SANDA MICULA

PROBABILITY AND STATISTICS
FOR
COMPUTATIONAL SCIENCES

CLUJ UNIVERSITY PRESS / PRESA UNIVERSITARĂ CLUJEANĂ

2009

Contents

PART II. Statistics 127

Preface

Mathematical probability is an old concept, that emerged from the necessity of people to deal with patterns that occur in random events. It has been called many names, from "the very guide of life" (by clergyman J. Butler, in his very influential Analogy of Religion (1736)) to "good sense reduced to calculation" (by P. S. de Laplace, in the early 19th century). Later on, it provided the foundations of modern Statistics. Nowadays, mathematical probability and statistics are found in virtually all fields of modern science. The interpretation of most of the research in all areas of science, from biology and health care to engineering and computing sciences, depends to some extent on probability and statistical methods. Furthermore, most of the times, a practicing computer scientist is expected to understand and help implement statistical techniques in the work place. For these reasons, a student in one of the fields of Computational Sciences should have early exposure to probability and statistical reasoning.

This book is a result of many years of experience in teaching Probability and Statistics courses at university level. The text is intended as an introduction to probability theory and mathematical statistics that emphasizes the probabilistic foundations required to understand probability models and statistical methods. It is especially aimed at students in Computational Sciences, but can also be used by those studying Engineering, Business, Management, Agricultural and Biological Sciences, etc. It does require a fairly solid background in Analysis in $\mathbb{R}^n$. The basic concepts are defined both mathematically, but also explained intuitively. Most of the theorems and properties stated come with a proof. However, those that are too technical, or too long, or require too extensive a mathematical knowledge of Analysis, are skipped, as they are considered to go beyond the purpose of this course. Many examples are provided to emphasize the notions and properties described, most of the examples being taken from real life computational problems.

The book comprises of two parts, Part I, *Probability Theory* and Part II, *Statistics*. Part I consists of six chapters, *Probability Space*, *Probabilistic Models*, *Random Variables and Random Vectors*, *Numerical Characteristics of Random Variables*, *Sequences of Random Variables* and *Laws of Large Numbers and Limit Theorems*. The notion of probability is described from three approaches, axiomatic, classical and geometrical. Random variables (and vectors) are introduced, as a measure of random phenomena, and their properties are analyzed. The main discrete and continuous distributions and their properties are studied in detail. The various types of convergence of sequences of random variables are defined, as well as the connections between them and the main results of convergence (laws of large numbers and limit theorems) are given, in preparation for statistical methods of inference.

Part II consists of four chapters, *Descriptive Statistics*, *Sample Theory*, *Estimation* and *Hypothesis Testing*. Data collection and analysis, sampling techniques, correlations between random variables are discussed (Descriptive Statistics). As a link towards Inferential Statistics, some notions of sample theory are given. Representative sampling assured, inferences and conclusions can be safely extended from the sample to the population as a whole. We discuss classical inferential techniques that have stood the test of time, as well as some new, modern ones. The methods of moments and maximum likelihood are described, with examples. Confidence intervals and hypothesis testing procedures are given for the main estimation problems (one population mean and variance, difference of two population means, ratio of two population variances). We conclude with two of the most widely used nonparametric tests.

The appendix contains notes on Euler's functions and statistical tables for the normal, Student, χ^2, Fisher and Kolmogorov distributions.

Many thanks to all of those who helped and supported me in this endeavor.

The Author

PART I. Probability Theory

Chapter 1

Probability Space

1.1 Experiments, Sample Spaces, Events

An **experiment (trial)** 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, $A_i,\ i \in \mathbb{N}$).

There are two special events associated with every experiment: the **impossible** event, denoted by $\emptyset$, which is the event that "never happens" and the **sure** (also called **certain**) event, denoted by S, which is the event that "always happens".

Since events are defined as sets, we can employ set theory in describing more events. For each event $A \subseteq S$, we define the event $\overline{A}$, the **complementary** event (the opposite), to mean that $\overline{A}$ occurs if and only if A does not occur. Obviously $\overline{\overline{A}} = A$.

We say that the event A **implies** the 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. Then A and B are **equal**, $A = B$, if A

implies B and B implies A.

For any two events $A, B \subseteq S$, we can define the following events:
– the **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,
– the **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,
– the **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 and
– the **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:

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

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

and **distributive**:

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

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

Two events A and B are said to be **mutually exclusive** (or **disjoint**) 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.

Example 1.1.1. 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, the events $\{A, B\}$ and $\{A, D, E\}$ are mutually exclusive. In fact, these last three are more than that, as will be seen from the next definition.

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 for all $i, j \in I,\ i \neq j$, the events A_i and A_j are mutually exclusive.
So, the events $\{A, D, E\}$ from our previous example form a partition of S.

Proposition 1.1.2.
*For every collection of events $\{A_i\}_{i\in I}$, **De Morgan's laws** hold:*

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

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

Proof.
(1) We have

$$\begin{aligned} e \in \overline{\bigcup_{i\in I} A_i} \quad &<=> \quad e \notin \bigcup_{i\in I} A_i \quad <=> \quad e \notin A_i,\ \forall i \in I \\ &<=> \quad e \in \overline{A_i},\ \forall i \in I \quad <=> \quad e \in \bigcap_{i\in I} \overline{A}_i. \end{aligned}$$

(2) Using part (1), we can write $\bigcup_{i\in I} \overline{A}_i = \overline{\overline{\bigcup_{i\in I} \overline{A}_i}} = \overline{\bigcap_{i\in I} \overline{\overline{A}}_i} = \overline{\bigcap_{i\in I} A_i}.$ □

1.2 Sigma Fields and Probability

Definition 1.2.1. *A collection $\mathcal{K}$ of events from S is said to be a **σ-field** 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 1.2.2. The power set $\mathcal{P}(S) = \{S' | S' \subseteq S\}$ is a σ-field over S.

Proposition 1.2.3. *Let* $\mathcal{K}$ *be a* σ*-field over* S*. Then the following properties hold:*

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

(2) *For all* $A, B \in \mathcal{K}$*,* $A \cap B$*,* $A \setminus B$*,* $A \Delta B \in \mathcal{K}$.

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

Proof.
(1) We know that $\mathcal{K} \neq \emptyset$, so there exists $A \in \mathcal{K}$. Then by definition $A, \overline{A} \in \mathcal{K}$ and $S = A \cup \overline{A} \in \mathcal{K}$. Also $\emptyset = \overline{S} \in \mathcal{K}$.

(2) Let $A, B \in \mathcal{K}$.
Then $\overline{A}, \overline{B} \in \mathcal{K}$ and so $\overline{A} \cup \overline{B} \in \mathcal{K}$. But $\overline{A} \cup \overline{B} = \overline{A \cap B}$, so $\overline{A \cap B} \in \mathcal{K}$.
Thus $A \cap B = \overline{\overline{A \cap B}} \in \mathcal{K}$.
Since $A, \overline{B} \in \mathcal{K}$ and $A \setminus B = A \cap \overline{B}$, it follows that $A \setminus B \in \mathcal{K}$.
$A \Delta B = (A \cup B) \setminus (A \cap B)$ and $A \cup B, A \cap B \in \mathcal{K}$, so $A \Delta B \in \mathcal{K}$.

(3) Let $A_n \in \mathcal{K}$, for all $n \in \mathbb{N}$. Then also $\overline{A}_n \in \mathcal{K}$, for all $n \in \mathbb{N}$ and by definition $\bigcup_{n=1}^{\infty} \overline{A}_n \in \mathcal{K}$. Hence $\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} \overline{\overline{A}}_n = \overline{\bigcup_{n=1}^{\infty} \overline{A}_n} \in \mathcal{K}$. □

Proposition 1.2.4. *Let $\mathcal{A}$ be a collection of subsets of S. Let*

$$\sigma(\mathcal{A}) = \bigcap \{\mathcal{F} \mid \mathcal{A} \subseteq \mathcal{F}, \mathcal{F} \text{ is a } \sigma\text{-field}\}.$$

Then $\sigma(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$, i.e.

(1) *$\sigma(\mathcal{A})$ is a σ-field,*

(2) *$\mathcal{A} \subseteq \sigma(\mathcal{A})$,*

(3) *for any σ-field $\mathcal{F}$ with $\mathcal{A} \subseteq \mathcal{F}$, $\sigma(\mathcal{A}) \subseteq \mathcal{F}$.*

Definition 1.2.5. *The σ-field $\sigma(\mathcal{A})$ defined in Proposition 1.2.4, is called the **σ-field generated by** A.*

Definition 1.2.6. *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,*

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n), \tag{1.1}$$

i.e. P is σ-additive.

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

Theorem 1.2.7. *Let $(S, \mathcal{K}, P)$ be a probability space, and let $A, B \in \mathcal{K}$. Then the following properties hold:*

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

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

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

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

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

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

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

i.e. $P(\overline{A}) = 1 - P(A)$.
Since $P(\overline{A}) \geq 0$, it follows that $P(A) \leq 1$, i.e. $0 \leq P(A) \leq 1$.
(2) $P(\emptyset) = P(\overline{S}) = 1 - P(S) = 0$.
(3) We have $A = (A \cap B) \cup (A \setminus B)$ and $A \cap B,\ A \setminus B$ are mutually exclusive. Thus

$$P(A) = P(A \cap B) + P(A \setminus B),$$

i.e. $P(A \setminus B) = P(A) - P(A \cap B)$.
(4) Since $A \subseteq B$, $A = A \cap B$. Then by (3), we have

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

i.e. $P(A) \leq P(B)$.
(5) 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),

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

□

Part (5) of Theorem 1.2.7 can be generalized to more than two events, as seen in the next theorem.

Theorem 1.2.8. *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*

$$\begin{aligned} P\Big(\bigcup_{i=1}^{n} A_i\Big) &= \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) \\ &+ \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k) \qquad (1.2) \\ &+ \ \dots \ + (-1)^{n-1} P(\bigcap_{i=1}^{n} A_i), \end{aligned}$$

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

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) \le \sum_{n=1}^{\infty} P(A_n), \qquad (1.3)$$

i.e P is subadditive.

Proof.
(1) The proof goes by induction on n. Case $n = 2$ corresponds to part (5) of Theorem 1.2.7. To simplify the calculations, we only present the proof for case $n = 3$ (the general case is similar). Let $A, B, C \in \mathcal{K}$. Using

Theorem 1.2.7(5) and de Morgan's laws, we have

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

(2) Define the sequence of events $(B_n)_{n \in \mathbb{N}}$ by

$$\begin{aligned}
B_1 &= A_1, \\
B_n &= A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i \right), \; n \in \mathbb{N}. \qquad (1.4)
\end{aligned}$$

This sequence defines the so-called *disjoint union* of $\bigcup_{n=1}^{\infty} A_n$. The name comes from the fact that the events $(B_n)_{n \in \mathbb{N}}$ are mutually exclusive and $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$. Also note that $B_n \subseteq A_n$, for all $n \in \mathbb{N}$, therefore $P(B_n) \leq P(A_n)$ for all $n \in \mathbb{N}$. Now, using the sequence $(B_n)_{n \in \mathbb{N}}$, we have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n) \leq \sum_{n=1}^{\infty} P(A_n).$$

□

1.3 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.

There are several approaches to the concept of probability. The statistical approach is based on the concept of *relative frequency* of an event and can be used whenever the experiment can be repeated many times and the results observed. In such cases, the probability of an event A, $P(A)$ is approximated by

$$P(A) \approx f(A) = \frac{f}{n} = \frac{\text{number of times event } A \text{ occurred}}{\text{number of times experiment was run}}.$$

One disadvantage of this approach is that the experiment must be repeatable. Another one, even more important, is that any probability obtained this way is an approximation, based on n independent trials. Further testing might lead to a different approximate value. However, there are random events for which a stability of the frequencies is observed; that is, as the number of trials n increases, the relative frequency of the event stabilizes and approaches a certain constant value.

This leads to the *classical approach*. This approach can be used only when it is reasonable to assume that the possible outcomes of an experiment are equally likely. The **classical definition of probability** was given independently by B. Pascal and P. Fermat. We consider an experiment whose outcomes are finite and equally likely. Then the **probability of the occurrence of the event** A is given by

$$P(A) = \frac{n(A)}{n(S)} = \frac{\text{number of favorable outcomes for the occurrence of } A}{\text{total number of possible outcomes of the experiment}}.$$

Example 1.3.1. 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 1.3.1: 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 $n(S) = 36$, the total number of possible outcomes of the experiment.
For event A, the favorable outcomes are

$$\{(i, i) \mid i = \overline{1,6}\},$$

so $n(A) = 6$ and $P(A) = \dfrac{1}{6}$.
For event B, the favorable outcomes are

$$\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\},$$

thus $n(B) = 10$ and $P(B) = \dfrac{5}{18}$. ■

1.4 Geometric Probability

The classical definition of probability involves the notion of "counting" number of outcomes that make up a certain event. Hence it is unusable in the case where the sample space is an uncountable set (e.g. the set of all points on a straight line). Poincaré gave a natural generalization for this case, substituting "counting" by "measuring". In what follows, for a set $E \subseteq S \subseteq \mathbb{R}^k$, $\mu(E)$ will denote the Lebesgue measure of E, i.e. the length of E, if $k = 1$, the area of E, if $k = 2$ and the volume of E, if $k = 3$.

Definition 1.4.1. *Let $S \subseteq \mathbb{R}^n$ be a Lebesgue measurable set, $\mu(S) < \infty$ and let $\mathcal{K}$ be a $\sigma-$field over S such that all elements of $\mathcal{K}$ are Lebesgue measurable. Then the* ***geometric probability*** *of $A \in \mathcal{K}$ is defined by*

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

Example 1.4.2. (Buffon's Needle Problem) In a plane consider a network of parallel, equidistant lines, at a distance $d > 0$ from each other. A needle of length l ($l < d$) is randomly placed in this plane. Find the probability that the needle intersects some line.

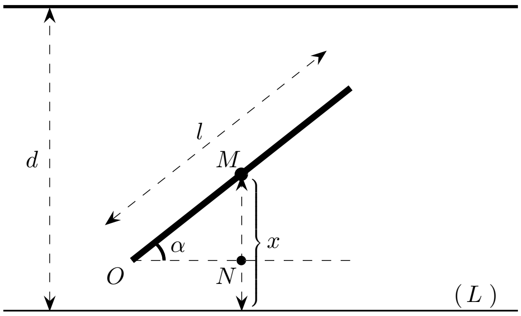


Fig. 1.1: Buffon's Needle Problem

Solution 1.4.2: First off, let us notice that since $l < d$, the needle can intersect at most one line.

Let x be the distance from the middle of the needle (denoted by M) to the closest line L, and let $\alpha \in [0, \pi]$ be the angle between the needle and the positive direction of L. Let A be the event that the needle intersects L.

Then the sample space is given by

$$S = \Big\{(\alpha, x) \in \mathbb{R}^2 \mid \alpha \in [0, \pi], x \in \Big[0, \frac{d}{2}\Big]\Big\}.$$

The needle intersects L if and only if $x \leq |MN|$ (see Figure 1.1). In the right triangle ΔOMN we have $|MN| = \frac{l}{2} \sin \alpha$, thus

$$A = \Big\{(\alpha, x) \in S \ \mid x \leq \frac{l}{2} \sin \alpha\Big\}.$$

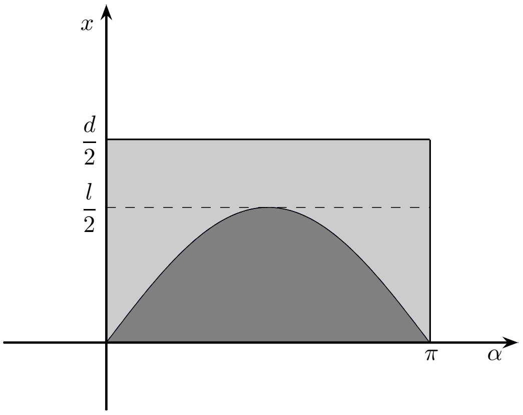


Fig. 1.2: Sample space S and event A

As seen in Figure 1.2, we have

$$\mu(S) = \pi \cdot \frac{d}{2}$$

and

$$\mu(A) = \int_0^\pi \frac{l}{2} \sin\alpha \, d\alpha = l.$$

Then

$$P(A) = \frac{\mu(A)}{\mu(S)} = \frac{2l}{\pi d}.$$

■

Remark 1.4.3. This problem can be used to approximate π the following way: We repeat the experiment n times and approximate the probability of A by its relative frequency

$$P(A) = \frac{2l}{\pi d} \approx \frac{f}{n}.$$

Then we obtain

$$\pi \approx \frac{2ln}{fd}.$$

1.5 Conditional Probability and Independence

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

Definition 1.5.1. *Let $B \in \mathcal{K}$ 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*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.5}$$

An immediate consequence is the following property:

Proposition 1.5.2. *Let $A, B \in \mathcal{K}$ with $P(A)P(B) \neq 0$. Then*

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B). \tag{1.6}$$

Proposition 1.5.3. *Let $A \in \mathcal{K}$ with $0 < P(A) < 1$. Then for all $B \in \mathcal{K}$,*

$$P(B) = P(A)P(B|A) + P\left(\overline{A}\right) P\left(B|\overline{A}\right). \tag{1.7}$$

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

$$B = B \cap S = B \cap \left(A \cup \overline{A}\right) = (B \cap A) \cup \left(B \cap \overline{A}\right).$$

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

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

Using (1.6) for both terms on the right hand side, we obtain (1.7). □

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

Proposition 1.5.4 (The Total Probability Rule)**.** *Let $(A_i)_{i \in I}$ be a partition of S and let $A \in \mathcal{K}$. Then*

$$P(A) = \sum_{i \in I} P\left(A_i\right) P\left(A | A_i\right) \tag{1.8}$$

Proof. Just as before, we have

$$A = A \cap S = A \cap \left(\bigcup_{i \in I} A_i \right) = \bigcup_{i \in I} \left(A \cap A_i\right),$$

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

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

□

Proposition 1.5.5 (The Multiplication Rule)**.** *Let $(A_i)_{i=\overline{1,n}} \subseteq \mathcal{K}$, with $P\left(A_1 \cap A_2 \cap \ldots \cap A_n\right) \neq 0$. Then*

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) P(A_2|A_1) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1}). \tag{1.9}$$

Proof. By Definition 1.5.1, the right hand side of (1.9) is equal to

$$P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot \ldots \cdot \frac{P\left(\bigcap\limits_{i=1}^{n} A_i \right)}{P\left(\bigcap\limits_{i=1}^{n-1} A_i \right)},$$

which after cancellations is $P(A_1 \cap \cdots \cap A_n)$. □

Proposition 1.5.6 (Bayes' formula)**.** *Let $(A_i)_{i\in I}$ be a partition of S with $P(A_i) \neq 0$ for all $i \in I$, and let $A \in \mathcal{K}$ with $P(A) \neq 0$. Then*

$$P(A_j|A) = \frac{P(A_j)P(A|A_j)}{\sum\limits_{i\in I} P(A_i)P(A|A_i)} \quad for\ all\ j \in I. \tag{1.10}$$

Proof. Let $j \in I$. Using (1.5), (1.6) and (1.8), we have

$$P(A_j|A) \;=\; \frac{P(A_j \cap A)}{P(A)} \;=\; \frac{P(A_j)P(A|A_j)}{\sum\limits_{i\in I} P(A_i)P(A|A_i)}.$$

□

Definition 1.5.7. *Two events $A, B \in \mathcal{K}$ are said to be **independent** if*

$$P(A \cap B) = P(A)P(B). \tag{1.11}$$

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

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

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

Remark 1.5.8. 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.

Remark 1.5.9. If $A = \emptyset$ (the impossible event) or $A = S$ (the certain event) and $B \in \mathcal{K}$ is any event, then A and B are independent. This follows easily from the definition.

Proposition 1.5.10. *Let $A, B \in \mathcal{K}$ be independent events. Then A and $\overline{B}$ are also independent.*

Proof. We have $A = (A \cap B) \cup \left(A \cap \overline{B}\right)$ a disjoint union, so

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

□

Remark 1.5.11.
1. A direct consequence of proposition 1.5.10 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 $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$.

Chapter 2

Probabilistic Models

In probability theory, one can notice that some experiments (and their outcomes) follow the same "patterns", so they are said to be in the same "class of experiments". Therefore, for each such class, we design a so-called **probabilistic model**. For each model, we then find the corresponding general computational formulas, which then are applied to each experiment from that class.

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 a ball is extracted, it is returned to the box, so it can be selected again and

$-$ **sampling without replacement**, which means that once a ball is extracted, it cannot be selected again.

If nothing else is specified, then the sampling is done with replacement.

2.1 Bernoulli Model (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" $(\overline{A})$; thus the sample space for each trial is $S = A \cup \overline{A}$,
(iii) the probability of success $p = P(A)$ stays 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**.

Bernoulli Model (Binomial 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.

