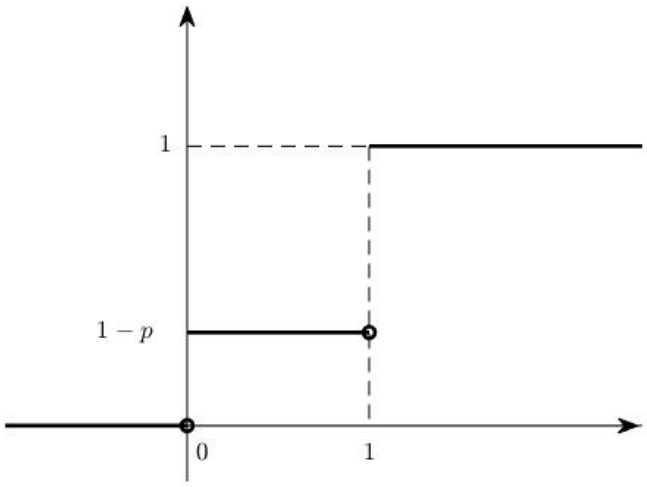


Fig. 1: Cumulative distribution function for the indicator random variable

the cdf is computed by

$$F(x) \;=\; \sum_{x_i \le x} p_i \tag{2.2}$$

and for every $A \subseteq \mathbb{R}$,

$$P(X \in A) \;=\; \sum_{x_i \in A} p_i. \tag{2.3}$$

Theorem 2.4. *Let X be a random variable with cdf $F : \mathbb{R} \to \mathbb{R}$. Then F has the following properties:*

a) *If $a < b$ are real numbers, then $P(a < X \le b) = F(b) - F(a)$.*

b) *F is monotonely increasing, i.e. if $a < b$, then $F(a) \le F(b)$.*

c) *F is right continuous, i.e. $F(x + 0) = F(x)$, for every $x \in \mathbb{R}$, where $F(x + 0) = \lim_{y \searrow x} F(y)$ is the limit from the right at x.*

d) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

e) *For every $x \in \mathbb{R}$, $P(X < x) = F(x - 0) = \lim_{y \nearrow x} F(y)$ and $P(X = x) = F(x) - F(x - 0)$.*

Proof. (Selected)
a) If $a < b$, then $X \leq a$ implies $X \leq b$, so, as events,

$$(X \leq a) \subseteq (X \leq b) \text{ and } (X \leq a) \cap (X \leq b) = (X \leq a).$$

Then, (by Theorem 2.5 c) in Chapter 1 (Lecture 1) and the fact that $A \cap \overline{B} = A \setminus B$),

$$\begin{aligned} P(a < X \leq b) &= P\Big((X \leq b) \cap (\overline{X \leq a})\Big) = P\Big((X \leq b) \setminus (X \leq a)\Big) \\ &= P(X \leq b) - P(X \leq a) = F(b) - F(a). \end{aligned}$$

b) If $a < b$, then $F(b) - F(a) = P(a < X \leq b) \geq 0$, since it is a probability.
d) We have

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} P(X \leq x) = P(\emptyset) = 0.$$

and

$$\lim_{x \to \infty} F(x) = \lim_{x \to \infty} P(X \leq x) = P(S) = 1.$$

e) Just the second part:

$$P(X = x) = P\Big((X \leq x) \setminus (X < x)\Big) = P(X \leq x) - P(X < x) = F(x) - F(x - 0).$$

□

3 Common Discrete Distributions

Bernoulli Distribution $Bern(p)$

A random variable X has a Bernoulli distribution with parameter $p \in (0, 1)$, if its pdf is

$$X \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}. \tag{3.1}$$

Notice that this is the pdf of the indicator random variable from Example 1.10 in Chapter 1 (Lecture 3). A Bernoulli r.v. models the occurrence or nonoccurrence of an event.

Discrete Uniform Distribution $U(m)$

A random variable X has a Discrete Uniform distribution (unid) with parameter $m \in \mathbb{N}$, if its pdf is

$$X \begin{pmatrix} k \\ \frac{1}{m} \end{pmatrix}_{k=\overline{1,m}}. \tag{3.2}$$

The random variable in Example 1.3 (and 1.9) (Lecture 3), the number shown on a die, has a Discrete Uniform distribution $U(6)$.

Binomial Distribution $B(n,p)$

A random variable X has a Binomial distribution (bino) with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ ($q = 1-p$), if its pdf is

$$X \begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}_{k=\overline{0,n}}. \tag{3.3}$$

This distribution corresponds to the Binomial model. Given n Bernoulli trials with probability of success p, let X denote the number of successes. Then $X \in B(n,p)$. Also, notice that the Bernoulli distribution is a particular case of the Binomial one, for $n = 1$, $Bern(p) = B(1,p)$.

Hypergeometric Distribution $H(N, n_1, n)$

A random variable X has a Hypergeometric distribution (hyge) with parameters $N, n_1, n \in \mathbb{N}$ ($n, n_1 \leq N$), if its pdf is

$$X \begin{pmatrix} k \\ \frac{C_{n_1}^k C_{N-n_1}^{n-k}}{C_N^n} \end{pmatrix}_{k=\overline{0,n}}. \tag{3.4}$$

This distribution corresponds to the Hypergeometric model. If X is the number of successes in a Hpergeometric model, then $X \in H(N, n_1, n)$.

Negative Binomial (Pascal) Distribution $NB(n,p)$

A random variable X has a Negative Binomial (Pascal) (nbin) distribution with parameters $n \in \mathbb{N}$ and $p \in (0,1)$, if its pdf is

$$X \begin{pmatrix} k \\ C_{n+k-1}^k p^n q^k \end{pmatrix}_{k=0,1,\ldots}. \tag{3.5}$$

This distribution corresponds to the Negative Binomial model. If X denotes the number of failures that occurred before the occurrence of the n^{th} success in a Negative Binomial model, then $X \in NB(n,p)$.

Geometric Distribution $Geo(p)$

As before (probabilistic models), we have an important special case for the Negative Binomial distribution; if $n = 1$ in the previous distribution, then we have a *Geometric distribution*. A random variable X has a Geometric distribution (geo) with parameter $p \in (0,1)$, if its pdf is given by

$$X \begin{pmatrix} k \\ pq^k \end{pmatrix}_{k=0,1,\ldots}. \tag{3.6}$$

If X denotes the number of failures that occurred before the occurrence of the 1^{st} success in a Geometric model, then $X \in Geo(p)$. Also, $Geo(p) = NB(1,p)$.

Poisson Distribution $\mathcal{P}(\lambda)$

A random variable X has a Poisson distribution (poiss) with parameter $\lambda > 0$, if its pdf is

$$X \begin{pmatrix} k \\ \dfrac{\lambda^k}{k!}e^{-\lambda} \end{pmatrix}_{k=0,1,\ldots} \tag{3.7}$$

A Poisson r.v. **does not** come from the Poisson model! Poisson random variables arise in connection with so-called Poisson *processes,* processes that involve observing discrete events in a continuous interval of time, length, space, etc. The variable of interest in a Poisson process, X, represents the number of occurrences of the discrete event in a fixed interval of time, length, space. For instance, the number of gas emissions taking place at a nuclear plant in a 3-month period, the number of earthquakes hitting a certain area in a year, the number of white blood cells in a drop of blood, all these are modeled by Poisson random variables. The parameter λ of a Poisson distribution represents the *average* number of occurrences of the event in that interval of time or other continuous medium (this will be discussed in more detail in the next chapter).

Poisson's distribution is also known as the "law of rare events", the name coming from the fact that

$$\lim_{k\to\infty} \frac{\lambda^k}{k!}e^{-\lambda} = 0,$$

i.e. as k gets larger, the event $(X = k)$ becomes less probable, more "rare". The discrete events that are counted in a Poisson process are also called "rare events".

4 Discrete Random Vectors; Joint Probability Distribution Function; Operations with Discrete Random Variables and Independent Discrete Random Variables

We will restrict our discussion to a two-dimensional discrete random vector $(X, Y) : S \to \mathbb{R}^2$.

Definition 4.1. *Let $(S, \mathcal{K}, P)$ be a probability space. A **discrete random vector** is a function $(X, Y) : S \to \mathbb{R}^2$ satisfying the following two conditions:*

(i) *for all $(x, y) \in \mathbb{R}^2$,*

$$(X \leq x, Y \leq y) = \{e \in S \mid X(e) \leq x, Y(e) \leq y\} \in \mathcal{K}$$

(ii) *the set of values that it takes $(X, Y)(S)$ is at most countable in $\mathbb{R}^2$;*

Definition 4.2. *Let $(X, Y) : S \to \mathbb{R}^2$ be a two-dimensional discrete random vector. The **joint probability distribution (function)** of (X, Y) is a two-dimensional array of the form*

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	
x_1					
$\vdots$			$\vdots$		
x_i		$\cdots$	p_{ij}	$\cdots$	p_i
$\vdots$			$\vdots$		
			q_j		

(4.1)

where $(x_i, y_j) \in \mathbb{R}^2$, $(i, j) \in I \times J$ are the values that (X, Y) takes and $p_{ij} = P(X = x_i, Y = y_j)$ is the probability that (X, Y) takes the value (x_i, y_j).

Proposition 4.3. *Let (X, Y) be a random vector with joint probability distribution given by (4.1). Then*

$$\sum_{j \in J} p_{ij} = p_i \text{ and } \sum_{i \in I} p_{ij} = q_j,$$

where $p_i = P(X = x_i),\ i \in I$ *and* $q_j = P(Y = y_j),\ j \in J$. *The probabilities* p_i *and* q_j *are called* ***marginal*** *pdf's.*

Let $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$ and $Y \begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}$ be two discrete random variables and let $\alpha \in \mathbb{R}$. As before, denote by $p_{ij} = P(X = x_i, Y = y_j)$. We can define the following operations:

Sum. The sum of X and Y is the random variable with pdf given by

$$X + Y \begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{4.2}$$

Product. The product of X and Y is the random variable with pdf given by

$$X \cdot Y \begin{pmatrix} x_i y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{4.3}$$

Scalar Multiple. The random variable αX, $\alpha \in \mathbb{R}$, with pdf given by

$$\alpha X \begin{pmatrix} \alpha x_i \\ p_i \end{pmatrix}_{i\in I}. \tag{4.4}$$

Quotient. The quotient of X and Y is the random variable with pdf given by

$$X/Y \begin{pmatrix} x_i/y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}, \tag{4.5}$$

provided that $y_j \neq 0$, for all $j \in J$.
In general, if $h : \mathbb{R} \to \mathbb{R}$ is a function, then we can define the random variable $h(X)$, with pdf given by

$$h(X) \begin{pmatrix} h(x_i) \\ p_i \end{pmatrix}_{i\in I}. \tag{4.6}$$

Definition 4.4. *Two discrete random variables* X *and* Y *with probability distribution functions*

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I} \quad \textit{and} \quad Y \begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}$$

*are said to be **independent** if*

$$p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(X = x_i\right) P\left(Y = y_j\right) = p_i q_j, \tag{4.7}$$

for all $(i, j) \in I \times J$.

Remark 4.5. If X and Y are independent discrete random variables, then in (4.2), (4.3) and (4.5), $p_{ij} = p_i q_j$, for all $(i, j) \in I \times J$.

Example 4.6. Let X be a random variable with pdf

$$X \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Find the pdf of $Y = 3X^2 - 1$.

Solution. Remember, we operate on the *values*, **never** on the probabilities!
If X takes the values $-1, 0$ and 1, then Y takes the values -1 (when $X = 0$) and 2 (when $X = -1$ or $X = 1$).

Now, we compute (carefully!) the probability for each value.

$$\begin{aligned} P(Y = -1) &= P(X = 0) \\ &= \frac{1}{4}, \\ P(Y = 2) &= P\Big((X = -1) \cup (X = 1)\Big) \\ &= P(X = -1) + P(X = 1) \\ &= \frac{1}{2} + \frac{1}{4} \;=\; \frac{3}{4}, \end{aligned}$$

since the events $(X = -1)$ and $(X = 1)$ are mutually exclusive.
Thus, the pdf of Y is

$$Y \begin{pmatrix} -1 & 2 \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

■

Example 4.7. Let X and Y be two independent random variables with pdf's

$$X \begin{pmatrix} -1 & 0 \\ 0.2 & 0.8 \end{pmatrix} \text{ and } Y \begin{pmatrix} 1 & 2 \\ 0.6 & 0.4 \end{pmatrix},$$

respectively. Find the pdf of $X+Y$.

Solution. First, let's find all the possible values of $X+Y$, by taking all the combinations of $x_i+y_j, i,j=1,2$. So, $X+Y$ can take the values $0,1$ and 2.
Then compute their corresponding probabilities:

$$\begin{aligned}
P(X+Y=0) &= P(X=-1,Y=1)\\
&\overset{\text{ind}}{=} P(X=-1)P(Y=1)\\
&= 0.2\cdot 0.6 \;=\; 0.12,\\
P(X+Y=1) &= P\Big((X=-1,Y=2)\cup(X=0,Y=1)\Big)\\
&\overset{\text{m.e.}}{=} P(X=-1,Y=2)+P(X=0,Y=1)\\
&\overset{\text{ind}}{=} P(X=-1)P(Y=2)+P(X=0)P(Y=1)\\
&= 0.2\cdot 0.4+0.8\cdot 0.6 \;=\; 0.56,\\
P(X+Y=2) &= P(X=0,Y=2)\\
&\overset{\text{ind}}{=} 0.8\cdot 0.4 \;=\; 0.32.
\end{aligned}$$

Alternatively, we could have computed the first and the third (which are easier) and found the second one by

$$P(X+Y=1)=1-\Big(P(X+Y=0)+P(X+Y=2)\Big)=1-0.44=0.56.$$

■

Remark 4.8.

1. The sum of n independent $Bern(p)$ random variables is a $B(n,p)$ variable.
2. The sum of n independent $Geo(p)$ random variables is a $NB(n,p)$ variable.

5 Continuous Random Variables and Probability Density Function

Recall the definition of a random variable (Definition 1.1 in Chapter 3, Lecture 4):

Let $(S, \mathcal{K}, P)$ be a probability space. A *random variable* is a function $X : S \to \mathbb{R}$ satisfying the property that for every $x \in \mathbb{R}$, the event

$$(X \leq x) \quad := \quad \{e \in S \mid X(e) \leq x\} \in \mathcal{K}.$$

Then, for every random variable X (not necessarily discrete), we defined the *cumulative distribution function* of X (Definition 2.1 in Chapter 3, Lecture 4): the function $F = F_X : \mathbb{R} \to \mathbb{R}$, defined by

$$F_X(x) \quad = \quad P(X \leq x),$$

Definition 5.1. *Let $(S, \mathcal{K}, P)$ be a probability space. A random variable $X : S \to \mathbb{R}$ is a **continuous random variable**, if the set of values $X(S)$ is any (finite or infinite) interval in $\mathbb{R}$.*

Proposition 5.2. *Let X be a continuous random variable with cdf $F : \mathbb{R} \to \mathbb{R}$. Then F is absolutely continuous, i.e. there exists a real function $f : \mathbb{R} \to \mathbb{R}$, such that*

$$F(x) \quad = \quad \int_{-\infty}^{x} f(t)\, dt, \tag{5.1}$$

for all $x \in \mathbb{R}$.

Definition 5.3. *Let X be a continuous random variable. Then the function f from Proposition 5.2 is called the **probability density function (pdf)** of X.*

Remark 5.4. So, we use "pdf" to describe any random variable, "probability *distribution* function" for discrete random variables and "probability *density* function" for the continuous case. The term "density" in the continuous case, extends in a natural way the notion of "distribution" from the discrete case, with summation being replaced by integration. Note that not all books or authors make that distinction (e.g. in Matlab, they are all called "densities").

Recall the properties of a cdf (Theorem 2.4., Chapter 3, Lecture 4).

Theorem (Properties of a cdf)**.** Let X be a random variable with cdf $F : \mathbb{R} \to \mathbb{R}$. Then F has the following properties:

a) If $a < b$ are real numbers, then $P(a < X \leq b) = F(b) - F(a)$.

b) F is monotonely increasing, i.e. if $a < b$, then $F(a) \leq F(b)$.

c) F is right continuous, i.e. $F(x + 0) = F(x)$, for every $x \in \mathbb{R}$, where $F(x + 0) = \lim_{y \searrow x} F(y)$ is the limit from the right at x.

d) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

e) For every $x \in \mathbb{R}$, $P(X < x) = F(x - 0) = \lim_{y \nearrow x} F(y)$ and $P(X = x) = F(x) - F(x - 0)$.

From them, some common properties of a density function can be derived.

Theorem 5.5. *Let X be a continuous random variable with cdf F and density function f. Then the following properties hold:*

a) $F'(x) = f(x)$, *for all* $x \in \mathbb{R}$.

b) $f(x) \geq 0$, *for all* $x \in \mathbb{R}$.

c) $\int_{\mathbb{R}} f(t)dt = 1$.

d) *For every* $x \in \mathbb{R}$, $P(X = x) = 0$ *and for every* $a, b \in \mathbb{R}$ *with* $a < b$,

$$\begin{aligned} P(a < X \leq b) &= P(a < X \leq b) = P(a < X < b) = P(a \leq X \leq b) \\ &= \int_a^b f(t)\, dt\,. \end{aligned} \tag{5.2}$$

Proof.
a) This property follows directly from the definition of a continuous random variable, by differentiating both sides of (5.1).
b) Recall from Theorem 2.4. that F is monotonely increasing. Thus, its derivative is nonnegative, for every $x \in \mathbb{R}$.
c) Recall that $\lim_{x \to \infty} F(x) = 1$ (Theorem 2.4.d)). So, we have

$$\int_{\mathbb{R}} f(t)\, dt = \int_{-\infty}^{\infty} f(t)\, dt = \lim_{x \to \infty} \int_{-\infty}^{x} f(t)\, dt = \lim_{x \to \infty} F(x) = 1.$$

(4) To prove the first part, let $x \in \mathbb{R}$ be fixed and recall from Theorem 2.4.e) that $P(X = x) = F(x) - F(x-0)$. But for a continuous random variable, F is absolutely continuous, so continuous at every point, thus $F(x) = F(x+0) = F(x-0)$. Hence, $P(X = x) = 0$.
Now let $a, b \in \mathbb{R}$ with $a < b$. By Theorem 2.4.a), we have

$$P(a < X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)\, dt \; - \int_{-\infty}^{a} f(t)\, dt = \int_{a}^{b} f(t)\, dt,$$

which, by the first part, is equal to all the other probabilities in (5.2).

□

6 Common Continuous Distributions

Uniform Distribution $U(a, b)$

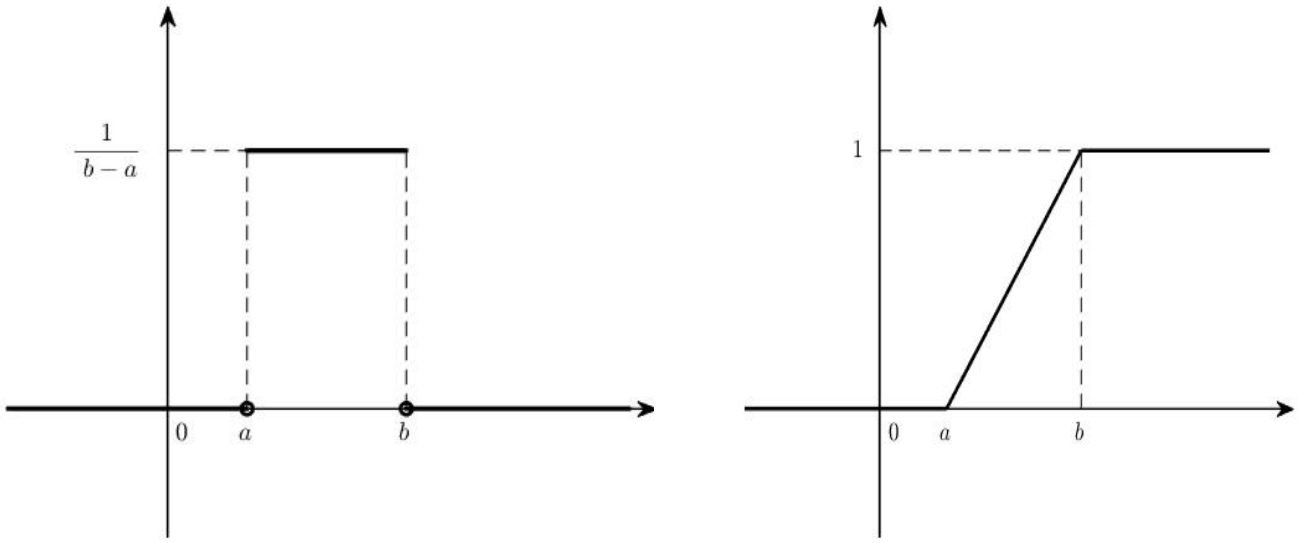
A random variable X has a Uniform distribution (unif) with parameters $a, b \in \mathbb{R},\ a < b$, if its pdf is

$$f(x) \;\; = \;\; \begin{cases} \dfrac{1}{b-a}, & \text{if } \; x \in [a, b] \\ 0, & \text{if } \; x \notin [a, b]. \end{cases} \tag{6.1}$$

Then, by (5.1), its cdf is

$$F(x) = \int_{-\infty}^{x} f(t)dt \;\; = \;\; \begin{cases} 0, & \text{if } \; x \leq a \\ \dfrac{x-a}{b-a}, & \text{if } \; a < x \leq b \\ 1, & \text{if } \; x \geq b\,. \end{cases} \tag{6.2}$$

Remark 6.1.
1. The Uniform distribution is used when a variable can take *any* value in a given interval, equally probable. For example, locations of syntax errors in a program, birthdays throughout a year, etc.
2. A special case is that of a **Standard Uniform Distribution**, where $a = 0$ and $b = 1$. The pdf and



(a) Density Function (pdf) (b) Cumulative Distribution Function (cdf)

Fig. 1: Uniform Distribution

cdf are given by

$$f_U(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}, \quad F_U(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x \leq 1 \\ 1, & x \geq 1\,. \end{cases} \tag{6.3}$$

Standard Uniform variables play an important role in stochastic modeling; in fact, *any* random variable, with any thinkable distribution (discrete or continuous) can be generated from Standard Uniform variables.

Normal Distribution $N(\mu, \sigma)$

The Normal distribution is, by far, the most important distribution, underlying many of the modern statistical methods used in data analysis. It was first described in the late 1700's by De Moivre, as a limiting case for the Binomial distribution (when n, the number of trials, becomes infinite), but did not get much attention. Half a century later, both Laplace and Gauss (independently of each other) rediscovered it in conjunction with the behavior of errors in astronomical measurements. It is also referred to as the "Gaussian" distribution.