Proposition 2.1.1. *The probability $P(n;k)$ in a binomial model is given by*

$$P(n;k) \;\;=\;\; C_n^k p^k q^{n-k},\; 0 \leq k \leq n. \tag{2.1}$$

Proof. For i=$\overline{1,n}$, let A_i denote the event: success occurred in the i^{th} trial. Let $B_{n,k}$ denote the event : k successes occurred in n trials.
Then we have

$$B_{n,k} = \bigcup_{(i_1,\ldots,i_k)\in I} \left((A_{i_1} \cap \cdots \cap A_{i_k}) \bigcap \left(\overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}\right)\right),$$

where

$$I = \{(i_1, \ldots, i_k)\,|1 \leq i_1 < \cdots < i_k \leq n,$$
$$i_{k+1}, \ldots, i_n \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}\}.$$

Now $B_{n,k}$ is a union of disjoint events and the events $\{A_i\}_{i=\overline{1,n}}$ are independent; for each $j = \overline{1,k}$, $P(A_{i_j}) = p$ and for every $j = \overline{k+1,n}$, $P(\overline{A}_{i_j}) = q$, so

$$P(B_{n,k}) = \sum_{(i_1,\ldots,i_k)\in I} P\left(A_{i_1} \cap \cdots \cap A_{i_k} \cap \overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}\right)$$

$$= \sum_{(i_1,\ldots,i_k)\in I} p^k q^{n-k} = \text{card } (I) p^k q^{n-k}$$

and card$(I) = \mid I \mid$ is the number of ways we can choose k objects out of n, where order does not matter, so $|I| = C_n^k$.
Thus

$$P(n;k) = P(B_{n,k}) = C_n^k p^k q^{n-k}.$$

□

Remark 2.1.2.
1. The number $P(n;k)$ in (2.1) is the coefficient of x^k in the binomial expansion

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

hence the name of this model.
2. As a consequence, $\sum\limits_{k=o}^{n} P(n;k) = 1$ (if $x = 1$ in the previous equality).
This also follows from the fact that the events $\{B_{n,k}\}_{k=\overline{0,n}}$ defined in the proof of proposition 2.1.1 form a partition of S.

Example 2.1.3. A die is rolled 5 times. Find the probability of the events
A : getting three 6's,
B : getting at least two even numbers.

Solution 2.1.3: Here a trial is a roll of the die. Therefore, $n = 5$.

For the first part, "success" means getting a 6. Hence, $p = \dfrac{1}{6}$ and we have

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

For part two, "success" means getting an even number, so $p = \dfrac{1}{2}$. 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

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

■

2.2 Bernoulli Model With r States (Multinomial Model)

This model is a generalization of the binomial model in the sense that a trial can have more than two possible outcomes.

Bernoulli Model With r States (Multinomial Model). Consider n repeated independent trials, where each trial has several possible outcomes

$E_1, \ldots, E_r$, which form a partition of the sample space and whose probabilities of occurrence $p_i = P(E_i),\ i = \overline{1, r}$ remain the same for each trial $(p_1 + \ldots + p_r = 1)$. Let $0 \leq n_1, \ldots, n_r \leq n$ with $n_1 + \ldots + n_r = n$. Find the probability $P(n; n_1, \ldots, n_r)$ that in n trials, E_i occurs n_i times, $i = \overline{1, r}$.

Proposition 2.2.1. *The probability* $P(n; n_1, \ldots, n_r)$ *in a multinomial model is given by*

$$P(n; n_1, \ldots, n_r) \quad = \quad \frac{n!}{n_1! n_2! \ldots n_r!} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}. \tag{2.2}$$

Proof. This proof is similar to the previous one.
Let $B_{n,n_1,\ldots,n_k}$ denote the event we are interested in.
For $j = \overline{1, n}$, $i = \overline{1, k}$ let A_j^i be the event: E_i occurs in the j^{th} trial.
Then we have

$$\begin{aligned} B_{n,n_1,\ldots,n_k} \quad = \quad & \bigcup \Big[\underbrace{(A^1_{i^1_1} \cap \cdots \cap A^1_{i^1_{n_1}})}_{E_1,\, n_1 \text{ times}} \bigcap \underbrace{(A^2_{i^2_1} \cap \cdots \cap A^2_{i^2_{n_2}})}_{E_2,\, n_2 \text{ times}} \bigcap \\ & \bigcap \cdots \bigcap \underbrace{(A^k_{i^k_1} \cap \cdots \cap A^k_{i^k_{n_k}})}_{E_k,\, n_k \text{ times}} \Big], \end{aligned}$$

the union being taken over all indices in the set

$$\begin{aligned} I = \Big\{ & \big(i^1_1, \ldots, i^1_{n_1}, i^2_1, \ldots, i^2_{n_2}, \ldots, i^k_1, \ldots, i^k_{n_k}\big) \ | \\ & 1 \leq i^j_1 < \cdots < i^j_{n_j} \leq n, j = \overline{1, k} \text{ and } i^l_k \neq i^s_t \text{ if } l \neq s \text{ or } k \neq t \Big\}. \end{aligned}$$

Since this is a disjoint union and $\{A_j^i\}_{\substack{i=\overline{1,k} \\ j=\overline{1,n}}}$ form a partition, we have

$$P(n; n_1, \ldots, n_k) = P(B_{n,n_1,\ldots,n_k}) = |I| p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}.$$

Now, what is $\mid I \mid$? It is the number of ways we can divide n objects into k sets such that the first one contains n_1 objects, the second contains $n_2, \ldots,$ the k^{th} one contains n_k.
For the first set of n_1 objects, there are $C_n^{n_1}$ possibilities of choosing them. Then for the second set there are $C_{n-n_1}^{n_2}$ ways, for the k^{th} set there are $C_{n-n_1-\cdots-n_{k-1}}^{n_k}$ ways.
All these are independent, so

$$\begin{aligned}
\mid I \mid &= C_n^{n_1}\, C_{n-n_1}^{n_2} \cdots\, C_{n-n_1-\cdots-n_{k-1}}^{n_k} \\
&= \frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdot \cdots \cdot \frac{(n-n_1-\cdots-n_{k-1})!}{n_k!\,\underbrace{(n-n_1-\cdots-n_k)!}_{0\,!}} \\
&= \frac{n!}{n_1!n_2!\ldots n_k!}.
\end{aligned}$$

So (2.2) is proved. □

Remark 2.2.2.
1. Let $I = \{(n_1, \ldots n_r) \mid 0 \leq n_1, \ldots, n_r \leq n, n_1 + \ldots + n_r = n\}$. The number $P(n; n_1, \ldots, n_r)$ in (2.2) is the coefficient of $x_1^{n_1} \ldots x_r^{n_r}$ in the multinomial expansion

$$(p_1x_1 + \ldots p_rx_r)^n = \sum_{(n_1,\ldots n_r)\in I} P(n; n_1, \ldots, n_r) x_1^{n_1} \ldots x_r^{n_r}.$$

2. The sum of the probabilities in (2.2), for all indices in I, satisfy

$$\sum_{(n_1,\ldots n_r)\in I} P(n; n_1, \ldots, n_r) = 1$$

(if we let $x_1 = \ldots = x_r = 1$ in the previous equality).

Example 2.2.3. A die is rolled 15 times. Find the probability of the event A: face 5 shows 3 times and one of the faces $\{1, 2, 6\}$ shows 8 times.

Solution 2.2.3: Again, a trial is a roll of the die, so $n = 15$. Now, in order for this to fit the multinomial model, we need a partition of the sample space. Looking at event A of interest, it is clear that the events "face 5 shows" and "one of the faces $\{1, 2, 6\}$ shows" need to be considered. However, these two only, do not form a partition. We have to consider all the possible outcomes. Therefore, a third event needs to be taken into account (covering the remaining possible outcomes), namely, "one of the faces $\{3, 4\}$ shows". This event will have to happen in the remaining $15 - 3 - 8 = 4$ trials. Let us make the following notations:
E_1: face 5 shows,
E_2: one of the faces $\{1, 2, 6\}$ shows,
E_3: one of the faces $\{3, 4\}$ shows.
Now $\{E_1, E_2, E_3\}$ form a partition of the sample space and $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{2}$, $p_3 = \frac{1}{3}$. Also, with our notations, now the event of interest is

A: E_1 occurs 3 times, E_2 occurs 8 times, E_3 occurs 4 times. So

$$P(A) = P(15; 3, 8, 4) = \frac{15!}{3!8!4!}\left(\frac{1}{6}\right)^3\left(\frac{1}{2}\right)^8\left(\frac{1}{3}\right)^4 \approx 0.0503.$$

■

2.3 Bernoulli Model Without Replacement (Hypergeometric Model)

To emphasize the sampling without replacement, we describe this model in the "box with balls" setup.

Hypergeometric Model. A box contains N balls, n_1 of which are white ($n_1 \leq N$) and the rest are black. A number of n $(0 \leq n \leq N)$ balls are extracted, one at a time, without replacement. Find the probability $P(n; k)$

that in n trials, k $(0 \leq k \leq min\{n, n_1\})$ white balls (and, implicitly, $n-k$ black balls) are extracted.

Proposition 2.3.1. *The probability $P(n;k)$ in a hypergeometric model is given by*

$$P(n;k) = \frac{C_{n_1}^k C_{N-n_1}^{n-k}}{C_N^n}. \tag{2.3}$$

Proof. Let $B_{n,k}$ be the event of interest.
For $i = \overline{1,n}$, let A_i be the event: the i^{th} ball extracted is white. Then

$$B_{n,k} = \bigcup \Big((A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \bigcap (\overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}) \Big),$$

where the union is over $(i_1, \dots, i_k) \in I$ (as before in the binomial model), with

$$I = \{(i_1, \dots, i_n) \mid$$

$$1 \leq i_1 < \cdots < i_k \leq n, i_{k+1}, \dots, i_n \in \{1, \dots, n\} \setminus \{i_1, \dots, i_k\}\}.$$

This is a disjoint union, so

$$P(B_{n,k}) = \sum P(A_{i_1} \cap \cdots \cap A_{i_k} \cap \overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}).$$

The events A_i are not independent anymore, but all the events $A_{i_1} \cap \cdots \cap A_{i_k} \cap \overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}$, $(i_1, \dots, i_n) \in I$ have the same probability as the event $A_1 \cap \cdots \cap A_k \cap \overline{A}_{k+1} \cap \cdots \cap \overline{A}_n$.
So, as in the binomial model,

$$P(B_{n,k}) = C_n^k P(A_1 \cap \cdots \cap A_k \cap \overline{A}_{k+1} \cap \cdots \cap \overline{A}_n).$$

To compute this probability, we use the multiplication rule:

$$\begin{aligned}
&P(A_1 \cap \cdots \cap A_k \cap \overline{A}_{k+1} \cap \cdots \cap \overline{A}_n) \\
&\quad = P(A_1) \cdot P(A_2|A_1) \cdot \;\cdots\; \cdot P\left(A_k | \bigcap_{i=1}^{k-1} A_i\right) \cdot P\left(\overline{A}_{k+1} | \bigcap_{i=1}^{k} A_i\right) \\
&\quad \cdot \;\cdots\; \cdot P\left(\overline{A}_n | \bigcap_{i=1}^{k} A_i \bigcap \bigcap_{j=k+1}^{n-1} \overline{A}_j\right) \\
&\quad = \frac{n_1}{N} \cdot \frac{n_1-1}{N-1} \cdot \;\cdots\; \cdot \frac{n_1-k+1}{N-k+1} \cdot \frac{N-n_1}{N-k} \cdot \frac{N-n_1-1}{N-k-1} \\
&\quad \cdot \;\cdots\; \cdot \frac{N-n_1-n+k+1}{N-n+1} = \frac{\dfrac{n_1!}{(n_1-k)!} \cdot \dfrac{(N-n_1)!}{(N-n_1-n+k)!}}{\dfrac{N!}{(N-n)!}}.
\end{aligned}$$

Then

$$P(B_{n,k}) = \frac{\dfrac{n_1!}{k!(n_1-k)!} \cdot \dfrac{(N-n_1)!}{(n-k)!(N-n_1-n+k)!}}{\dfrac{N!}{n!(N-n)!}} = \frac{C_{n_1}^k \cdot C_{N-n_1}^{n-k}}{C_N^n}.$$

□

Remark 2.3.2.
1. The probability $P(n;k)$ given by (2.3) can be also computed using the classical definition of probability. Indeed, the total number of possible outcomes for the experiment is C_N^n. There are $C_{n_1}^k$ ways of choosing the k white balls and $C_{N-n_1}^{n-k}$ ways of choosing the $n-k$ black balls (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. Since the events $B_{n,k}$ defined in the proof of proposition 2.3.1 form a

partition of the sample space, we have

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

i.e.

$$\sum_{k=0}^{n} C_{n_1}^{k} C_{N-n_1}^{n-k} = C_N^n .$$

Example 2.3.3. 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 2.3.3: For event A, an equal number of boys and girls out of 10 people, means 5 boys and 5 girls. Therefore

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

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

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

■

2.4 Bernoulli Model Without Replacement With r States

Just as before, we now generalize this model, by allowing more than two outcomes for each trial, i.e. the balls in the box will have more than two colors. The sampling is still without replacement.

Bernoulli Model Without Replacement With r States. A box contains M balls of r colors, m_i balls of color i, $i = \overline{1,r}$ $(M = m_1 + \cdots + m_r)$. A number of n $(0 \leq n \leq M)$ balls are extracted, one at a time, without replacement. Find the probability $P(n; n_1, \ldots, n_r)$ that in n trials, n_i $(0 \leq n_i \leq m_i,\ i = \overline{1,r})$ balls of color i are extracted $(n = n_1 + \ldots + n_r)$.

Proposition 2.4.1. *The probability $P(n; n_1, \ldots, n_r)$ in a Bernoulli model without replacement with r states is given by*

$$P(n; n_1, \ldots, n_r) = \frac{C_{m_1}^{n_1} \ldots C_{m_r}^{n_r}}{C_M^n}. \tag{2.4}$$

The proof is similar to that of proposition 2.3.1 and we skip it.

Remark 2.4.2. Let $I = \{(n_1, \ldots, n_r) \mid 0 \leq n_i \leq m_i,\ n_1 + \ldots + n_r = n\}$. The sum of the probabilities in (2.4) for all indices in I satisfy

$$\sum_{(n_1, \ldots n_r) \in I} P(n; n_1, \ldots, n_r) = 1.$$

Example 2.4.3. In a game of bridge, find the probability that a player gets
a) 4 spades, 4 hearts, 3 diamonds and 2 clubs (event A);
b) a uniform $4-4-3-2$ distribution of cards (event B).

Solution 2.4.3:
a) There are 52 cards in a deck, 13 of each suit and a player gets 13 cards. Hence

$$P(A) = P(13; 4, 4, 3, 2) = \frac{C_{13}^4 C_{13}^4 C_{13}^3 C_{13}^2}{C_{52}^{13}} \approx 0.0179\,.$$

b) Now the suits are not specified, only their distribution, $4-4-3-2$. So, they can be interchanged and the number of ways they can be interchanged (order does matter!) is $A_4^4 = P_4 = 4!$. However, since two of the numbers in the distribution are the same, 4, when those two suits are interchanged, we don't get a new distribution, so those cases are counted twice. Thus

$$P(B) = \frac{4!}{2} P(A) = 12 \cdot P(13; 4, 4, 3, 2) \approx 0.2155.$$

■

2.5 Poisson's Model

This model is also a generalization of the binomial model, in the sense that it allows the probability of success to vary at each trial.

Poisson's Model. Consider an experiment where in each trial there are only 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 repeated trials, exactly k $(0 \leq k \leq n)$ successes occur.

Proposition 2.5.1. *The probability $P(n; k)$ in a Poisson's model is given by*

$$P(n; k) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} p_{i_1} \ldots p_{i_k} q_{i_{k+1}} \cdots q_{i_n}, \tag{2.5}$$

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

Proof. Let $B_{n,k}$ denote the event of interest. For $i = \overline{1, n}$ let A_i denote the event: success occurred in the i^{th} trial. Then

$$B_{n,k} = \bigcup_{I} \left((A_{i_1} \cap \cdots \cap A_{i_k}) \cap (\overline{A}_{i_{k+1}} \cap \cdots \cap \overline{A}_{i_n}) \right),$$

$$I = \{(i_1, \cdots, i_k) \mid$$
$$1 \leq i_1 < \cdots < i_k \leq n, i_{k+1}, \cdots, i_n \in \{1, \cdots, n\} \setminus \{i_1, \cdots, i_k\}\}.$$

This is a disjoint union and the events $\{A_i\}_{i=\overline{1,n}}$ are independent. Thus

$$P(n,k) = P(B_{n,k}) = \sum_{1\le i_1<\cdots<i_k\le n} p_{i_1}\ldots p_{i_k} q_{i_{k+1}}\ldots q_{i_n}.$$

□

Remark 2.5.2.
1. The number $P(n;k)$ in (2.5) is the coefficient of x^k in the polynomial

$$(p_1x+q_1)\ldots(p_nx+q_n) = \sum_{k=0}^{n} P(n;k)x^k.$$

In fact, this will provide an easier computational formula for $P(n;k)$ than (2.5).
2. If $p_i = p$ (and consequently, $q_i = q$), $\forall i = \overline{1,n}$, then this becomes the binomial model and (2.5) is reduced to (2.1).
2. As a consequence, again $\sum\limits_{k=o}^{n} P(n;k) = 1$ (if $x = 1$ in the above equality).

Example 2.5.3. (The Three Shooters Problem) Three shooters aim at a target and they hit it with probability 0.4, 0.5 and 0.7, respectively. Each of them shoots once. Find the probability p that the target is hit once.

Solution 2.5.3: Define "success" as "the target is hit". Then we have $n = 3$ 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 the remark above, it is equal to the coefficient of x in the polynomial

$$(0.4x+0.6)(0.5x+0.5)(0.7x+0.3),$$

i.e. $p = 0.36$. ■

2.6 Pascal's Model (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. Another novelty is that in this model we have an infinite number of trials.

Pascal's Model (Negative Binomial 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 ($n \in \mathbb{N},\ k \in \mathbb{N} \cup \{0\}$).

Proposition 2.6.1. *The probability $P(n, k)$ in a Pascal's model is given by*

$$P(n, k) = C^{k}_{n+k-1} p^n q^k. \tag{2.6}$$

Proof. Consider the following events:

E: the n^{th} success occurs in the $(n + k)^{\text{th}}$ trial,
A: $n - 1$ successes occur in $n + k - 1$ trials,
B: success occurs in the $(n + k)^{\text{th}}$ trial.

Then $E = A \cap B$ (the event of interest) and A, B are independent, so $P(E) = P(A)P(B)$. Obviously $P(B) = p$.
To compute $P(A)$, we use the binomial model:

$$P(A) = P(n + k - 1; n - 1) = C^{\,n-1}_{n+k-1} p^{n-1} q^k.$$

So

$$P(E) = C^{\,n-1}_{n+k-1} p^n q^k.$$

□

Remark 2.6.2.
1. The number $P(n, k)$ in (2.6) 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 negative binomial.

2. As before, we have $\sum_{k=0}^{\infty} P(n,k) = 1.$

2.7 Geometric Model

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

Geometric Model. In an infinite sequence of Bernoulli trials with probability of success p (and probability of failure $q = 1-p$), find the probability p_k of the 1st success occurring after k failures ($k \in \mathbb{N} \cup \{0\}$).

All the properties derived in Section 2.6 translate for the geometric model, as well.

Proposition 2.7.1. *The probability p_k in a geometric model is given by*

$$p_k = pq^k. \tag{2.7}$$

Remark 2.7.2.
1. The number p_k in (2.7) is the coefficient of x^k in the geometric expansion

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

hence, the name geometric.

2. We have $\sum_{k=0}^{\infty} p_k = 1.$

Example 2.7.3. When a die is rolled, find the probability of the following events:
A: the first 6 appears after 5 throws;
B: the 3^{rd} even appears after 5 throws.

Solution 2.7.3:
For event A, success means that face 6 appears, hence $p = \frac{1}{6}$. We want the first success to occur after 5 failures, so this is a geometric model. By (2.7), we have

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

For event B, success means that an even number shows, so $p = \frac{1}{2}$. The 3rd even appears after 5 throws, which means after 3 odds, i.e. after 3 failures. Thus, using (2.6), we have

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

■

Chapter 3

Random Variables and Random Vectors

In order to describe random phenomena in a more rigorous way, we need to take a different look at σ-fields associated with an experiment. That is, to describe them quantitatively, in the form of *random variables* $-$ variables whose observed values are determined by chance. The concept of a random variable is one of the most significant in probability theory, setting the fundaments of modern statistics. Random variables fall into one of two categories: *discrete* or *continuous*.

3.1 Discrete Random Variables and Probability Distribution Function

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

Definition 3.1.1. *A* ***discrete random variable*** *is a function $X : S \to \mathbb{R}$ satisfying the following two conditions:*

(i) *the set of values that X takes is at most countable;*

(ii) $X^{-1}(x) = \{e \in S \mid X(e) = x\} = (X = x) \in \mathcal{K}$, *for all* $x \in \mathbb{R}$.

Example 3.1.2. 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 number i appeared, $i = \overline{1,6}$.

a) Let $\mathcal{K} = P(S)$ and let $X(e_i) = i,\ i = \overline{1,6}$.
The set of values that X takes is $\{1, \ldots, 6\}$, which is finite, so condition (i) of definition 3.1.1 is satisfied.
To verify the second condition, let $x \in \mathbb{R}$. If $x = i \in \{1, \ldots, 6\}$, then $X^{-1}(x) = \{e_i\} \in \mathcal{K}$, otherwise $X^{-1}(x) = \emptyset \in \mathcal{K}$, so (ii) is satisfied as well and X defines a discrete random variable.

b) Let $\mathcal{K} = \{\emptyset, \{e_1, e_3, e_5\}, \{e_2, e_4, e_6\}, S\}$. It is easy to check that $\mathcal{K}$ is a σ-field over S. Let again $X(e_i) = i,\ i = \overline{1,6}$.
Condition (i) is obviously still satisfied. But for $x = i \in \{1, \ldots, 6\}$, $X^{-1}(x) = \{e_i\} \notin \mathcal{K}$, thus X is **not** a discrete random variable.

Example 3.1.3. (The **indicator** of an event) Let $A \in \mathcal{K}$ and $X_A : S \to \mathbb{R}$ be defined by

$$X_A(e) \quad = \quad \begin{cases} 1, & \text{if } e \in A \\ 0, & \text{if } e \in \overline{A}. \end{cases} \tag{3.1}$$

Then $X_A(S) = \{0, 1\}$, which is finite, so (i) is satisfied and

$$\begin{aligned} X_A^{-1}(0) &= \overline{A} \in \mathcal{K}, \\ X_A^{-1}(1) &= A \in \mathcal{K}, \\ X_A^{-1}(x) &= \emptyset \in \mathcal{K},\ \forall x \in \mathbb{R} \setminus \{0, 1\}, \end{aligned}$$

thus, (ii) is also fulfilled.

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

Example 3.1.4. Let $n \in \mathbb{N}$ and $\{A_i\}_{i=\overline{1,n}} \subseteq \mathcal{K}$ be a partition of S. Let $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}$ and define $X : S \to \mathbb{R}$ by

$$X(e) = \sum_{i=1}^{n} x_i X_{A_i}(e), \tag{3.2}$$

i.e. if $e \in A_i$, then $X(e) = x_i$. Now, $X(S) = \{x_1, \ldots, x_n\}$, finite and

$$\begin{aligned} X^{-1}(x) &= A_i \in \mathcal{K}, \forall x = x_i \in \{x_1, \ldots, x_n\}, \\ X^{-1}(x) &= \emptyset \in \mathcal{K}, \ \forall x \in \mathbb{R} \setminus \{x_1, \ldots, x_n\}. \end{aligned}$$

Definition 3.1.5. *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. In fact, example 3.1.4 provides the general form of a simple discrete random variable.

Proposition 3.1.6. *Any simple discrete random variable has an expression of the form (3.2).*

Proof. The simple discrete random variable X takes only a finite set of distinct values, say $X(S) = \{x_1, \ldots, x_n\}$. For $i = \overline{1,n}$, define the events A_i by

$$A_i = (X = x_i) = \{e \in S \mid X(e) = x_i\}.$$

Then it is easy to check that the events $\{A_i\}_{i=\overline{1,n}}$ so defined form a partition of S and that

$$X(e) = \sum_{i=1}^{n} x_i X_{A_i}(e), \ \forall e \in S.$$

□

Furthermore, (3.2) leads us to the general form of a discrete random variable.

Proposition 3.1.7. *Let $X : S \to \mathbb{R}$ be a discrete random variable with $X(S) = \{x_i \mid i \in I\}$, $I \subseteq \mathbb{N}$ a countable set. Then X can be expressed in the form*

$$X(e) = \sum_{i \in I} x_i X_{A_i}(e),\ \forall e \in S, \tag{3.3}$$

where the set $\{A_i\}_{i \in I} \subseteq \mathcal{K}$ forms a partition of S.

The proof is analogous to the one of proposition 3.1.6.

Definition 3.1.8. *Let $X : S \to \mathbb{R}$ be a discrete random variable. The **probability distribution (function)** of X is an array of the form*

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}, \tag{3.4}$$

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

Remark 3.1.9.
1. All values $x_i,\ i \in I$, in (3.4) 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 probability distribution function (3.4).
3. If X is a discrete random variable with probability distribution function (3.4), then

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

(a necessary and sufficient condition for such an array to represent a probability distribution function 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 probability distribution function and use (3.4) to describe it.

Example 3.1.10.
1. Consider the discrete random variable defined in example 3.1.2. Its probability distribution function is

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

2. For the indicator random variable defined in example 3.1.3, the probability distribution function is

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

where $p = P(A)$.

3.2 Cumulative Distribution Function

Definition 3.2.1. *Let X be a random variable. The function $F_X : \mathbb{R} \to \mathbb{R}$, defined by*

$$F_X(x) = P(X < x), \tag{3.5}$$

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

Example 3.2.2. Consider the indicator random variable from Example 3.1.3. It only takes the two values 0 and 1. So, for $x \le 0$, $F_A(x) = P(X_A < x) = 0$. If $0 < x \le 1$, the only value less than x that X can take is 0, and that happens with probability $1-p$. Hence, for this case, $F_A(x) = P(X_A < x) = 1-p$. Finally, for $x > 1$, all the values that X takes are less than x, so, for this case, $F_A(x) = P(X_A < x) = 1-p+p = 1$. Thus

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

The graphic representation of F_A is given in Figure 3.1.

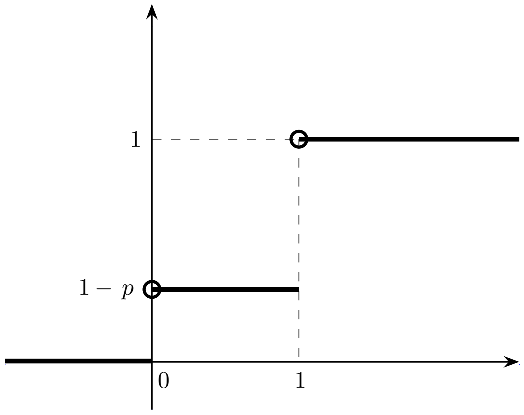


Fig. 3.1: Cumulative distribution function for the indicator of an event

Example 3.2.3. It easily follows from the previous example that for a discrete random variable

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

the cumulative distribution function is computed by

$$F(x) = \sum_{x_i<x} p_i. \tag{3.6}$$

Remark 3.2.4. We can see from Figure 3.1 some important properties of F_A, such as the fact that all its values are in $[0,1]$ (since it represents a probability), it is nondecreasing, continuous from the left at all points and continuous at all points except 0 and 1 (which are the values that the random variable X_A takes) and the end behavior, i.e. $\lim\limits_{x\to-\infty} F_A(x) = 0$ and $\lim\limits_{x\to\infty} F_A(x) = 1$. These properties are common to all cumulative distribution functions, as can be seen from the next theorem.

Theorem 3.2.5. *Let X be a random variable and let $F : \mathbb{R} \to \mathbb{R}$ be its cumulative distribution function . Then F has the following properties:*

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

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

(3) *F is left continuous, i.e. $F(x-0) = F(x)$, for every $x \in \mathbb{R}$, where $F(x-0) = \lim\limits_{y \nearrow x} F(y)$ is the limit from the left at x.*

(4) *$\lim\limits_{x \to -\infty} F(x) = 0$ and $\lim\limits_{x \to \infty} F(x) = 1$.*

(5) *$P(X \leq x) = F(x+0) = \lim\limits_{y \searrow x} F(y)$ and $P(X = x) = F(x+0) - F(x)$, for every $x \in \mathbb{R}$.*

(6) *The set of all points of jump discontinuity of F is at most countable.*

Proof.
(1) By theorem 1.2.7(3) and since the events $(X < a) \subseteq (X < b)$, we have

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

(2) Let $a < b$. Then $0 \leq P(a \leq X < b) = F(b) - F(a)$.
(3) Let $x \in \mathbb{R}$ be fixed and let $\{x_n\}_{n \in \mathbb{N}} \subseteq \mathbb{R}$ be an increasing sequence converging to x, $x_n \nearrow x$. Define the sequence of events $\{A_n\}_{n \in \mathbb{N}}$ by

$$A_n = (x_n \leq X < x), \ \forall n \in \mathbb{N}.$$

Then $A_{n+1} \subseteq A_n$ and $\lim\limits_{n \to \infty} P(A_n) = P\left(\bigcap\limits_{n \in \mathbb{N}} A_n\right) = P(\emptyset) = 0$.

On the other hand,

$$\begin{aligned} \lim_{n \to \infty} P(A_n) &= \lim_{n \to \infty} P((x_n \leq X < x) \\ &= \lim_{n \to \infty} (F(x) - F(x_n)) \\ &= F(x) - F(x-0). \end{aligned}$$

(4) Let $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$ be an increasing sequence going to ∞. Consider the sequence of events $\{A_n\}_{n\in\mathbb{N}}$ given by

$$A_n = (X < x_n), \forall n \in \mathbb{N}.$$

We have $A_n \subseteq A_{n+1}$ and so $\lim\limits_{n\to\infty} P(A_n) = P\left(\bigcup\limits_{n\in\mathbb{N}} A_n\right) = P(S) = 1$.
But

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} P((X < x_n) = \lim_{n\to\infty} F(x_n) = \lim_{x\to\infty} F(x).$$

Similarly, using a decreasing sequence of real numbers going to $-\infty$, we obtain that $\lim\limits_{x\to-\infty} F(x) = 0$.
(5) Let $x \in \mathbb{R}$ be fixed and define the sequence of events $\{A_n\}_{n\in\mathbb{N}}$ by

$$A_n = \left(X < x + \frac{1}{n}\right), \forall n \in \mathbb{N}.$$

Then $A_{n+1} \subseteq A_n$ and so

$$\begin{aligned} P(X \leq x) &= P\left(\bigcap_{n\in\mathbb{N}} \left(X < x + \frac{1}{n}\right)\right) = P\left(\bigcap_{n\in\mathbb{N}} A_n\right) \\ &= \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} P\left(X < x + \frac{1}{n}\right) \\ &= \lim_{n\to\infty} F\left(x + \frac{1}{n}\right) = F(x+0). \end{aligned}$$

For the second part, we have

$$(X = x) = (X \leq x) \setminus (X < x)$$

and by the first part and theorem 1.2.7(3),

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

(6) We will take a closer look at the jump discontinuities of F and find a way to "count" them. Namely, we consider the length of such a jump.
Since $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$ and F is monotonely increasing, it follows that F can have at most $1 = 2^1 - 1$ jump of length greater than $\frac{1}{2}$, otherwise, their lengths put together would make up a value greater than 1, which is the total length of the interval in which F takes values.
By the same reasoning, F can have at most $3 = 2^2 - 1$ jumps of length greater than $\frac{1}{2^2}$ and in general, for any $n \in \mathbb{N}$, F can have at most $2^n - 1$ jumps of length greater than $\frac{1}{2^n}$.
Since a jump of F of any length would have to fall into one of these categories (i.e. its length would have to be greater than $\frac{1}{2^n}$ for some $n \in \mathbb{N}$), the number of jump discontinuities is at most $\sum_{n=1}^{\infty}(2^n - 1)$, and thus, countable. □

Remark 3.2.6.
1. Any function $F : \mathbb{R} \to \mathbb{R}$ satisfying conditions (2), (3) and (4) of Theorem 3.2.5 is a cumulative distribution function , i.e. for any such function, there exists a random variable whose cumulative distribution function is the function F.
2. The cumulative distribution function does not uniquely identify a random variable; there are distinct random variables having the same cumulative distribution function.

3.3 Discrete Distributions

We give a list of some of the most common discrete probability distributions.

Bernoulli Distribution $Bern(p)$

We say that a random variable X follows a *Bernoulli distribution* with parameter $p \in (0,1)$ if its probability distribution function is

$$X \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}. \tag{3.7}$$

Notice that this is the probability distribution function of the indicator random variable from Example 3.1.10(2).

Discrete Uniform Distribution $U(m)$

We say that a random variable X follows a *discrete uniform distribution* with parameter $m \in \mathbb{N}$, if its probability distribution function is

$$X \begin{pmatrix} k \\ \frac{1}{m} \end{pmatrix}_{k=\overline{1,m}}. \tag{3.8}$$

The random variable in Example 3.1.10(1) follows a discrete uniform distribution $U(6)$.

Binomial Distribution $B(n,p)$

We say that a random variable X follows a *binomial distribution* with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ ($q = 1-p$), if its probability distribution function is

$$X \begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}_{k=\overline{0,n}}. \tag{3.9}$$

This distribution corresponds to the binomial model. Given n Bernoulli trials with probability of success p, let X denote the number of successes. Then $X \in B(n,p)$. (see Proposition 2.1.1).

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

We say that a random variable X follows a *hypergeometric distribution* with parameters $N, n_1, n \in \mathbb{N}$ ($n, n_1 \leq N$), if its probability distribution function is

$$X\begin{pmatrix} k \\ \dfrac{C_{n_1}^k C_{N-n_1}^{n-k}}{C_N^n} \end{pmatrix}_{k=\overline{0,n}} . \tag{3.10}$$

This distribution corresponds to the hypergeometric model. If X denotes the number of successes in a hypergeometric model, then $X \in H(N, n_1, n)$ (see Proposition 2.3.1).

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

We say that a random variable X follows a *Poisson distribution* with parameter $\lambda > 0$, if its probability distribution function is

$$X\begin{pmatrix} k \\ \dfrac{\lambda^k}{k!}e^{-\lambda} \end{pmatrix}_{k=0,1,\ldots} \tag{3.11}$$

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 process represents the average number of occurrences of the event in question per measurement unit (this will be discussed in more detail in the next chapter).

Another important property of the Poisson distribution is that, under certain conditions (n large, p small and np approaching a constant), it provides a good approximation for the binomial distribution with parameters n and p. We state this result (first presented by Poisson in 1837) without proof (for the proof, see e.g. [5]).

Theorem 3.3.1. *Let X be a random variable following a binomial distribution with parameters n and p_n (depending on n). If $\lim\limits_{n\to\infty} np_n = \lambda > 0$, then for $n \to \infty$, X follows a Poisson distribution with parameter λ.*

Remark 3.3.2. 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".

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

We say that a random variable X follows a *negative binomial (Pascal) distribution* with parameters $n \in \mathbb{N}$ and $p \in (0, 1)$, if its probability distribution function is

$$X \begin{pmatrix} k \\ C^k_{n+k-1} p^n q^k \end{pmatrix}_{k=0,1,\dots}. \tag{3.12}$$

This distribution corresponds to the negative binomial model. If X denotes the number of failures that occurred before the occurrence of the nth success in a negative binomial model, then $X \in NB(n, p)$. (see Proposition 2.6.1).

Geometric Distribution $Geo(p)$

As before, we have an important particular case for the negative binomial distribution; if $n = 1$ in the previous distribution, then we have a *geometric distribution*. We say that a random variable X follows a geometric distribution with parameter $p \in (0, 1)$, if its probability distribution function is given by

$$X \begin{pmatrix} k \\ pq^k \end{pmatrix}_{k=0,1,\dots}. \tag{3.13}$$

If X denotes the number of failures that occurred before the occurrence of the 1st success in a geometric model, then $X \in Geo(p)$. (see Proposition 2.7.1).

3.4 Discrete Random Vectors and Joint Probability Distribution Function

Definition 3.4.1. *Let $(S, \mathcal{K}, P)$ be a probability space. A* ***discrete random vector*** *is a function $X = (X_1, \dots, X_n) : S \to \mathbb{R}^n$ satisfying the following two conditions:*

(i) *the set of values that X takes is at most countable;*

(ii) *for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$,*

$$\begin{aligned} X^{-1}(x) &= \{e \in S \mid X_1(e) = x_1, \dots, X_n(e) = x_n\} \\ &= (X_1 = x_1, \dots, X_n = x_n) \in \mathcal{K}. \end{aligned}$$

For the remainder of this section, we will restrict our discussion to a two-dimensional discrete random vector $(X, Y) : S \to \mathbb{R}^2$.

Definition 3.4.2. *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*

$$\begin{array}{c|cccc|c} X \setminus Y & y_1 & \dots & y_j & \dots & \\ \hline x_1 & & & & & \\ \vdots & & & \vdots & & \\ x_i & & \cdots & p_{ij} & \cdots & p_i \\ \vdots & & & \vdots & & \\ & & & & & \\ \hline & & & q_j & & \end{array} \tag{3.14}$$

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 3.4.3. *Let (X,Y) be a random vector with joint probability distribution given by (3.14). Then*

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

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

Proof. We will prove only the first assertion, the second one following by symmetry. First, let us note that the sets of events $\{(X = x_i)\}_{i\in I}$ and $\{(Y = y_j)\}_{j\in J}$ both form a partition of the sample space. Then we have

$$\begin{aligned}
\sum_{j\in J} p_{ij} &= \sum_{j\in J} P(X = x_i, Y = y_j) \\
&= P\left(\bigcup_{j\in J}\Big((X = x_i)\bigcap(Y = y_j)\Big)\right) \\
&= P\left((X = x_i)\bigcap\left(\bigcup_{j\in J}(Y = y_j)\right)\right) \\
&= P\Big((X = x_i)\bigcap S\Big) \ = \ P\left(X = x_i\right) \ = \ p_i.
\end{aligned}$$

□

Definition 3.4.4. *Two discrete random variables X and Y with probability distribution functions*

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

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

$$p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(X = x_i\right)P\left(Y = y_j\right) = p_i q_j, \qquad (3.15)$$

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

3.5 Operations with Discrete Random Variables

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 probability distribution function given by

$$X + Y\begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{3.16}$$

Product. The product of X and Y is the random variable with probability distribution function given by

$$X \cdot Y\begin{pmatrix} x_i y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}. \tag{3.17}$$

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

$$\alpha X\begin{pmatrix} \alpha x_i \\ p_i \end{pmatrix}_{i\in I}. \tag{3.18}$$

Quotient. The quotient of X and Y is the random variable with probability distribution function given by

$$X/Y\begin{pmatrix} x_i/y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}, \tag{3.19}$$

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 probability distribution function given by

$$h(X) \begin{pmatrix} h(x_i) \\ p_i \end{pmatrix}_{i \in I}. \tag{3.20}$$

Remark 3.5.1. If X and Y are independent discrete random variables, then in (3.16), (3.17) and (3.19), $p_{ij} = p_i q_j$, for all $(i, j) \in I \times J$.

3.6 Continuous Random Variables and Probability Density Function

So far, we mentioned random variables, in general, and defined discrete random variables. The only other type is a *continuous* random variable. Before we give its precise definition, let us extend Definition 3.1.1 to a random variable, in general, and then particularize it for the continuous case.

Definition 3.6.1. *Let $(S, \mathcal{K}, P)$ be a probability space. A* ***random variable*** *is a function $X : S \to \mathbb{R}$ satisfying the condition*

$$X^{-1}((-\infty, x)) = \{e \in S \mid X(e) < x\} = (X < x) \in \mathcal{K}, \tag{3.21}$$

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

Remark 3.6.2. Notice that the definition of the cumulative distribution function given in Definition 3.2.1 still works for a general random variable (since the event $(X < x) \in \mathcal{K}$, $P(X < x)$ is well defined). Also, the properties of a cumulative distribution function stated in Theorem 3.2.5 still hold for a general random variable.

Definition 3.6.3. *Let X be a random variable, and let $F : \mathbb{R} \to \mathbb{R}$ be its cumulative distribution function . We say that X is a **continuous random variable** if F is absolutely continuous, i.e. there exists a real function $f : \mathbb{R} \to \mathbb{R}$, such that*

$$F(x) = \int\limits_{-\infty}^{x} f(t)\, dt, \tag{3.22}$$

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

Definition 3.6.4. *Let X be a continuous random variable. Then the function f from the previous definition is called the **(probability) density function** of X.*

Unlike discrete random variables, continuous random variables assume an uncountable set of values, usually an interval. Also, 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.

Theorem 3.6.5. *Let X be a continuous random variable with cumulative distribution function F and density function f. Then the following properties hold:*

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

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

(3) $\int\limits_{\mathbb{R}} f(t)dt = 1.$

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

$$\begin{aligned} P(a < X < b) &= P(a \leq X < b) = P(a \leq X \leq b) \\ &= P(a < X \leq b) = \int\limits_{a}^{b} f(t)\, dt\,. \end{aligned} \tag{3.23}$$

Proof.
(1) This property follows directly from the definition of a continuous random variable, by differentiating both sides of (3.22).

(2) Recall from Theorem 3.2.5(2) that F is monotonely increasing. Thus its derivative is nonnegative, for every $x \in \mathbb{R}$.

(3) Using (3.22) and Theorem 3.2.5(4), we have

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

(4) To prove the first part, let $x \in \mathbb{R}$ be fixed and recall from Theorem 3.2.5(5) that $P(X = x) = F(x+0) - F(x)$. 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 (3.22) and Theorem 3.2.5(1), we have

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

which, by the first part, is equal to all the other probabilities in (3.23).

□

Remark 3.6.6. Any integrable function $f : \mathbb{R} \to \mathbb{R}$ satisfying conditions (2) and (3) of Theorem 3.6.5 is a density function, i.e. for any such function, there exists a random variable whose density is f.