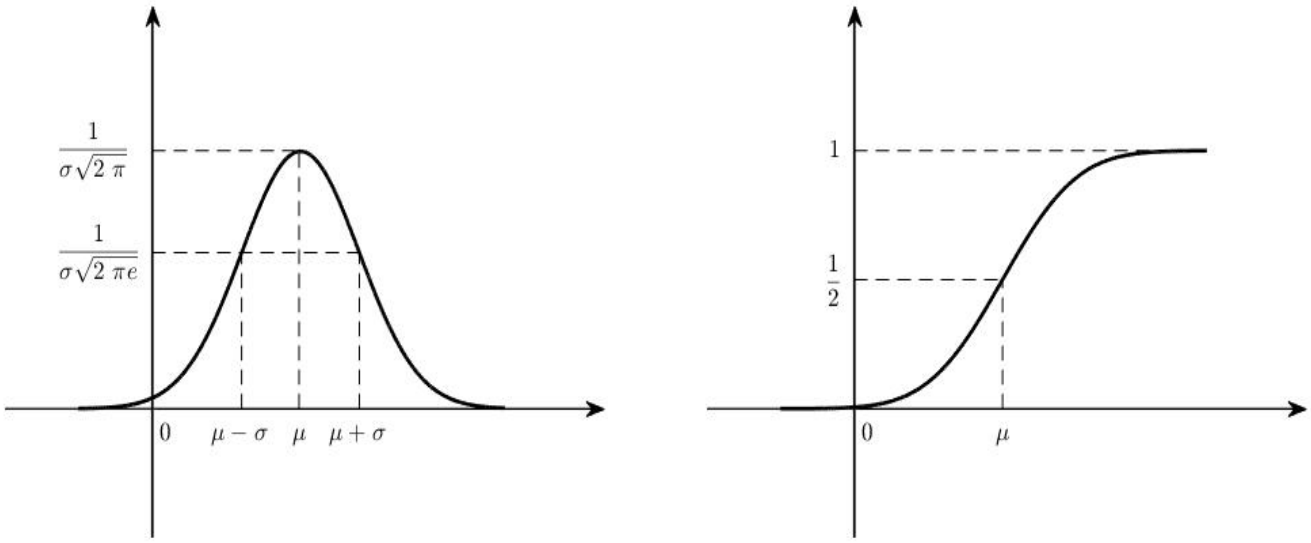
A random variable X has a Normal distribution (norm) with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if

its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{6.4}$$

The cdf of a Normal variable is then given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}} dt. \tag{6.5}$$



(a) Density Function (pdf) (b) Cumulative Distribution Function (cdf)

Fig. 2: Normal Distribution

The graph of the Normal density is a symmetric, bell-shaped curve (known as "Gauss's bell" or "Gauss's bell curve") centered at the value of the first parameter μ, as can be seen in Figure 2(a). The graph of the cdf of a Normally distributed random variable is given in Figure 2(b) and this is approximately what the graph of the cdf of *any* continuous random variable looks like.

Remark 6.2.

1. There is an important particular case of a Normal distribution, namely $N(0,1)$, called the **Standard (or Reduced) Normal Distribution**. A variable having a Standard Normal distribution is usually denoted by Z. The density and cdf of Z are given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R} \quad \text{and} \quad F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{6.6}$$

The function given in (6.6) is known as *Laplace's function* (or *the error function*) and its values can be found in tables or can be computed by most mathematical software.

3. As noticed from (6.5) and (6.6), there is a relationship between the cdf of any Normal $N(\mu, \sigma)$ variable X and that of a Standard Normal variable Z, namely

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right).$$

Exponential Distribution $Exp(\lambda)$

A random variable X has an Exponential distribution (exp) with parameter $\lambda > 0$, if its density function and cdf are given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}, \tag{6.7}$$

respectively.

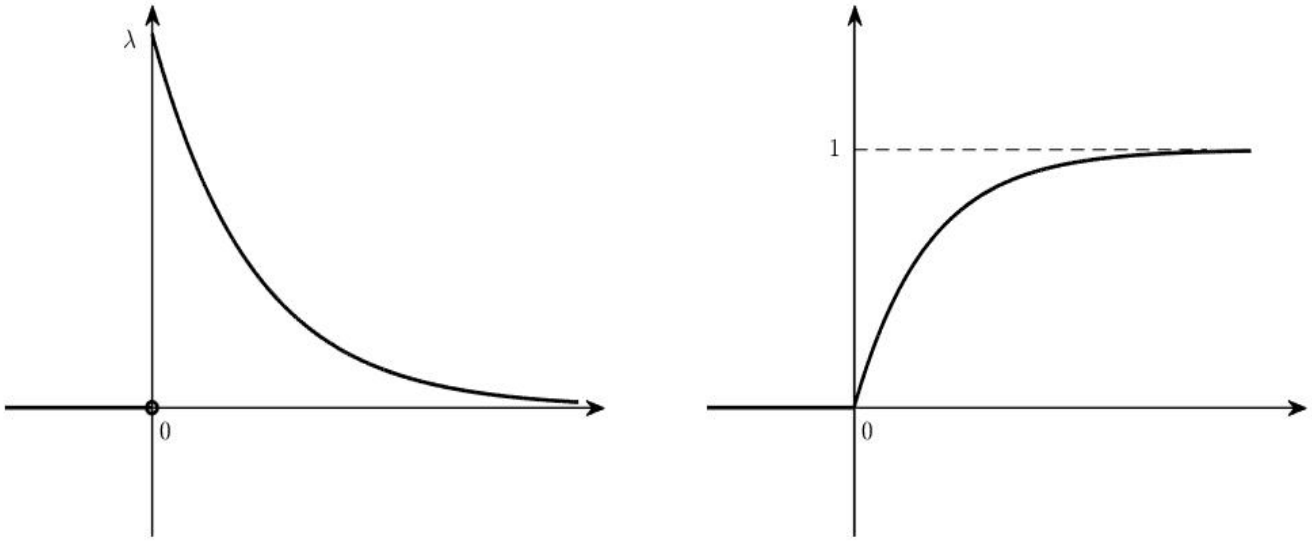
Remark 6.3.

1. The Exponential distribution is often used to model *time*: lifetime, waiting time, halftime, interarrival time, failure time, time between rare events, etc. In a sequence of rare events (where the number of rare events has a Poisson distribution), the time between two consecutive rare events is Exponential. The parameter λ represents the frequency of rare events, measured in time^{-1}.

2. A word of **caution** here: The parameter μ in Matlab (where the Exponential pdf is defined as $\frac{1}{\mu} e^{-\frac{1}{\mu}x}, x \geq 0$) is actually $\mu = 1/\lambda$. It all comes from the different interpretation of the "frequency". For instance, if the frequency is "2 per hour", then $\lambda = 2/\text{hr}$, but this is equivalent to " one every half an hour", so $\mu = 1/2$ hours. The parameter μ is measured in time units.

3. The Exponential distribution is a special case of a more general distribution, namely the $Gamma(a,b)$, $a, b > 0$, distribution (gam). The Gamma distribution models the *total* time of a multistage scheme.

4. If $\alpha \in \mathbb{N}$, then the sum of α independent $Exp(\lambda)$ variables has a $Gamma(\alpha.1/\lambda)$ distribution.



(a) Density Function (pdf) (b) Cumulative Distribution Function (cdf)

Fig. 3: Exponential Distribution

Remark 6.4. In Statistics, the most widely used distributions are the following:
– the Normal distribution, $N(\mu, \sigma)$, especially $N(0, 1)$,
– the Student (T) distribution, $T(n)$,
– the χ^2 distribution, $\chi^2(n)$,
– the Fisher (F) distribution, $F(m, n)$.

7 Continuous Random Vectors, Joint Density Function and Marginal Densities

Again, we will restrict our study to the two-dimensional case.

Definition 7.1. *Let $(S, \mathcal{K}, P)$ be a probability space.*

- *A two-dimensional **random vector** is a function $(X, Y) : S \to \mathbb{R}^2$ satisfying the condition*

$$(X \leq x, Y \leq y) = \{e \in S \mid X(e) \leq x, Y(e) \leq y\} \in \mathcal{K}, \tag{7.1}$$

for all $(x, y) \in \mathbb{R}^2$.

- *The function* $F : \mathbb{R}^2 \to \mathbb{R}$ *defined by*

$$F(x, y) \quad = \quad P(X \leq x, Y \leq y) \tag{7.2}$$

is called the ***joint cumulative distribution function (joint cdf)*** *of the vector* (X, Y).

The properties of the cdf of a random variable translate very naturally for a random vector, as well.

Theorem 7.2. *Let* (X, Y) *be a random vector with joint cdf* $F : \mathbb{R}^2 \to \mathbb{R}$ *and let* $F_X, F_Y : \mathbb{R} \to \mathbb{R}$ *be the cdf's of* X *and* Y*, respectively. Then following properties hold:*

a) *If* $a_k < b_k,\ k = \overline{1, 2}$*, then*

$$\begin{aligned} P(a_1 < X \leq b_1, a_2 < Y \leq b_2) = \quad & F(b_1, b_2) \ - \ F(b_1, a_2) \\ - \quad & F(a_1, b_2) \ + \ F(a_1, a_2). \end{aligned} \tag{7.3}$$

b) F *is monotonically increasing in each variable.*

c) F *is right continuous in each variable.*

d) $\lim\limits_{x,y\to\infty} F(x, y) = 1,$
$\lim\limits_{y\to-\infty} F(x, y) = \lim\limits_{x\to-\infty} F(x, y) = 0,\ \forall x, y \in \mathbb{R},$
$\lim\limits_{y\to\infty} F(x, y) = F_X(x),\ \forall x \in \mathbb{R},$
$\lim\limits_{x\to\infty} F(x, y) = F_Y(y),\ \forall y \in \mathbb{R}.$

Proof. (Selected)
a) This proof is similar to the proof for random variables.
d) Let $x \in \mathbb{R}$. We have

$$\lim_{y\to\infty} F(x, y) = P\left(X \leq x, Y < \infty\right) = P\left(X \leq x\right) = F_X(x)$$

and by symmetry, $\lim\limits_{x\to\infty} F(x, y) = F_Y(y),\ \forall y \in \mathbb{R}$. Then, it follows that

$$\lim_{x,y\to\infty} F(x, y) = \lim_{y\to\infty} F_Y(y) = \lim_{x\to\infty} F_X(x) = 1.$$

For any $x \in \mathbb{R}$,

$$\lim_{y \to -\infty} F(x,y) = P\left(X \leq x, Y \leq -\infty\right) = P(\emptyset) \ = \ 0$$

and by symmetry, $\lim_{x \to -\infty} F(x,y) = 0, \ \forall y \in \mathbb{R}$, also.

□

Definition 7.3. *Let $(S, \mathcal{K}, P)$ be a probability space. A random vector $(X,Y) : S \to \mathbb{R}^2$ is a **continuous random vector**, if the set of values $(X,Y)(S)$ is a (finite or infinite) continuous subset of $\mathbb{R}^2$.*

Proposition 7.4. *Let (X,Y) be a continuous random vector with joint cdf $F : \mathbb{R}^2 \to \mathbb{R}$. Then F is absolutely continuous, i.e. there exists a real function $f : \mathbb{R}^2 \to \mathbb{R}$, such that*

$$F(x,y) \ = \ \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v) \, du \, dv, \tag{7.4}$$

for all $x, y \in \mathbb{R}$.

Definition 7.5. *Let (X,Y) be a continuous random vector. Then the function f from Proposition 7.4 is called the **joint probability density function (joint pdf)** of (X,Y).*

Theorem 7.6. *Let (X,Y) be a continuous random vector with joint cdf $F$and joint density function f. Let $F_X, F_Y : \mathbb{R} \to \mathbb{R}$ be the cdf's of X and Y and $f_X, f_Y : \mathbb{R} \to \mathbb{R}$ be the pdf's of X and Y, respectively. Then the following properties hold:*

a) $\dfrac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$, *for all* $(x,y) \in \mathbb{R}^2$.

b) $f(x,y) \geq 0$, *for all* $(x,y) \in \mathbb{R}^2$.

c) $\displaystyle\iint_{\mathbb{R}^2} f(x,y) \, dxdy = 1.$

d) *For any domain* $D \subseteq \mathbb{R}^2$, $P\big((X,Y) \in D\big) = \displaystyle\iint_D f(x,y) \, dxdy.$

e) $f_X(x) = \displaystyle\int_{\mathbb{R}} f(x,y) \, dy, \ \forall x \in \mathbb{R}$ *and* $f_Y(y) = \displaystyle\int_{\mathbb{R}} f(x,y) \, dx, \ \forall y \in \mathbb{R}.$

Proof. These properties follow easily from Proposition 7.4 and the properties of the joint cdf stated in Theorem 7.2. □

Remark 7.7. When obtained from the vector (X, Y), the pdf's f_X and f_Y are called *marginal* densities.

Definition 7.8. *Two continuous random variables X and Y are **independent** if*

$$f_{(X,Y)}(x, y) \quad = \quad f_X(x) f_Y(y), \tag{7.5}$$

for all $(x, y) \in \mathbb{R}^2$.

8 Functions of Continuous Random Variables

Proposition 8.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotone and differentiable function, with $g'(x) \neq 0, \forall x \in \mathbb{R}$. Let X be a continuous random variable with pdf f_X and let $Y = g(X)$. Then for $y \in \mathbb{R}$, the pdf of Y is given by*

$$f_Y(y) = \begin{cases} \dfrac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}, & \text{if } y \in g(\mathbb{R}) \\ \\ 0, & \text{if } y \notin g(\mathbb{R})\,. \end{cases} \tag{8.1}$$

Proof.
Case I. Assume g is strictly increasing. We have

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

$$= \begin{cases} P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)), & \text{if } y \in g(\mathbb{R}) \\ 0, & \text{if } y < \inf\ g(\mathbb{R}) \\ 1, & \text{if } y > \sup\ g(\mathbb{R}). \end{cases}$$

We differentiate to obtain

$$f_Y(y) = \begin{cases} F_X'(g^{-1}(y)) \cdot (g^{-1}(y))', & \text{if } y \in g(\mathbb{R}) \\ 0, & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \dfrac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}\,, & \text{if } y \in g(\mathbb{R}) \\ \\ 0, & \text{otherwise} \end{cases}$$

Case II. If g is strictly decreasing, in a similar way, we get

$$F_Y(y) = \begin{cases} P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)), & \text{if } y \in g(\mathbb{R}) \\ 0, & \text{if } y \leq \inf\ g(\mathbb{R}) \\ 1, & \text{if } y \geq \sup\ g(\mathbb{R}) \end{cases}$$

$$f_Y(y) = \begin{cases} -\dfrac{f_X(g^{-1}(y))}{g'(g^{-1}(y))}\,, & \text{if } y \in g(\mathbb{R}) \\ \\ 0, & \text{else} \end{cases}$$

Since if g is increasing, then $g' > 0$ and if g is decreasing, then $g' < 0$, in both cases, we have (8.1).

□

Chapter 4. Numerical Characteristics of Random Variables

The distribution of a random variable or a random vector, the full collection of related probabilities, contains the entire information about its behavior. This detailed information can be summarized in a few vital numerical characteristics describing the average value, the most likely value of a random variable, its spread, variability, etc. These are numbers that will provide some information about a random variable or about the relationship between random variables.

1 Expectation

Definition 1.1.

(i) If $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ *is a discrete random variable, then the* ***expectation (expected value, mean value)*** *of* X *is the real number*

$$E(X) \;=\; \sum_{i \in I} x_i P(X = x_i) \;=\; \sum_{i \in I} x_i p_i, \tag{1.1}$$

if it exists (the series is absolutely convergent).

(ii) If X *is a continuous random variable with density function* f*, then its* ***expectation (expected value, mean value)*** *is the real number*

$$E(X) \;=\; \int_{\mathbb{R}} x f(x) dx, \tag{1.2}$$

if it exists (the integral is absolutely convergent).

Remark 1.2.

1. The expected value can be thought of as a "long term" average value, a number that we *expect* the values of a random variable to stabilize on.

2. It can also be interpreted as a point of equilibrium, a center of gravity. In the discrete case, if we imagine the probabilities p_i to be weights distributed in the points x_i, then $E(X)$ would be the point

that holds the whole thing in equilibrium. In fact, notice that formula (1.1) is *actually* a weighted mean. Consider a random variable with pdf

$$X \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}.$$

Observing this variable many times, we shall see $X = 0$ about 50% of times and $X = 1$ about 50% of times. The average value of X will then be close to 0.5, so it is reasonable to have $E(X) = 0.5$, which we get by (1.1).

Now, suppose that $P(X = 0) = 0.75$ and $P(X = 1) = 0.25$, i.e its pdf is now

$$X \begin{pmatrix} 0 & 1 \\ 0.75 & 0.25 \end{pmatrix}.$$

Then, in a long run, X is equal to 1 only $1/4$ of times, otherwise it equals 0. Therefore, in this case, $E(X) = 0.25$.

The expected value as a center of gravity is illustrated in Figure 1.

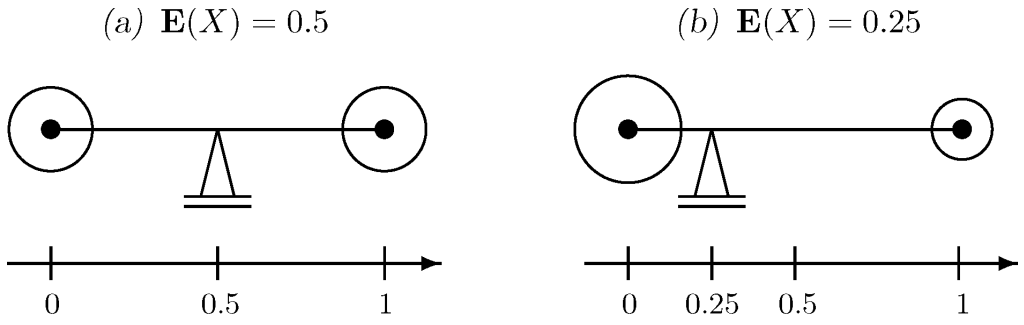


Fig. 1: Expectation as a center of gravity

The same interpretation would go for the continuous case, only there the "weight" would be continuously distributed, according to the density function f.

3. If $f : \mathbb{R} \to \mathbb{R}$ is a measurable function, then

$$E(h(X)) \quad = \quad \sum_{i \in I} h(x_i) p_i, \tag{1.3}$$

if X is discrete and

$$E(h(X)) \;=\; \int_{\mathbb{R}} h(x)f(x)\,dx, \tag{1.4}$$

if X is continuous.

Example 1.3. Let us start with a simple, intuitive example. Let X be the random variable that denotes the number shown when a die is rolled. What would be the "expected average value" of X, if the die was rolled over and over?

Since any of the 6 numbers is equally probable to show on the die, we would expect that, in the long run, we would roll as many 1's as 6's. These would average out at $\dfrac{1+6}{2} = 3.5$. Also, we would expect to roll the same number of 2's as 5's, which would also average at $\dfrac{2+5}{2} = 3.5$. Finally, about the same number of 3's and 4's would be expected to show and their average is again, 3.5. So, the "long term average" should be, intuitively, 3.5.

On the other hand, we know that X has a Discrete Uniform $U(6)$ distribution, with pdf

$$X \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

Then, by (1.1),

$$E(X) \;=\; \sum_{i\in I} x_i p_i \;=\; \sum_{i=1}^{6} i \cdot \frac{1}{6} \;=\; \frac{1}{6}\cdot\frac{6\cdot 7}{2} \;=\; \frac{7}{2},$$

the value we obtained intuitively.

Example 1.4. Consider now a (continuous) Uniform variable $X \in U(a,b)$. That means X can take any value in the interval $[a,b]$, equally probable (recall Problem 3 in Seminar 2, about a spyware breaking passwords). In the long run, it is just as likely to take values at the beginning of the interval, as it is to take the ones towards the end of $[a,b]$. So they would average out at the value right in the middle, i.e. the midpoint of the interval, $\dfrac{a+b}{2}$.

Indeed, since the pdf of X is $f(x) = \dfrac{1}{b-a},\ x \in [a,b]$ (and 0 everywhere else), by (1.2), its

expected value is

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} xf(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b xdx \\ &= \frac{1}{b-a}\cdot\frac{1}{2}x^2\Big|_a^b = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}. \end{aligned}$$

Example 1.5. The expected value of a $Bern(p), p \in (0,1)$ variable with pdf

$$X\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

is

$$E(X) = 0\cdot(1-p)+1\cdot p = p. \tag{1.5}$$

Theorem 1.6. *(Properties of the expected value)*
If X and Y are either both discrete or both continuous random variables, then the following properties hold:

a) $E(aX+b) = aE(X)+b$, *for all* $a, b \in \mathbb{R}$.

b) $E(X+Y) = E(X)+E(Y)$.

c) *If X and Y are independent, then* $E(X\cdot Y) = E(X)E(Y)$.

d) *If $X \le Y$, i.e. $X(e) \le Y(e)$, for all $e \in S$, then* $E(X) \le E(Y)$.

Proof. (Selected, only the discrete case)
a) If X is discrete, with pdf

$$X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I},$$

then $Y = aX+b$ has pdf

$$Y\begin{pmatrix} ax_i+b \\ p_i \end{pmatrix}_{i\in I}.$$

So, its expectation is

$$E(aX+b) = \sum_{i\in I}(ax_i+b)p_i = a\sum_{i\in I}x_ip_i + b\sum_{i\in I}p_i = aE(X)+b.$$

b) For X and Y both discrete, recall that their sum has pdf

$$X+Y \left(\begin{array}{c} x_i+y_j \\ p_{ij} \end{array} \right)_{(i,j)\in I\times J}, p_{ij} = P(X=x_i, Y=y_j)$$

and that

$$\sum_{j\in J} p_{ij} = p_i, \ \sum_{i\in I} p_{ij} = q_j$$

where $p_i = P(X=x_i),\ i\in I$ and $q_j = P(Y=y_j),\ j\in J$. Then

$$\begin{aligned} E(X+Y) &= \sum_{i\in I}\sum_{j\in J}(x_i+y_j)p_{ij} \\ &= \sum_{i\in I}\sum_{j\in J} x_i p_{ij} + \sum_{j\in J}\sum_{i\in I} y_j p_{ij} \\ &= \sum_{i\in I} x_i \underbrace{\sum_{j\in J} p_{ij}}_{p_i} + \sum_{j\in J} y_j \underbrace{\sum_{i\in I} p_{ij}}_{q_j} \\ &= \sum_{i\in I} x_i p_i + \sum_{j\in J} y_j q_j \\ &= E(X)+E(Y). \end{aligned}$$

c) For X and Y discrete and independent, we have

$$\begin{aligned} E(XY) &= \sum_{i\in I}\sum_{j\in J} x_i y_j p_{ij} \overset{\text{ind}}{=} \sum_{i\in I}\sum_{j\in J} x_i y_j p_i q_j \\ &= \sum_{i\in I} x_i \Big(\underbrace{\sum_{j\in J} y_j q_j}_{E(Y)} \Big) p_i \\ &= E(Y)\cdot \sum_{i\in I} x_i p_i \\ &= E(X)\cdot E(Y). \end{aligned}$$

d) We show that if $Z \geq 0$, then $E(Z) \geq 0$. Then by a) and b) applied to $Z = Y - X,$ the property follows.
If Z is discrete, $Z \geq 0$ means its values $z_i \geq 0,\ \forall i \in I$ and then

$$E(Z) = \sum_{i \in I} z_i P(Z = z_i) \geq 0.$$

□

Remark 1.7.
1. Property b) in Theorem 1.6 can be generalized to

$$E\Big(\sum_{i=1}^{n} X_i\Big) \ = \ \sum_{i=1}^{n} E(X_i).$$

2. Property c) in Theorem 1.6 can also be generalized: If $X_1, \ldots, X_n$ are independent, then

$$E\Big(\prod_{i=1}^{n} X_i\Big) \ = \ \prod_{i=1}^{n} E(X_i).$$

Example 1.8. Find the expectation of a Binomial variable $X \in B(n,p), n \in \mathbb{N}, p \in (0,1)$.

Solution. Recall (Remark 4.8, Lecture 4) that a Binomial variable $X \in B(n,p)$ is the sum of n independent $X_i \in Bern(p)$ random variables. All variables X_i have the same expected value $E(X_i) = p$, since they have the same distribution. Then, by the previous theorem,

$$E(X) \ = \ E\Big(\sum_{i=1}^{n} X_i\Big) \ = \ \sum_{i=1}^{n} E(X_i) \ = \ \sum_{i=1}^{n} p \ = \ np.$$

■

Remark 1.9. For a Normal variable $X \in N(\mu, \sigma)$, the expected value is $E(X) = \mu$.

2 Variance and Standard Deviation

Expectation shows where the average value of a random variable is located, or where the variable is expected to be, plus or minus some error. How large could this "error" be, and how much can a variable vary around its expectation? The answer to these questions can give important information about a random variable.

Knowledge of the mean value of a random variable is important, but that knowledge *alone* can be misleading. Suppose two patients in a hospital, X and Y, have their pulse (number of heartbeats per minute) checked every day. Over the course of time, they each have a mean pulse of 75, which is considered healthy. But, for patient X the pulse ranges between 70 and 80, while for patient Y, it oscillates between 40 and 110. Obviously, the second patient might have some serious health problems, which the *expected value alone* would not show.

So, next, we define some measures of variability.

Definition 2.1. *Let X be a random variable. The **variance (dispersion)** of X is the number*

$$V(X) \;=\; E\left[(X - E(X))^2\right], \tag{2.1}$$

*if it exists. The value $\sigma(X) = \mathrm{Std}(X) = \sqrt{V(X)}$ is called the **standard deviation** of X.*

Theorem 2.2. *(Properties of the variance) Let X and Y be random variables. Then the following properties hold:*

a) $V(X) = E(X^2) - E(X)^2$.

b) $V(aX + b) = a^2 V(X)$, *for all* $a, b \in \mathbb{R}$.

c) *If X and Y are independent, then*

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

d) *If X and Y are independent, then*

$$\begin{aligned} V(X \cdot Y) &= V(X)V(Y) + E(X)^2 V(Y) + E(Y)^2 V(X) \\ &= E(X^2)E(Y^2) - E(X)^2 E(Y)^2. \end{aligned}$$

Proof. (Selected)
a) By definition (2.1) and the properties of expectation in Theorem 1.6, we have

$$\begin{aligned} V(X) &= E\left[X^2 - 2E(X)X + (E(X))^2\right] \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned}$$

b)

$$\begin{aligned} V(aX+b) &= E\left[(aX+b-E(aX+b))^2\right] \\ &= E\left[(aX+b-aE(X)-b)^2\right] \\ &= a^2 E\left[(X-E(X)^2\right] \\ &= a^2 V(X). \end{aligned}$$

c) If X, Y are independent, then so are $X - E(X), Y - E(Y)$, so

$$\begin{aligned} V(X+Y) &= E\left[(X+Y-E(X+Y))^2\right] \\ &= E\left[(X-E(X)+Y-E(Y))^2\right] \\ &= E\left[(X-E(X))^2\right] + 2E\left[(X-E(X))(Y-E(Y))\right] + E\left[(Y-E(Y))^2\right] \\ &\overset{\text{ind}}{=} V(X) + 2E\left[(X-E(X)\right] \cdot E\left[(Y-E(Y)\right] + V(Y) \\ &= V(X) + V(Y), \end{aligned}$$

since $E\left[(X - E(X)\right] = 0$. □

Remark 2.3.

1. Part a) of Theorem 2.2 provides a more practical computational formula for the variance than the definition. Thus, if $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ is discrete, then

$$V(X) = \sum_{i \in I} x_i^2 p_i - \left(\sum_{i \in I} x_i p_i \right)^2$$

and if X is continuous with density function f, then

$$V(X) = \int_{\mathbb{R}} x^2 f(x)\, dx - \left(\int_{\mathbb{R}} x f(x)\, dx \right)^2 .$$

2. A direct consequence of Theorem 2.2a) (since $V(X) \geq 0$) is the following inequality:

$$|E(X)| \leq \sqrt{E(X^2)},$$

which will be discussed later on in this chapter.

3. If $X = b$ is a constant random variable (i.e. it only takes that one value with probability 1), then by Theorem 2.2a), $V(X) = 0$, which is to be expected (the variable X does not vary *at all*).
4. Part c) of Theorem 2.2 can be generalized to any number of random variables: If $X_1, \ldots, X_n$ are independent, then

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i).$$

5. A consequence of parts b) and c) of Theorem 2.2 is the following property: If X and Y are independent, then

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

Example 2.4. Find the variance of a random variable X having
a) a Bernoulli $Bern(p)$ distribution;
b) a Binomial $B(n, p)$ distribution.

Solution.
a) We have

$$X\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \; X^2\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix},$$

so both $E(X) = E(X^2) = p$ and thus

$$V(X) = p - p^2 = pq.$$

b) If X is Binomial, again we use the fact that it can be written as

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

where $X_1, \ldots, X_n$ are independent and identically distributed with a $Bern(p)$ distribution. Then by part a), $V(X_i) = pq$, for each $i = \overline{1, n}$ and by the previous remarks,

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = npq.$$

■

Remark 2.5. For a Normal variable $X \in N(\mu, \sigma)$, the variance is $V(X) = \sigma^2$ (and its standard

deviation is $\sigma(X) = \mathrm{Std}(X) = \sigma$. So, the parameters of a Normal variable $X \in N(\mu, \sigma)$ are its mean value and its standard deviation.

3 Moments

The idea of expected value and variance can be generalized.

Definition 3.1. *Let X be a random variable and let $k \in \mathbb{N}$.*
*The **(initial) moment of order k** of X is (if it exists) the number*

$$\nu_k = E(X^k). \tag{3.1}$$

*The **absolute moment of order k** of X is (if it exists) the number*

$$\underline{\nu}_k = E(|X|^k). \tag{3.2}$$

*The **central (centered) moment of order k** of X is (if it exists) the number*

$$\mu_k = E\left[\Big(X - E(X)\Big)^k\right]. \tag{3.3}$$

Remark 3.2.
1. If X is a discrete random variable with pdf $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, then for every $k \in \mathbb{N}$,

$$\begin{aligned} \nu_k &= \sum_{i\in I} x_i^k p_i, \\ \underline{\nu}_k &= \sum_{i\in I} |x_i|^k p_i, \\ \mu_k &= \sum_{i\in I} (x_i - E(X))^k p_i. \end{aligned}$$

If X is a continuous random variable with density function f, then for every $k \in \mathbb{N}$,

$$\nu_k = \int_{\mathbb{R}} x^k f(x)\, dx,$$

$$\begin{aligned} \underline{\nu}_k &= \int_{\mathbb{R}} |x|^k f(x)\, dx, \\ \mu_k &= \int_{\mathbb{R}} (x - E(X))^k f(x)\, dx. \end{aligned}$$

2. The expectation of a random variable X is the moment of order 1,

$$E(X) = \nu_1.$$

The variance of a random variable X is the central moment of order 2,

$$V(X) = \mu_2 = \nu_2 - \nu_1^2.$$

For any random variable X, the central moment of order 1 is 0,

$$\mu_1 = E\left(X - E(X)\right) = E(X) - E(X) = 0.$$

3. An important property of the moments of a random variable X, which we just state, without proof, is the following: If $\underline{\nu}_n = E(|X|^n)$ exists for some $n \in \mathbb{N}$, then $\nu_k,\ \underline{\nu}_k$ and μ_k also exist, for all $k = \overline{1, n}$.

4 Quantiles

Definition 4.1. *Let X be a random variable with cumulative distribution function $F : \mathbb{R} \to \mathbb{R}$ and let $\alpha \in (0, 1)$. A **quantile (percentile) of order α** is a number q_α satisfying the conditions*

$$\begin{aligned} P(X < q_\alpha) &\leq \alpha \\ P(X > q_\alpha) &\leq 1 - \alpha, \end{aligned} \tag{4.1}$$

or, equivalently,

$$P(X < q_\alpha) \leq \alpha \leq P(X \leq q_\alpha),$$

i.e.

$$F(q_\alpha - 0) \leq \alpha \leq F(q_\alpha). \tag{4.2}$$

To interpret (4.1), a quantile is a number with the property that it exceeds at most $100\alpha\%$ of the data, and is exceeded by at most $100(1 - \alpha)\%$ of the data.

Of all quantiles, the most important are:
The **median**, the number $m = q_{1/2}$; there are at most 50% of the data to the left of the median and at most 50% to its right.
The **quartiles** are the numbers

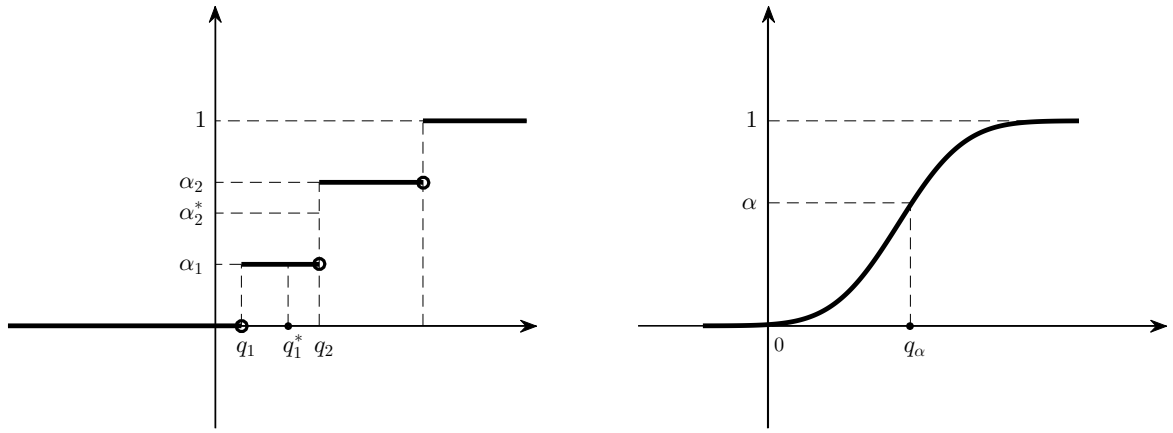
$$Q_1 = q_{1/4},\ Q_2 = m = q_{1/2},\ Q_3 = q_{3/4}.$$

Remark 4.2.
1. Quantiles are useful in statistical analysis of data. The median roughly locates the "middle" of a set of data, while the quartiles approximately locate every 25 $\%$ of a set of data. These will be discussed again in the next chapter.
2. If X is discrete, then a quantile can take an infinite number of values, if the line $y = \alpha$ and the curve $y = F(x)$ have in common a segment line (see Figure 1a).

The case when X is continuous is more interesting and the one we will use in Statistics. If X is continuous, then for each $\alpha \in (0, 1)$, there is a *unique* quantile q_α, given by

$$F(q_\alpha) = \alpha,$$

since F is a continuous function, $F(q_\alpha - 0) = \alpha = F(q_\alpha)$. In this case, for $F : \mathbb{R} \to \mathbb{R}$ there always exists $A \subset \mathbb{R}$ such that $F : A \to [0, 1]$ is both injective and surjective, hence invertible. Thus, in



(a) Discrete cdf

(b) Continuous cdf

Fig. 1: Quantiles

this case the unique quantile q_α is found by

$$q_\alpha \;=\; F^{-1}(\alpha). \tag{4.3}$$

Now, as an interpretation, let us recall that for continuous random variables, the cdf is expressed as an integral, which means as an area. So we have

$$\alpha \;=\; F(q_\alpha) \;=\; \int_{-\infty}^{q_\alpha} f(x)\,dx,$$

which is the area underneath the graph of the pdf f, above the x-axis and to the left of q_α. This is illustrated in Figure 2.

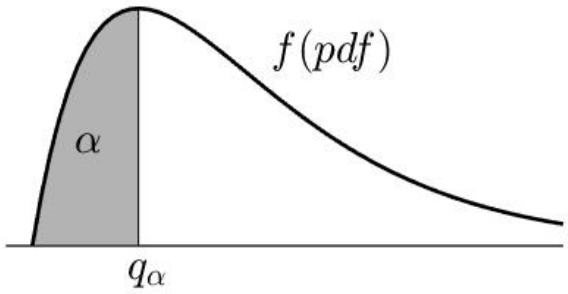


Fig. 2: Quantile of order α

5 Covariance and Correlation Coefficient

So far we have discussed numerical characteristics associated with one random variable. But oftentimes it is important to know if there is some kind of relationship between two (or more) random variables. So we need to define numerical characteristics that somehow measure that relationship.

Definition 5.1. *Let X and Y be random variables. The **covariance** of X and Y is the number*

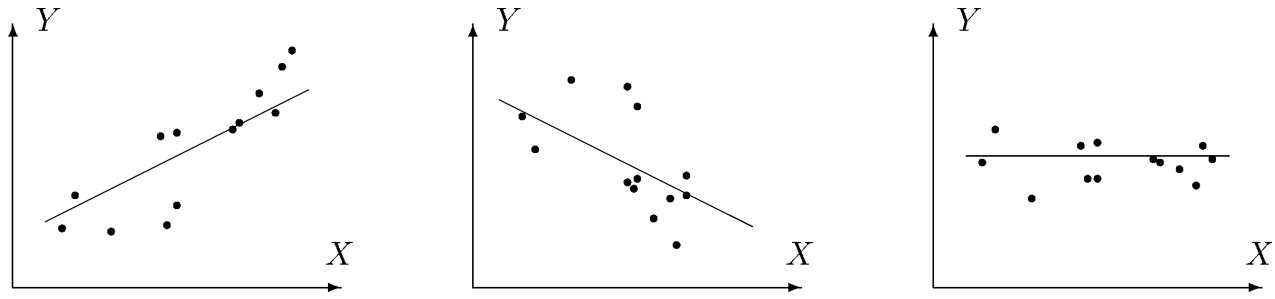
$$\operatorname{cov}(X,Y) \;=\; E\Big((X-E(X))\cdot(Y-E(Y))\Big), \tag{5.1}$$

*if it exists. The **correlation coefficient** of X and Y is the number*

$$\rho(X,Y) \;=\; \frac{\operatorname{cov}(X,Y)}{\sqrt{V(X)V(Y)}} \;=\; \frac{\operatorname{cov}(X,Y)}{\sigma(X)\sigma(Y)}, \tag{5.2}$$

if $\operatorname{cov}(X,Y), V(X), V(Y)$ exist and $V(X)\neq 0, V(Y)\neq 0$.

Notice the similarity between the definition of the covariance and that of the variance. The covariance measures the variation of two random variables with respect to each other. Just like with variance, large values (in absolute value) of the covariance show a strong relationship between X and Y, while small absolute values suggest a weak relationship. Unlike variance, covariance can also be negative. A negative value means that as the values of one variable increase, the values of the other decrease (see Figure 3).



(a) $\mathrm{Cov}(X,Y) > 0$ (b) $\mathrm{Cov}(X,Y) < 0$ (c) $\mathrm{Cov}(X,Y) = 0$

Fig. 3: Covariance

Theorem 5.2. *(Properties of covariance) Let X, Y and Z be random variables. Then the following properties hold:*

a) $\mathrm{cov}(X,X) = V(X)$.

b) $\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y)$.

c) *If X and Y are independent, then* $\mathrm{cov}(X,Y) = \rho(X,Y) = 0$ (*we say that X and Y are* ***uncorrelated***).

d) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2ab\,\mathrm{cov}(X,Y)$, *for all* $a,b \in \mathbb{R}$.

e) $\mathrm{cov}(X+Y,Z) = \mathrm{cov}(X,Z) + \mathrm{cov}(Y,Z)$.

Proof.
a) This follows directly from Definition 5.1.
b) A straightforward computation leads to

$$\begin{aligned}\mathrm{cov}(X,Y) &= E\Big(XY - E(X)Y - E(Y)X + E(X)E(Y)\Big)\\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)\\ &= E(XY) - E(X)E(Y)\end{aligned}$$

c) This follows from b), keeping in mind that X and Y are independent, so $E(XY) = E(X)E(Y)$.

d)

$$\begin{aligned} V(aX+bY) &= E\Big(aX+bY-aE(X)-bE(Y)\Big)^2 \\ &= E\Big[a\Big(X-E(X)\Big)+b\Big(Y-E(Y)\Big)\Big]^2 \\ &= E\Big[a^2\Big(X-E(X)\Big)^2+2ab\Big(X-E(X)\Big)\Big(Y-E(Y)\Big) \\ &+ b^2\Big(Y-E(Y)\Big)^2\Big] \\ &= a^2V(X)+b^2V(Y)+2ab\,\mathrm{cov}(X,Y). \end{aligned}$$

e)

$$\begin{aligned} \mathrm{cov}(X+Y,Z) &= E\Big((X+Y-E(X)-E(Y))(Z-E(Z))\Big) \\ &= E\Big((X-E(X))(Z-E(Z)) \\ &+ (Y-E(Y))(Z-E(Z))\Big) \\ &= \mathrm{cov}(X,Z)+\mathrm{cov}(Y,Z). \end{aligned}$$

□

Remark 5.3.

1. Property d) of Theorem 5.2 can be generalized to any number of variables:

$$V\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_i^2V(X_i)+2\sum_{1\le i<j\le n}a_ia_j\,\mathrm{cov}(X_i,X_j).$$

2. A consequence of a) and e) of Theorem 5.2 is the following property:

$$\mathrm{cov}(aX+b,X)=aV(X),\ \text{for all}\ \ a,b\in\mathbb{R}.$$

3. The converse of Theorem 5.2c) is *not* true. Independence is a much stronger condition.

Theorem 5.4. *Let X and Y be random variables. Then the following properties hold:*

a) $|\rho(X,Y)|\le 1$ *(i.e.* $-1\le\rho(X,Y)\le 1$*).*

b) $|\rho(X,Y)|=1$ *if and only if there exist* $a,b\in\mathbb{R}$*,* $a\neq 0$*, such that* $Y=aX+b$*.*

Remark 5.5. As Theorem 5.4 states, the correlation coefficient $\rho(X,Y)$ measures the linear trend between the variables X and Y. When $\rho=\pm1$, there is "perfect linear correlation", so all the points

(X, Y) are on a straight line (see Figure 4). The closer its value is to ± 1, the "more linear" the relationship between X and Y is. This notion will be revisited in the next chapter.

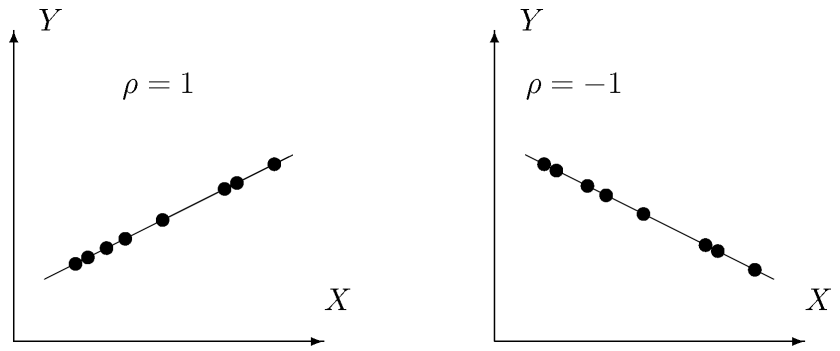


Fig. 4: Perfect correlation

6 Inequalities

Inequalities can be useful in estimation theory, for approximating probabilities or numerical characteristics associated with a random variable.

Proposition 6.1 (**Hölder's Inequality**). *Let X and Y be random variables and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$E(|XY|) \leq (E(|X|^p))^{\frac{1}{p}} \cdot (E(|Y|^q))^{\frac{1}{q}}. \tag{6.1}$$

Remark 6.2.

1. One important particular case of Hölder's inequality is for $p = q = 2$,

$$E(|XY|) \leq \sqrt{E(X^2)} \cdot \sqrt{E(Y^2)}, \tag{6.2}$$

known as **Schwarz's inequality**.

2. A particular case of the above inequality is for $Y = 1$,

$$E(|X|) \leq \sqrt{E(X^2)}, \tag{6.3}$$

known as **Cauchy-Buniakowsky's inequality**.

Proposition 6.3 (**Minkowsky's Inequality**). *Let X and Y be random variables and let $p > 1$. Then*

$$(E(|X+Y|^p))^{\frac{1}{p}} \leq (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}. \tag{6.4}$$

Proposition 6.4 (**Lyapunov's Inequality**). *Let X be a random variable, let $0 < a < b$ and $c \in \mathbb{R}$. Then*

$$(E(|X-c|^a))^{\frac{1}{a}} \leq \left(E(|X-c|^b)\right)^{\frac{1}{b}}. \tag{6.5}$$

The next two inequalities are *specific* to random variables. They have many applications in statistical analysis.

Proposition 6.5 (**Markov's Inequality**). *Let X be a random variable and let $a > 0$. Then*

$$P(|X| \geq a) \leq \frac{1}{a} E(|X|). \tag{6.6}$$

Proof. Let $A = \{e \in S \mid |X(e)| \geq a\}$, with the indicator function

$$I_A(e) \;=\; \begin{cases} 0, & |X(e)| < a \\ 1, & |X(e)| \geq a. \end{cases}$$

Then

$$a\, I_A(e) \;=\; \begin{cases} 0, & |X(e)| < a \\ a, & |X(e)| \geq a. \end{cases}$$

Now, if $|X(e)| < a$, then

$$aI_A(e) = 0 \leq |X(e)|$$

and if $|X(e)| \geq a$, then

$$aI_A(e) = a \leq |X(e)|.$$

So, either way, $aI_A(e) \leq |X(e)|$, $\forall e \in S$. That means, as random variables, $aI_A \leq |X|$, which means the same thing is true for their expected values, $E(aI_A) \leq E(|X|)$. Now, the pdf of aI_A is

$$aI_A \begin{pmatrix} 0 & a \\ 1 - P(|X| \geq a) & P(|X| \geq a) \end{pmatrix},$$

so

$$E(aI_A) = aP(|X| \geq a).$$

Thus,

$$aP(|X| \geq a) \leq E(|X|),$$

i.e.

$$P(|X| \geq a) \leq \frac{1}{a}E(|X|).$$

□

Proposition 6.6 (**Chebyshev's Inequality**). *Let X be a random variable and let $\varepsilon > 0$. Then*

$$P\Big(|X - E(X)| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}V(X), \tag{6.7}$$

or, equivalently,

$$P\Big(|X - E(X)| < \varepsilon\Big) \geq 1 - \frac{1}{\varepsilon^2}V(X), \tag{6.8}$$

Proof.
Apply Markov's inequality (6.6) to $\Big(X - E(X)\Big)^2$ and $a = \varepsilon^2$, to get

$$P\Big((X - E(X))^2 \geq \varepsilon^2\Big) \leq \frac{1}{\varepsilon^2}\, E\Big((X - E(X))^2\Big),$$

i.e.

$$P\Big(|X - E(X)| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}\, V(X),$$

and, equivalently,

$$1 - P\Big(|X - E(X)| < \varepsilon\Big) \leq \frac{1}{\varepsilon^2}V(X),$$

$$P\Big(|X - E(X)| < \varepsilon\Big) \geq 1 - \frac{1}{\varepsilon^2}V(X).$$

□

Example 6.7. Suppose the number of errors in a new software, X, has expectation $E(X) = 20$. Find a bound for the probability that there are at least 30 errors if the standard deviation is
a) $\sigma(X) = 2$;
b) $\sigma(X) = 5$.