3.7 Continuous Distributions

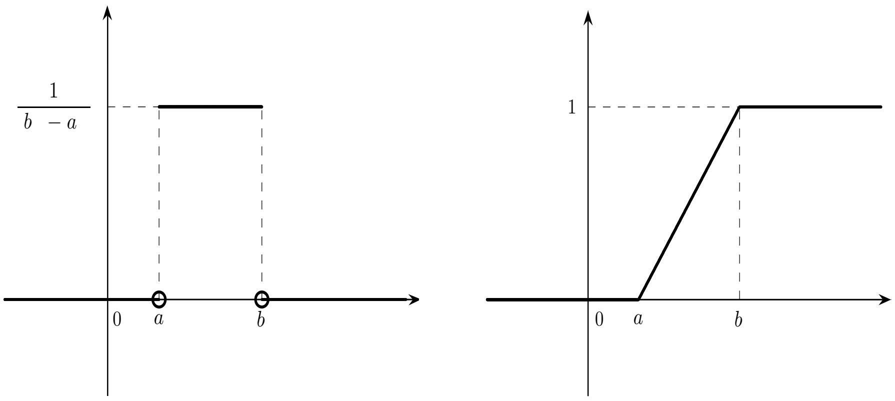
Uniform Distribution on an Interval $\mathcal{U}(a,b)$

We say that a random variable X follows a *uniform distribution* with parameters $a, b \in \mathbb{R},\ a < b,$ if its density function is

$$f(x) \quad = \quad \begin{cases} \dfrac{1}{b-a}\,, & \text{if } \ x \in [a,b] \\ 0, & \text{if } \ x \notin [a,b]. \end{cases} \tag{3.24}$$

Then by (3.22), its cumulative distribution function is

$$F(x) = \int\limits_{-\infty}^{x} f(t)dt \quad = \quad \begin{cases} 0, & \text{if } \ x < a \\ \dfrac{x-a}{b-a}\,, & \text{if } \ a \le x \le b \\ 1, & \text{if } \ x > b\,. \end{cases} \tag{3.25}$$



(a) Density Function

(b) Cumulative Distribution Function

Fig. 3.2: Uniform Distribution

The density function and cumulative distribution function of a random variable uniformly distributed on the interval $[a, b]$ are given in Figure 3.2.

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

We say that a random variable X follows a *normal distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, if its density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \;\; x \in \mathbb{R}. \tag{3.26}$$

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 (independent of each other) rediscovered it in conjunction with the behavior of errors in astronomical measurements. It is also referred to as the "Gaussian" distribution.

Let us take a closer look at its density. Taking the derivative in (3.26), we have

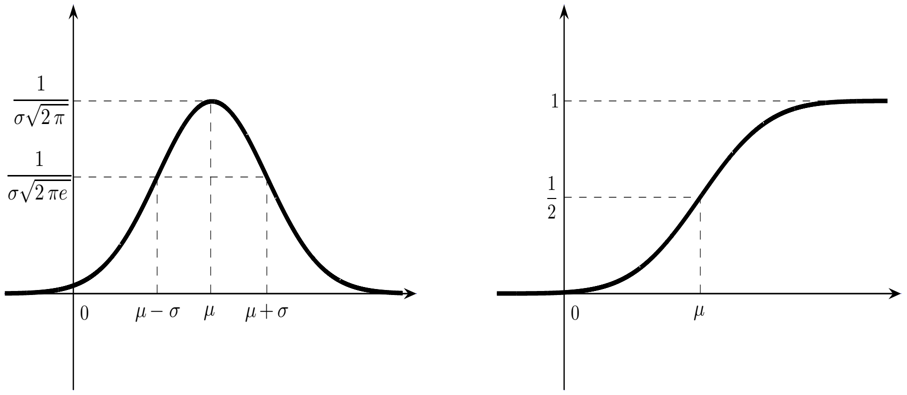
$$f'(x) = 0 \;\; <=> \;\; x = \mu \text{ and } f(\mu) = \frac{1}{\sigma\sqrt{2\pi}},$$

so there is one critical point at $\left(\mu, \dfrac{1}{\sigma\sqrt{2\pi}}\right)$. Differentiating again, we get

$$f''(x) = 0 \;\; <=> \;\; x = \mu \pm \sigma \text{ and } f(\mu \pm \sigma) = \frac{1}{\sigma\sqrt{2\pi e}},$$

so there are two inflection points, $\left(\mu - \sigma, \dfrac{1}{\sigma\sqrt{2\pi e}}\right)$ and $\left(\mu + \sigma, \dfrac{1}{\sigma\sqrt{2\pi e}}\right)$.

The graph of the normal density is a symmetric, bell-shaped curve (known as "Gauss's bell") centered at the value of the first parameter μ, as can be seen in Figure 3.3(a).



(a) Density Function (b) Cumulative Distribution Function

Fig. 3.3: Normal Distribution

The cumulative distribution function of a normal variable is then given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}}\, dt. \qquad (3.27)$$

The graph of the cumulative distribution function of a normally distributed random variable is given in Figure 3.3(b).

Remark 3.7.1.

1. There is an important particular case of a normal distribution, namely $N(0,1)$, called the *standard (or reduced) normal distribution*. A variable following the standard normal distribution is usually denoted by Z. The density and cumulative distribution function of Z are given by

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \;\; x \in \mathbb{R} \qquad (3.28)$$

and

$$F_Z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt. \tag{3.29}$$

The function given in (3.29) is known as *Laplace's function*.
2. The values of Laplace's function can be found in tables (see Appendix B.1) or can be computed by most mathematical software.
3. As noticed from (3.27) and (3.29), there is a relationship between the cumulative distribution function of any normal $N(\mu, \sigma)$ variable X and that of a standard normal variable Z, namely

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

Student (T) Distribution $T(n)$

We say that a random variable X follows a *Student (or T) distribution* with parameter $n \in \mathbb{N}$ (number of degrees of freedom), if its density function is

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}. \tag{3.30}$$

The graphs of the density and cumulative distribution functions of the Student distribution look similar to those of the normal distribution.

Gamma Distribution $Gamma(a, b)$

We say that a random variable X follows a *Gamma distribution* with parameters $a, b > 0$, if its density function is

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(a)\, b^a}\, x^{a-1} e^{-\frac{x}{b}}, & \text{if } x > 0 \\ 0, & \text{if } x \le 0, \end{cases} \tag{3.31}$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

is Euler's Gamma function (for more properties of Γ, see Appendix A.1).

Exponential Distribution $Exp(\lambda)$

We say that a random variable X follows an *exponential distribution* with parameter $\lambda > 0$, if its density function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{3.32}$$

Remark 3.7.2. The exponential distribution is a particular case for the Gamma distribution, with $a = 1$ and $b = \dfrac{1}{\lambda}$.

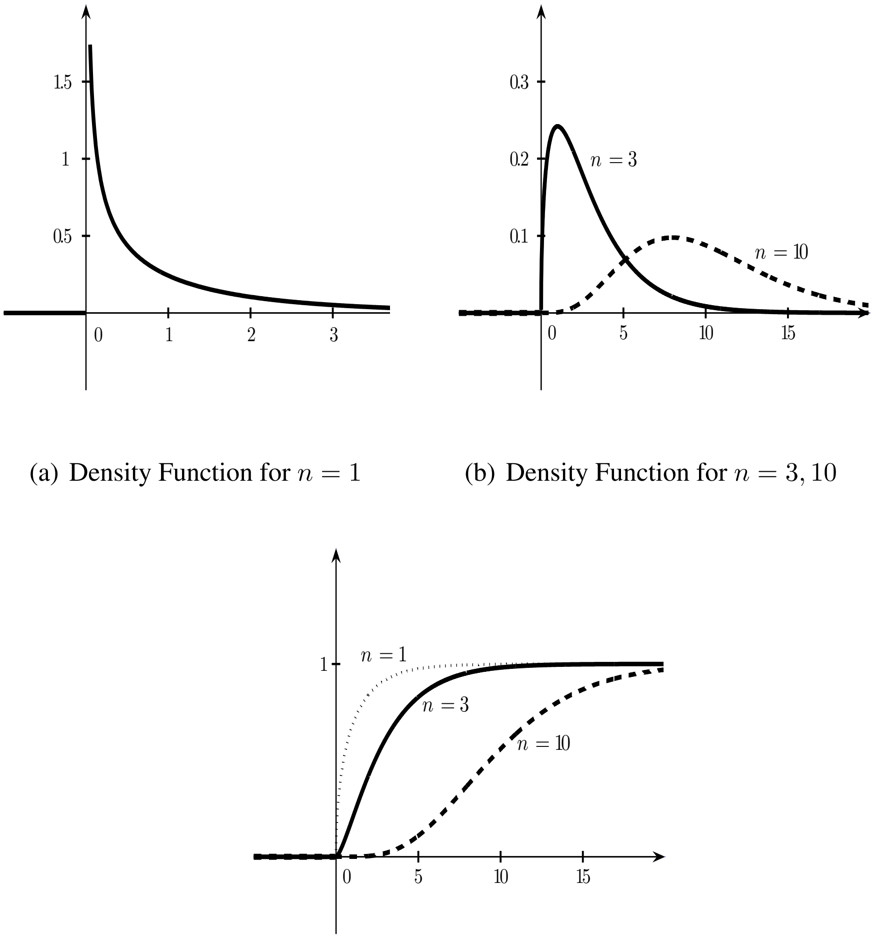
Chi-squared Distribution $\chi^2(n, \sigma)$

We say that a random variable X follows a χ^2 *distribution* with parameters $n \in \mathbb{N}$ and $\sigma > 0$, if its density function is

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{n}{2})\, 2^{\frac{n}{2}} \sigma^n} x^{\frac{n}{2}-1} e^{-\dfrac{x}{2\sigma^2}}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{3.33}$$

The graph of the χ^2 density function is no longer symmetric. A few examples, for $\sigma = 1$ and $n = 1, 3, 10$ are presented in Figure 3.4(a) and (b). Also, the graph of the cumulative distribution function is given for the same values of the parameters in Figure 3.4(c).

Remark 3.7.3. The χ^2 distribution is a particular case for the Gamma distribution, with $a = \dfrac{n}{2}$ and $b = 2\sigma^2$.



(a) Density Function for $n = 1$

(b) Density Function for $n = 3, 10$

(c) Cum. Distribution Function for $n = 1, 3, 10$

Fig. 3.4: χ^2 Distribution

Erlang Distribution $Erl(a, r)$

We say that a random variable X follows an *Erlang distribution* with parameters $a, r > 0$, if its density function is

$$f(x) = \begin{cases} \dfrac{ar}{\Gamma(a)} \, (arx)^{a-1} \, e^{-arx}, & \text{if } x > 0 \\ \quad 0, & \text{if } x \leq 0. \end{cases} \tag{3.34}$$

Remark 3.7.4. The Erlang distribution is a particular case for the Gamma distribution, with $b = \dfrac{1}{ar}$.

Beta Distribution $B(a, b)$

We say that a random variable X follows a *Beta distribution* with parameters $a, b > 0$, if its density function is

$$f(x) = \begin{cases} \dfrac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, & \text{if } x \in [0,1] \\ \quad 0, & \text{if } x \notin [0,1], \end{cases} \tag{3.35}$$

where

$$\beta(a,b) = \int\limits_0^1 x^{a-1}(1-x)^{b-1} dx$$

is Euler's Beta function (for more properties of β, see Appendix A.2).

Beta Distribution of the Second Kind $B_2(a, b)$

We say that a random variable X follows a *Beta distribution of the second kind* with parameters $a, b > 0$, if its density function is

$$f(x) = \begin{cases} \dfrac{1}{\beta(a,b)} x^{a-1}(1+x)^{-(a+b)}, & \text{if } x > 0 \\ \quad 0, & \text{if } x \leq 0. \end{cases} \tag{3.36}$$

With the change of variables $1+t=\dfrac{1}{1-x}$, we have that

$$\beta(a,b)=\int\limits_0^\infty t^{a-1}(1+t)^{-(a+b)}dt.$$

Cauchy Distribution $Cauchy(a,b)$

We say that a random variable X follows a *Cauchy distribution* with parameters $a\in\mathbb{R}$ and $b>0$, if its density function is

$$f(x)=\frac{1}{\pi b\left[1+\left(\dfrac{x-a}{b}\right)^2\right]},\quad x\in\mathbb{R}. \tag{3.37}$$

Remark 3.7.5. Notice that $Cauchy(0,1)=T(1)$, an important particular case, which comes up in many applications. The density function of a random variable following this distribution is given by

$$f(x)=\frac{1}{\pi(1+x^2)},\quad x\in\mathbb{R}.$$

Fisher (F) Distribution $F(m,n)$

We say that a random variable X follows a *Fisher (or F) distribution* with parameters $m,n\in\mathbb{N}$ (degrees of freedom), if its density function is

$$f(x)=\begin{cases}\dfrac{1}{\beta(\frac{m}{2},\frac{n}{2})}\left(\dfrac{m}{n}\right)^{\frac{m}{2}}x^{\frac{m}{2}-1}\left(1+\dfrac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{if } x>0\\ \qquad 0, & \text{if } x\le 0\,.\end{cases} \tag{3.38}$$

3.8 Random Vectors and Joint Distribution Function

In accordance with Definition 3.6.1, we can now define a general random vector.

Definition 3.8.1. *Let $(S, \mathcal{K}, P)$ be a probability space. An n-dimensional* ***random vector*** *is a function $X = (X_1, \ldots, X_n) : S \to \mathbb{R}^n$ satisfying the condition*

$$X^{-1}(x_1, \ldots, x_n) = (X_1 < x_1, \ldots, X_n < x_n) \in \mathcal{K}, \tag{3.39}$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

Definition 3.8.2. *Let $X = (X_1, \ldots, X_n)$ be a random vector. Then the function $F : \mathbb{R}^n \to \mathbb{R}$, defined by*

$$F(x_1, \ldots, x_n) = P(X_1 < x_1, \ldots, X_n < x_n), \tag{3.40}$$

is called the ***joint distribution function*** *of the random vector $(X_1, \ldots, X_n)$.*

The properties of the cumulative distribution function of a random variable translate very naturally for a random vector, as well. As with the case of a discrete random vector, we will restrict our discussion to the case of a 2-dimensional vector.

Theorem 3.8.3. *Let $X = (X_1, X_2)$ be a random vector with joint distribution function $F : \mathbb{R}^2 \to \mathbb{R}$ and for each $k = \overline{1,2}$, let $F_k = F_{X_k} : \mathbb{R} \to \mathbb{R}$ be the cumulative distribution function of X_k. Then F has the following properties:*

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

$$\begin{aligned} P(a_1 \le X_1 < b_1, a_2 \le X_2 < b_2) = {} & F(b_1, b_2) \ - \ F(b_1, a_2) \\ - {} & F(b_1, a_2) \ + \ F(a_1, a_2). \end{aligned} \tag{3.41}$$

(2) F *is monotonically increasing in each variable.*

(3) F *is left continuous in each variable.*

(4) $\lim\limits_{x_1,x_2\to\infty} F(x_1,x_2) = 1$ *and*

$$\begin{aligned}\lim_{x_2\to\infty} F(x_1,x_2) &= F_1(x_1),\ \forall x_1 \in \mathbb{R},\\ \lim_{x_1\to\infty} F(x_1,x_2) &= F_2(x_2),\ \forall x_2 \in \mathbb{R}.\end{aligned}$$

(5) $\lim\limits_{x_2\to-\infty} F(x_1,x_2) = \lim\limits_{x_1\to-\infty} F(x_1,x_2) = 0,\ \forall x_1,x_2 \in \mathbb{R}.$

Proof.
(1) This proof is similar to the proof for random variables.

$$\begin{aligned} & P(a_1 \le X_1 < b_1, a_2 \le X_2 < b_2)\\ = & P(a_1 \le X_1 < b_1, X_2 < b_2) - P(a_1 \le X_1 < b_1, X_2 < a_2)\\ = & P(X_1 < b_1, X_2 < b_2) - P(X_1 < a_1, X_2 < b_2)\\ - & (P(X_1 < b_1, X_2 < a_2) - P(X_1 < a_1, X_2 < a_2))\\ = & F(b_1,b_2) - F(a_1,b_2) - F(b_1,a_2) + F(a_1,a_2).\end{aligned}$$

The proofs of (2) and (3) are similar to the ones for random variables and we skip them.
(4) Let $x_1 \in \mathbb{R}$. We have

$$\begin{aligned}\lim_{x_2\to\infty} F(x_1,x_2) &= P\left(X_1 < x_1, X_2 < \infty\right)\\ &= P\left(X_1 < x_1\right)\\ &= F_1(x_1)\end{aligned}$$

and by symmetry, $\lim\limits_{x_1\to\infty} F(x_1,x_2) = F_2(x_2),\ \forall x_2 \in \mathbb{R}$. Then, it follows that

$$\lim_{x_1,x_2\to\infty} F(x_1,x_2) = \lim_{x_2\to\infty} F_2(x_2) = 1.$$

(5) For any $x_1 \in \mathbb{R}$,

$$\begin{aligned} \lim_{x_2 \to -\infty} F(x_1, x_2) &= P\left(X_1 < x_1, X_2 < -\infty\right) \\ &= P(\emptyset) \;=\; 0, \end{aligned}$$

and by symmetry, $\lim\limits_{x_1 \to -\infty} F(x_1, x_2) = 0,\ \forall x_2 \in \mathbb{R}$, also. □

3.9 Joint Density Function and Marginal Densities

Again, our discussion will be restricted to the 2-dimensional case. It is easy to generalize it to the n-dimensional case.

Definition 3.9.1. *Let $X = (X_1, X_2)$ be a random vector with joint distribution function $F : \mathbb{R}^2 \to \mathbb{R}$. We say that X is a 2-dimensional **continuous random vector** if F is absolutely continuous, i.e. there exists a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$F(x_1, x_2) = \int\limits_{-\infty}^{x_1} \int\limits_{-\infty}^{x_2} f(t_1, t_2)\, dt_1 dt_2. \tag{3.42}$$

*The function f is called the **joint density function** of the random vector X.*

Theorem 3.9.2. *Let $X = (X_1, X_2)$ be a continuous random vector with joint distribution function F and joint density function f. For $k = \overline{1, 2}$, let $f_k = f_{X_k} : \mathbb{R} \to \mathbb{R}$ be the density function of X_k. Then the following properties hold:*

(1) $\dfrac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} = f(x_1, x_2)$, *for all* $(x_1, x_2) \in \mathbb{R}^2$.

(2) $f(x_1, x_2) \geq 0$, *for all* $(x_1, x_2) \in \mathbb{R}^2$.

(3) $\displaystyle\iint\limits_{\mathbb{R}^2} f(t_1, t_2)\, dt_1 dt_2 = 1$.

(4) *For any domain* $D \subseteq \mathbb{R}^2$,

$$P\left((X_1, X_2) \in D\right) = \iint\limits_{D} f(t_1, t_2)\, dt_1 dt_2 \,.$$

(5) $f_1(x) = \displaystyle\int\limits_{\mathbb{R}} f(x, y)\, dy,\ \forall x \in \mathbb{R}$ *and* $f_2(y) = \displaystyle\int\limits_{\mathbb{R}} f(x, y)\, dx,\ \forall y \in \mathbb{R}.$

Proof.
These properties follow easily from Definition 3.9.1 and the properties of the joint distribution function stated in Theorem 3.8.3. □

3.10 Independent Random Variables

Definition 3.10.1. *Two random variables X and Y are* ***independent*** *if*

$$F_{(X,Y)}(x, y) = F_X(x) F_Y(y), \ \ for\ all\ x, y \in \mathbb{R}. \tag{3.43}$$

The ***random variables*** *$X_1, \ldots, X_n$ are* ***independent*** *if*

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n),$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

Remark 3.10.2.
(1) If X and Y are discrete random variables with probability distribution functions $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$ and $Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}$, respectively, then the condition of independence (3.43) is equivalent to the previous condition (3.15)

$$P\left(X = x_i, Y = y_j\right) = P\left(X = x_i\right) P\left(Y = y_j\right) = p_i q_j, \forall (i,j) \in I \times J.$$

(2) If X and Y are continuous random variables with probability density functions f_X, f_Y, respectively, then the condition of independence (3.43) is equivalent to

$$f_{(X,Y)}(x,y) = f_X(x) f_Y(y)\,. \tag{3.44}$$

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

3.11 Functions of Continuous Random Variables

Let (X,Y) be a continuous random vector with joint probability density function $f = f_{(X,Y)}$ and let f_X, f_Y be the probability density functions of X and Y, respectively.