Solution. According to (6.7),

$$P(|X - 20| \geq \varepsilon) \leq \frac{(\sigma(X))^2}{\varepsilon^2}.$$

So,

$$\begin{aligned} P(X \geq 30) &= P(X - 20 \geq 10) \\ &\leq P\Big((X - 20 \geq 10) \cup (X - 20 \leq -10)\Big) \\ &= P(|X - 20| \geq 10) \\ &\leq \frac{(\sigma(X))^2}{100} \end{aligned}$$

a) If $\sigma(X) = 2$, we can estimate that

$$P(X \geq 30) \leq 0.04.$$

b) However, for a larger standard deviation of $\sigma(X) = 5$, the estimation is

$$P(X \geq 30) \leq 0.25.$$

■

7 Central Limit Theorem

Central Limit Theorems are also results that can help approximate characteristics of random variables. First, a little bit of preparation.

Given the special nature of random variables, as opposed to numerical variables, there are various types of convergence that can be defined for sequences of such variables, having to do with probability-related notions (convergence in probability, convergence in mean, convergence almost surely, etc.).

Definition 7.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with cumulative distribution functions $F_n = F_{X_n}$, $n \in \mathbb{N}$ and let X be a random variable with cdf $F = F_X$. Then X_n* ***converges in distribution*** *to X, denoted by $X_n \xrightarrow{d} X$, if*

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{7.1}$$

for every $x \in \mathbb{R}$, a point of continuity of F.

A statement about the limit in distribution of a sequence of random variable in called a **limit theorem**. If the limit variable has a Normal distribution, then such a result is called a **central limit**

theorem. So, there are *many* such results, the name "Central Limit Theorem" is just generic.

We want to discuss a central limit theorem that applies to the following case: Suppose $X_1, X_2, \ldots,$ X_n are **independent, identically distributed (iid)** random variables. Having the same pdf, they have the same expectation $\mu = E(X_i)$ and the same standard deviation $\sigma = \mathrm{Std}(X_i) = \sqrt{V(X_i)}$. We are interested in the random variable

$$S_n \;=\; X_1 + \ldots + X_n.$$

This case appears in many applications and in many statistical procedures. We see right away that

$$\begin{aligned} E(S_n) &= n\mu, \\ V(S_n) &= n\sigma^2. \end{aligned}$$

How does S_n behave for large n?

The *pure* sum S_n diverges. In fact, this should be anticipated because

$$V(S_n) = n\sigma^2 \to \infty,$$

so the variability of S_n grows unboundedly as n goes to infinity.

The *average* S_n/n converges. Indeed, in this case, we have

$$V(S_n/n) = \frac{1}{n^2} V(S_n) = \frac{\sigma^2}{n} \to 0,$$

so the variability of S_n/n vanishes as $n \to \infty$.

An interesting case is the variable $S_n/\sqrt{n}$, which neither diverges nor converges. In fact, it behaves like some random variable. The following theorem (CLT) states that this variable has approximately Normal distribution for large n. In fact, the result is for its *reduced (standardized)* variable

$$\frac{S_n/\sqrt{n} - E(S_n/\sqrt{n})}{\mathrm{Std}(S_n/\sqrt{n})} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Theorem 7.2. [Central Limit Theorem (CLT)]
Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with expectation $\mu = E(X_i)$ and standard deviation $\sigma = \sigma(X_i)$ and let

$$S_n \;=\; X_1 + \ldots + X_n. \tag{7.2}$$

Then, as $n \to \infty$, the reduced sum

$$Z_n \;=\; \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \in N(0,1), \tag{7.3}$$

which means

$$F_{Z_n} \to F_{N(0,1)}, \;\text{ i.e. } P(Z_n \le x) \to P(Z \le x),\; \forall x \in \mathbb{R}, \;\text{ as }\; n \to \infty.$$

Remark 7.3.
1. This result can be very helpful, since $F_{N(0,1)}(x) = \Phi(x)$, Laplace's function (see equation (6.6) in Lecture 5), whose values are known.
2. The CLT can be used as an approximation tool for n "large". In practice, it has been determined that that means $n > 30$.

Example 7.4. A disk has free space of 330 megabytes. Is it likely to be sufficient for 300 independent images, if each image has expected size of 1 megabyte with a standard deviation of 0.5 megabytes?

Solution. For each $i = 1, 2, \ldots, n$ (i.e. for each image), let X_i denote the space it takes, in megabytes. Then the *total* space taken by all 300 images will be $S_n = X_1 + X_2 + \ldots + X_n$ and there will be sufficient space on the disk if $S_n \le 330$.
We have $n = 300, \mu = 1, \sigma = 0.5$. The number of images n is large enough, so the CLT applies to their total size S_n. Then

$$\begin{aligned}
P(\text{sufficient space}) &= P(S_n \le 330) \\
&= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{330 - n\mu}{\sigma\sqrt{n}}\right) \\
&= P\left(Z_n \le \frac{330 - 300 \cdot 1}{0.5 \cdot 10\sqrt{3}}\right) \\
&= P(Z_n \le 3.46) \\
&\overset{\text{CLT}}{\approx} P(Z \le 3.46) \;=\; \Phi(3.46) \;=\; 0.9997,
\end{aligned}$$

a very high probability, hence, the available disk space is very likely to be sufficient.

■

PART II. STATISTICS

Chapter 5. Descriptive Statistics

Statistics is a branch of Mathematics that deals with the collection, analysis, display and interpretation of numerical data. It consists of two main areas:
Descriptive Statistics includes the collection, presentation and description of numerical data. It is what most people think of when they hear the word "Statistics".
Inferential Statistics consists of the techniques of interpretation, of modeling the results from descriptive Statistics and then using them to make inferences.

1 Analysis and Display of Data

1.1 Basic Concepts

A **population** is a set of individuals, objects, items or measurements whose properties are to be analyzed.
In order to form a population, a set must have a common feature. The population of interest must be carefully defined and is considered so when its membership list is specified.

A subset of the population (a set of observed units collected from the population) is called a **sample**, or a **selection**. A sample must be **random** (each element of the population must have the same chance of being chosen) and representative for the population it was drawn from (the structure of the sample must be similar to the structure of the population).

A **characteristic** or **variable** is a certain feature of interest of the elements of a population or a sample, that is about to be analyzed statistically. Characteristics can be *quantitative* (numerical) or *qualitative* (a certain trait). From the probabilistic point of view, a numerical characteristic is a random variable. Further, numerical variables can be *discrete* (if they can be counted) or *continuous* (if they can be measured). A numerical characteristic is called a **parameter**, if it refers to an entire population and a **statistic**, if it refers just to a sample.

The outcomes of an experiment yield a set of **data**, i.e. the values that a variable takes for all the

elements of a population or a sample.

1.2 Data Collection, Sampling

An important first step in any statistical analysis is the **sampling technique**, i.e. the collection of methods and procedures used to gather data. There are several ways of collecting data: If every element of a population is selected, then a **census** is compiled. However, this technique is hardly ever used these days, because it can be expensive, time consuming or just plain impossible. Instead, only a **sample** is selected, which is analyzed and based on the findings, inferences (estimates) are made about the entire population, as well as measurements of the degree of accuracy of the estimates.
A sample is chosen based on a **sampling design**, the process used to collect sample data. If elements are chosen on the basis of being "typical", then we have a **judgment sample**, whereas if they are selected based on probability rules, we have a **probability sample**. Statistical inference requires probability samples. The most familiar probability sample is a **random sample**, in which each possible sample of a certain size has the same chance of being selected and every element in the population has an equal probability of being chosen.
Other types of samples may be considered:

- *systematic* sample
- *stratified* sample
- *quota* sample
- *cluster* sample

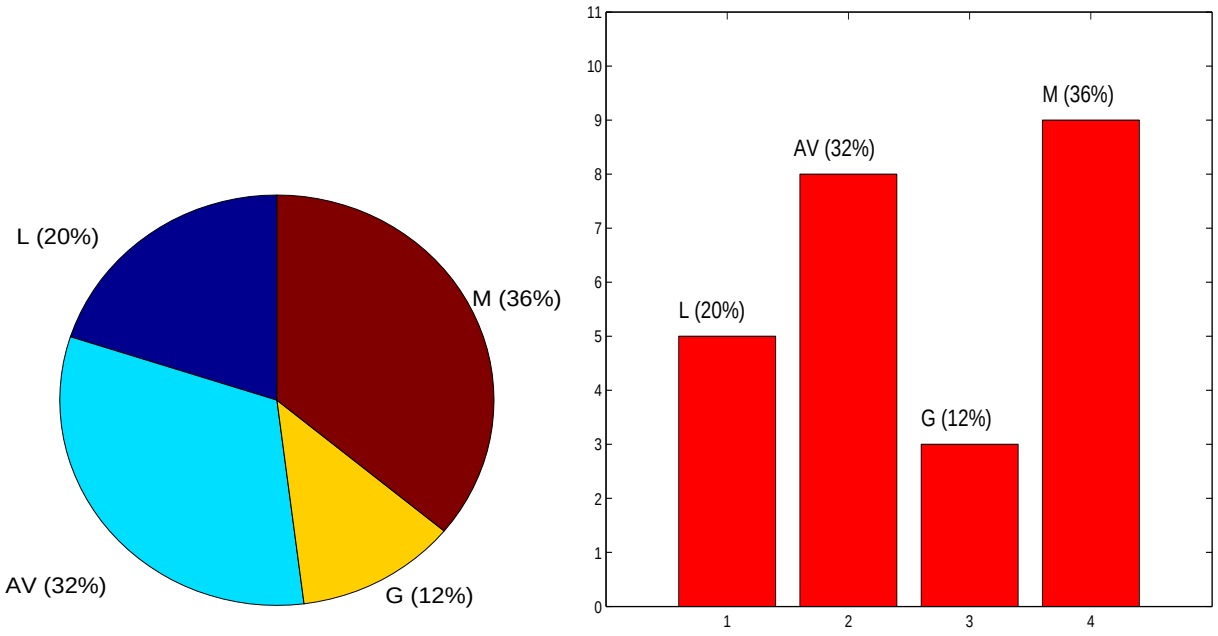
Throughout the remaining chapters, we will only consider **random samples**.
Sometimes discrepancies occur between a sample and its underlying population.
Sampling errors are caused simply by the fact that only a portion of the entire population is observed. For most statistical procedures, sampling errors decrease (and converge to zero) if the sample size is appropriately increased.
Non-sampling errors are produced by inappropriate sampling designs or wrong statistical techniques. No statistical procedures can save a poorly collected sample!

1.3 Graphical Display of Data, Frequency Distribution Tables, Histograms

"A picture is worth a thousand words!"
Once the sample data is collected, it must be represented in a relevant, "easy to read" way, one that hopefully reveals important features, patterns of behavior, connections, etc.
Circle graphs ("pie" charts) and **bar graphs** are popular ways of displaying data, that use the proportions of each type of data and represent them as percentages.

Example 1.1. Suppose that a software company is having 25 items on sale, 5 of which are learning programs (L), 8 are antivirus programs (AV), 3 are games (G) and the rest (9) are miscellaneous (M).

The pie chart and the bar graph are shown in Figure 1.



(a) Pie Chart (b) Bar Graph

Fig. 1: Example 1.1

Frequency Distribution Tables

Once collected, the raw data must be "organized" in a relevant and meaningful manner. One way to do that is to write it in a **frequency distribution table**, which contains the values $x_i, i = \overline{1,k}$, sorted in increasing order, together with their **(absolute) frequencies**, $f_i, i = \overline{1,k}$, i.e. the number of times each value occurs in the sample data, as seen in Table 1.

Value	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k

Table 1: Frequency Distribution Table

If needed, the table can also contain the **relative frequencies**

$$rf_i = \frac{f_i}{N},\ \forall i = \overline{1,k},$$

usually expressed as percentages, the **cumulative frequencies**

$$F_i = \sum_{j=1}^{i} f_j,\ \forall i = \overline{1,k},$$

or **relative cumulative frequencies**

$$rF_i = \frac{1}{N}\sum_{j=1}^{i} f_j,\ \forall i = \overline{1,k},$$

where $N = \sum_{i=1}^{k} f_i$ is the sample size.

However, when the data volume is large and the values are non-repetitive, the frequency distribution is not of much help. Every value is listed with a frequency of 1. In this case, it is better to *group* the data into *classes* and construct a **grouped frequency distribution table**. So, first we decide on a reasonable number of classes n, small enough to make our work with the data easier, but still large enough to not lose the relevance of the data. Then for each class $i = \overline{1,n}$, we have

– the **class limits** c_{i-1}, c_i,

– the **class mark** $x_i = \frac{c_{i-1} + c_i}{2}$, the midpoint of the interval, as an identifier for the class,

– the **class width (length)** $l_i = c_i - c_{i-1}$,

– the **class frequency** f_i, the sum of the frequencies of all observations x in that class.

Notice that we used the same notation x_i for primary data and for class marks. This is by choice, since in the case of grouped data, the class mark plays the role of a "representative" for that class and the class frequency is taken as being the frequency of that one value. The double notation should not

cause confusion throughout the text, since N is the sample size, so $x_1, \ldots, x_N$ denotes the primary data, while n is the number of classes and thus,

$$\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$$

denotes the grouped frequency distribution of the data.

The grouped frequency distribution table will look similar to the one in Table 1, only it will contain classes instead of individual values, each with their corresponding features.

Remark 1.2.

1. Relative or cumulative frequencies can also be computed for grouped data, as well, using the same formulas as for ungrouped data.

2. In general, the classes are taken to be of the same length l.

3. When all classes have the same length, the number of classes, n, and the class length l determine each other (if one is known, so is the other). In this case, there are two customary procedures (empirical formulas) of determining the number of classes:

One is a formula for n, known as *Sturges' rule*

$$n = 1 + \frac{10}{3}\log_{10} N, \tag{1.1}$$

where N is the sample size. Then it follows that $l = \dfrac{x_{\max} - x_{\min}}{n}$.

The other is a formula for the class width

$$l = \frac{8}{100}\left(x_{\max} - x_{\min}\right). \tag{1.2}$$

Then $n = \dfrac{x_{\max} - x_{\min}}{l}$.

Once we determined n and l, we have $c_i = x_{\min} + i \cdot l,\ i = \overline{0, n}$.

Histograms and Frequency Polygons

When data is grouped into classes, the best way to visualize the frequency distribution is by constructing a **histogram** (hist). A histogram is a type of bar graph, where classes are represented by rectangles whose bases are the class lengths and whose heights are chosen so that the areas of the rectangles are proportional to the class frequencies. If the classes have all the same length, then the heights will be proportional to the class frequencies. If relative frequencies are considered (so the

proportionality factor is N, the total number of observations), then the total areas of all rectangles will be equal to 1. For a large volume of data grouped into a reasonably large number of classes, the histogram gives a rough approximation of the density function (pdf) of the population from which the sample data was drawn.

An alternative in that sense (the sense of roughly approximating the shape of the density function) to histograms are **frequency polygons**, obtained by joining the points with coordinates $(x_i, f_i),\ i = \overline{1, n}$ ($x-$coordinates are the class marks and $y-$coordinates are the class frequencies).

Example 1.3. The following represents the grades distribution in a Probability and Statistics exam, for a group of 2^{nd} year students:

$$\begin{array}{cccccccccccccccccc} 7 & 8 & 10 & 5 & 4 & 5 & 5 & 6 & 5 & 8 & 9 & 9 & 1 & 4 & 5 & 5 & 7 & 10 \\ 5 & 9 & 2 & 2 & 10 & 10 & 8 & 3 & 8 & 7 & 5 & 6 & 7 & 8 & 9 & 9 & 9 & 4. \end{array}$$

Let us analyze these data. First, we sort them in increasing order:

$$\begin{array}{cccccccccccccccccc} 1 & 2 & 2 & 3 & 4 & 4 & 4 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 6 & 6 & 7 \\ 7 & 7 & 7 & 8 & 8 & 8 & 8 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 10 & 10 & 10 & 10 \end{array}$$

There are $N = 36$ observations, with $x_{\min} = 1$ and $x_{\max} = 10$.
Since the sample size is not too large and there are repetitions, we can construct the ungrouped frequency distribution table:

Value	Frequency
1	1
2	2
3	1
4	3
5	8
6	2
7	4
8	5
9	6
10	4

Table 2: Frequency Distribution Table

Let us group the data into classes of the same length. With Sturges' rule, we get

$$n = 6.1877 \approx 6,\ \ l = 1.5,$$

while if using formula (1.2), we have

$$l = 0.72, \;\; n \approx 12.$$

The grouped frequency tables are shown in Tables 3 and 4. We have also included the relative and cumulative frequencies.
Figure 2 shows the corresponding histogram and frequency polygon for grouped data.

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 2.50)	1.75	3	3	8%	8%
2	[2.50 , 4.00)	3.25	4	7	11%	19%
3	[4.00 , 5.50)	4.75	8	15	22%	41%
4	[5.50 , 7.00)	6.25	6	21	17%	58%
5	[7.00 , 8.50)	7.75	5	26	14%	72%
6	[8.50 , 10.00]	9.25	10	36	28%	100%

Table 3: Grouped Frequency Distribution Table With $n = 6$ Classes

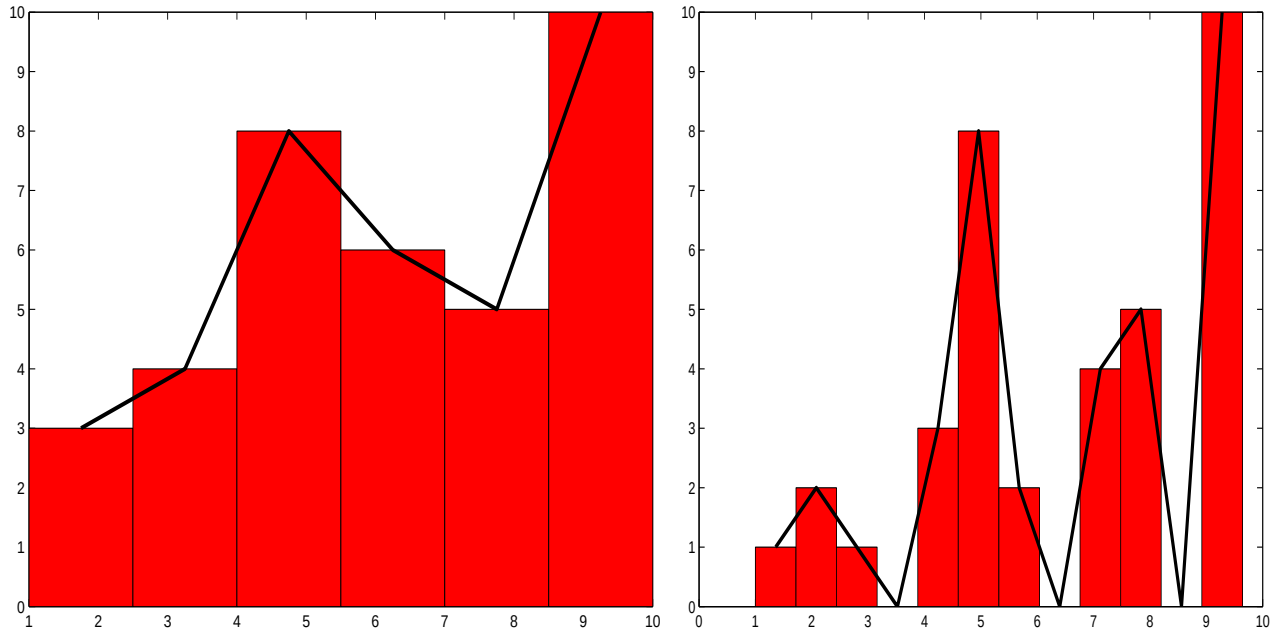
No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 1.72)	1.36	1	1	3%	3%
2	[1.72 , 2.44)	2.08	2	3	6%	9%
3	[2.44 , 3.16)	2.80	1	4	3%	12%
4	[3.16 , 3.88)	3.52	0	4	0%	12%
5	[3.88 , 4.60)	4.24	3	7	8%	20%
6	[4.60 , 5.32)	4.96	8	15	22%	42%
7	[5.32 , 6.04)	5.68	2	17	6%	48%
8	[6.04 , 6.76)	6.40	0	17	0%	48%
9	[6.76 , 7.48)	7.12	4	21	11%	59%
10	[7.48 , 8.20)	7.84	5	26	14%	73%
11	[8.20 , 8.92)	8.56	0	26	0%	73%
12	[8.92 , 10]	9.46	10	36	27%	100%

Table 4: Grouped Frequency Distribution Table With $n = 12$ Classes

Remark 1.4. Due to rounding errors, the length of the last class may be slightly different than the rest of them, even when we group data into classes of the same width.

2 Calculative Descriptive Statistics

In the last section, we have considered some graphical methods for getting an idea of the shape of the density function of the population from which the sample data was drawn. Some characteristics,



(a) $n = 6$ (b) $n = 12$

Fig. 2: Histogram and Frequency Polygon

such as symmetry, regularity can be observed from these graphical displays of the data. Next, we consider some statistics that allow us to summarize the data set analytically. It is hoped that these will give us some idea of the values of the parameters that characterize the entire population. We are looking mainly at two types of statistics: *measures of central tendency*, i.e. values that locate the observations with highest frequencies (so, where most of the data values lie) and *measures of variability* that indicate how much the values are spread out.

2.1 Measures of Central Tendency

These are values that tend to locate in some sense the "middle" of a set of data. The term "average" is often associated with these values. Each of the following measures of central tendency can be called the "average" value of a set of data.

Definition 2.1. *The **(arithmetic) mean*** (mean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_a = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{2.1}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_a = \frac{1}{N}\sum_{i=1}^{n} f_i x_i.$$

Remark 2.2. Some immediate properties of the arithmetic mean are the following:
1. The sum of all deviations from the mean is equal to 0. Indeed,

$$\sum_{i=1}^{N}(x_i - \overline{x}_a) = \sum_{i=1}^{N} x_i - N\overline{x}_a = 0.$$

2. The mean minimizes the mean square deviation, i.e. for every $a \in \mathbb{R}$,

$$\sum_{i=1}^{N}(x_i - a)^2 \geq \sum_{i=1}^{N}(x_i - \overline{x}_a)^2.$$

A straightforward computation leads to

$$\begin{aligned}\sum_{i=1}^{N}(x_i - a)^2 &= \sum_{i=1}^{N}\left[(x_i - \overline{x}_a) - (a - \overline{x}_a)\right]^2 \\ &= \sum_{i=1}^{N}(x_i - \overline{x}_a)^2 - 2(a - \overline{x}_a)\sum_{i=1}^{N}(x_i - \overline{x}_a) \\ &+ N\sum_{i=1}^{N}(a - \overline{x}_a)^2 \\ &\geq \sum_{i=1}^{N}(x_i - \overline{x}_a)^2,\end{aligned}$$

since the second term is 0 and the third term is always nonnegative.

Definition 2.3. *The **geometric mean*** (geomean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_g = \sqrt[N]{x_1 \ldots x_N}. \tag{2.2}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_g = \sqrt[N]{x_1^{f_1} \ldots x_n^{f_n}}.$$

The geometric mean is used in Economics Statistics for price study. One of its distinctive features is that it emphasizes the relative deviations from central tendency, as opposed to the absolute deviations, emphasized by the arithmetic mean.

Definition 2.4. *The **harmonic mean*** (harmmean) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}. \tag{2.3}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x}_h = \frac{N}{\sum_{i=1}^{n} \frac{f_i}{x_i}}.$$

The harmonic mean has applications in Economics Statistics in the study of time norms.

Remark 2.5.
1. For any set of data $x_1, \ldots, x_N$, the well-known *means inequality* holds:

$$\overline{x}_h \leq \overline{x}_g \leq \overline{x}_a,$$

with equality holding if and only if $x_1 = \ldots = x_N$.
2. The most widely used is the arithmetic mean. When nothing else is mentioned, we simply say *mean*, instead of *arithmetic mean*, and use the simplified notation $\overline{x}$.

Definition 2.6. *The **median*** (median) *is the value x_{me} that divides a set of ordered data X into two equal parts, i.e. the value with the property*

$$P(X < x_{me}) \leq \frac{1}{2} \leq P(X \leq x_{me}). \tag{2.4}$$

Remark 2.7. The median may or may not be one of the values in the data. If the sorted primary data is

$$x_1 \leq \ldots \leq x_N,$$

then

$$x_{me} = \begin{cases} x_{k+1}, & \text{if } N = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & \text{if } N = 2k \end{cases}.$$

Definition 2.8. *A **mode**, x_{mo}, of a set of data is a most frequent value.*

Remark 2.9.
1. Notice from the wording of the definition that the mode may not be unique. A set of data can have one mode, two modes − *bimodal data*, three modes − *trimodal data*, or more − *multimodal data*. If every value occurs only once, we say that there is *no mode*.
2. For perfectly symmetric distributions, we have

$$\overline{x} = x_{me} = x_{mo}.$$

This is true, for instance, for the Normal distribution. In general,

$$x_{mo} \approx \overline{x} - 3(\overline{x} - x_{me}).$$

2.2 Measures of Variability

Once we have located the "middle" of a set of data, it is important to measure the variability of the data, how unstable the data can be and how much the data values can differ from its expectation or from other middle values. These values will help us assess reliability of our estimates and accuracy of our forecasts.These measures of variation will have small values for closely grouped data (little variation) and larger values for more widely spread out data (large variation).

Consider the primary data $X = \{x_1, \ldots, x_N\}$. The first two measures of variation give a very general idea of the spread in the data values.

Definition 2.10. *The **range*** (range) *of X is the difference*

$$x_{max} - x_{min}.$$

If the values of X are sorted in increasing order, then the range is $x_N - x_1$.

Definition 2.11. *The **mean absolute deviation*** (mad) *of X is the value*

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |x_i - \overline{x}|.$$

Next, following the idea behind the definition of the median, we define values that divide the data into certain percentages.

Definition 2.12. *Let X be a set of data sorted increasingly.*

(1) *The **percentiles*** (prctile) *of X are the values $P_1, P_2, \ldots, P_{99}$ that divide the data into 100 equal parts, i.e. for $k = \overline{1, 99}$, P_k has the property*

$$P(X < P_k) \leq \frac{k}{100}, \ \frac{100 - k}{100} \leq P(X \leq P_k). \tag{2.5}$$

(2) *The **quartiles** of X are the values*

$$Q_1 = P_{25}, \ Q_2 = P_{50} = x_{me} \ \text{ and } \ Q_3 = P_{75}, \tag{2.6}$$

that divide the data into 4 equal parts.

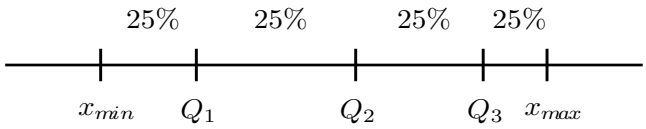


Fig. 3: Quartiles

Definition 2.13. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range*** (iqr) *is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{2.7}$$

(2) *The **interquartile deviation** or the **semi interquartile range** is the value*

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{2.8}$$

(3) *The **interquartile deviation coefficient** or the **relative interquartile deviation** is the value*

$$IQDC = \frac{IQD}{x_{me}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{2.9}$$

Remark 2.14.
1. The interquartile deviation is an absolute measure of variation and it has an important property: the range $x_{me} \pm IQD$ contains approximately 50% of the data.
2. The interquartile deviation coefficient $IQDC$ varies between -1 and 1, taking values close to 0 for symmetrical distributions, with little variation and values close to ± 1 for skewed data with large variation.

Outliers

The interquartile range is also involved in another important aspect of statistical analysis, namely the detection of outliers. An *outlier*, as the name suggests, is basically an atypical value, "far away" from the rest of the data, that does not seem to belong to the distribution of the rest of the values in the data set.
For example, in set of data where all values but one are between 0 and 1, a value of 1000 would surely seem out of place!

```
>> x=[rand(10,1); 1000]
x =

   1.0e+03 *

    0.0007
    0.0000
    0.0003
    0.0000
    0.0001
    0.0008
    0.0007
    0.0003
    0.0010
    0.0000
    1.0000
```

Outliers can arise for two reasons: either they are legitimate observations whose values are simply unusually large or unusually small, compared to the rest of the values in the data set, or they are the result of an error in measurement, of poor experimental techniques, or of mistakes in recording or

entering the data. Whichever the reason, they can adversely affect some values of the measures of central tendency and of variation, thus leading to erroneous inferential results.

```
>> mean(x)

ans =

   91.2707

>> median(x)

ans =

    0.3171
```

Once the presence of such outliers is detected, it is suggested that sample statistics be computed both with and without the outliers.

```
>> x = x(1:end-1)

x =

    0.7000
         0
    0.3000
         0
    0.1000
    0.8000
    0.7000
    0.3000
    1.0000
         0

>> mean(x)
```

```
ans =

    0.3900

>> median(x)

ans =

    0.3000
```

Thus the problem of detecting and locating an outlier is an important part of any statistical data analysis process.

For instance, one simple procedure would be to consider an outlier any value that is more than 2.5 standard deviations away from the mean, and an extreme outlier a value more than 3 standard deviations away from the mean. This procedure is justified by the "3σ rule" (the "3σ rule" is an application of Chebyshev's inequality and states that most of the values that any random variable takes, at least 89%, lie within 3 standard deviations away from the mean) and would work well for unimodal and symmetrical distributions.

A more general approach, that works for skewed data, is to consider an outlier any observation that is outside the range

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right] = \left[Q_1 - 3IQD,\ Q_3 + 3IQD\right].$$

Example 2.15. Consider the following set of data

$$\begin{array}{rrrrr} 0.5973 & 0.3624 & 0.8304 & 1.7347 & 1.2499 \\ 0.1104 & 0.8082 & 0.6039 & 0.3046 & 0.6183 \\ 0.0065 & 0.8748 & 1.3528 & 1.6458 & 1.5117 \\ 0.3253 & -2.0000 & -1.3000 & 1.7500 & 3.8500 \end{array}$$

We sort them in increasing order:

$$\begin{array}{rrrrr} -2.0000 & -1.3000 & 0.0065 & 0.1104 & 0.3046 \\ 0.3253 & 0.3624 & 0.5973 & 0.6039 & 0.6183 \\ 0.8082 & 0.8304 & 0.8748 & 1.2499 & 1.3528 \\ 1.5117 & 1.6458 & 1.7347 & 1.7500 & 3.8500 \end{array}$$

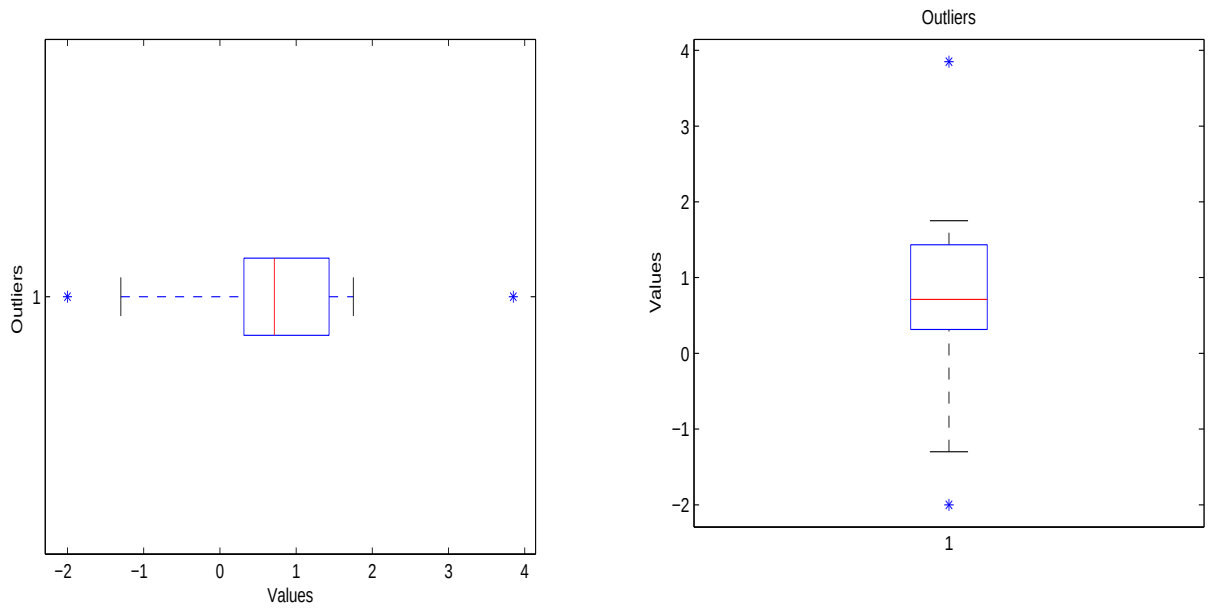
We have:

$$\begin{aligned} Q1 &= 0.3150, \\ Q2 &= 0.7133, \\ Q3 &= 1.4323, \end{aligned}$$

$$Q_1 - \frac{3}{2}IQR = -1.3610,$$

$$Q_3 + \frac{3}{2}IQR = 3.1082.$$

The data (boxplot) is displayed graphically in Figure 4.



(a) horizontally (b) vertically

Fig. 4: Quartiles, Interquartile Range, Outliers

Definition 2.16.

(1) *The **moment of order k** is the value*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k, \quad \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{2.10}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order k*** (moment) *is the value*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^k, \;\; \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^k \tag{2.11}$$

for primary and for grouped data, respectively.

(3) *The **variance*** (var) *is the value*

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2, \;\; \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2 \tag{2.12}$$

*for primary and for grouped data, respectively. The quantity $\overline{\sigma} = \sqrt{\overline{\sigma}^2}$ is the **standard deviation*** (std).

Remark 2.17.

1. We will see later that when the data represents a sample (not the entire population), a better formula for the variance is

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \;\; s^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2, \tag{2.13}$$

for the sample variance for primary or grouped data. The reason for that will have to do with the "bias" and will be explained later on in the next chapter. For now, we will just agree to use (2.12) to compute the variance of a set of data that represents a population and (2.13) for the variance of a sample.

2. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right), \tag{2.14}$$

which follows straight from the definition.

Definition 2.18. *The **coefficient of variation** is the value*

$$CV = \frac{\overline{\sigma}}{\overline{x}}.$$

Remark 2.19.
1. The coefficient of variation can be expressed as a ratio or as a percentage. It is useful in comparing the degrees of variation of two sets of data, even when their means are different.
2. The coefficient of variation is widely used in Biostatistics and Business Statistics. For example, in the investing world, the coefficient of variation helps brokers determine how much volatility (risk) they are assuming in comparison to the amount of return they can expect from a certain investment. The lower the value of the CV, the better the risk-return trade off.

3 Correlation and Regression

So far we have been discussing a number of descriptive techniques for describing one variable only. However, a very important part of Statistics is describing the association between two (or more) variables, whether or not they are independent, and if they are not, what is the nature of their dependence. One of the most fundamental concepts in statistical research is the concept of correlation.
Correlation is a measure of the relationship between one dependent variable, called the *response* variable and one or more independent variables, called *predictor* variables (or, simply, predictors). If two variables are correlated, this means that one can use information about one variable to predict the values of the other variable. **Regression** is then the method or statistical procedure that is used to establish that relationship.

3.1 Correlation, Curves of Regression

We will restrict our discussion to the case of two characteristics, X and Y. If X and Y have the same length, we can get a first idea of the relationship between the two, by plotting them in a **scattergram**, or **scatterplot**, which is a plot of the points with coordinates $(x_i, y_i)_{i=\overline{1,k}},\ x_i \in X,\ y_i \in Y,\ i = \overline{1,k}$. We group the N primary data into mn classes and denote by (x_i, y_j) the class mark and by f_{ij} the absolute frequency of the class (i, j), $i = \overline{1,m}$, $j = \overline{1,n}$. Then we represent the two-dimensional characteristic (X, Y) in a *correlation table*, or *contingency table*, as shown in Table 5.
Notice that

$$\sum_{j=1}^{n} f_{ij} = f_{i.},\ \sum_{i=1}^{m} f_{ij} = f_{.j},\ \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{n} f_{.j} = f_{..} = N.$$

Now we can define numerical characteristics associated with (X, Y).

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\dots$	f_{mj}	$\dots$	f_{mn}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

Table 5: Correlation Table

Definition 3.1. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5 and let $k_1, k_2 \in \mathbb{N}$.*

(1) *The **(initial) moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{3.1}$$

(2) *The **central moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{3.2}$$

where $\overline{x} = \overline{\nu}_{10} = \frac{1}{N} \sum_{i=1}^{m} f_{i.} x_i$ and $\overline{y} = \overline{\nu}_{01} = \frac{1}{N} \sum_{j=1}^{n} f_{.j} y_j$ are the means of X and Y, respectively.

Remark 3.2. Just as the means of the two characteristics X and Y can be expressed as moments of (X, Y), so can their variances:

$$\begin{aligned} \overline{\sigma}_X^2 &= \overline{\mu}_{20} = \overline{\nu}_{20} - \overline{\nu}_{10}^2, \\ \overline{\sigma}_Y^2 &= \overline{\mu}_{02} = \overline{\nu}_{02} - \overline{\nu}_{01}^2. \end{aligned}$$

Definition 3.3. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5.*

(1) *The **covariance*** ($\boxed{\text{cov}}$) *of (X, Y) is the value*

$$\text{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{3.3}$$

(2) *The **correlation coefficient*** ($\boxed{\text{corrcoef}}$) *of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{3.4}$$

These two notions have been mentioned before, for two random variables. They are defined similarly for sets of data and they have the same properties. The covariance gives a rough idea of the relationship between X and Y. As before, if X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0. If large values of X are associated with large values of Y, then the covariance will have a positive value, if, on the contrary, large values of X are associated with small values of Y, then the covariance will have a negative value. Also, an easier computational formula for the covariance is $\text{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y}$$

and, as before, it satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{3.5}$$

and, by its variation between -1 and 1, its value measures the linear relationship between X and Y. If $\overline{\rho}_{XY} = 1$, there is a *perfect positive correlation* between X and Y, if $\overline{\rho}_{XY} = -1$, there is a *perfect negative correlation* between X and Y. In both cases, the linearity is "perfect", i.e there exist $a, b \in \mathbb{R},\ a \neq 0$, such that $Y = aX + b$. If $\overline{\rho}_{XY} = 0$, then there is no linear correlation between X and Y, they are said to be *(linearly) uncorrelated*. However, in this case, they may not be independent, some other type of relationship (not linear) may exist between them.

In our task of finding a relationship between X and Y, we may go the following path: knowing the value of one of the characteristics, try to find a probable, an "expected" value for the other. If the two characteristics are related in any way, then there should be a pattern developing, that is the expected

value of one of them, conditioned by the other one taking a certain value, should be a function of that value that the other variable assumes. In other words, we should consider *conditional means*, defined similarly to regular means, only taking into account the condition.

Definition 3.4. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 5.*

(1) *The **conditional mean** of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{3.6}$$

(2) *The **conditional mean** of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{3.7}$$

Definition 3.5. *Let (X, Y) be a two-dimensional characteristic.*

(1) *The curve $y = f(x)$ formed by the points with coordinates $(x_i, \overline{y}_i)$, $i = \overline{1, m}$, is called the **curve of regression** of Y on X.*

(2) *The curve $x = g(y)$ formed by the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1, n}$, is called the **curve of regression** of X on Y.*

Remark 3.6. The curve of regression of a characteristic Y with respect to another characteristic X is then the mean value of Y, $\overline{y}(x)$, given $X = x$. The curve of regression is determined so that it approximates best the scatterplot of (X, Y).

3.2 Least Squares Estimation, Linear Regression

One of the most popular ways of finding curves of regression is the *least squares method*.
Assume the curve of regression of Y on X is of the form

$$y = y(x) = f(x; a_1, \ldots, a_s).$$

We determine the unknown parameters $a_1, \ldots, a_s$ so that the *sum of squares error* (SSE) (the sum of the squares of the differences between the responses y_j and their fitted values $y(x_i)$, each counted

with the corresponding frequency)

$$S = SSE = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - y(x_i)\Big)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)^2$$

is minimum (hence, the name of the method).
We find the point of minimum $(\overline{a}_1, \ldots, \overline{a}_s)$ of S by solving the system

$$\frac{\partial S}{\partial a_k} = 0,\ k = \overline{1, s},$$

i.e.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \tag{3.8}$$

for every $k = \overline{1, s}$.
Then the equation of the curve of regression of Y on X is

$$y = f\left(x; \overline{a}_1, \ldots, \overline{a}_s\right).$$

Let us consider the case of *linear regression* and find the equation of the *line of regression* of Y on X. We are finding a curve

$$y = ax + b,$$

for which

$$S(a, b) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - ax_i - b\Big)^2$$

is minimum. The system (3.8) becomes

$$\begin{cases} \left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}x_i^2\right)a + \left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}x_i\right)b = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}x_iy_j \\ \left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}x_i\right)a + \left(\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}\right)b = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} f_{ij}y_j \end{cases}$$

and after dividing both equations by N,

$$\begin{cases} \overline{\nu}_{20}a & + & \overline{\nu}_{10}b & = & \overline{\nu}_{11} \\ \overline{\nu}_{10}a & + & \overline{\nu}_{00}b & = & \overline{\nu}_{01}. \end{cases}$$

Its solution is

$$\overline{a} \;=\; \frac{\overline{\nu}_{11} - \overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\nu}_{20} - \overline{\nu}_{10}^2} \;=\; \frac{\overline{\nu}_{11} - \overline{x}\cdot\overline{y}}{\overline{\sigma}_X^2} \;=\; \frac{\overline{\nu}_{11} - \overline{x}\cdot\overline{y}}{\overline{\sigma}_X\overline{\sigma}_Y}\cdot\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} \;=\; \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X},$$

$$\overline{b} \;=\; \overline{\nu}_{01} - \overline{\nu}_{10}\overline{a} \;=\; \overline{y} - \overline{a}\cdot\overline{x}.$$

So the equation of the line of regression of Y on X is

$$y - \overline{y} = \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x - \overline{x}) \tag{3.9}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x - \overline{x} = \overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y - \overline{y})\,. \tag{3.10}$$

Remark 3.7.

1. The point of intersection of the two lines of regression, $(\overline{x}, \overline{y})$, is called the *centroid* of the distribution of the characteristic (X, Y).

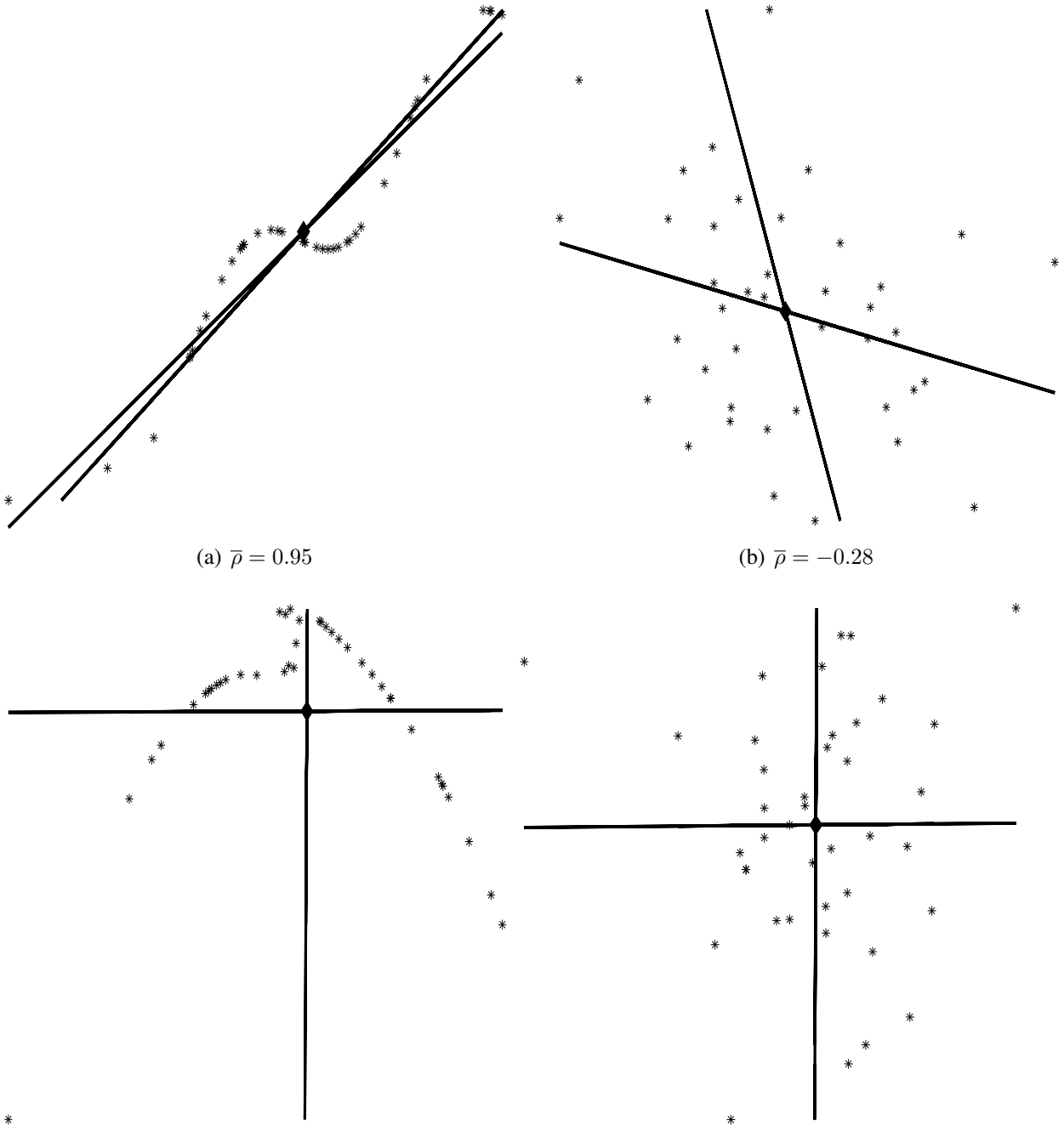
2. The slope $\overline{a}_{Y|X} = \overline{\rho}\,\dfrac{\overline{\sigma}_Y}{\overline{\sigma}_X}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y} = \overline{\rho}\,\dfrac{\overline{\sigma}_X}{\overline{\sigma}_Y}$ is the coefficient of regression of X on Y and

$$\overline{\rho}^2 = \overline{a}_{Y|X}\;\overline{a}_{X|Y}.$$

3. For the angle α between the two lines of regression, we have

$$\tan\alpha \;=\; \frac{1 - \overline{\rho}^2}{\overline{\rho}^2}\cdot\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2 + \overline{\sigma}_Y^2}.$$

So, if $|\,\overline{\rho}\,| = 1$, then $\alpha = 0$, i.e. the two lines coincide. If $|\,\overline{\rho}\,| = 0$ (for instance, if X and Y are independent), then $\alpha = \dfrac{\pi}{2}$, i.e. the two lines are perpendicular.