Proposition 3.11.1. *The density function of the* ***sum*** $Z = X + Y$ *of the random variables* X *and* Y *is given by*

$$f_Z(z) = \int\limits_{\mathbb{R}} f(u, z-u)\, du, \; z \in \mathbb{R}, \tag{3.45}$$

which, if X *and* Y *are independent, becomes*

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(u) f_Y(z-u)\, du, \; z \in \mathbb{R}.$$

Proof. First we determine the cumulative distribution function of the random variable Z. For a fixed $z \in \mathbb{R}$, we have

$$F_Z(z) = P(z < Z) = P(X + Y < z) = \iint\limits_{D_{(x,y)}} f(x,y)dxdy,$$

where $D_{(x,y)} = \{(x,y) \in \mathbb{R}^2 \mid x + y < z\}$.
We make the change of variables

$$\begin{aligned} x &= u \\ y &= v - u. \end{aligned}$$

Then the domain of integration $D_{(x,y)}$ becomes

$$D_{(u,v)} = \{(u,v) \in \mathbb{R}^2 \mid v < z\}$$

and the determinant of the Jacobian of the transformation is

$$\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

So

$$\begin{aligned} F_Z(z) &= \iint\limits_{D_{(u,v)}} f(u, v-u) \cdot 1 \, dudv \\ &= \int\limits_{\mathbb{R}} \left(\int\limits_{-\infty}^{z} f(u, v-u) \, dv \right) du \\ &= \int\limits_{-\infty}^{z} \left(\int\limits_{\mathbb{R}} f(u, v-u) \, du \right) dv. \end{aligned}$$

We differentiate[1] with respect to z to obtain

$$f_Z(z) = F'_Z(z) = \int\limits_{\mathbb{R}} f(u, z-u)du,\ z \in \mathbb{R}.$$

If X and Y are independent, then by (3.44), the formula becomes

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(u) f_Y(z-u)\, du,\ z \in \mathbb{R}.$$

□

Proposition 3.11.2. *The density function of the* ***product*** $Z = X \cdot Y$ *of the random variables* X *and* Y *is given by*

$$f_Z(z) = \int\limits_{\mathbb{R}} f\Big(u, \frac{z}{u}\Big)\frac{1}{|u|}\, du,\ z \in \mathbb{R}, \tag{3.46}$$

which, if X *and* Y *are independent, becomes*

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(u) f_Y\Big(\frac{z}{u}\Big)\frac{1}{|u|}\, du,\ z \in \mathbb{R}.$$

Proof. Again, we start with the cumulative distribution function of Z. For a fixed $z \in \mathbb{R}$, we have

$$F_Z(z) = P(z < Z) = P(X \cdot Y < z) = \iint\limits_{D_{(x,y)}} f(x,y)\, dxdy,$$

[1] Differentiation formula: $\frac{d}{dz}\Bigg(\int\limits_{a(z)}^{b(z)} g(t)dt\Bigg) = g(b(z))b'(z) - g(a(z))a'(z).$

where $D_{(x,y)} = \{(x,y) \in \mathbb{R}^2 \mid xy < z\}$. With the change of variables

$$\begin{aligned} x &= u \\ y &= \frac{v}{u}, \end{aligned}$$

the domain of integration $D_{(x,y)}$ becomes $D_{(u,v)} = \{(u,v) \in \mathbb{R}^2 \mid v < z\}$ and the determinant of the Jacobian of the transformation is

$$\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -\frac{v}{u^2} & \frac{1}{u} \end{vmatrix} = \frac{1}{u} \, .$$

Then

$$\begin{aligned} F_Z(z) &= \iint\limits_{D_{(u,v)}} f\left(u, \frac{v}{u}\right) \cdot \frac{1}{|u|} \, dudv \\ &= \int\limits_{-\infty}^{z} \left(\int\limits_{\mathbb{R}} f\left(u, \frac{v}{u}\right) \frac{1}{|u|} \, du \right) dv. \end{aligned}$$

By differentiation with respect to z, we get

$$f_Z(z) = F_Z'(z) = \int\limits_{\mathbb{R}} f\left(u, \frac{z}{u}\right) \frac{1}{|u|} \, du, \; z \in \mathbb{R}.$$

Obviously, if X and Y are independent, then

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(u) f_Y\left(\frac{z}{u}\right) \frac{1}{|u|} \, du, \; z \in \mathbb{R}.$$

□

Proposition 3.11.3. *The density function of the* ***quotient*** $Z = \dfrac{X}{Y}$ *of the random variables* X *and* Y *(*$Y \neq 0$*) is given by*

$$f_Z(z) = \int\limits_{\mathbb{R}} f(uz, u)|u|\, du,\ z \in \mathbb{R}, \tag{3.47}$$

which, if X *and* Y *are independent, becomes*

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(uz) f_Y(u)|u|\, du,\ z \in \mathbb{R}.$$

Proof. For a fixed $z \in \mathbb{R}$, the cumulative distribution function of Z is given by

$$F_Z(z) = P(z < Z) = P(X \cdot Y < z) = \iint\limits_{D_{(x,y)}} f(x,y)\, dxdy,$$

where $D_{(x,y)} = \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; \dfrac{x}{y} < z\right\}$. The change of variables

$$\begin{aligned} x &= uv \\ y &= u, \end{aligned}$$

maps the domain of integration $D_{(x,y)}$ into $D_{(u,v)} = \{(u,v) \in \mathbb{R}^2 \mid v < z\}$ and the determinant of the Jacobian of the transformation is

$$\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1 & 0 \end{vmatrix} = -u.$$

Thus

$$\begin{aligned} F_Z(z) &= \iint\limits_{D_{(u,v)}} f(uv,u)\cdot|u|\,dudv \\ &= \int\limits_{-\infty}^{z}\left(\int\limits_{\mathbb{R}} f(uv,u)|u|\,du\right)dv. \end{aligned}$$

By differentiation with respect to z, we get

$$f_Z(z) = F'_Z(z) = \int\limits_{\mathbb{R}} f(uv,u)|u|\,du,\ z\in\mathbb{R},$$

which, if X and Y are independent, becomes

$$f_Z(z) = \int\limits_{\mathbb{R}} f_X(uv)f_Y(u)|u|\,du,\ z\in\mathbb{R}.$$

□

Proposition 3.11.4. *Let $g:\mathbb{R}\to\mathbb{R}$ be a strictly monotone and differentiable function. Let X be a continuous random variable with density function f_X and let $Y=g(X)$. Then for $y\in\mathbb{R}$, the density function of Y is given by*

$$f_Y(y) = \begin{cases} \dfrac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}, & \text{if } y\in g(\mathbb{R}) \\ \\ 0, & \text{if } y\notin g(\mathbb{R})\,. \end{cases} \tag{3.48}$$

Proof. Note first, that since g is strictly monotone and differentiable, its derivative can never be 0, so the term on the right hand side of (3.48) exists.

Case I. Assume g is strictly increasing. We have

$$F_Y(y) = P(Y < y) = P(g(X) < y)$$

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

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

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

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

Since if g is increasing, then $g' > 0$ and if g is decreasing, then $g' < 0$, in both cases, we have (3.48). □

Proposition 3.11.5. *Let $X_1, \dots, X_n$ be independent random variables having a standard normal $N(0,1)$ distribution. Then the variable $Y_n = \sum_{i=1}^{n} X_i^2$ follows a $\chi^2(n,1)$ distribution.*

Proof.
We use induction on n. For $n = 1$, we have $Y_1 = X_1^2$ and

$$F_{Y_1}(y) = P(Y_1 < y) = P(X_1^2 < y).$$

So, obviously, if $y \le 0$, $F_{Y_1}(y) = 0$ and thus $f_{Y_1}(y) = 0$. If $y > 0$, then

$$F_{Y_1}(y) = P(-\sqrt{y} < X_1 < \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f_{X_1}(x)dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{\sqrt{y}} e^{-\frac{x^2}{2}} dx$$

and

$$f_{Y_1}(y) = \frac{2}{\sqrt{2\pi}} \cdot e^{-\frac{y}{2}} \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} \cdot y^{\frac{1}{2}-1} \cdot e^{-\frac{y}{2}}.$$

Hence $Y_1 \in \chi^2(1,1)$.
Also, note that the same is true for every i, $X_i^2 \in \chi^2(1,1)$.
Now assume $Y_n \in \chi^2(n,1)$. Since $X_{n+1}^2 \in \chi^2(1,1)$, $Y_{n+1} = Y_n + X_{n+1}^2$ and Y_n, X_{n+1}^2 are independent, it follows that $Y_{n+1} \in \chi^2(n+1,\sigma)$.
Note: We used the property that if $X \in \chi^2(m,\sigma)$, $Y \in \chi^2(n,\sigma)$ are independent, then $X + Y \in \chi^2(m+n,\sigma)$ (see [11]).

□

We state two more results that are used in Statistics.

Proposition 3.11.6. *Let X and Y be independent random variables, having an $N(0,1)$ and a $\chi^2(n,1)$ distribution, respectively. Then the random variable*

$$Z = \frac{X}{\sqrt{\frac{1}{n}Y}}$$

follows a Student $T(n)$ distribution.

Proposition 3.11.7. *Let X and Y be independent random variables, with $\chi^2(m,1)$, $\chi^2(n,1)$ distributions, respectively. Then the random variable*

$$Z = \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

follows a Fisher $F(m,n)$ distribution.

3.12 Conditional Distribution and Conditional Density

Definition 3.12.1. *The **conditional distribution function** $F_{X|B} : \mathbb{R} \to \mathbb{R}$ of a random variable X, given the event $B \in \mathcal{K}$ with $P(B) > 0$ is defined by*

$$F_{X|B}(x|B) = P(X < x|B). \tag{3.49}$$

*If X and Y are discrete random variables such that $P(Y = y) > 0$, for some $y \in \mathbb{R}$, then the **conditional distribution function** of X, given the event $\{Y = y\}$ is*

$$F_{X|Y}(x|y) = P(X < x|Y = y) = \frac{P(X < x, Y = y)}{P(Y = y)}. \tag{3.50}$$

*If X and Y are continuous random variables with $f_Y(y) > 0$, for some $y \in \mathbb{R}$, then the **conditional distribution function** of X, given the event $\{Y = y\}$ is*

$$F_{X|Y}(x|y) = \frac{1}{f_Y(y)} \int\limits_{-\infty}^{x} f_{(X,Y)}(u,y)du. \tag{3.51}$$

Proposition 3.12.2. *If the set of events $\{B_i\}_{i\in I}$ forms a partition of S, then the distribution function of a random variable X can be written as the weighted sum of conditional distribution functions*

$$F_X(x) = \sum_{i\in I} P(B_i) F_{X|B}(x|B_i).$$

Definition 3.12.3. *Let X be a random variable and let $F_{X|B}$ be its conditional distribution function, given the event $B \in \mathcal{K}$ with $P(B) > 0$. If there exists a function $f_{X|B} : \mathbb{R} \to \mathbb{R}$ such that*

$$F_{X|B}(x|B) = \int\limits_{-\infty}^{x} f_{X|B}(t|B)dt, \tag{3.52}$$

*then $f_X(\cdot|B)$ is said to be the **conditional density function** of the random variable X, given the event B.*
*If X and Y are continuous random variables, then the **conditional density function** of X, given $\{Y = y\}$ ($f_Y(y) > 0$), is defined by*

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}. \tag{3.53}$$

Proposition 3.12.4. *The conditional distribution and the conditional density function of X, given $\{Y = y\}$ satisfy the following relations:*

(1) $F_{X|Y}(x|y) = \int\limits_{-\infty}^{x} f_{X|Y}(u|y)\, du$, *for all* $x \in \mathbb{R}$.

(2) $F_X(x) = \int\limits_{\mathbb{R}} f_Y(y) F_{X|Y}(x|y)\, dy$, *for all* $x \in \mathbb{R}$.

(3) *Bayes' formula for conditional density functions*

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{\int\limits_{\mathbb{R}} f_{Y|X}(y|u) f_X(u)\, du}, \text{ for all } x, y \in \mathbb{R}.$$

Chapter 4

Numerical Characteristics of Random Variables

In order to better understand and describe random variables, we associate some numerical values that characterize their distributions.

4.1 Expectation

Definition 4.1.1.
(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*

$$E(X) = \sum_{i \in I} x_i P(X = x_i) = \sum_{i \in I} x_i p_i, \tag{4.1}$$

if the series is absolutely convergent.

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

$$E(X) = \int\limits_{\mathbb{R}} x f_X(x) dx, \tag{4.2}$$

if the integral is absolutely convergent.

Remark 4.1.2.
1.The expectation of X can also be written in short as

$$E(X) = \int\limits_{\mathbb{R}} x \, dF_X(x),$$

where F_X is the cumulative distribution function of X.
2. Let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function and X be a random variable, either discrete with probability distribution function $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ or continuous with density function f_X . Then the expectation of the random variable $h(X)$ is given by

$$E\Big(h(X)\Big) = \sum_{i \in I} h(x_i) P(X = x_i) = \sum_{i \in I} h(x_i) p_i,$$

if X is discrete, or by

$$E\Big(h(X)\Big) = \int\limits_{\mathbb{R}} x f_{h(X)}(x) \, dx = \int\limits_{\mathbb{R}} h(x) f_X(x) \, dx,$$

if X is continuous.

Example 4.1.3.
1. Let X be a random variable with a Bernoulli $Bern(p)$ distribution

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

Then

$$E(X) = 0 \cdot (1-p) + 1 \cdot p = p.$$

2. Let X be a uniformly distributed $\mathcal{U}(a,b)$ random variable. Then by (3.24), its expectation is

$$E(X) = \int\limits_{\mathbb{R}} x f_X(x)\, dx = \frac{1}{b-a} \int\limits_a^b x\, dx = \frac{a+b}{2}.$$

Theorem 4.1.4. *If X and Y are either both discrete or both continuous random variables, then the following properties hold:*

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

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

(3) *If X and Y are independent, then*

$$E(X \cdot Y) = E(X)E(Y).$$

(4) *If $X(e) \leq Y(e)$ for all $e \in S$, then $E(X) \leq E(Y)$.*

Proof.

(1) If X is discrete, with probability distribution function $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, then $Y = aX + b$ has distribution $Y\begin{pmatrix} ax_i + b \\ p_i \end{pmatrix}_{i\in I}$, with expectation

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

Now, let us consider the case where X is continuous. If $a = 0$, then $aX+b$ is a discrete distribution $\begin{pmatrix} b \\ 1 \end{pmatrix}$ and, hence, its expectation is

$$E(aX+b) = b \cdot 1 = b = 0 \cdot E(X) + b.$$

Now we can assume $a \neq 0$ and use Proposition 3.11.4. We have

$$g(x) = ax + b,\ g'(x) = a \neq 0,\ g^{-1}(x) = \frac{x-b}{a},$$

so

$$f_{g(X)}(x) = \frac{f_X\left(\frac{x-b}{a}\right)}{|a|}, \forall x \in \mathbb{R}$$

and thus

$$E(g(X)) = \frac{1}{|a|} \int\limits_{\mathbb{R}} x\ f_X\left(\frac{x-b}{a}\right) dx.$$

We make the change the variables $t = \dfrac{x-b}{a},\ x = at + b,\ dx = a\ dt$ and get

$$\begin{aligned} E(g(X)) &= \frac{a}{|a|} \cdot \text{ sign } (a) \int\limits_{\mathbb{R}} (at+b)\ f_X(t)dt \\ &= a \int\limits_{\mathbb{R}} t\ f_X(t)dt + b \int\limits_{\mathbb{R}} f_X(t)dt = aE(X) + b. \end{aligned}$$

(2) If both X and Y are discrete, then

$$\begin{aligned} E(X+Y) &= \sum_{i\in I} \sum_{j \in J} (x_i + y_j) P(X = x_i, Y = y_j) \\ &= \sum_{i \in I} x_i \sum_{j \in J} P(X = x_i, Y = y_j) \\ &+ \sum_{j \in J} y_j \sum_{i \in I} P(X = x_i, Y = y_j) \\ &= \sum_{i \in I} x_i P(X = x_i) + \sum_{j \in J} y_j P(Y = y_j) \\ &= E(X) + E(Y). \end{aligned}$$

If X and Y are both continuous with densities f_X, f_Y, respectively, then by Proposition 3.11.1, $f_{X+Y}(z) = \int\limits_{\mathbb{R}} f_{(X,Y)}(u, z-u)du$. We have

$$
\begin{aligned}
E(X+Y) &= \int\limits_{\mathbb{R}} z \left[\int\limits_{\mathbb{R}} f_{(X,Y)}(u, z-u) du \right] dz \\
&= \int\limits_{\mathbb{R}} \left[\int\limits_{\mathbb{R}} z f_{(X,Y)}(u, z-u) dz \right] du
\end{aligned}
$$

We change the variables $(z, u) \to (v, t)$, where $u = v$, $z = v + t$. Then $z - u = t$ and the determinant of the Jacobian is

$$J = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1,$$

We have

$$
\begin{aligned}
E(X+Y) &= \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} (v+t) \ f_{(X,Y)}(v,t) dv dt \\
&= \int\limits_{\mathbb{R}} v \left[\int\limits_{\mathbb{R}} f_{(X,Y)}(v,t) dt \right] dv \\
&+ \int\limits_{\mathbb{R}} t \left[\int\limits_{\mathbb{R}} f_{(X,Y)}(v,t) dv \right] dt \\
&= \int\limits_{\mathbb{R}} v \ f_X(v) dv + \int\limits_{\mathbb{R}} t \ f_Y(t) dt \\
&= E(X) + E(Y).
\end{aligned}
$$

(3) For X and Y discrete and independent, we have

$$\begin{aligned} E(XY) &= \sum_{i\in I}\sum_{j\in J} x_i y_j P(X=x_i, Y=y_j) \\ &= \sum_{i\in I}\sum_{j\in J} x_i y_j P(X=x_i)P(Y=y_j) \\ &= \sum_{i\in I} x_i \left(\sum_{j\in J} y_j P(Y=y_j)\right) P(X=x_i) \\ &= E(Y)\cdot\sum_{i\in I} x_i P(X=x_i) \\ &= E(X)\cdot E(Y). \end{aligned}$$

If X and Y are continuous and independent, then by Proposition 3.11.2,

$$f_{XY}(z) = \int\limits_{\mathbb{R}} f_X(u) f_Y(\frac{z}{u}) \cdot \frac{du}{|u|}$$

and thus $E(X\cdot Y) = \int\limits_{\mathbb{R}} z \int\limits_{\mathbb{R}} f_X(u) f_Y(\frac{z}{u}) \cdot \frac{1}{|u|} du dz.$

We change the variables $u = v, \;\; z = vt, \;\; J = \begin{vmatrix} 1 & 0 \\ t & v \end{vmatrix} = v, \;\; \frac{z}{u} = t.$ Then

$$\begin{aligned} E(X\cdot Y) &= \iint\limits_{\mathbb{R}^2} vt\ f_X(v)\ f_Y(t)\cdot\frac{1}{|v|}\cdot|v| dv dt \\ &= \int\limits_{\mathbb{R}} v \left(\int\limits_{\mathbb{R}} t\ f_Y(t) dt\right)\cdot f_X(v) dv = E(Y)\cdot\int\limits_{\mathbb{R}} v\ f_X(v) dv \\ &= E(X)\cdot E(Y) \end{aligned}$$

(4) We will show that if $Z \geq 0$, then $E(Z) \geq 0$. Then by (1) and (2) applied to $Z = Y - X,$ the property follows.

If Z is discrete, $Z \geq 0$ means $z_i \geq 0, \ \ \forall i \in I$ and then

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

For Z continuous, $E(Z) = \int\limits_{\mathbb{R}} z\ f_Z(z)dz = \int\limits_{-\infty}^{0} z\ f_Z(z)dz + \int\limits_{0}^{\infty} z\ f_Z(z)dz.$
Now, for values $z \leq 0$, $F_Z(z) = P(Z < z) = 0$, so $f_Z(z) = F_Z'(z) = 0$. Hence

$$E(Z) = \int\limits_{0}^{\infty} z\ f_Z(z)dz \geq 0.$$

□

Example 4.1.5. Let $X_1, \ldots, X_n$ be independent random variables, identically distributed with a $Bern(p)$ distribution and let $Y = \sum\limits_{i=1}^{n} X_i$. Find $E(Y)$.

Solution 4.1.5: We know from Example 4.1.3 that $E\left(X_i\right) = p$, for all $i = \overline{1, n}$. Then by Theorem 4.1.4(1),

$$E(Y) = \sum_{i=i}^{n} E\left(X_i\right) = np.$$

■

Remark 4.1.6. Note that under the conditions described in Example 4.1.5, the variable Y follows a binomial $B(n, p)$ distribution. Thus, we have found the expectation of a binomial $B(n, p)$ variable to be np.

4.2 Variance, Standard Deviation, Moments

Definition 4.2.1. *Let X be a random variable. The **variance** (**dispersion**) of X is the number*

$$V(X) = E\Big(X - E(X)\Big)^2, \tag{4.3}$$

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

Theorem 4.2.2. *Let X and Y be random variables. Then the following properties hold:*

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

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

(3) *If X and Y are independent, then*

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

(4) *If X and Y are independent, then*

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

Proof.
(1) By definition of the variance (4.3) and using the properties of expectation, we have

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

(2)

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

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

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

(4) Since X and Y are independent, so are X^2, Y^2 and by Theorem 4.1.4, $E(X \cdot Y) = E(X) \cdot E(Y)$, $E(X^2Y^2) = E(X^2)E(Y^2)$. We have

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

while the term

$$V(X)V(Y) + E(X)^2V(Y) + E(Y)^2V(X)$$

is equal to

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

□

Remark 4.2.3.
1. Part (1) of Theorem 4.2.2 provides a more practical computational formula for the variance than (4.3). 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

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

2. A direct consequence of Theorem 4.2.2(1) (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 4.2.2(1), $V(X) = 0$, which is to be expected (the variable X does not vary at all).
4. Part (3) of Theorem 4.2.2 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 (2) and (3) of Theorem 4.2.2 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 4.2.4. Find the variance of a random variable X following
a) a Bernoulli $Bern(p)$ distribution;
b) a binomial $B(n,p)$ distribution;
c) a uniform $\mathcal{U}(a,b)$ distribution.