(a) $\overline{\rho} = 0.95$

(b) $\overline{\rho} = -0.28$

(c) $\overline{\rho} = 0$

(d) $\overline{\rho} = 0$

Fig. 5: Scattergram, Lines of Regression and Centroid

Example 3.8. Let us examine the situations graphed in Figure 5.

- In Figure 5(a) $\overline{\rho} = 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 5(b) $\overline{\rho} = -0.28$, negative and fairly small, close to 0. If a relationship exists between X and Y, it does not seem to be linear. In fact, they are very close to being independent, since the points are scattered around the plane, no pattern being visible. The two lines of regression are very distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y.

- In Figure 5(c) $\overline{\rho} = 0$, so the two characteristics are uncorrelated, no linear relationship exists between them. However they are not independent, they were chosen so that $Y = -X^2 + \sin\left(\frac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.

- Finally, in Figure 5(d) $\overline{\rho} = 0$, again, so no linear relationship exists. In fact the two characteristics are independent, which is suggested by their random scatter inside the plane.

Remark 3.9. Other types of curves of regression that are fairly frequently used are
– *exponential* regression $y = ab^x$,
– *logarithmic* regression $y = a \log x + b$,
– *logistic* regression $y = \dfrac{1}{ae^{-x} + b}$,
– *hyperbolic* regression $y = \dfrac{a}{x} + b$.

Chapter 6. Statistical Inference

1 Sample Theory

In inferential Statistics, we will have the following situation: we are interested in studying a characteristic (a random variable) X, relative to a population P of (known or unknown) size N. The difficulty or even the impossibility of studying the entire population, as well as the merits of choosing and studying a random sample from which to make inferences about the population of interest, have already been discussed in the previous chapter. Now, we want to give a more rigorous and precise definition of a random sample, in the framework of random variables, one that can then employ probability theory techniques for making inferences.

We choose n objects from the population and actually study $X_i,\ i = \overline{1,n}$, the characteristic of interest *for the i^{th} object selected*. Since the n objects were randomly selected, it makes sense that for $i = \overline{1,n}$, X_i is a random variable, one that has *the same* distribution (pdf) as X, the characteristic relative to the entire population. Furthermore, these random variables are independent, since the value assumed by one of them has no effect on the values assumed by the others. Once the n objects have been selected, we will have n numerical values available, $x_1, \ldots, x_n$, the observed values of $X_1, \ldots, X_n$.

Definition 1.1. *A **random sample of size n** from the distribution of X, a characteristic relative to a population P, is a collection of n independent random variables $X_1, \ldots, X_n$, having the same distribution as X. The variables $X_1, \ldots, X_n$, are called **sample variables** and their observed values $x_1, \ldots, x_n$, are called **sample data**.*

We are able now to define sample functions, or statistics, in the more precise context of random variables.

Definition 1.2. *A **sample function** or **statistic** is a random variable*

$$Y_n = h_n(X_1, \ldots, X_n),$$

where $h_n : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. The value of the sample function Z_n is $y_n = h_n(x_1, \ldots, x_n)$.

We will revisit now some sample numerical characteristics discussed in the previous chapter and define them as sample functions. That means they will have a pdf, a cdf, a mean value, variance,

standard deviation, etc. A sample function will, in general, be an approximation for the corresponding population characteristic.

In what follows, $\{X_1, \ldots, X_n\}$ denotes a sample of size n drawn from the distribution of some population characteristic X.

Definition 1.3. *The **sample mean** is the sample function defined by*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.1}$$

and its value is $\overline{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$.

Now that the sample mean is defined as a random variable, we can discuss its distribution and its numerical characteristics.

Proposition 1.4. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$. Then*

$$E\left(\overline{X}\right) = \mu \text{ and } V\left(\overline{X}\right) = \frac{\sigma^2}{n}. \tag{1.2}$$

Moreover, if $X \in N(\mu, \sigma)$, then $\overline{X} \in N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Proof. Since $X_1, \ldots, X_n$ are identically distributed, with the same distribution as X, $E(X_i) = E(X) = \mu$ and $V(X_i) = V(X) = \sigma^2$, $\forall i = \overline{1, n}$. Then, by the usual properties of expectation, we have

$$E\left(\overline{X}\right) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu.$$

Further, since $X_1, \ldots, X_n$ are also independent, by the properties of variance, it follows that

$$V\left(\overline{X}\right) = V\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}.$$

The last part follows from the fact that $\overline{X}$ is a linear combination of independent, Normally distributed random variables.

□

Corollary 1.5. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$ and for $n \in \mathbb{N}$ let*

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

Then the variable Z_n converges in distribution to a Standard Normal variable, as $n \to \infty$, i.e. $F_{Z_n} \overset{n\to\infty}{\longrightarrow} F_Z = \Phi$. Moreover, if $X \in N(\mu, \sigma)$, then the statement is true for every $n \in \mathbb{N}$.

Definition 1.6. *The statistic*

$$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{1.3}$$

*is called the **sample moment of order k** and its value is $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.*

The statistic

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \tag{1.4}$$

*is called the **sample central moment of order k** and its value is $\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k$.*

Remark 1.7. Just like for theoretical (population) moments, we have

$$\begin{aligned} \overline{\nu}_1 &= \overline{X}, \\ \overline{\mu}_1 &= 0, \\ \overline{\mu}_2 &= \overline{\nu}_2 - \overline{\nu}_1^2. \end{aligned}$$

Proposition 1.8. *Let X be a characteristic with the property that for $k \in \mathbb{N}$, the theoretical moment $\nu_{2k} = \nu_{2k}(X) = E\left(X^{2k}\right)$ exists. Then*

$$E\left(\overline{\nu}_k\right) = \nu_k \text{ and } V\left(\overline{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \tag{1.5}$$

Corollary 1.9. *Let X be a characteristic as in Proposition 1.8 and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{\nu}_k - \nu_k}{\sqrt{\frac{\nu_{2k} - \nu_k^2}{n}}}.$$

Then $Z_n \xrightarrow{d} Z$, as $n \to \infty$.

Proposition 1.10. *Let X be a characteristic with $V(X) = \mu_2 = \sigma^2$ and for which the theoretical moment $\nu_4 = E\left(X^4\right)$ exists. Then*

$$\begin{aligned} E\left(\overline{\mu}_2\right) &= \frac{n-1}{n}\,\sigma^2, \\ V\left(\overline{\mu}_2\right) &= \frac{n-1}{n^3}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big]. \end{aligned} \tag{1.6}$$

Remark 1.11. Notice that the sample central moment of order 2 is the first statistic whose expected value *is not* the corresponding population function, in this case the theoretical variance. This is the motivation for the next definition.

Definition 1.12. *The statistic*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.7}$$

*is called the **sample variance** and its value is* $\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$.

*The statistic $s = \sqrt{s^2}$ is called the **sample standard deviation**.*

Remark 1.13. With this definition, we have for the sample variance

$$E\left(s^2\right) = \mu_2 = \sigma^2. \tag{1.8}$$

So, for the rest of this chapter, we will use these notations:

Function	Population (theoretical)	Sample
Mean	$\mu = E(X)$	$\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$
Variance	$\sigma^2 = V(X)$	$s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$
Standard deviation	$\sigma = \sqrt{V(X)}$	$s = \sqrt{s^2}$
Moment of order k	$\nu_k = E\left(X^k\right)$	$\overline{\nu}_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k$
Central moment of order k	$\mu_k = E\left[(X - E(X))^k\right]$	$\overline{\mu}_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^k$

Table 1: Notations

2 Estimation; Basic Notions

We will refer to the parameter to be estimated as the **target parameter** and denote it by θ.
Two types of estimation will be considered: **point estimate**, when the result of the estimation is one single value and **interval estimate**, when the estimate is an interval enclosing the value of the target parameter. In either case, the actual estimation is accomplished by an **estimator**, a rule, a formula, or a procedure that leads us to the value of an estimate, based on the data from a sample.

Throughout this chapter, we consider a characteristic X (relative to a population), whose pdf $f(x;\theta)$ depends on the parameter θ, which is to be estimated.

As before, we consider a random sample of size n, i.e. sample variables $X_1, \ldots, X_n$, which are **independent and identically distributed (iid)**, having the same pdf as X.

A **point estimator** for (the estimation of) the target parameter θ is a sample function (statistic)

$$\overline{\theta} = \overline{\theta}(X_1, X_2, \ldots, X_n).$$

Other notations may be used, such as $\hat{\theta}$ or $\tilde{\theta}$.

Each statistic is a random variable because it is computed from random data. It has a so-called *sampling distribution* (a pdf). Each statistic estimates the corresponding population parameter and adds certain information about the distribution of X, the variable of interest. The value of the point estimator, the **point estimate**, is the actual approximation of the unknown parameter.

Many different point estimators may be obtained for the same target parameter. Some are considered "good", others "bad", some "better" than others. We need some criteria to decide on one estimator versus another.

For one thing, it is highly desirable that the sampling distribution of an estimator $\overline{\theta}$ to be "clustered" around the target parameter. In simple terms, we *expect* that the value the point estimator provides to be the actual value of the parameter it estimates. This justifies the following notion.

Definition 2.1. *A point estimator $\overline{\theta}$ is called an **unbiased** estimator for θ if*

$$E(\overline{\theta}) \;\; = \;\; \theta. \tag{2.1}$$

*The **bias** of $\overline{\theta}$ is the value $B = E(\overline{\theta}) - \theta$.*

Unbiasedness means that in the long-run, collecting a large number of samples and computing $\overline{\theta}$ from each of them, on the average we hit the unknown parameter θ exactly. In other words, in a long run, unbiased estimators neither underestimate nor overestimate the parameter.

Example 2.2.
1. Recall from Proposition 1.4. that for the sample mean, as a random variable, we have $E(\overline{X}) = \mu$. Thus the sample mean is an unbiased estimator for the population mean.
2. Also, by Proposition 1.8., the sample moment of order k is an unbiased estimator for the theoretical moment of order k. 3. By Proposition Proposition 1.10., the sample central moment of order 2 *is not* an unbiased estimator for the population central moment of order 2 (or it is a *biased* estimator), since

$$E(\overline{\mu}_2) = \frac{n-2}{n}\mu_2 \neq \mu_2 = \sigma^2.$$

3. However, the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator for the population variance, since $E(s^2) = \sigma^2$. That was the main reason for the way the sample variance was defined.

Another desirable trait for a point estimator is that its values do not vary too much from the value of the target parameter. So we need to evaluate variability of computed statistics and especially parameter estimators. That can be accomplished by computing the following statistic.

Definition 2.3. *The **standard error** of an estimator $\overline{\theta}$, denoted by $\sigma_{\overline{\theta}}$, is its standard deviation*

$$\sigma_{\overline{\theta}} = \sigma(\overline{\theta}) = \mathrm{Std}(\overline{\theta}) = \sqrt{V(\overline{\theta})}.$$

Both population and sample variances are measured in squared units. Therefore, it is convenient to have standard deviations that are comparable with our variable of interest, X. As a measure of variability, standard errors show precision and reliability of estimators. They show how much estimators of the same target parameter θ can vary if they are computed from different samples. Ideally, we would like to deal with unbiased or nearly unbiased estimators that have *low* standard error.

3 Estimation by Confidence Intervals

3.1 Confidence Intervals, General Framework

Point estimators provide one single value, $\overline{\theta}$, to estimate the value of an unknown parameter θ, but little measure of the accuracy of the estimate. In contrast, an **interval estimator** specifies a *range* of values, within which the parameter is estimated to lie. More specifically, the sample will be used to produce *two* sample functions, $\overline{\theta}_L(X_1, \ldots, X_n) < \overline{\theta}_U(X_1, \ldots, X_n)$, with values $\overline{\theta}_L = \overline{\theta}_L(x_1, \ldots, x_n)$, $\overline{\theta}_U = \overline{\theta}_U(x_1, \ldots, x_n)$, respectively, such that for a given $\alpha \in (0, 1)$,

$$P(\overline{\theta}_L \leq \theta \leq \overline{\theta}_U) \quad = \quad 1 - \alpha. \tag{3.1}$$

Then
– the range $[\overline{\theta}_L, \overline{\theta}_U]$ is called a **confidence interval (CI)**, more specifically, a $100(1-\alpha)\%$ confidence interval,
– the values $\overline{\theta}_L, \overline{\theta}_U$ are called (lower and upper) **confidence limits**,
– the quantity $1 - \alpha$ is called **confidence level** or **confidence coefficient** and
– the value α is called **significance level**.

Remark 3.1.
1. It may seem a little peculiar that we use $1 - \alpha$ instead of simply α in (3.1), since both values are in $(0, 1)$, but the reasons are in close connection with *hypothesis testing* and will be revealed in the next sections.
2. The condition (3.1) *does not* uniquely determine a $100(1-\alpha)\%$ CI.
3. Evidently, the smaller α and the length of the interval $\overline{\theta}_U - \overline{\theta}_L$ are, the better the estimate for θ. Unfortunately, as we will see, as the confidence level increases, so does the length of the CI, thus, reducing accuracy.

To produce a CI estimate for θ, we need a *pivotal quantity*, i.e. a statistic S that satisfies two conditions:
– $S = S(X_1, \ldots, X_n; \theta)$ is a function of the sample measurements and the unknown parameter θ, this being the *only* unknown,
– the distribution of S is known and does not depend on θ.

We will use the pivotal method to find $100(1-\alpha)\%$ CI's. We start with the case where the pivot has a $N(0, 1)$ distribution, so we can better understand the ideas.

Let θ be a target parameter and let $\overline{\theta}$ be an unbiased estimator for θ ($E(\overline{\theta}) = \theta$), with standard

error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is known that

$$Z \;=\; \frac{\overline{\theta}-\theta}{\sigma_{\overline{\theta}}} \;\left(=\; \frac{\overline{\theta}-E(\overline{\theta})}{\sigma(\overline{\theta})}\right) \tag{3.2}$$

has an approximately Standard Normal $N(0,1)$ distribution. We can use Z as a pivotal quantity to construct a $100(1-\alpha)\%$ CI for estimating θ. Since the pdf of Z is known, we can choose two values, Z_L, Z_U such that for a given $\alpha \in (0,1)$,

$$P(Z_L \leq Z \leq Z_U) \;=\; 1-\alpha. \tag{3.3}$$

How to choose them? Of course, there are infinitely many possibilities. Recall that for continuous random variables, the probability in (3.3) is an *area*, namely the area under the graph of the pdf and above the x-axis, between the values Z_L and Z_U. Basically, the values Z_L and Z_U should be chosen so that that area is $1-\alpha$. We will take advantage of the symmetry of the Standard Normal pdf and choose the two values so that the area $1-\alpha$ is in "the middle". That means (since the total area under the graph is 1) the two portions left on the two sides, both should have an area of $\dfrac{\alpha}{2}$, as seen in Figure 1.

So what should the values be? Recall *quantiles*. A quantile of a given order $\beta \in (0,1)$ for a random variable X, is a value q_β with the property that

$$F(q_\beta) \;=\; P(X \leq q_\beta) \;=\; \beta,$$

i.e., that the area under the graph of the pdf, to the *left* of q_β is β.

Since for Z_L we want the area to its left to be $\alpha/2$, we choose it to be the quantile of order $\alpha/2$ for Z,

$$Z_L \;=\; z_{\alpha/2}.$$

For the value Z_U, the area to its *right* should be $\alpha/2$, which means the area to the left is $1-\alpha/2$. Thus, we choose

$$Z_U \;=\; z_{1-\alpha/2}.$$

Indeed, now we have

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) \;=\; 1-\alpha,$$

as in (3.3).

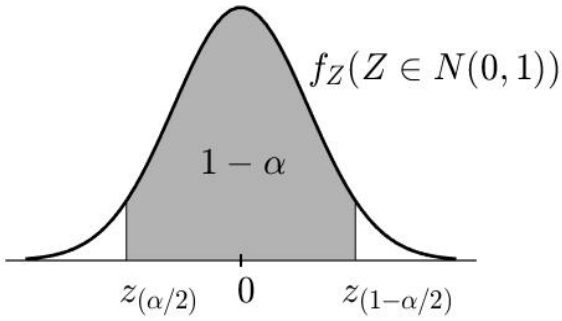


Fig. 1: Confidence Interval for $N(0,1)$ distribution

From here, we proceed to rewrite the inequality inside, until we get the limits of the CI for θ. We have

$$\begin{aligned} 1-\alpha &= P\left(z_{\frac{\alpha}{2}} \leq \frac{\overline{\theta}-\theta}{\sigma_{\overline{\theta}}} \leq z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}} \leq \overline{\theta}-\theta \leq \sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}\right) \\ &= P\left(-\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \leq \theta-\overline{\theta} \leq -\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right) \\ &= P\left(\overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}} \leq \theta \leq \overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right), \end{aligned}$$

so the $100(1-\alpha)\%$ CI for θ is given by

$$\left[\overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}},\ \overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right]. \tag{3.4}$$

Remark 3.2.

1. Since the Standard Normal distribution is symmetric about the origin, $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$ and the CI can be written as

$$\left[\overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}},\ \overline{\theta}+\sigma_{\overline{\theta}} \cdot z_{1-\frac{\alpha}{2}}\right] \quad \text{or} \quad \left[\overline{\theta}+\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}},\ \overline{\theta}-\sigma_{\overline{\theta}} \cdot z_{\frac{\alpha}{2}}\right].$$

2. The CI we determined is a **two-sided CI**, because it gives bounds on both sides. A two-sided CI is not always the most appropriate for the estimation of a parameter θ. It may be more relevant to make a statement simply about how *large* or how *small* the parameter might be, i.e. to find confidence intervals of the form $(-\infty, \overline{\theta}_U]$ and $[\overline{\theta}_L, \infty)$, respectively, such that the probability that

θ is in the CI is $1-\alpha$. These are called **one-sided confidence intervals** and they can be found the same way, using quantiles of an appropriate order.
3. In what follows, for estimating various population parameters, the pivot will be different, but the procedure of finding the CI will be the same, even when the distribution of the pivot *is not* symmetric.

3.2 Confidence Intervals for the Mean and Variance of One Population

Let X be a population characteristic, with mean $\mu = E(X)$ and variance $V(X) = \sigma^2$, whose pdf depends on a parameter θ, $f(x;\theta)$. Let $X_1, X_2, \ldots, X_n$ be a sample drawn from the pdf of X.
The formulas for finding confidence intervals for the mean μ and variance σ^2 are based on the following results (which follow either from properties of random variables, or are the consequence of some CLT).

Proposition 3.3. *Assume $X \in N(\mu, \sigma)$. Then*

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1),\ T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1) \text{ and}$$
$$V = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)\,s^2}{\sigma^2} \in \chi^2(n-1).$$

Proposition 3.4. *If the sample size is large enough ($n > 30$), then*

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1) \text{ and } T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1).$$

CI for the mean, known variance
If either $X \in N(\mu, \sigma)$ or the sample is large enough ($n > 30$) and σ is known, then by Propositions 3.3 and 3.4, we can use the pivot

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in N(0,1).$$

The procedure will go *exactly* as described in the previous section, with $\theta = \mu, \overline{\theta} = \overline{X}, \sigma_{\overline{\theta}} = \frac{\sigma}{\sqrt{n}}$.

The $100(1-\alpha)\%$ CI for the mean is given by

$$\mu \in \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \tag{3.5}$$

Since $N(0,1)$ is symmetric (and one quantile is the negative of the other), we can write it in short as

$$\overline{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}. \tag{3.6}$$

CI for the mean, unknown variance

In practice, it is somewhat unreasonable to expect to know the value of σ, if the value of μ is unknown. We can find CI's for the mean, without knowing the variance. If either $X \in N(\mu,\sigma)$ or the sample is large enough ($n > 30$), then by Propositions 3.3 and 3.4, we can use the pivot

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \in T(n-1).$$

The same computations as before will lead to the $100(1-\alpha)\%$ CI for the mean:

$$\mu \in \left[\overline{X} - t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\ \overline{X} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]. \tag{3.7}$$

Notice that we change the notations for the quantiles, according to the pdf of the pivot (z for $N(0,1)$, t for $T(n-1)$, etc.). The Student $T(n-1)$ is also symmetric, so again, we can write the CI in short as

$$\overline{X} \pm t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \quad \text{or} \quad \overline{X} \mp t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}. \tag{3.8}$$

CI for the variance

By Proposition 3.3, if $X \in N(\mu,\sigma)$, then we can use the pivot

$$V = \frac{(n-1)\,s^2}{\sigma^2} \in \chi^2(n-1).$$

Let us see how to do that. Even though the $\chi^2(n-1)$ is not symmetric (see Figure 2), so we cannot really talk about the "middle" for the area, we can still use the quantiles as before. So, we have:

$$1-\alpha = P\left(\chi^2_{\frac{\alpha}{2}} \le V \le \chi^2_{1-\frac{\alpha}{2}}\right)$$

$$
\begin{aligned}
&= P\left(\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)\,s^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right) \\
&= P\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2}}} \leq \frac{\sigma^2}{(n-1)\,s^2} \leq \frac{1}{\chi^2_{\frac{\alpha}{2}}}\right) \\
&= P\left(\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}\right).
\end{aligned}
$$

Thus, a $100(1-\alpha)\%$ CI for the variance is

$$
\sigma^2 \in \left[\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}},\ \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}\right] \tag{3.9}
$$

and one for the standard deviation is

$$
\sigma \in \left[\sqrt{\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}}},\ \sqrt{\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}}}\right] \tag{3.10}
$$

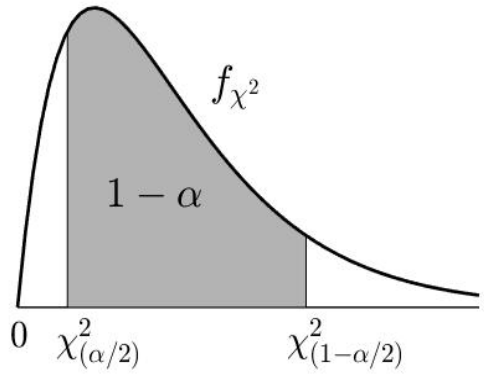


Fig. 2: Confidence Interval for χ^2 distribution

Remark 3.5.

1. Remember, "χ^2_α" is just a notation for the quantile of order α for the $\chi^2(n-1)$ distribution, it *does not* mean you have to take the square of it!

2. Since the $\chi^2(n-1)$ is no longer symmetric, there is no relationship between the two quantiles, we have to use *both* and there is no shorter writing for the CI for the variance than the one in (3.9)

(or (3.10) for the standard deviation).

Example 3.6. The time spent for finding a parking space downtown Cluj-Napoca during the week was recorded for 64 drivers. The average and variance were found to be 15 minutes and 256 minutes, respectively. Find a 95% confidence interval for the true average time spent to find a parking spot during the week in downtown Cluj-Napoca.

Solution. The population here is the set of times spent to find a parking space downtown Cluj by *all* people who need to park downtown. We want to estimate its *average*, so the mean μ.
For our sample, $n = 64$, $\overline{X} = 15$ and $s^2 = 256$. To attain a confidence level of $1 - \alpha = 0.95$, we need $\alpha = 0.05$ and $\alpha/2 = 0.025$. Since σ is not known, we use formula (3.7) (or, actually, (3.8)). The quantiles for the $T(63)$ distribution are

$$t_{0.025} \ = \ -1.9983, \ t_{0.975} \ = \ 1.9983$$

and the 95% CI for the mean is

$$\left[\overline{X} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] \ = \ [11.0034, \ 18.9966].$$

So

$$\mu \in [11.0034, \ 18.9966] \, ,$$

with probability 0.95. The interpretation is that 95% of the drivers spend, on average, between 11.0034 and 18.9966 minutes trying to find a parking space downtown Cluj-Napoca during the week.

■

3.3 Confidence Intervals for Comparing Means and Variances of Two Populations

It will be necessary sometimes to compare characteristics of two populations. For that, we will need results on sample functions referring to both collections.

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, relative to two populations, with means $\mu_1 = E(X_{(1)}), \mu_2 = E(X_{(2)})$ and variances $\sigma_1^2 = V(X_{(1)}), \sigma_2^2 = V(X_{(2)})$, respectively.
We draw from both populations random samples of sizes n_1 and n_2, respectively, that are **independent**. Denote the two sets of random variables by

$$X_{11}, \ldots, X_{1n_1} \text{ and } X_{21}, \ldots, X_{2n_2}.$$

Then we have two sample means and two sample variances, given by

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

and

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2, \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2,$$

respectively. In addition, denote by

$$s_p^2 = \frac{\sum_{i=1}^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j} - \overline{X}_2\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the **pooled variance** of the two samples, i.e. a variance that considers the sample data from both samples.

In inferential Statistics, when comparing the means of two populations, we estimate their *difference* and when comparing the variances, we estimate their *ratio*.
The formulas for finding confidence intervals for the difference of means $\mu_1 - \mu_2$ and for the ratio of variances $\frac{\sigma_1^2}{\sigma_2^2}$ are based on the following results (which follow either from properties of random variables, or are the consequence of some CLT).

Proposition 3.1. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$. Then*

a) $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \in N(0,1)$;

b) $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \in T(n_1 + n_2 - 2)$;

c) $T^* = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \in T(n)$, *where* $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ *and* $c = \dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$;

d) $F = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1)$.

Proposition 3.2. *If the samples are large enough ($n_1 + n_2 > 40$), then parts a), b) and c) of Proposition 3.1 still hold.*

CI for the difference of means

Case σ_1, σ_2 known

If either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$) and σ_1, σ_2 are known, then by Propositions 3.1 and 3.2, we can use the pivot

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \in N(0,1).$$

With the same line of computations as before, we find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \ \in \ \left[\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{3.1}$$

or, using symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right], \tag{3.2}$$

where the quantiles $z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}$ refer to the $N(0,1)$ distribution.

Case $\sigma_1 = \sigma_2$ unknown

Assume that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$). The population variances are *not* known anymore, but they are known to be equal. Then each is approximated by the pooled variance s_p^2. Then by Propositions 3.1 and 3.2, we use the pivot

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in T(n_1 + n_2 - 2).$$

A $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\mu_1 - \mu_2 \quad \in \quad \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right], \tag{3.3}$$

where the quantiles $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n_1 + n_2 - 2)$ distribution. Again, by symmetry we can write the CI in short as

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{3.4}$$

Case σ_1, σ_2 unknown

Assuming that either $X_{(1)} \in N(\mu_1, \sigma_1),\ X_{(2)} \in N(\mu_2, \sigma_2)$ or the samples are large enough ($n_1 + n_2 > 40$), by Propositions 3.1 and 3.2, we use the pivot

$$T^* = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \in T(n),$$

where $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ and $c = \dfrac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

We find a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ as

$$\mu_1 - \mu_2 \quad \in \quad \left[\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right], \tag{3.5}$$

or, by symmetry,

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \tag{3.6}$$

where the quantile $t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}$ refer to the $T(n)$ distribution, with n given above.

CI for the ratio of variances

Assume the two independent samples were drawn from approximately Normal distributions $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively. By Proposition 3.2, we use the pivot

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \in F(n_1 - 1, n_2 - 1).$$

A $100(1-\alpha)\%$ CI for $\dfrac{\sigma_1^2}{\sigma_2^2}$ is given by

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{1}{f_{1-\frac{\alpha}{2}}} \cdot \frac{s_1^2}{s_2^2}, \ \frac{1}{f_{\frac{\alpha}{2}}} \cdot \frac{s_1^2}{s_2^2}\right] \tag{3.7}$$

and, from here, a $100(1-\alpha)\%$ CI for $\dfrac{\sigma_1}{\sigma_2}$ is

$$\frac{\sigma_1}{\sigma_2} \in \left[\sqrt{\frac{1}{f_{1-\frac{\alpha}{2}}}} \cdot \frac{s_1}{s_2}, \ \sqrt{\frac{1}{f_{\frac{\alpha}{2}}}} \cdot \frac{s_1}{s_2}\right], \tag{3.8}$$

where the quantiles $f_{\frac{\alpha}{2}}, f_{1-\frac{\alpha}{2}}$ refer to the $F(n_1 - 1, n_2 - 1)$ distribution.

Example 3.3. An account on server A is more expensive than an account on server B. However, server A is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is. A certain computer algorithm is executed 30 times on server A and 20 times on server B with the following results:

Server A			Server B		
n_1	$=$	30	n_2	$=$	20
$\overline{X}_1$	$=$	6.7 min	$\overline{X}_2$	$=$	7.5 min
s_1	$=$	0.6 min	s_2	$=$	1.2 min

a) Construct a 95% confidence interval for the difference $\mu_1 - \mu_2$ between the mean execution times on server A and server B.

b) Assuming that the observed times are approximately Normal, find a 95% confidence interval for the ratio of the two population standard deviations.

Solution.

a) The samples are large enough ($n_1 + n_2 = 50$), that we can use Proposition 3.2. Nothing is said about the population variances (that they might be known, or known to be equal). Also, the second sample standard deviation is twice as large as the first one, therefore, equality of population variances can hardly be assumed. We use the general case for unknown, unequal variances and use formula (3.6).
We want confidence level $1 - \alpha = 0.95$, so $\alpha = 0.05$ and $\alpha/2 = 0.025$.
The parameter n in (3.6) is found to be

$$n \;=\; 25.3989 \;\approx\; 25.$$

For the $T(25)$ distribution, we find the quantile

$$t_{0.025} \;=\; -2.0595.$$

Them the 95% CI for the difference of means is

$$\left[\overline{X}_1 - \overline{X}_2 \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right] \;=\; \left[6.7 - 7.5 \pm 2.06\sqrt{\frac{0.6^2}{30} + \frac{1.2^2}{20}}\right] \;=\; [-0.8 \pm 0.505],$$

so,

$$\mu_1 - \mu_2 \;\in\; [-1.305,\; -0.295]$$

with probability 0.95. Since *all* values in the CI are negative, with high probability, it seems that $\mu_1 - \mu_2 < 0$, so indeed the first server seems to be faster, on average.

b) Since now the times are assumed to be approximately Normal, we can use formula (3.8). For the $F(29, 19)$ distribution, the quantiles are

$$f_{0.025} \;=\; 0.4482,\;\; f_{0.975} \;=\; 2.4019.$$

Now,

$$\begin{aligned}\frac{s_1}{s_2} \;&=\; \frac{0.6}{1.2} \;=\; 0.5,\\ \frac{s_1^2}{s_2^2} \;&=\; \frac{0.36}{1.44} \;=\; 0.25.\end{aligned}$$

Then, the 95% CI for the ratio of variances is

$$\left[\frac{1}{f_{1-\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2},\ \frac{1}{f_{\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2}\right] = \left[\frac{1}{2.4019}\cdot 0.25,\ \frac{1}{0.4482}\cdot 0.25\right] = [0.104,\ 0.558]$$

and the 95% CI for the ratio of standard deviations is

$$\left[\sqrt{\frac{1}{f_{1-\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2},\ \sqrt{\frac{1}{f_{\frac{\alpha}{2}}}}\cdot\frac{s_1}{s_2}\right] = \left[\sqrt{\frac{1}{2.4019}}\cdot 0.5,\ \sqrt{\frac{1}{0.4482}}\cdot 0.5\right] = [0.323,\ 0.747].$$

■

4 Hypothesis Testing

In the previous sections we have considered the basic ideas of parameter estimation in some detail. We attempted to approximate the value of some population parameter θ, based on a sample, *without* having any predetermined notion concerning the actual value of this parameter. We simply tried to ascertain its value, to the best of our ability, from the information given by a random sample. In contrast, **statistical hypothesis testing** is a method of making statistical inferences on some unknown population characteristic, when *there is* a preconceived notion concerning its value or its properties.

Based on a random sample, we can use Statistics to verify a various number of statements, such as:

− the average connection speed is as claimed by the internet service provider,

− a system has not been infected,

− the proportion of defective products is at most a certain percentage, as promised by the manufacturer,

− a hardware upgrade was efficient,

− service times have a certain distribution,

− the average number of customers has increased by a certain number this year, etc.

Testing statistical hypotheses has wide applications far beyond Mathematics or Computer Science. These methods can be used to prove efficiency of a new medical treatment, safety of a new automobile brand, innocence of a defendant, authorship of a document; to establish cause-and-effect relationships; to identify factors that can significantly improve performance; to detect information leaks; and so forth.

4.1 Basic Concepts

So, we will work with **statistical hypotheses**, about some characteristic X (relative to a population), whose pdf $f(x;\theta)$ depends on the parameter θ, which is to be estimated.
The method(s) used to decide whether a hypothesis is true or not (in fact, to decide whether to *reject* a hypothesis or not) make up the **hypothesis test**. To begin with, we need to state *exactly* what we are testing. Any hypothesis test will involve two theories, two hypotheses,
$-$ the **null hypothesis**, denoted by H_0 and
$-$ the **alternative (research) hypothesis**, denoted by H_1 (or H_a).
A null hypothesis is always an equality, showing absence of an effect or relation, some "normal" usual statement that people have believed in for years. The alternative is the opposite (in some way) of the null hypothesis, a "new" theory proposed by the researcher to "challenge" the old one. In order to overturn the common belief and to reject the null hypothesis, *significant* evidence is needed. Such evidence can only be provided by data. Only when such evidence is found, and when it *strongly* supports the alternative H_1, can the hypothesis H_0 be rejected in favor of H_1. The purpose of each test is to determine whether the data provides sufficient evidence *against* H_0 in favor of H_1. This is similar to a criminal trial. The jury are required to determine if the presented evidence against the defendant is sufficient and convincing. By default, the *presumption of innocence*, insufficient evidence leads to acquittal.

To determine the truth value of a hypothesis, we use a sample function called
$-$ the **test statistic (TS)**.
The set of values of the test statistic for which we decide to *reject* H_0 is called
$-$ the **rejection region (RR)** or **critical region (CR)**.
The purpose of the experiment is to decide if the evidence (the data from a sample) tends to rebut the null hypothesis (if the value of the test statistic is in the rejection region) or not (if that value falls outside the rejection region).
If the statistical hypothesis refers to the parameter(s) of the distribution of the characteristic X, then we have a **parametric** test, otherwise, a **nonparametric** test. For parametric tests, we will consider that the target parameter

$$\theta \in A = A_0 \cup A_1,\ A_0 \cap A_1 = \emptyset,$$

and then the two hypotheses will be set as

$$\begin{aligned} H_0 &: \ \theta \in A_0 \\ H_1 &: \ \theta \in A_1. \end{aligned}$$

If the set A_0 consists of one single value, $A_0 = \{\theta_0\}$, which completely specifies the population distribution, then the hypothesis is called **simple**, otherwise, it is called a **composite** hypothesis (and the same is true for A_1 and the alternative hypothesis). The null hypothesis will *always* be taken to be simple. Then the null hypothesis

$$H_0 : \theta = \theta_0$$

will have one of the alternatives

$$\begin{aligned} H_1 :&\quad \theta < \theta_0 \text{ (left-tailed test)},\\ H_1 :&\quad \theta > \theta_0 \text{ (right-tailed test)},\\ H_1 :&\quad \theta \neq \theta_0 \text{ (two-tailed test)}. \end{aligned}$$

Remark 4.1. The first and one of the most important tasks in a hypothesis testing problem is to state the *relevant* null and alternative hypotheses to be tested. The null hypothesis is usually taken to be a simple hypothesis, but the *appropriate* alternate has to be *understood from the context*. We mentioned that H_1 is the opposite "in some way" of H_0. Let us clarify this.

1. Consider a problem in which the effectiveness of a fever medicine is tested. It is supposed to reduce the fever to the normal value of 37^oC or below. If the temperature values of a number of patients taking this medicine are considered, then for the mean temperature the relevant hypotheses would be

$$\begin{aligned} H_0 :&\quad \mu = 37\\ H_1 :&\quad \mu > 37, \end{aligned}$$

since an average lower than or equal to 37^oC would mean the same thing in this context, the patients are fine. A problem would be a mean temperature *greater* than 37^oC. In this sense, H_0 and H_1 are "opposites" of each other.

2. To verify that the average connection speed is 54 Mbps, we test the hypothesis

$$\begin{aligned} H_0 :&\quad \mu = 54\\ H_1 :&\quad \mu \neq 54. \end{aligned}$$

However, if we worry about a *low* connection speed only, we can conduct a one-sided test of

$$\begin{aligned} H_0 :&\quad \mu = 54\\ H_1 :&\quad \mu < 54. \end{aligned}$$

In this case, we only measure the amount of evidence supporting the one-sided alternative $H_1 : \mu <$

54. In the absence of such evidence, we gladly accept the null hypothesis.

Designing a hypothesis test means constructing the rejection region RR, such that for a given $\alpha \in (0, 1)$, the conditional probability, conditioned by H_0 being true,

$$P(TS \in RR \mid H_0) = \alpha. \tag{4.1}$$

The value α is called **significance level** or **risk probability**.
For any given hypothesis testing problem, we have the following possibilities:

Decision	Actual situation	
	H_0 true	H_1 true
Reject H_0	Type I error (prob. α)	Right decision
Not reject H_0	Right decision	Type II error (prob. β)

Table 1: Decisions and errors

In two of the cases, we make the right decision, in the other two, we make an error.
A **type I error** occurs when we reject a true null hypothesis and by (4.1), the probability of making such an error is the significance level

$$P(\text{type I error}) \;=\; P(\text{ reject } H_0 \mid H_0) \;=\; P(TS \in RR \mid H_0) \;=\; \alpha, \tag{4.2}$$

while a **type II error** happens when we fail to reject a false null hypothesis, and its probability is denoted by β,

$$P(\text{ type II error}) \;=\; P(\text{ not reject } H_0 \mid H_1) \;=\; P(TS \notin RR \mid H_1) \;=\; \beta. \tag{4.3}$$

Remark 4.2.
1. The rejection region and hence, the hypothesis test, are *not* uniquely determined by (4.1), as was the case with confidence intervals.
2. Since both α and β represent risks of making an error, we would like to design tests such that both of their values are small. Unfortunately, making one of them very small will result in the other being unreasonably large. But, for almost all statistical tests, α and β will both decrease as the

sample size increases.
3. In general, α is preset and a procedure is given for finding an appropriate rejection region.

4.2 General Framework, Z-Tests

Just like with confidence intervals, we start with the case where the test statistic has a $N(0,1)$ distribution, so we can better understand the ideas.

Let θ be a target parameter and let $\overline{\theta}$ be an unbiased estimator for θ ($E(\overline{\theta}) = \theta$), with standard error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is known that

$$Z \;=\; \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \;\left(= \;\frac{\overline{\theta} - E(\overline{\theta})}{\sigma(\overline{\theta})}\right) \tag{4.4}$$

has an approximately Standard Normal $N(0,1)$ distribution. We design a hypothesis testing procedure for θ the following way: for a given level of significance $\alpha \in (0,1)$, consider the hypotheses

$$H_0: \;\; \theta = \theta_0,$$

with one of the alternatives

$$H_1: \;\; \begin{cases} \theta < \theta_0 \\ \theta > \theta_0 \\ \theta \neq \theta_0. \end{cases} \tag{4.5}$$

We will use the test statistic $TS = Z$ given by (4.4).
The **observed value of the test statistic** from the sample data is

$$TS_0 \;=\; TS(\theta = \theta_0). \tag{4.6}$$

In our case, this is

$$Z_0 \;=\; TS(\theta = \theta_0) \;=\; \frac{\overline{\theta} - \theta_0}{\sigma_{\overline{\theta}}}.$$

How to design the rejection region RR? Let us start with the left-tailed case. We need to determine the RR such that (4.1) holds. Intuitively, we reject H_0 if the observed value of the test statistic is *far* from the value specified in H_0, "far" in the sense of the alternative H_1, in this case *far to the*

left of θ_0. So, we determine a rejection region of the form

$$RR \;=\; \{Z_0 \mid Z_0 \leq k_1\} \;=\; (-\infty, k_1].$$

We have

$$\begin{aligned} \alpha &= P(Z_0 \in RR \mid H_0) \\ &= P(Z_0 \leq k_1 \mid \theta = \theta_0) \\ &= P(Z_0 \leq k_1 \mid Z_0 \in N(0,1)). \end{aligned}$$

Now, we know that if $Z_0 \in N(0,1)$, $P(Z_0 \leq z_\alpha) = \alpha$, where z_α is the quantile of order α for the $N(0,1)$ distribution. Thus, we choose $k_1 = z_\alpha$ and

$$RR_{\text{left}} \;=\; \{Z_0 \leq z_\alpha\}. \tag{4.7}$$

Similarly, for a right-tailed test, we want to find a rejection region of the form

$$RR \;=\; \{Z_0 \mid Z_0 \geq k_2\} \;=\; [k_2, \infty),$$

so that

$$\begin{aligned} \alpha &= P(Z_0 \in RR \mid H_0) \\ &= P(Z_0 \geq k_2 \mid \theta = \theta_0) \\ &= P(Z_0 \geq k_2 \mid Z_0 \in N(0,1)) \\ &= 1 - P(Z_0 < k_2 \mid Z_0 \in N(0,1)). \end{aligned}$$

Since $P(Z_0 < z_{1-\alpha}) = 1 - \alpha$, then $P(Z_0 \geq z_{1-\alpha}) = \alpha$ and so we choose $k_2 = z_{1-\alpha}$, the quantile of order $1 - \alpha$ for the $N(0,1)$ distribution and

$$RR_{\text{right}} \;=\; \{Z_0 \geq z_{1-\alpha}\}. \tag{4.8}$$

Finally, for a two-tailed test, we reject the null hypothesis if the observed value of the test statistic is far away from θ_0 *on either side*. That is, the rejection region should be of the form $RR = \{Z_0 \mid Z_0 \leq k_1 \text{ or } Z_0 \geq k_2\} = (-\infty, k_1] \cup [k_2, \infty)$. The rejection region should be chosen such that

$$P(Z_0 \leq k_1 \text{ or } Z_0 \geq k_2 \mid \theta = \theta_0) \;=\; \alpha,$$

or, equivalently,

$$P(k_1 < Z_0 < k_2 \mid Z_0 \in N(0,1)) \;=\; 1-\alpha.$$

We encountered such problems before in the previous section, when finding (two-sided) confidence intervals. As we did then, we will choose $k_1 = z_{\frac{\alpha}{2}}$ and $k_2 = z_{1-\frac{\alpha}{2}}$, so

$$RR_{\text{two}} = \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } \; Z_0 \geq z_{1-\frac{\alpha}{2}}\}, \tag{4.9}$$

or, since the distribution of Z is symmetric and $z_{1-\frac{\alpha}{2}} > 0$,

$$\begin{aligned} RR_{\text{two}} &= \{Z_0 \leq -z_{1-\frac{\alpha}{2}} \text{ or } \; Z_0 \geq z_{1-\frac{\alpha}{2}}\} \\ &= \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}. \end{aligned}$$

To summarize, the rejection regions for the three alternatives (4.5) are given by

$$RR: \; \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } \; Z_0 \geq z_{1-\frac{\alpha}{2}}\} = \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.10}$$

Remark 4.3.

1. Since a test statistic $Z \in N(0,1)$ was used, these are commonly known as ***Z*-tests**.

2. We will derive hypothesis tests for all the common parameters (mean, variance, difference of means, ratio of variances). The test statistics and their distributions will change, but the ideas and the principles will remain the same, as for the case we just described.

3. Notice from our derivation of the rejection region for a two-tailed test, that there is a strong relationship between confidence intervals and rejection regions: The values θ_0 of a target parameter θ in a $100(1-\alpha)\%$ CI $(\alpha \in (0,1))$, are precisely the values for which the test statistic falls *outside* the RR, and hence, for which the null hypothesis $\theta = \theta_0$ is not rejected at the significance level α. We say that the $100(1-\alpha)\%$ two-sided CI consists of all the *acceptable* values of the parameter, at the significance level α.

4. **Caution!** This is **not** saying that the rejection region is the complement of the confidence interval! The RR contains values for the *test statistic* TS, while the CI consists of values of the *parameter* θ.

Example 4.4. The number of monthly sales at a firm is known to have a mean of 20 and a standard deviation of 4 and all salary, tax and bonus figures are based on these values. However, in times of economical recession, a sales manager fears that his employees do not average 20 sales per month, but less, which could seriously hurt the company. For a number of 36 randomly selected salespeople,

it was found that in one month they averaged 19 sales. At the 5% significance level, does the data confirm or contradict the manager's suspicion?

Solution. Recall that for n large, we have that

$$Z \;=\; \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

has an approximately $N(0,1)$ distribution. Since the sample size $n = 36 > 30$ and σ is known, we can use a Z-test. The test is on the *average* number of sales per month, so for the mean μ. The manager's suspicion is that the average is *less* than 20, which is supposed to be, so the two relevant hypotheses for this problem are

$$\begin{aligned} H_0: &\quad \mu = 20 \\ H_1: &\quad \mu < 20, \end{aligned}$$

a left-tailed test.

A type I error would mean concluding that the average number of monthly sales is less than 20, when in fact, it is not; a type II error would be deciding that the average number of monthly sales is 20 (or higher), but it actually is not. We allow for the probability of a type I error (the significance level) to be $\alpha = 0.05$. The population standard deviation is known, $\sigma = 4$ and the sample mean is $\overline{X} = 19$.

The value of the test statistic is

$$Z_0 = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \;=\; \frac{19-20}{\frac{4}{6}} \;=\; -1.5.$$

The rejection region is, by (4.10),

$$RR \;=\; (-\infty, z_\alpha] \;=\; (-\infty, -1.645].$$

Since $Z_0 \notin RR$, we *do not reject* H_0. The evidence obtained from the data is not sufficient to reject it. In the absence of sufficient evidence, by default, we accept the null hypothesis. So, at the 5% significance level, the data *does not* confirm the manager's suspicion.

■

Short review

Let us recall: we have a population characteristic X, whose pdf $f(x;\theta)$ depends on θ, the target parameter to be estimated. The estimation is done based on a sample of size n, i.e. sample variables $X_1, X_2, \ldots, X_n$ that are *iid*, with the same pdf as X.