Solution 4.2.4:
a) We have

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

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

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

b) If X is binomial, then it can be written as

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

where $X_1,\dots,X_n$ are independent and identically distributed with a $Bern(p)$ distribution (see Remark 4.1.6). Then by part a), $V(X_i)=pq$, for each $i=\overline{1,n}$ and by the previous remarks,

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

c) We know from Example 4.1.3 that $E(X)=\dfrac{a+b}{2}$. Further, we have

$$E(X^2)=\int\limits_{\mathbb{R}} x^2 f_X(x)\,dx=\frac{1}{b-a}\int\limits_a^b x^2\,dx=\frac{a^2+ab+b^2}{3}$$

and

$$V(X)=\frac{a^2+ab+b^2}{3}-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}.$$

■

Definition 4.2.5. *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*

$$\nu_k = E(X^k). \tag{4.4}$$

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

$$\overline{\nu}_k = E(|X|^k). \tag{4.5}$$

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

$$\mu_k = E\Big(X - E(X)\Big)^k. \tag{4.6}$$

Remark 4.2.6.
1. If X is a discrete random variable with probability distribution function $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, then for every $k \in \mathbb{N}$,

$$\begin{aligned} \nu_k &= \sum_{i\in I} x_i^k p_i, \\ \overline{\nu}_k &= \sum_{i\in I} |x_i|^k p_i, \\ \mu_k &= \sum_{i\in I} (x_i - \nu_1)^k p_i. \end{aligned}$$

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

$$\begin{aligned}
\nu_k &= \int\limits_{\mathbb{R}} x^k f(x)\,dx, \\
\overline{\nu}_k &= \int\limits_{\mathbb{R}} |x|^k f(x)\,dx, \\
\mu_k &= \int\limits_{\mathbb{R}} (x-\nu_1)^k f(x)\,dx.
\end{aligned}$$

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 $E(|X|^n)$ exists for some $n \in \mathbb{N}$, then $E(X^k)$ and $E(|X|^k)$ also exist, for all $k = \overline{1,n}$.

Example 4.2.7. Find the k^{th} order central moments for a random variable having a normal distribution $N(m,\sigma)$.

Solution 4.2.7:

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{\mathbb{R}} x e^{-\dfrac{(x-m)^2}{2\sigma^2}}\,dx.$$

Make the change of variables $x = \sigma t + m$. Then

$$E(X) = \frac{\sigma}{\sqrt{2\pi}} \int\limits_{\mathbb{R}} te^{-\frac{t^2}{2}} dt + m\sqrt{\frac{2}{\pi}} \int\limits_0^\infty e^{-\frac{t^2}{2}} dt.$$

The first integral is 0 (since the integrand is an odd function) and by letting $u = \frac{t^2}{2}$, the second integral is $\frac{1}{\sqrt{2}}\,\Gamma\left(\frac{1}{2}\right) = \sqrt{\frac{\pi}{2}}$ (see Appendix A.1). Hence

$$E(X) = m.$$

Now, the k^{th} order central moments are

$$\mu_k = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{\mathbb{R}} (x-m)^k e^{-\frac{(x-m)^2}{2\sigma^2}} dx.$$

Change the variables $t = x - m$. Then

$$\mu_k = \frac{1}{\sigma\sqrt{2\pi}} \int\limits_{\mathbb{R}} t^k e^{-\frac{t^2}{2\sigma^2}} dt.$$

The integrand is either an odd function (if k is odd) or an even function (if k is even). So

$$\begin{aligned} \mu_{2l+1} &= 0, \\ \mu_{2l} &= \frac{2}{\sigma\sqrt{2\pi}} \int\limits_0^\infty t^{2l} e^{-\frac{t^2}{2\sigma^2}} dt. \end{aligned}$$

Change the variables $u = \dfrac{t^2}{2\sigma^2}$, to get

$$\begin{aligned}
\mu_{2l} &= \frac{2^l\sigma^{2l}}{\sqrt{\pi}}\int\limits_0^{\infty} u^{l-\frac{1}{2}}e^{-u}du = \frac{2^l\sigma^{2l}}{\sqrt{\pi}}\Gamma\left(l+\frac{1}{2}\right) \\
&= \frac{2^l\sigma^{2l}}{\sqrt{\pi}}\left(l-\frac{1}{2}\right)\left(l-\frac{3}{2}\right)\dots\frac{1}{2}\Gamma\left(\frac{1}{2}\right) \\
&= \frac{\sigma^{2l}}{\sqrt{\pi}}(2l-1)(2l-3)\dots1\cdot\sqrt{\pi} = \sigma^{2l}(2l-1)!!
\end{aligned}$$

In particular, $V(X) = \mu_2 = \sigma^2$. ■

4.3 Covariance, Correlation Coefficient, Quantiles

Definition 4.3.1. *Let X and Y be random variables. The* ***covariance*** *of X and Y is the number*

$$\operatorname{cov}(X,Y) = E\Big((X-E(X))\cdot(Y-E(Y))\Big), \tag{4.7}$$

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

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{V(X)V(Y)}}, \tag{4.8}$$

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

Theorem 4.3.2. *Let X, Y and Z be random variables. Then the following properties hold:*

(1) $\operatorname{cov}(X,X) = V(X)$.

(2) $\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y)$.

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

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

(5) $\operatorname{cov}(X+Y,Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$.

Proof.
(1) This follows directly from Definition 4.3.1.
(2) A straightforward computation leads to

$$\begin{aligned}\operatorname{cov}(X,Y) = {} & E\left[XY - E(X)Y - E(Y)X + E(X)E(Y)\right] \\ = {} & E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y).\end{aligned}$$

(3) This follows from (2), keeping in mind that X and Y are independent, so $E(XY) = E(X)E(Y)$.
(4)

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

(5)

$$\begin{aligned}\operatorname{cov}(X+Y,Z) &= E\left[(X+Y-E(X)-E(Y))(Z-E(Z))\right] \\ &= E\left[(X-E(X))(Z-E(Z))\right. \\ &+ \left.(Y-E(Y))(Z-E(Z))\right] \\ &= \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z).\end{aligned}$$

□

Remark 4.3.3.
1. Property (4) of Theorem 4.3.2 can be generalized to any number of variables:

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

2. A consequence of (1) and (5) of Theorem 4.3.2 is the following property:

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

3. The converse of Theorem 4.3.2(3) is not true! Independence is a stronger condition. The next example will illustrate that.

Example 4.3.4. Consider the random variables

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

We have

$$E(X) = 0,\ E(X^2) = \frac{1}{2},\ V(X) = \frac{1}{2},$$
$$E(Y) = \frac{1}{2},\ ,E(Y^2) = \frac{1}{2},\ V(Y) = \frac{1}{4},$$
$$E(XY) = 0,\ \operatorname{cov}(X,Y) = 0.$$

So X and Y are uncorrelated, but $Y = X^2$, thus clearly they are not independent.

Theorem 4.3.5. *Let X and Y be random variables. Then the following properties hold:*

(1) $|\rho(X,Y)| \le 1.$

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

Proof.
(1) Let $Z = t(X - E(X)) + (Y - E(Y))$, for all $t \in \mathbb{R}$. Then clearly

$$E(Z^2) \geq 0, \ \forall t \in \mathbb{R}.$$

We have

$$\begin{array}{rcl} E(Z^2) & = & E\left[t^2(X - E(X))^2 + 2t(X - E(X))(Y - E(Y))\right. \\ & + & \left.(Y - E(Y))^2\right] \\ & = & t^2 V(X) + 2t \text{ cov}(X,Y) + V(Y). \end{array}$$

Since the above quadratic equation takes nonnegative values for all $t \in \mathbb{R}$ and the leading coefficient $V(X) \geq 0$, it follows that the discriminant $\Delta \leq 0$. But

$$\Delta = (\text{ cov}(X,Y))^2 - V(X)V(Y) \leq 0,$$

i.e.

$$\rho^2(X,Y) \leq 1.$$

(2) Suppose $Y = aX + b$, $a, b \in \mathbb{R}$, $a \neq 0$. Then by Remark 4.3.3(2) and Theorem 4.2.2(2),

$$\text{cov}(X,Y) = aV(X) \text{ and } V(Y) = a^2 V(X).$$

Then

$$\rho(X,Y) = \frac{a\, V(X)}{\sqrt{V(X)} \cdot |a| \cdot \sqrt{V(X)}} = \text{sign } (a) = \begin{cases} \ \ 1, & a > 0 \\ -1, & a < 0, \end{cases}$$

so $|\rho(X,Y)| = 1$.

Now assume $\rho\,(X,Y) \;=\; \pm\,1$.
Let $U \;=\; \dfrac{X-E(X)}{\sqrt{V(X)}}$, $W \;=\; \dfrac{Y-E(Y)}{\sqrt{V(Y)}}$, the reduced variables.
We have

$$\begin{aligned} E(U) = &\;\; E(W) = 0, \\ V(U) = &\;\; E(U^2) = V(W) = E(W^2) = 1, \end{aligned}$$

and

$$E(UW) = \frac{\text{cov}(X,Y)}{\sqrt{V(X)\,V(Y)}} = \rho\,(X,Y) = \pm 1.$$

Then

$$E\left[(U \mp W)^2\right] = E(U^2) \mp 2E(UW) + E(V^2) = 0.$$

Since $(U \;\mp\; W)^2 \;\geq\; 0$, $U \;\mp\; W \;=\; 0$, so $U \;=\; \pm\,W$. Hence

$$\frac{Y-E(Y)}{\sqrt{V(Y)}} = \pm\,\frac{X-E(X)}{\sqrt{V(X)}},$$

i.e.

$$Y = \pm\sqrt{\frac{V(Y)}{V(X)}}\,(X-E(X)) + E(Y),$$

so

$$Y = aX + b,$$

with

$$a = \pm\sqrt{\frac{V(Y)}{V(X)}},\quad b = E(Y) - aE(X).$$

□

Remark 4.3.6. As seen from Theorem 4.3.5, the correlation coefficient $\rho(X,Y)$ measures the linear trend between the variables X and Y; the closer its value is to ± 1, the “more linear” the relationship between X and Y is. This notion will be revisited in Chapter 7.

Definition 4.3.7. *Let X be a random variable with cumulative distribution function F_X and let $\alpha \in (0,1)$. A* ***quantile (percentile) of order $\boldsymbol{\alpha}$*** *is a number q_α satisfying the condition*

$$P(X < q_\alpha) \le \alpha \le P(X \le q_\alpha), \tag{4.9}$$

or, equivalently,

$$F_X(q_\alpha) \le \alpha \le F_X(q_\alpha + 0). \tag{4.10}$$

The ***median*** *is the number $m = q_{\frac{1}{2}}$. The* ***quartiles*** *are the numbers*

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

Remark 4.3.8.
1. The 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 Chapter 7.
2. If X is continuous, then for each $\alpha \in (0,1)$, there is a unique quantile q_α , given by $F_X(q_\alpha) = \alpha$, or equivalently, $q_\alpha = F_X^{-1}(\alpha)$.
3. If X is discrete, then the median (or any other quantile) can take an infinite number of values, if the line $y = \dfrac{1}{2}$ (or $y = \alpha$) and the curve $y = F_X(x)$ have in common a segment line.

4.4 Conditional Expectation

The notion of *conditional expectation* is used in Statistics, in finding curves of regression. We define it and briefly discuss some of its properties, to be used in Chapter 7.

Definition 4.4.1. *Let X be a random variable, $B \in \mathcal{K}$ an event with $P(B) > 0$ and $F_{X|B}$ the conditional distribution function of X, given B.*

*The **conditional expectation** of X, given the event B is defined by*

$$E(X|B) = \int_{\mathbb{R}} x \, dF_{X|B}(x|B). \tag{4.11}$$

*The **conditional variance** of X, given the event B is defined by*

$$V(X|B) = E\Big((X - E(X|B))^2|B\Big). \tag{4.12}$$

Remark 4.4.2.
1. If X is a discrete random variable with probability distribution function $\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, then

$$E(X|B) = \sum_{i\in I} x_i P(X = x_i|B),$$

if the series is absolutely convergent.
2. If X is a continuous random variable, then

$$E(X|B) = \int_{\mathbb{R}} x f_{X|B}(x|B) \, dx,$$

where $f_{X|B}$ is the conditional density function of X given B, if the integral is absolutely convergent.
3. The properties of the expectation and variance stated in Theorems 4.1.4 and 4.2.2, respectively, also hold for conditional expectation and variance.

Theorem 4.4.3. *Let X be a random variable and let $\{B_i\}_{i\in I}$ be a partition of S. The expectation of X can be written as the weighted sum of conditional expectations*

$$E(X) = \sum_{i\in I} P(B_i)E(X|B_i). \tag{4.13}$$

Proof. Since the set $\{B_i\}_{i\in I}$ forms a partition of S, by Proposition 3.12.2,

$$F_X(x) = \sum_{i\in I} P(B_i) F_{X|B}(x|B_i).$$

Then, by Remark 4.1.2,

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} x\, dF_X(x) = \sum_{i\in I} P(B_i) \int_{\mathbb{R}} x\, dF_{X|B}(x|B_i) \\ &= \sum_{i\in I} P(B_i) E(X|B_i). \end{aligned}$$

□

Proposition 4.4.4. *Let X and Y be random variables. Then the following hold:*

(1) $E\left(E(X|Y)\right) = E(X)$,

(2) $V(X) = E\left(V(X|Y)\right) + V\left(E(X|Y)\right)$.

4.5 Inequalities

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

We start by reminding the reader the following result from number theory, known as *Hölder's inequality* for numbers:

Lemma 4.5.1. *Let $x, y \in \mathbb{R}$ and $p, q > 1$ with $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Then*

$$|xy| \le \frac{x^p}{p} + \frac{y^q}{q}.$$

Proposition 4.5.2 (Hölder's Inequality). *Let X and Y be random variables and $p, q > 1$ with $\dfrac{1}{p} + \dfrac{1}{q} = 1$. Then*

$$E(|XY|) \leq (E(|X|^p))^{\frac{1}{p}} \cdot (E(|Y|^q))^{\frac{1}{q}} . \tag{4.14}$$

Proof. We can assume $E(|X|^p),\ \ E(|Y|^2) > 0$. Apply Lemma 4.5.1 to

$$\begin{aligned} x &= \frac{X}{(E(|X|^p))^{\frac{1}{p}}}, \\ y &= \frac{Y}{(E(|Y|^q))^{\frac{1}{q}}}, \end{aligned}$$

to get

$$\frac{|X\,Y|}{(E(|X|^p))^{\frac{1}{p}}\,(E(|Y|^q))^{\frac{1}{q}}} \leq \frac{|X|^p}{pE(|X|^p)} + \frac{|Y|^q}{qE(|Y|^q)}.$$

Take the expected value on both sides. We have

$$\frac{E\,(|X\,Y|)}{(E(|X|^p))^{\frac{1}{p}}\,(E(|Y|^q))^{\frac{1}{q}}} \leq \frac{E\,(|X|^p)}{pE(|X|^p)} + \frac{E\,(|Y|^q)}{qE(|Y|^q)} = \frac{1}{p} + \frac{1}{q} = 1,$$

i.e.

$$E\,(|XY|) \leq (E(|X|^p))^{\frac{1}{p}}\,(E(|Y|^q))^{\frac{1}{q}}$$

□

Remark 4.5.3.

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

$$E(|XY|) \leq \sqrt{E(X^2)} \cdot \sqrt{E(Y^2)}, \tag{4.15}$$

known as **Schwarz's inequality**.

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

$$E(|X|) \leq \sqrt{E(X^2)}, \tag{4.16}$$

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

Proposition 4.5.4 (Minkowsky's Inequality). *Let X and Y be random variables and let $p > 1$. Then*

$$(E(|X+Y|^p))^{\frac{1}{p}} \leq (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}. \tag{4.17}$$

Proof. The case $p = 1$ is obvious. Assume $p > 1$.
We have

$$\begin{aligned} E\left(|X+Y|^p\right) &= E\Big(|X+Y|\,|X+Y|^{p-1}\Big) \\ &\leq E\Big((|X|+|Y|)\,|X+Y|^{p-1}\Big) \\ &= E\Big(|X|\,|X+Y|^{p-1}\Big) + E\Big(|Y|\,|X+Y|^{p-1}\Big). \end{aligned}$$

To each of these we apply Hölder's inequality. Then

$$\begin{aligned} E\left(|X+Y|^p\right) \leq &\ (E(|X|^p))^{\frac{1}{p}} \left(E\left(|X+Y|^{(p-1)q}\right)\right)^{\frac{1}{q}} \\ + &\ (E(|Y|^p))^{\frac{1}{p}} \left(E\left(|X+Y|^{(p-1)q}\right)\right)^{\frac{1}{q}}. \end{aligned}$$

Now $(p - 1)q = pq - q = p + q - q = p$. So

$$E\left(|X+Y|^p\right) \leq \left(E\left(|X+Y|^p\right)\right)^{\frac{1}{q}} \Big[(E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}\Big]$$

and hence

$$\left(E\left(|X+Y|^p\right)\right)^{\frac{1}{p}} \leq (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}$$

□

Proposition 4.5.5 (Markov's Inequality). *Let X be a random variable and let $a > 0$. Then*

$$P\Big(|X| \geq a\Big) \leq \frac{1}{a}E(|X|). \tag{4.18}$$

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

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

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

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 $aI_A(e) \leq |X(e)|, \forall e \in S$.
Take the expected value on both sides to get $E(aI_A) \leq E(|X|)$. Now, the distribution of aI_A is

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

so

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

Thus

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

i.e.

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

□

Proposition 4.5.6 (Chebyshev's Inequality). *Let X be a random variable and let $\varepsilon > 0$. Then*

$$P\Big(|X - E(X)| \geq \varepsilon\Big) \leq \frac{1}{\varepsilon^2}V(X), \tag{4.19}$$

or, equivalently,

$$P\Big(|X - E(X)| < \varepsilon\Big) \geq 1 - \frac{1}{\varepsilon^2} V(X), \tag{4.20}$$

Proof.

Apply Markov's inequality (4.18) to $\Big(X - E(X)\Big)^2$ and $a = \epsilon^2$, to get

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

i.e.

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

and, equivalently,

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

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

□

Remark 4.5.7. Chebyshev's inequality (4.19) can be used to roughly estimate the distribution range of a random variable X whose expectation $E(X)$ and variance $V(X)$ are known. If we take $\varepsilon = k\sqrt{V(X)} = k\,\sigma(X)$, an integer multiple of the standard deviation, we get

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

For $k = 3$, we have

$$P\Big(|X - E(X)| < 3\sigma(X)\Big) \geq 1 - \frac{1}{9} \approx .89. \tag{4.21}$$

This is known as *the* 3σ *rule* : For any random variable, most of its values (at least 89%) lie within three standard deviations from the mean.

Proposition 4.5.8 (Lyapunov's Inequality). *Let X be a random variable, let $0 < a < b$ and $c \in \mathbb{R}$. Then*

$$(E(|X - c\,|^a))^{\frac{1}{a}} \leq \left(E(|X - c\,|^b)\right)^{\frac{1}{b}}. \tag{4.22}$$

Proof. Let $p = \dfrac{b}{a} > 1$. We apply Hölder's inequality (4.13) to the variables

$$\widetilde{X} = \frac{|X - c|^a}{(E(|X - c\,|^b))^{\frac{1}{p}}}, \quad \widetilde{Y} = 1.$$

Then

$$E\left(\widetilde{X}\widetilde{Y}\right) \;=\; E\left(\widetilde{X}\right) \;=\; \frac{E\left(|X - c|^a\right)}{(E(|X - c\,|^b))^{\frac{a}{b}}}.$$

We have

$$|\widetilde{X}|^p \;=\; \frac{|X - c|^{ap}}{E(|X - c\,|^b)} \;=\; \frac{|X - c|^b}{E(|X - c\,|^b)},$$

so $E\left(|\widetilde{X}|\right)^p = 1$. Obviously, $E\left(|\widetilde{Y}|\right)^q = 1$, also. So Hölder's inequality yields

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

Raising both sides to the power $\dfrac{1}{a}$, we get

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

□

Proposition 4.5.9 (Kolmogorov's Inequality). *Let $X_1, \ldots, X_n$ be independent random variables and for each $k = \overline{1, n}$, let*

$$S_k = \sum_{j=1}^{k} (X_j - E(X_j)) .$$

Then, for every $\varepsilon > 0$,

$$P\left(\max_{1 \le k \le n} |S_k| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2} \sum_{k=1}^{n} V(X_k). \tag{4.23}$$

Remark 4.5.10. Kolmogorov's inequality is, in a way, a generalization of Chebyshev's inequality, for two or more random variables. For the proof, see [5]. We stated it since it is in strong relationship with notions from Chapter 6.

Chapter 5

Sequences of Random Variables

5.1 Types of Convergence of Sequences of Random Variables

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. We present some of these.

In what follows, 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 cumulative distribution function $F = F_X$.

Definition 5.1.1. *We say that*

(1) X_n ***converges in probability*** *to* X*, denoted by* $X_n \xrightarrow{p} X$*, if*

$$\lim_{n\to\infty} P\Big(|X_n - X| < \varepsilon\Big) = 1, \quad for\ every\ \varepsilon > 0\,. \tag{5.1}$$

(2) X_n ***converges strongly*** *to* X*, denoted by* $X_n \xrightarrow{s} X$*, if*

$$\lim_{n\to\infty} P\Big(\bigcap_{k\geq n} \{|X_k - X| < \varepsilon\} \Big) = 1, \quad for\ every\ \varepsilon > 0\,. \tag{5.2}$$

(3) X_n ***converges almost surely*** *to* X*, denoted by* $X_n \xrightarrow{a.s.} X$*, if*

$$P\Big(\lim_{n\to\infty} X_n \neq X \Big) = 0\,. \tag{5.3}$$

(4) X_n ***converges in mean of order r*** *(* $0 < r < \infty$*) to* X*, denoted by* $X_n \xrightarrow{L^r} X$*, if*

$$\lim_{n\to\infty} E\Big(|X_n - X|^r \Big) = 0\,. \tag{5.4}$$

(5) X_n ***converges in distribution*** *to* X*, denoted by* $X_n \xrightarrow{d} X$*, if*

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{5.5}$$

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

Remark 5.1.2.
1. The formulas in Definition 5.1.1(1), (2) and (3) can be written equivalently, using $P(\overline{A}) = 1 - P(A)$.
2. The special case when $r = 2$ in Definition 5.1.1(4) is called **mean square convergence**.

Some of the conditions defining various types of convergence are rather difficult to check, which is why, we first give this equivalence result, without proof (for details, see [5]).

Theorem 5.1.3. *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of random variables, and let* X *be a random variable. Then the following are equivalent:*

(1) $X_n \xrightarrow{s} X$.

(2) $X_n \overset{a.s.}{\to} X$.

(3) $\lim\limits_{n\to\infty} P\Big(\bigcup\limits_{k\geq n}\Big\{|X_k - X| \geq \varepsilon\Big\}\Big) = 0, \quad for\ every\ \varepsilon > 0.$

(4) $\lim\limits_{n\to\infty} P\Big(\sup\limits_{k\geq n}|X_k - X| < \varepsilon\Big) = 1, \quad for\ every\ \varepsilon > 0.$

Therefore, from here on in, we identify strong convergence by almost surely convergence, since the latter is easier to verify.

Example 5.1.4. Let $S = [0,1]$, $\mathcal{K} = \mathcal{B}[0,1]$ (the set of all open subsets of $[0,1]$ and let $P = \mu$ be the Lebesgue measure on $[0,1]$ (length, distance). In the probability space $(S, \mathcal{K}, P)$, consider the sequence of random variables given by

$$X_n(e) = \begin{cases} n^{\frac{2}{r}}, & 0 \leq e \leq \frac{1}{n} \\ 0, & e > \frac{1}{n}, \end{cases}$$

$r > 0$. Study the various types of convergence of X_n to $X = 0$.

Solution 5.1.4: The Lebesgue measure gives us the probability. So the probability distribution function s of X_n and X are

$$X_n\left(\begin{array}{cc} 0 & n^{\frac{2}{r}} \\ 1-\frac{1}{n} & \frac{1}{n} \end{array}\right), \; X\left(\begin{array}{c} 0 \\ 1 \end{array}\right).$$

Let $\varepsilon > 0$ be fixed. We have

$$\begin{array}{rcl} P\Big(|X_n - X| < \varepsilon\Big) & = & P(X_n < \varepsilon) \;\; = \;\; F_n(\varepsilon) \\ & = & \begin{cases} 1 - \dfrac{1}{n}, & 0 < \varepsilon \leq n^{\frac{2}{r}} \\ 1, & \varepsilon > n^{\frac{2}{r}}, \end{cases} \end{array}$$

so $\lim\limits_{n\to\infty} P\Big(|X_n - X| < \varepsilon\Big) = 1,\ \forall\varepsilon > 0$ and $X_n \xrightarrow{p} 0$.
Notice that for $e \in [0,1]$, we have

$$\lim_{n\to\infty} X_n(e) = 0 = X(e),\ \forall e \neq 0 \ \text{ and } \lim_{n\to\infty} X_n(0) = \lim_{n\to\infty} n^2 = \infty.$$

Since the Lebesgue measure (length) of one point in $\mathbb{R}$ is 0, it follows that

$$P\Big(e \in [0,1] \,\Big|\, \lim_{n\to\infty} X_n(e) = X(e)\Big) = \mu((0,1]) = 1$$

and thus, $X_n \xrightarrow{a.s.} 0$ (and, implicitly, $X_n \xrightarrow{s} 0$).
For convergence in mean of order r, we have

$$|X_n - X|^r \begin{pmatrix} 0 & n^2 \\ 1-\frac{1}{n} & \frac{1}{n} \end{pmatrix}$$

and

$$E\left(|X_n - X|^r\right) = 0 \cdot \left(1 - \frac{1}{n}\right) + n^2 \cdot \frac{1}{n} = n \to \infty.$$

Hence $X_n \overset{L^r}{\nrightarrow} 0$.
Finally, for convergence in distribution, we have

$$F_n(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \dfrac{1}{n}, & 0 < x \leq n^{\frac{2}{r}} \\ 1, & x > n^{\frac{2}{r}} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0. \end{cases}$$

For every point of continuity of F, i.e. for every $x \neq 0$,

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

so

$$X_n \xrightarrow{d} 0.$$

■

Having defined and described various types of convergence, the question of uniqueness of the limit naturally arises. We address this issue next.

Definition 5.1.5. *Two random variables X and Y are said to be* ***almost surely equal****, denoted by $X \overset{a.s.}{=} Y$, if*

$$P\,(X \neq Y) = 0, \tag{5.6}$$

or, equivalently,

$$P(X = Y) = 1.$$

Theorem 5.1.6. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, X and Y random variables such that $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$. Then $X \overset{a.s.}{=} Y$.*

Proof. We know that for every $n \in \mathbb{N}$,

$$|X - Y| \leq |X - X_n| + |X_n - Y|.$$

Let $\varepsilon > 0$ be fixed. If both

$$|X_n - X| < \frac{\varepsilon}{2} \text{ and } |X_n - Y| < \frac{\varepsilon}{2},$$

it follows then that $|X - Y| < \varepsilon$. Thus, as events,

$$\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \cap \left\{|X_n - Y| < \frac{\varepsilon}{2}\right\} \subseteq \{|X - Y| < \varepsilon\}$$

and we have the reverse inequality for the contrary events

$$\{|X - Y| \geq \varepsilon\} \subseteq \left\{|X_n - X| \geq \frac{\varepsilon}{2}\right\} \cup \left\{|X_n - Y| \geq \frac{\varepsilon}{2}\right\}.$$

Taking the probability of these events, it follows that

$$0 \;\leq\; P\,(|X - Y| \geq \varepsilon) \;\leq\; P\left(|X_n - X| \geq \frac{\varepsilon}{2}\right) + P\left(|X_n - Y| \geq \frac{\varepsilon}{2}\right).$$

Since both $X_n \overset{p}{\to} X$ and $X_n \overset{p}{\to} Y$, we have

$$\lim_{n\to\infty} P\Big(|X_n - X| \geq \frac{\varepsilon}{2}\Big) = \lim_{n\to\infty} P\Big(|X_n - Y| \geq \frac{\varepsilon}{2}\Big) = 0.$$

Hence

$$0 \leq P\left(|X - Y| \geq \varepsilon\right) \leq 0$$

and that holds for every fixed $\varepsilon > 0$. Thus $P\left(X \neq Y\right) = 0,$ so

$$X \overset{a.s.}{=} Y.$$

□

Remark 5.1.7. The same property holds true for almost surely convergence and convergence in mean of order r, but we skip the proofs.

5.2 Properties

We know that strong and almost surely convergence are equivalent. Next we want to look closer at the relationship between different types of convergence.

Theorem 5.2.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, and let X be a random variable. Then the following hold:*

(1) $X_n \overset{a.s}{\to} X \Longrightarrow X_n \overset{p}{\to} X.$

(2) $X_n \overset{L^r}{\to} X \Longrightarrow X_n \overset{p}{\to} X.$

(3) $X_n \overset{p}{\to} X \Longrightarrow X_n \overset{d}{\to} X.$

Proof.
(1) Assume $X_n \overset{a.s}{\to} X$. Then by Theorem 5.1.3,

$$\lim_{n\to\infty} P\Big(\sup_{k\geq n} |X_k - X| < \epsilon\Big) = 1, \ \forall\, \epsilon > 0.$$

But
$$P\Big(\sup_{k\geq n} |X_k - X| < \epsilon\Big) \leq P\Big(|X_n - X| < \epsilon\Big),$$
so
$$\lim_{n\to\infty} P\Big(|X_n - X| < \epsilon\Big) = 1,\ \forall\, \epsilon > 0,$$
i.e.
$$X_n \overset{p}{\to} X.$$

(2) Let $\epsilon > 0$ be fixed. By Markov's inequality (4.18), we have
$$P\Big(|X_n - X| \geq \epsilon\Big) = P\Big(|X_n - X|^r \geq \epsilon^r\Big) \leq \frac{1}{\epsilon^r}\, E\Big(|X_n - X|^r\Big).$$
Since $X_n \overset{L^r}{\to} X$, i.e. $\lim\limits_{n\to\infty} E\Big(|X_n - X|^r\Big) = 0$, it follows that
$$\lim_{n\to\infty} P\Big(|X_n - X| \geq \epsilon\Big) = 0,\ \ \forall\, \epsilon\ >\ 0,$$
i.e. $X_n \overset{p}{\to} X$.

(3) For the proof, see [5] or [14].

□

In general, the reverse implications are not true, as seen, for instance, in Example 5.1.4. However, there is a special case when the implication in Theorem 5.2.1(3) can be reversed, which we present next.

Proposition 5.2.2. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables such that $X_n \overset{d}{\to} a \in \mathbb{R}$. Show that $X_n \overset{p}{\to} a$.*

Proof. Here the limit is $X\begin{pmatrix} a \\ 1 \end{pmatrix}$. Then its cumulative distribution function is
$$F(x) = \begin{cases} 0, & x \leq a \\ 1, & x > a, \end{cases}$$

which is continuous everywhere except at $x = a$. Thus, since $X_n \overset{d}{\to} X$, $\lim\limits_{n\to\infty} F_n(x) = F(x),\ \forall x \neq a$.
Now let $\varepsilon > 0$ be fixed. We have

$$\begin{aligned} P\left(|X_n - X| < \varepsilon\right) &= P\left(|X_n - a| < \varepsilon\right) \\ &= P\left(a - \varepsilon < X_n < a + \varepsilon\right) \\ &= F_n(a+\varepsilon) - F_n(a-\varepsilon). \end{aligned}$$

Then $\lim\limits_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = F(a+\varepsilon) - F(a-\varepsilon) = 1 - 0 = 1$, (since $a + \varepsilon \neq a \neq a - \varepsilon$), so $X_n \overset{p}{\to} a$. □

We mentioned that the reverse of Theorem 5.2.1(1) is, in general, not true. The most that can be said is the following result, which we present without proof.

Proposition 5.2.3. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, and let X be a random variable such that $X_n \overset{p}{\to} X$. Then there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}} \subseteq \{X_n\}_{n\in\mathbb{N}}$ such that $X_{n_k} \overset{a.s.}{\to} X$.*

We conclude this section by considering functions of sequences of random variables.

Theorem 5.2.4. *Let $G : \mathbb{R}^2 \to \mathbb{R}^2$ be a continuous function. Then the following are true:*

(1) $X_n \overset{a.s}{\to} X,\ Y_n \overset{a.s}{\to} Y \Longrightarrow G(X_n, Y_n) \overset{a.s}{\to} G(X, Y)$;

(2) $X_n \overset{p}{\to} X,\ Y_n \overset{p}{\to} Y \Longrightarrow G(X_n, Y_n) \overset{p}{\to} G(X, Y)$.

The proof is similar to that of Theorem 5.1.6 and we skip it.

Chapter 6

Laws of Large Numbers and Limit Theorems

When working with sequences of random variables, there are two types of problems that we are interested in, namely to establish the stability of the average for a large number of observations $-$ these results are stated in the so-called *laws of large numbers*, and based on that, to determine the form and shape of the limit cumulative distribution function of a sequence of random variables $-$ the *limit theorems*. These results are very important, as they lie at the core of modern Mathematical Statistics, but, in general, their proofs are lengthy and complicated and will therefore, be skipped for the most part. Details can be found in [5], [14].

6.1 Laws of Large Numbers

The basis of these theorems relies upon the observation that there is a strong relationship between the frequency of the occurrence of an event and its probability. The first results belonged to Bernoulli and Poisson, with others following later on.

6.1.1 Weak Law of Large Numbers

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with $E(|X_n|) < \infty$, for all $n \in \mathbb{N}$.

Definition 6.1.1. *The sequence $\{X_n\}$ is said to* ***follow (obey) the weak law of large numbers*** $($***WLLN***$)$ *if*

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \xrightarrow{p} 0. \tag{6.1}$$

Theorem 6.1.2 (Hincin)**.** *If $\{X_n\}_{n\in\mathbb{N}}$ are independent and identically distributed with $E(X_n) = a$, for every $n \in \mathbb{N}$, then $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.*

Theorem 6.1.3 (Markov)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables satisfying Markov's condition*

$$\lim_{n\to\infty}\frac{1}{n^2}V\Big(\sum_{k=1}^{n}X_k\Big) = 0. \tag{6.2}$$

Then $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.

Proof.
Recall Chebyshev's inequality (4.19)

$$P\Big(|X - E(X)| < \varepsilon\Big) \geq 1 - \frac{V(X)}{\varepsilon^2}\ ,\ \ \forall\varepsilon > 0.$$

Apply it to $X = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}X_k$. We have

$$1 \geq P\Big(\frac{1}{n}\Big|\sum_{k=1}^{n}(X_k - E(X_k))\Big| < \varepsilon\Big) \geq 1 - \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\,V\Big(\sum_{k=1}^{n}X_k\Big).$$

Take the limit as $n \to \infty$, to get

$$\lim_{n\to\infty} P\Big(\frac{1}{n}\Big|\sum_{k=1}^{n}(X_k - E(X_k))\Big| < \varepsilon\Big) = 1, \ \ \forall\, \varepsilon > 0,$$

i.e. $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.

□

Theorem 6.1.4 (Chebyshev). *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables with $V\{X_n\} \le L < \infty$ for all $n \in \mathbb{N}$, where $L > 0$ is a constant. Then $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.*

Proof. Since $\{X_n\}_{n\in\mathbb{N}}$ are independent,

$$V\Big(\sum_{k=1}^{n} X_k\Big) = \sum_{k=1}^{n} V\Big(X_k\Big) \le nL$$

and thus

$$\lim_{n\to\infty} \frac{1}{n^2}\, V\Big(\sum_{k=1}^{n} X_k\Big) \le \lim_{n\to\infty} \frac{L}{n} = 0.$$

So $\{X_n\}_{n\in\mathbb{N}}$ satisfies Markov's condition. By Theorem 6.1.3, $\{X_n\}_{n\in\mathbb{N}}$ follows WLLN. □

Theorem 6.1.5 (Poisson). *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables with probability distribution functions*

$$X_n \begin{pmatrix} 0 & 1 \\ q_n & p_n \end{pmatrix}, p_n \in [0,1], q_n = 1 - p_n, \forall n \in \mathbb{N}.$$

Then $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.

Proof. We have

$$\begin{aligned} E(X_n) &= \ p_n, \\ V(X_n) &= E(X_n^2) - (E(X_n))^2 = p_n - p_n^2 \leq \frac{1}{4} \ . \end{aligned}$$

Then we can apply Theorem 6.1.4 with $L = \dfrac{1}{4}$ (the maximum of $x(1-x)$, for $x \in [0,1]$). Thus $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN. □

Remark 6.1.6. Poisson's Theorem can also be stated in the following equivalent form: If in a sequence of independent trials of an experiment, the probability of an event A occurring in the i^{th} trial is p_i, then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{k_n}{n} - \frac{p_1 + \cdots + p_n}{n}\right| < \varepsilon\right) = 1,$$

where k_n is the number of occurrences of the event A in n trials (the frequency of the event A).
This follows easily since

$$E(X_n) = p_n, \ \ \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k_n}{n}, \ \ \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} p_i.$$

Theorem 6.1.7 (Bernoulli)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables, with probability distribution functions*

$$X_n \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, p \in [0,1], q = 1-p, \forall n \in \mathbb{N}.$$

Then $\{X_n\}_{n\in\mathbb{N}}$ follows the WLLN.

Proof.
This is a particular case for Poisson's Theorem 6.1.5, for $p_n = p$, $n \in \mathbb{N}$. □

Remark 6.1.8. Bernoulli's Theorem can also be stated in the following equivalent form: If in a sequence of independent trials of an experiment, the probability of an event A occurring is p for each trial, then for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{k_n}{n} - p\right| < \varepsilon\right) = 1,$$

where k_n is the number of occurrences of the event A in n trials (the frequency of the event A).
So the sequence of relative frequencies of the event A converges to the probability of A.

6.1.2 Strong Law of Large Numbers

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with $E(|X_n|) < \infty$, for all $n \in \mathbb{N}$.

Definition 6.1.9. *The sequence $\{X_n\}$ is said to* ***follow (obey) the strong law of large numbers*** $($***SLLN***$)$ *if*

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \overset{a.s.}{\to} 0. \tag{6.3}$$

Using Kolmogorov's inequality (Proposition 4.5.9), the following theorem holds.

Theorem 6.1.10. *If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of independent random variables with $\sum\limits_{n=1}^{\infty} V(X_n) < \infty$, then the series $\sum\limits_{k=1}^{\infty}(X_k - E(X_k))$ is a.s. convergent, i.e.*

$$P\Big(\sum_{k=1}^{\infty}(X_k - E(X_k)) < \infty\Big) = 1. \tag{6.4}$$

We give next a result from Analysis, which will be used in the proof of the next theorem.

Lemma 6.1.11 (Kronecker)**.** *Let* $\{x_n\}_{n\in\mathbb{N}}$ *be a sequence of real numbers, such that the series* $\sum\limits_{n=1}^{\infty}\dfrac{x_n}{n}<\infty$. *Then*

$$\lim_{n\to\infty}\frac{x_1+\cdots+x_n}{n}=0.$$

Theorem 6.1.12 (Kolmogorov 1^{st})**.** *Let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of independent random variables with* $\sum\limits_{n=1}^{\infty}\dfrac{1}{n^2}V(X_n)<\infty$. *Then* $\{X_n\}_{n\in\mathbb{N}}$ *follows the SLLN.*

Proof.
We have

$$\sum_{n=1}^{\infty}V\Big(\frac{X_n}{n}\Big)=\sum_{n=1}^{\infty}\frac{1}{n^2}V(X_n)<\infty.$$

Then by Theorem 6.1.10 applied to $\dfrac{X_n}{n}$, we have that $\sum\limits_{n=1}^{\infty}\dfrac{1}{n}\Big(X_n-E(X_n)\Big)$ is a.s. convergent. So

$$P\Big(\sum_{n=1}^{\infty}\frac{1}{n}\Big(X_n-E(X_n)\Big)<\infty\Big)=1.$$

But by Lemma 6.1.11 applied to $x_n=X_n-E(X_n)$, since

$$\sum_{n=1}^{\infty}\frac{1}{n}\Big(X_n-E(X_n)\Big)<\infty,$$

it follows that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{\infty}\Big(X_k - E(X_k)\Big) = 0.$$

Then

$$\begin{aligned} 1 &= P\Big(\sum_{n=1}^{\infty}\frac{1}{n}\Big(X_n - E(X_n)\Big) < \infty\Big) \\ &\leq P\Big(\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{\infty}\Big(X_k - E(X_k)\Big) = 0\Big). \end{aligned}$$

Thus $\{X_n\}_{n\in\mathbb{N}}$ follows the SLLN.

□

Theorem 6.1.13 (Kolmogorov 2$^{\text{nd}}$)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $E(X_n) = a < \infty$, for all $n \in \mathbb{N}$. Then $\{X_n\}_{n\in\mathbb{N}}$ follows the SLLN.*

For the proof, see [5], [14].

6.2 Limit Theorems

Limit theorems are applications of laws of large numbers that find the limit in distribution of a sequence of random variables, i.e. they give the limit cumulative distribution function for such a sequence.
If the limit cumulative distribution function is the normal distribution, then we have a *central limit theorem*.