We set up two hypotheses, the *null* hypothesis, always simple, i.e.

$$H_0 : \theta = \theta_0$$

and one of the *alternative* hypotheses

$$\begin{aligned} H_1 :&\quad \theta < \theta_0 \ \text{(left-tailed test)}, \\ H_1 :&\quad \theta > \theta_0 \ \text{(right-tailed test)}, \\ H_1 :&\quad \theta \neq \theta_0 \ \text{(two-tailed test)}. \end{aligned} \tag{4.1}$$

We want to decide if H_0 is *rejected* (in favor of H_1) or *not rejected* (accepted). We use a *test statistic* TS (with the same properties as the pivot in CI's) and a *rejection (critical) region* RR, such that for a given *significance level* $\alpha \in (0,1)$,

$$P(\text{type I error}) \;=\; P(\text{ reject } H_0 \mid H_0) \;=\; P(TS \in RR \mid H_0) \;=\; \alpha. \tag{4.2}$$

The probability of a *type II error* is

$$P(\text{ type II error}) \;=\; P(\text{ not reject } H_0 \mid H_1) \;=\; P(TS \notin RR \mid H_1) \;=\; \beta.$$

In general, the significance level α is preset and a procedure is given for finding an appropriate rejection region, such that β is also reasonably small.

We considered the case where for a target parameter θ, $\overline{\theta}$ is an unbiased estimator ($E(\overline{\theta}) = \theta$), with standard error $\sigma_{\overline{\theta}}$, such that, under certain conditions, it is known that

$$Z \;=\; \frac{\overline{\theta} - \theta}{\sigma_{\overline{\theta}}} \;\left(= \;\frac{\overline{\theta} - E(\overline{\theta})}{\sigma(\overline{\theta})}\right) \tag{4.3}$$

has an approximately Standard Normal $N(0,1)$ distribution. Using Z as a test statistic, we found

the rejection regions for the three alternatives (4.1) as

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{Z_0 \leq z_{\frac{\alpha}{2}} \text{ or } Z_0 \geq z_{1-\frac{\alpha}{2}}\} = \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.4}$$

4.3 Significance Testing, P-Values

There is a problem that might occur in hypothesis testing: We preset α, the probability of a type I error and henceforth determine a rejection region. We get a value of the test statistic that *does not belong* to it, so we cannot reject the null hypothesis H_0, i.e. we accept it as being true. However, when we compute the probability of getting that value of the test statistic under the assumption that H_0 is true, we find it is *very small*, comparable with our preset α. So, we accept H_0, yet considering it to be true, we find that it is *very unlikely* (very improbable) that the test statistic takes the observed value we found for it. That makes us wonder if we set our RR right and if we didn't "accept" H_0 too easily, by hastily dismissing values of the test statistic that did not fall into our RR. So we should take a look at how "far-fetched" does the value of the test statistic seem, under the assumption that H_0 is true. If it seems really implausible to occur by chance, i.e. if its probability is *small*, then maybe we should reject the null hypothesis H_0.

To avoid this situation, we perform what is called a **significance test**: for a given random sample (i.e. sample variables $X_1, \ldots, X_n$), we still set up H_0 and H_1 as before and we choose an appropriate test statistic. Then, we compute the probability of observing a value *at least as extreme* (in the sense of the test conducted) of the test statistic TS as the value observed from the sample, TS_0, under the assumption that H_0 is true. This probability is called the critical value, the descriptive significance level, the probability of the test, or, simply the ***P*-value** of the test. If it is small, we reject H_0, otherwise we do not reject it. The P-value is a numerical value assigned to the test, it depends only on the sample data and its distribution, but *not* on α.
In general, for the three alternatives (4.1), if TS_0 is the value of the test statistic TS under the assumption that H_0 is true and F is the cdf of TS, the P-value is computed by

$$P = \begin{cases} P(TS \leq TS_0 \mid H_0) & = & F(TS_0) \\ P(TS \geq TS_0 \mid H_0) & = & 1 - F(TS_0) \\ 2 \cdot \min\{P(TS \leq TS_0 \mid H_0), P(TS \geq TS_0 \mid H_0)\} & = & 2 \cdot \min\{F(TS_0), 1 - F(TS_0)\}. \end{cases} \tag{4.5}$$

Then the decision will be

$$\begin{aligned} &\text{if } \ P \leq \alpha, \text{ reject } \ H_0, \\ &\text{if } \ P > \alpha, \text{ do not reject } \ H_0. \end{aligned} \tag{4.6}$$

So, more precisely, the P-value of a test is the smallest level at which we could have preset α and still have been able to reject H_0, or the lowest significance level that *forces* rejection of H_0, i.e. the *minimum rejection level*.

Remark 4.1.

1. Thus, we can avoid the costly computation of the rejection region (costly because of the quantiles) and compute the P-value instead. Then, we simply compare it to the significance level α. If α is above the P-value, we reject H_0, but if it is below that minimum rejection level, we can no longer reject the null hypothesis.
2. Hypothesis testing (determining the rejection region) and significance testing (computing the P-value) are two methods for testing *the same* thing (the same two hypotheses), so, of course, the outcome (the decision of rejecting or not H_0) will be *the same*, for the same data.

Example 4.2. Recall the problem in Example 4.4 (Lecture 10): The number of monthly sales at a firm is known to have a mean of 20 and a standard deviation of 4 and all salary, tax and bonus figures are based on these values. However, in times of economical recession, a sales manager fears that his employees do not average 20 sales per month, but less, which could seriously hurt the company. For a number of 36 randomly selected salespeople, it was found that in one month they averaged 19 sales. At the 5% significance level, does the data confirm or contradict the manager's suspicion? Now, let us perform a significance test.

Solution. We tested a left-tailed alternative for the mean

$$\begin{aligned} H_0: &\ \ \mu = 20 \\ H_1: &\ \ \mu < 20. \end{aligned}$$

The population standard deviation was given, $\sigma = 4$, and for a sample of size $n = 36$, the sample mean was $\overline{X} = 19$. For the test statistic

$$Z \ = \ \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \ \in \ N(0,1),$$

the observed value was

$$Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \ = \ \frac{19 - 20}{\frac{4}{6}} \ = \ -1.5.$$

Now, we compute the P-value

$$P \;=\; P(Z \leq Z_0) \;=\; P(Z \leq -1.5) \;=\; 0.0668.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.0668 \;=\; P,$$

(is below the minimum rejection level), we do not reject H_0, so, at the 5% significance level, we conclude that the data contradicts the manager's suspicion. But, for example, at the 7% significance level, we would have rejected it.

■

4.4 Tests for the Parameters of One Population

Let X be a population characteristic, with pdf $f(x;\theta)$, mean $E(X) = \mu$ and variance $V(X) = \sigma^2$. Let $X_1, X_2, \ldots, X_n$ be sample variables.

Tests for the mean of a population, $\theta = \mu$

We test the hypotheses

$$\begin{array}{ll} H_0: & \mu = \mu_0, \text{ versus one of} \\ H_1: & \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \neq \mu_0, \end{cases} \end{array} \tag{4.7}$$

under the assumption that either X is approximately Normally $N(\mu, \sigma)$ distributed or that the sample is large ($n > 30$).

Case σ known (**ztest**)
We use the test statistic

$$TS \;=\; Z \;=\; \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \;\in\; N(0,1), \tag{4.8}$$

with observed value

$$Z_0 \;=\; \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \tag{4.9}$$

Then, as before, at the $\alpha \in (0,1)$ significance level, the rejection region for each test will be given by

$$RR: \begin{cases} \{Z_0 \leq z_\alpha\} \\ \{Z_0 \geq z_{1-\alpha}\} \\ \{|Z_0| \geq z_{1-\frac{\alpha}{2}}\} \end{cases} \tag{4.10}$$

and the P-value will be computed as

$$P = \begin{cases} P(Z \leq Z_0 \mid H_0) &= \Phi(Z_0) \\ P(Z \geq Z_0 \mid H_0) &= 1 - \Phi(Z_0) \\ P(|Z| \geq |Z_0| \mid H_0) &= 2\left(1 - \Phi(|Z_0|)\right), \end{cases} \tag{4.11}$$

since $N(0,1)$ is symmetric, where

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

is Laplace's function, the cdf for the Standard Normal $N(0,1)$ distribution.

Case σ unknown (**ttest**)
In this case, we use the test statistic

$$TS \;=\; T \;=\; \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \;\in\; T(n-1), \tag{4.12}$$

with observed value

$$T_0 \;=\; \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}. \tag{4.13}$$

Similarly to the previous case, we find the rejection region for the three alternatives as

$$RR: \begin{cases} \{T_0 \leq t_\alpha\} \\ \{T_0 \geq t_{1-\alpha}\} \\ \{|T_0| \geq t_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{4.14}$$

and compute the P-value will by

$$P = \begin{cases} P(T \leq T_0 \mid H_0) & = & F(T_0) \\ P(T \geq T_0 \mid H_0) & = & 1 - F(T_0) \\ P(|T| \geq |T_0| \mid H_0) & = & 2\left(1 - F(|T_0|)\right), \end{cases} \tag{4.15}$$

where the cdf F and the quantiles refer to the $T(n-1)$ distribution.

Tests for the variance of a population, $\theta = \sigma^2$ (vartest)

Assuming that X has a Normal $N(\mu, \sigma)$ distribution, we test the hypotheses

$$\begin{array}{lll} H_0: & \sigma^2 = \sigma_0^2, & H_0: \quad \sigma = \sigma_0, \\ H_1: & \begin{cases} \sigma^2 < \sigma_0^2 \\ \sigma^2 > \sigma_0^2 \\ \sigma^2 \neq \sigma_0^2, \end{cases} \quad \text{equivalent to} & H_1: \begin{cases} \sigma < \sigma_0 \\ \sigma > \sigma_0 \\ \sigma \neq \sigma_0. \end{cases} \end{array} \tag{4.16}$$

The test statistic will be

$$TS \;=\; V \;=\; \frac{(n-1)s^2}{\sigma^2} \;\in\; \chi^2(n-1), \tag{4.17}$$

with observed value

$$V_0 \;=\; \frac{(n-1)s^2}{\sigma_0^2}. \tag{4.18}$$

Even though the $\chi^2(n-1)$ distribution is not symmetric, we use the same line of reasoning and computations to find the rejection region for the three alternatives:

$$RR: \begin{cases} \{V_0 \leq \chi^2_{\alpha}\} \\ \{V_0 \geq \chi^2_{1-\alpha}\} \\ \{V_0 \leq \chi^2_{\frac{\alpha}{2}} \text{ or } V_0 \geq \chi^2_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.19}$$

Same goes for the computation of the P-values:

$$P = \begin{cases} P(V \leq V_0 \mid H_0) & = & F(V_0) \\ P(V \geq V_0 \mid H_0) & = & 1 - F(V_0) \\ 2 \cdot \min\{P(V \leq V_0 \mid H_0), P(V \geq V_0 \mid H_0)\} & = & 2 \cdot \min\{F(V_0), 1 - F(V_0)\}, \end{cases} \tag{4.20}$$

where the cdf F and the quantiles refer to the $\chi^2(n-1)$ distribution.

Example 4.3. Let us consider again the problem in Example 4.2, where $\sigma = 4$ was given. Suppose that for the sample considered, the standard deviation is found to be $s = 4.5$. Assuming that the number of monthly sales at that firm is Normally distributed, at the 5% significance level, does the assumption on σ seem to be correct?

Solution. We are now testing the variance. We want to know if the value $\sigma = 4$ is correct *or not*, so, this will be a *two-tailed* test.

$$\begin{aligned} H_0: &\quad \sigma = 4 \\ H_1: &\quad \sigma \neq 4, \end{aligned}$$

i.e.,

$$\begin{aligned} H_0: &\quad \sigma^2 = 16 = \sigma_0^2 \\ H_1: &\quad \sigma^2 \neq 16 = \sigma_0^2, \end{aligned}$$

We have $n = 36$ and $s^2 = (4.5)^2 = 20.25$. The observed value of the test statistic is

$$V_0 \;=\; \frac{(n-1)s^2}{\sigma_0^2} \;=\; \frac{35 \cdot 20.25}{16} \;=\; 44.2969.$$

The significance level is $\alpha = 0.05$ and the two quantiles for the $\chi^2(35)$ distribution are

$$\begin{aligned} \chi^2_{0.025} &= 20.5694, \\ \chi^2_{0.975} &= 53.2033. \end{aligned}$$

Then the rejection region is

$$RR \;=\; (-\infty, 20.5694] \,\cup\, [53.2033, \infty),$$

which *does not* include the value V_0. Therefore, the decision is to *not reject* the null hypothesis, i.e. to conclude that the assumption $\sigma = 4$ is correct.

On the other hand, the P-value is

$$P \;=\; 2 \cdot \min\{P(V \leq V_0), P(V \geq V_0)\} \;=\; 2 \cdot \min\{0.8652, 0.1348\} \;=\; 0.2697.$$

Since

$$\alpha \;=\; 0.05 \;<\; 0.2697 \;=\; P,$$

the decision is to *not reject* the null hypothesis.

Notice that the significance test tells us more! Since the P-value is so large (remember, it is comparable to a probability of an *error*, so a *small* quantity), not only at the 5% significance level we decide to accept H_0, but at *any* reasonable significance level the decision would be the same. That means that the data *strongly* suggests that H_0 is true and should not be rejected. Even though the *sample* standard deviation *is not* equal to 4, still, statistically, the data strongly suggests that the *population* standard deviation *is* 4. We should be careful not to extrapolate the property of one sample to the entire population (data from a sample may be misleading, if it is not used properly ...)

■

4.5 Tests for Comparing the Parameters of Two Populations

Assume we have two population characteristics $X_{(1)}$ and $X_{(2)}$, with means and variances $E(X_{(1)}) = \mu_1, V(X_{(1)}) = \sigma_1^2$ and $E(X_{(2)}) = \mu_2, V(X_{(2)}) = \sigma_2^2$, respectively. We draw two independent random samples $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2n}$, with sample means $\overline{X}_1, \overline{X}_2$, sample variances s_1^2, s_2^2, respectively and *pooled* variance

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

Tests for the difference of means, $\theta = \mu_1 - \mu_2$

We test the hypotheses

$$\begin{array}{lll} H_0: & \mu_1 - \mu_2 = 0, & \\ H_1: & \left\{ \begin{array}{l} \mu_1 - \mu_2 < 0 \\ \mu_1 - \mu_2 > 0 \\ \mu_1 - \mu_2 \neq 0, \end{array} \right. & \text{equivalent to} \end{array} \quad \begin{array}{lll} H_0: & \mu_1 = \mu_2, & \\ H_1: & \left\{ \begin{array}{l} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2, \end{array} \right. & \end{array} \tag{4.21}$$

under the assumption that either $X_{(1)}$ and $X_{(2)}$ have approximately Normal distributions or that the samples are large enough ($n_1 + n_2 > 40$).

Case σ_1, σ_2 known

We use the test statistic

$$TS = Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1), \tag{4.22}$$

with observed value

$$Z_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}. \tag{4.23}$$

The rejection regions and P-values for the three alternatives are then given by equations (4.10)-(4.11), with Z_0 from (4.23).

Case $\sigma_1 = \sigma_2$ unknown (**ttest2**)

The test statistic is

$$TS \;=\; T \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \;\in\; T(n_1 + n_2 - 2), \tag{4.24}$$

with observed value

$$T_0 \;=\; \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}. \tag{4.25}$$

The rejection regions and P-values for the three alternatives are then given by equations (4.14)-(4.15), where T_0 is given in (4.25) and the cdf F and the quantiles refer to the $T(n_1 + n_2 - 2)$ distribution.

Case σ_1, σ_2 unknown (**ttest2**)

We now use the test statistic

$$TS \;=\; T^* \;=\; \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \;\in\; T(n), \tag{4.26}$$

where $\dfrac{1}{n} = \dfrac{c^2}{n_1 - 1} + \dfrac{(1-c)^2}{n_2 - 1}$ and $c = \dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$.

The observed value of the test statistic is

$$T_0^* \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}. \tag{4.27}$$

The rejection regions and P-values for the three alternatives are again as in equations (4.14)-(4.15), with T_0 replaced by T_0^* from (4.27). The cdf F and the quantiles refer to the $T(n)$ distribution.

Remark 4.4. The same Matlab command **ttest2** performs a T-test for the difference of two population means, when the variances are *not* assumed equal, with the option *vartype* set on "unequal" (the default being "equal", when it can be omitted).

Tests for the ratio of variances, $\theta = \dfrac{\sigma_1^2}{\sigma_2^2}$ (vartest2)

Assuming that both $X_{(1)}$ and $X_{(1)}$ have Normal distributions, we test the hypotheses

$$\begin{array}{lll} H_0: \ \dfrac{\sigma_1^2}{\sigma_2^2} = 1, & & \\ H_1: \ \begin{cases} \dfrac{\sigma_1^2}{\sigma_2^2} < 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} > 1 \\ \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1, \end{cases} & <=> & \begin{array}{ll} H_0: & \sigma_1^2 = \sigma_2^2, \\ H_1: & \begin{cases} \sigma_1^2 < \sigma_2^2 \\ \sigma_1^2 > \sigma_2^2 \\ \sigma_1^2 \neq \sigma_2^2, \end{cases} \end{array} \quad <=> \quad \begin{array}{ll} H_0: & \sigma_1 = \sigma_2, \\ H_1: & \begin{cases} \sigma_1 < \sigma_2 \\ \sigma_1 > \sigma_2 \\ \sigma_1 \neq \sigma_2. \end{cases} \end{array} \end{array} \tag{4.28}$$

The test statistic used is

$$TS \ = \ F \ = \ \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \ \in \ F(n_1 - 1, n_2 - 1), \tag{4.29}$$

with observed value

$$F_0 \ = \ \frac{s_1^2}{s_2^2}. \tag{4.30}$$

Again, just like in the case of one population variance, the $F(n_1 - 1, n_2 - 1)$ distribution is not symmetric, but proceeding as before, we find the rejection region for the three alternatives as

$$RR: \ \begin{cases} \{F_0 \leq f_\alpha\} \\ \{F_0 \geq f_{1-\alpha}\} \\ \{F_0 \leq f_{\frac{\alpha}{2}} \text{ or } \ F_0 \geq f_{1-\frac{\alpha}{2}}\}. \end{cases} \tag{4.31}$$

and the P-values given by

$$P = \begin{cases} P(F \leq F_0 \mid H_0) & = & F(F_0) \\ P(F \geq F_0 \mid H_0) & = & 1 - F(F_0) \\ 2 \cdot \min\{P(F \leq F_0 \mid H_0), P(F \geq F_0 \mid H_0)\} & = & 2 \cdot \min\{F(F_0), 1 - F(F_0)\}, \end{cases} \tag{4.32}$$

where the cdf F and the quantiles refer to the $F(n_1 - 1, n_2 - 1)$ distribution.

Example 4.5. Suppose the strengths to a certain load of two types of material, $M1$ and $M2$, are studied, knowing that they are approximately Normally distributed. The more weight they can resist to, the stronger they are. Two independent random samples are drawn and they yield the following data.

$M1$			$M2$		
n_1	$=$	25	n_2	$=$	16
$\overline{X}_1$	$=$	380	$\overline{X}_2$	$=$	370
s_1^2	$=$	537	s_2^2	$=$	196

a) At the 5% significance level, do the variances of the two populations seem to be equal or not?
b) At the same significance level, does the data suggest that on average, $M1$ is stronger than $M2$?
(In both parts, perform both hypothesis and significance testing).

Solution.
a) First, we compare the variances of the two populations, so we know which way to proceed for comparing the means. We want to know if they are equal or not, so it is a two-tailed test. Hence, our hypotheses are

$$\begin{aligned} H_0 &: \ \sigma_1^2 = \sigma_2^2 \\ H_1 &: \ \sigma_1^2 \neq \sigma_2^2. \end{aligned}$$

The observed value of the test statistic is

$$F_0 \ = \ \frac{s_1^2}{s_2^2} \ = \ \frac{537}{196} = 2.7398.$$

For $\alpha = 0.05$, $n_1 = 25$ and $n_2 = 16$, the quantiles for the $F(24, 15)$ distribution are

$$\begin{aligned} f_{\frac{\alpha}{2}} \quad &= \quad f_{0.025} \quad = \quad 0.4103 \\ f_{1-\frac{\alpha}{2}} \quad &= \quad f_{0.975} \quad = \quad 2.7006. \end{aligned}$$

Thus, the rejection region for our test is

$$RR \;=\; (-\infty,\; 0.4103] \;\cup\; [2.7006, \infty)$$

and clearly, $F_0 \in RR$. Thus we reject H_0 in favor of H_1, i.e. we conclude that the data suggests that the population variances are *different*.
Let us also perform a significance test. The P-value of this (two-tailed) test is

$$P \;\;=\;\; 2 \cdot \min\{P(F \leq F_0), P(F \geq F_0)\} \;\;=\;\; 2 \cdot \min\{0.9765, 0.0235\} \;\;=\;\; 0.0469.$$

Since our $\alpha > P$, the "minimum rejection significance level", we reject H_0.
Note. We now know that for instance, at 1% significance level (or any level less than 4.69%), we would have *not* rejected the null hypothesis. This goes to show that the data can be "misleading". Simply comparing the values of the sample functions does not necessarily mean that the same thing will be true for the corresponding population parameters. Here, s_1^2 is *much* larger than s_2^2, yet at 1% significance level, we would have concluded that the population variances seem to be equal.

b) Next we want to compare the population means. If $M1$ is to be *stronger* than $M2$ on average, than we must perform a *right*-tailed test:

$$\begin{aligned} H_0 :&\quad \mu_1 = \mu_2 \\ H_1 :&\quad \mu_1 > \mu_2 \end{aligned}$$

Which one of the tests for the difference of means should we use? The answer is in part a). At this significance level, the variances are unknown and *different*.
Then the value of the test statistic is, by (4.27)

$$T_0^* \;=\; \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \;=\; \frac{380 - 370}{\sqrt{\dfrac{537}{25} + \dfrac{196}{16}}} \;=\; 1.7218.$$

To find the rejection region, we compute

$$c = 0.6368, \quad n = 38.9244 \approx 39$$

and the quantile for the $T(39)$ distribution

$$t_{1-\alpha} \ = \ t_{0.95} \ = \ 1.6849.$$

Then the rejection region of the test is

$$RR \ = \ [1.6849, \infty),$$

which includes the value T_0^*, so we *reject* H_0 in favor of H_1. So we conclude that yes, the data suggests that material $M1$ is, on average, stronger than material $M2$.
On the other hand, the P-value of this test is

$$P \ = \ P(T^* \geq T_0^*) \ = \ 1 - F(T_0^*) = 1 - F(1.7218) \ = \ 0.0465,$$

where F is the cdf of the $T(39)$ distribution. Again, the P-value is lower than $\alpha = 0.05$, which forces the rejection of H_0.

■

Remark 4.6.
1. As mentioned before, both hypothesis and significance testing lead to the same conclusion. From the implementation point of view, significance testing is more efficient, since it avoids the inversion of a cdf (i.e. computation of quantiles), which is often a complicated improper integral. This is the reason why, although the main tests *are* implemented in Matlab, the rejection regions *are not* computed.
2. Many tests (and formulas for CI's) work under the assumption of Normality of the population from which the sample was drawn. In practice, when there are outliers in the data, that is rarely the case. How important is this assumption of Normality and how affected are the results of these tests by small departures from model assumptions? Z-tests and T-tests work well even when the underlying population is not quite Normally distributed. From this point of view, they are called **robust** tests. χ^2-tests and F-tests, however, are *not* robust, they perform very poorly when the assumption of Normality is breached. In modern Statistics there is an ongoing search for finding robust methods of estimation for variances.