6.2.1 Central Limit Theorems

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables. We make the following notations:

$$\begin{aligned} m_k &= E(X_k), \\ \sigma_k^2 &= V(X_k), \\ \sigma_{(n)}^2 &= \sum_{k=1}^n \sigma_k^2, \ \text{so } \sigma_{(n)} = \sqrt{\sum_{k=1}^n \sigma_k^2}, \end{aligned}$$

for each $k, n \in \mathbb{N}$ and we consider the sequence

$$X_{(n)} = \frac{1}{\sigma_{(n)}} \sum_{k=1}^n (X_k - m_k). \tag{6.5}$$

For $\left\{X_{(n)}\right\}_{n\in\mathbb{N}}$, we have

$$\begin{aligned} E\left(X_{(n)}\right) &= \frac{1}{\sigma_{(n)}} \sum_{k=1}^n (E(X_k) - m_k) &= 0, \\ V\left(X_{(n)}\right) &= \frac{1}{\sigma_{(n)}^2} \sum_{k=1}^n V\left(X_k - m_k\right) &= \frac{1}{\sigma_{(n)}} \sum_{k=1}^n \sigma_k^2 = 1. \end{aligned}$$

As usually, we denote by $F_k, F_{(n)}$ the cumulative distribution functions of $X_k,\ X_{(n)},$ respectively. The problem we want to solve is finding a random variable X such that

$$X_{(n)} \xrightarrow{d} X.$$

Definition 6.2.1. *We say that the sequence $\{X_n\}_{n\in\mathbb{N}}$ satisfies* ***Lindeberg's condition*** *if*

$$\lim_{n\to\infty} \frac{1}{\sigma_{(n)}^2} \sum_{k=1}^n \int\limits_{\{|x-m_k|\geq\varepsilon\sigma_{(n)}\}} (x-m_k)^2 \, dF_k(x) = 0, \tag{6.6}$$

for every $\varepsilon > 0$.

Remark 6.2.2.
1. If for $k \in \mathbb{N}$, $X_k \begin{pmatrix} x_{i_k} \\ p_{i_k} \end{pmatrix}_{i_k \in I_k}$ are discrete random variables, then (6.6) is equivalent to

$$\lim_{n\to\infty} \frac{1}{\sigma^2_{(n)}} \sum_{k=1}^{n} \sum_{\substack{i_k \in I_k \\ \{|x_{i_k} - m_k| \geq \varepsilon\sigma_{(n)}\}}} (x_{i_k} - m_k)^2 p_{i_k} = 0, \quad ,$$

while if X_k are continuous random variables with densities f_k, $k \in \mathbb{N}$, (6.6) becomes

$$\lim_{n\to\infty} \frac{1}{\sigma^2_{(n)}} \sum_{k=1}^{n} \int\limits_{\{|x-m_k| \geq \varepsilon\sigma_{(n)}\}} (x - m_k)^2 f_k(x)\, dx = 0.$$

2. Lindeberg's condition (6.6) implies the following:

$$\lim_{n\to\infty} P\left(\max_{1\leq k\leq n} |X_k - m_k| \geq \varepsilon\sigma_{(n)}\right) = 0, \ \forall\varepsilon > 0,$$

which is saying that as the number of trials n becomes large, the probability of the maximum error between the value of X_k and the observed value m_k approaches 0.

Theorem 6.2.3 (Lindeberg)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables satisfying Lindeberg's condition* (6.6)*. Then*

$$\lim_{n\to\infty} F_{(n)}(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt,$$

for all $x \in \mathbb{R}$, i.e. the sequence

$$X_{(n)} \xrightarrow{d} X \in N(0, 1)$$

converges in distribution to a standard normal random variable.

An important immediate consequence is the following result:

Proposition 6.2.4. *If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of independent, identically distributed random variables, with $E(X_k) = m < \infty$, $V(X_k) = \sigma^2 < \infty$, for all $k \in \mathbb{N}$, then*

$$X_{(n)} \overset{d}{\to} X \in N(0,1).$$

Proof. Here we have $\sigma^2_{(n)} = n\sigma^2$, $\sigma_{(n)} = \sigma\sqrt{n}$. Then (6.6) becomes

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n\sigma^2} \sum_{k=1}^{n} \int\limits_{\{|x-m|>\varepsilon\sigma\sqrt{n}\}} (x-m)^2 dF(x) \\ = &\lim_{n\to\infty} \frac{1}{\sigma^2} \int\limits_{\{|x-m|>\varepsilon\sigma\sqrt{n}\}} (x-m)^2 dF(x) \quad = \quad 0. \end{aligned}$$

So condition (6.6) is satisfied. Then by Theorem 6.2.3,

$$X_{(n)} \overset{d}{\to} X \in N(0,1).$$

□

Definition 6.2.5. *We say that the sequence $\{X_n\}_{n\in\mathbb{N}}$ satisfies **Lyapunov's condition** if there exists $\delta > 0$ such that*

$$\lim_{n\to\infty} \frac{1}{\sigma^{2+\delta}_{(n)}} \sum_{k=1}^{n} E\left(|X_k - m_k|^{2+\delta}\right) = 0. \tag{6.7}$$

Theorem 6.2.6 (Lyapunov)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent random variables satisfying Lyapunov's condition* (6.7)*. Then*

$$X_{(n)} \overset{d}{\to} X \in N(0,1).$$

Proof. We will show that Lyapunov's condition (6.7) implies Lindeberg's condition (6.6) and then use Theorem 6.2.3.
Assume (6.7) holds and let $\varepsilon > 0$, $n \in \mathbb{N}$, $k \in \{1, \dots, n\}$. Then

$$\begin{aligned} E\Big(|X_k - m_k|^{2+\delta}\Big) &= \int\limits_{\mathbb{R}} |x - m_k|^{2+\delta}\, dF_k(x) \\ &\geq \int\limits_{\{|x-m_k|\geq \varepsilon\sigma_{(n)}\}} |x - m_k|^{2+\delta} dF_k(x) \\ &\geq \varepsilon^{\delta}\sigma_{(n)}^{\delta} \int\limits_{\{|x-m_k|\geq \varepsilon\sigma_{(n)}\}} |x - m_k|^{2} dF_k(x) \end{aligned}$$

Then

$$\begin{aligned} &\frac{1}{\sigma_{(n)}^2}\sum_{k=1}^{n} \int\limits_{\{|x-m_k|>\varepsilon\sigma_{(n)}\}} \Big(x - m_k\Big)^2 dF(x) \\ \leq\ & \frac{1}{\varepsilon^2\sigma_{(n)}^{2+\delta}} \sum_{k=1}^{n} E\Big(|X_k - m_k|^{2+\delta}\Big) \end{aligned}$$

Let $n \to \infty$. Then $\{X_n\}$ satisfies (6.6). So by Theorem 6.2.3,

$$X_{(n)} \xrightarrow{d} X \in N(0, 1).$$

□

6.2.2 Moivre-Laplace Theorems

These theorems give approximations related to the binomial probability.

Let us consider a sequence of independent Bernoulli trials with probability of success $p \in (0, 1)$. Recall from Proposition 2.1.1 the probability of k

successes occurring in n trials

$$P(n;k) = C_n^k p^k q^{n-k},$$

with $k \in \{0, \dots, n\}$ and $q = 1 - p$.

Theorem 6.2.7 (Local Moivre-Laplace). *If the sequence $\{x_n\}_{n\in\mathbb{N}}$ defined by*

$$x_n = \frac{k_n - np}{\sqrt{npq}}$$

is bounded, then

$$\lim_{n\to\infty} \frac{\sqrt{npq}\; P(n;k_n)}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_n^2}{2}}} = 1, \tag{6.8}$$

i.e. the binomial probability $P(n;k_n)$ is asymptotically equal to

$$\frac{1}{\sqrt{npq}}\phi(x_n),$$

where $\phi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the density function of a standard normal variable.

Theorem 6.2.8 (Lindeberg-Lévy). *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables with $E(X_n) = \mu$ and $V(X_n) = \sigma^2 > 0,$ for all $n \in \mathbb{N}$. Let $S_n = \sum\limits_{k=1}^{n} X_k,$ for all $n \in \mathbb{N}$. Then*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}} \int\limits_a^b e^{-\frac{t^2}{2}}\, dt, \tag{6.9}$$

for all $a, b \in \mathbb{R}, a < b.$

Proof. We have $m_k = E(X_k) = \mu$, $\sigma^2_{(n)} = n\sigma^2$. So

$$X_{(n)} = \frac{1}{\sigma_{(n)}} \sum_{k=1}^{n} \Big(X_k - m_k\Big) = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

By Proposition 6.2.4, $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} \overset{d}{\to} X \in N(0,1)$. Then

$$\begin{aligned} \lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} < b\right) &\le b \lim_{n\to\infty} \Big(F_{(n)}(b) - F_{(n)}(a)\Big) \\ &= \frac{1}{\sqrt{2\pi}} \int\limits_a^b e^{-\frac{t^2}{2}} dt, \end{aligned}$$

for all $a, b \in \mathbb{R}, a < b$. □

Theorem 6.2.9 (Integral Moivre-Laplace). *Let $\{S_n\}_{n\in\mathbb{N}}$ be a sequence of independent, binomially distributed random variables, with probability distribution functions*

$$S_n \begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}_{k=\overline{0,n}},$$

where $n \in \mathbb{N}$, $p \in (0,1)$ and $q = 1 - p$. Then

$$\lim_{n\to\infty} P\left(a \le \frac{k - np}{\sqrt{npq}} < b\right) = \frac{1}{\sqrt{2\pi}} \int\limits_a^b e^{-\frac{t^2}{2}} dt, \qquad (6.10)$$

for all $a, b \in \mathbb{R},\ a < b$.

Proof. For each $j = \overline{1,n}$, let X_j be independent random variables, each with a $Bern(p)$ distribution

$$X_j \begin{pmatrix} 0 & 1 \\ q & 1-p \end{pmatrix}.$$

Then by Proposition 6.2.4,

$$\lim_{n\to\infty} F_{(n)}(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \ \ , x \in \mathbb{R},$$

where $F_{(n)}$ is the cumulative distribution function of $X_{(n)}$.
On the other hand,

$$\begin{array}{lllll} m_k & = & E(X_k) & = & p, \\ \sigma^2 & = & V(X_k) & = & pq, \\ S_n & = & \sum\limits_{k=1}^{n} X_k & \in & B(n,p). \end{array}$$

So $X_{(n)} = \sum\limits_{k=1}^{n} \frac{X_k - p}{\sqrt{npq}} = \frac{S_n - np}{\sqrt{npq}}$ takes the values $\frac{k-np}{\sqrt{npq}}$, for $k = \overline{0,n}$.
Then

$$\lim_{n\to\infty} P\Big(a \leq \frac{k-np}{\sqrt{npq}} \leq b\Big) = \lim_{n\to\infty} \Big(F_{(n)}(b) - F_{(n)}(a)\Big) = \frac{1}{\sqrt{2\pi}} \int\limits_{a}^{b} e^{-\frac{t^2}{2}} dt.$$

□

Remark 6.2.10.
1. The values of the functions $\Phi, \Psi : \mathbb{R} \to \mathbb{R}$, defined by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-\frac{t^2}{2}} dt, \ \ \Psi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

can be found in tables (see for Ψ Appendix B.1). They are both known as *Laplace's function* and they are related by the formula

$$\Phi(x) = \Psi(x) - \frac{1}{2}.$$

The function $\Psi(x)$ is in fact the cumulative distribution function of a standard normal variable. Using these functions, formula (6.10) can be rewritten as

$$P\left(a \leq \frac{k-np}{\sqrt{npq}} < b\right) \approx \Phi(b) - \Phi(a) = \Psi(b) - \Psi(a),$$

or, equivalently,

$$\begin{aligned} P(a \leq k < b) &\approx \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right) \\ &= \Psi\left(\frac{b-np}{\sqrt{npq}}\right) - \Psi\left(\frac{a-np}{\sqrt{npq}}\right). \end{aligned}$$

2. We can use the previous remark to compare the frequency of the occurrence of a success, $\frac{k}{n}$, to the probability of success, p, in a sequence of n Bernoulli trials. We have

$$\begin{aligned} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) &= P\left(-\varepsilon + p < \frac{k}{n} < \varepsilon + p\right) \\ &= P(np - n\varepsilon < k < np + n\varepsilon) \\ &\approx \Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) - \Phi\left(\frac{-n\varepsilon}{\sqrt{npq}}\right) = 2\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right), \end{aligned}$$

the last equality following from the fact that Φ is an odd function.
3. A direct application of the previous remark, known as *the 3σ rule* :

$$P\left(|k - np| < 3\sqrt{npq}\right) = P\left(\frac{|k-np|}{\sqrt{npq}} < 3\right) \approx 2\Phi(3) > 0.99,$$

follows if we take $\varepsilon = \frac{3\sqrt{npq}}{n}$. This is saying that almost all values (more than 99%) that a binomially distributed random variable with parameters n and p (and hence, variance $V(X) = npq = \sigma^2$) takes, are in the interval $[np - 3\sigma, np + 3\sigma]$.

6.2.3 Other Limit Theorems

We state a few limit theorems for sequences of random variables following other distributions. The first result was actually already stated as Theorem 3.3.1. We restate it now as a limit theorem.

Theorem 6.2.11 (Poisson)**.** *Let* $X_n \in B(n, p_n)$, *for every* $n \in \mathbb{N}$, *with* $p_n \in (0,1)$. *If* $\lim\limits_{n\to\infty} np_n = \lambda > 0$, *then*

$$X_n \xrightarrow{d} X \in P(\lambda).$$

Theorem 6.2.12. *Let* $X_n \in \chi^2(n,1)$, *for every* $n \in \mathbb{N}$. *Then*

$$Y_n = \frac{X_n - E\{X_n\}}{\sqrt{V\{X_n\}}} \xrightarrow{d} X \in N(0,1).$$

Theorem 6.2.13. *Let* $\{\lambda_n\}_{n\in\mathbb{N}}$ *be a sequence of strictly positive real numbers with* $\lim\limits_{n\to\infty} \lambda_n = \infty$ *and let* $X_n \in P(\lambda_n)$, *for every* $n \in \mathbb{N}$. *Then*

$$Y_n = \frac{X_n - E\{X_n\}}{\sqrt{V\{X_n\}}} \xrightarrow{d} X \in N(0,1).$$

PART II. Statistics

Chapter 7

Descriptive Statistics

Statistics is the universal language of sciences. It enables us to accurately describe the findings of scientific research, make decisions, find estimates. Although considered a relatively new science, it has been used in fact since ancient times, for practical, administrative, military, or social purposes, when compiling a census, evaluating or planning production, passing laws, etc. With the development of Probability Theory starting with the eighteenth century, statistical methods blossomed and diversified. In the last few decades, it has become increasingly evident that the interpretation of much of the research in virtually all scientific areas, depends more or less on statistical methods.

Statistics has different meanings to different individuals: to some it represents a way of collecting and displaying large amounts of numerical information, to others a way of making decisions in the face of uncertainty. In fact, each point of view is correct. Statistics is all that and much more.

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.
Historically, descriptive Statistics was developed first, dealing with the "raw" data that people had to handle every day. As that task became increasingly difficult, a scientific and more rigorous approach of Statistics was needed. The transition to inferential Statistics started at the beginning of last century, with the heavier employment of probabilistic methods. Nowadays, especially after the revolution in technology that we have witnessed in the last few decades, it is inferential Statistics that fulfills the needs of modern scientific research.

7.1 Analysis and Display of Data

7.1.1 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 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.

7.1.2 Data Collection, Sampling

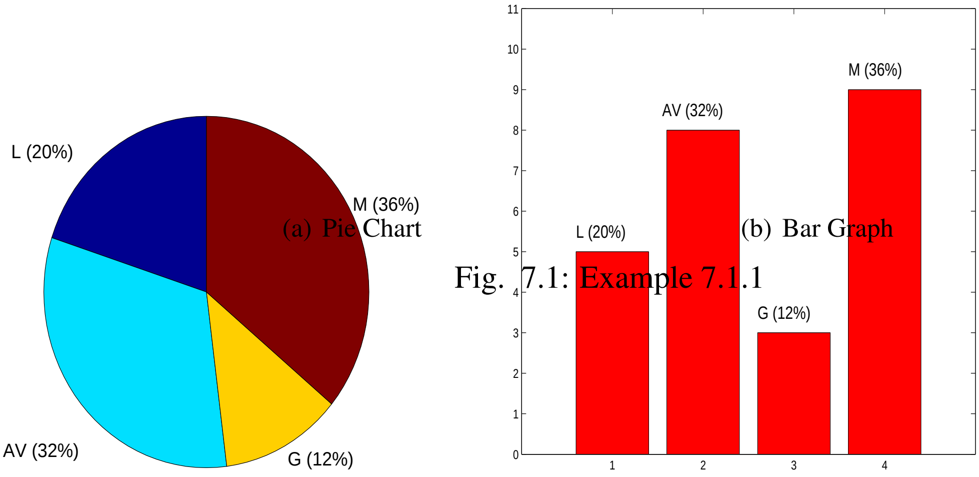
An important first step in any statistical analysis is the **sampling technique**, i.e. the collection of methods and procedures used to gather data. There are several ways of collecting data: If every element of a population is selected, then a **census** is compiled. However, this technique is hardly 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 are made about the entire population.
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, but are of little importance to the purpose of this course and will not be mentioned. Throughout the remaining chapters, we will only consider random samples.

7.1.3 Graphical Display of Data, Frequency Distribution Tables, Histograms

“A picture is worth a thousand words”, a saying still valid in Statistics!
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 7.1.1. 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).

Solution 7.1.1: The pie chart and the bar graph are shown in Figure 7.1. ■



(a) Pie Chart

(b) Bar Graph

Fig. 7.1: Example 7.1.1

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 7.1.

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

Table 7.1: 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 $N = \sum\limits_{i=1}^{k} f_i$ is the sample size.

However, when the data volume is large and the values are nonrepetitive, 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 $m_i = \dfrac{c_{i-1} + c_i}{2}$, 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 $x_1, \dots, x_N$ 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 7.1, only it will contain classes instead of individual values, each with their corresponding features.

Remark 7.1.2.
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*

$$n = 1 + \frac{10}{3}\log_{10}N, \tag{7.1}$$

where N is the sample size. Then it follows that $l = \dfrac{x_{\text{max}} - x_{\text{min}}}{n}$.
The other is a formula for the class width

$$l = \frac{8}{100}\left(x_{\text{max}} - x_{\text{min}}\right). \tag{7.2}$$

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

Histograms and Frequency Polygons
When data is grouped into classes, the best way to visualize the frequency distribution is by constructing a **histogram**. 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 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 7.1.3. The following represents the grades distribution in a Probability and Statistics exam, for one section 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 7.2: 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 (7.2), we have

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

The grouped frequency tables are shown in Tables 7.3 and 7.4. We have also included the relative and cumulative frequencies.
Figure 7.2 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	0.08%	0.08%
2	[2.50 , 4.00)	3.25	4	7	0.11%	0.19%
3	[4.00 , 5.50)	4.75	8	15	0.22%	0.41%
4	[5.50 , 7.00)	6.25	6	21	0.17%	0.58%
5	[7.00 , 8.50)	7.75	5	26	0.14%	0.72%
6	[8.50 , 10.00]	9.25	10	36	0.28%	1.00%

Table 7.3: Grouped Frequency Distribution Table With $n = 6$ Classes

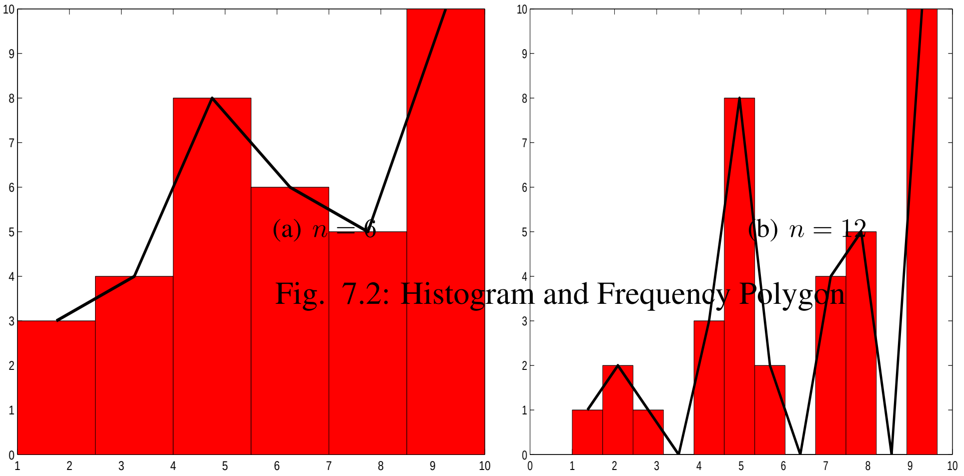
Remark 7.1.4. Due to rounding errors, the length of the last class may be slightly different than the rest of them, even when we group data into classes of the same width.

7.2 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, 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.

No	Class	Mark	Freq.	C. Freq.	R. Freq.	R. C. Freq.
1	[1.00 , 1.72)	1.36	1	1	0.03%	0.03%
2	[1.72 , 2.44)	2.08	2	3	0.06%	0.09%
3	[2.44 , 3.16)	2.80	1	4	0.03%	0.12%
4	[3.16 , 3.88)	3.52	0	4	0.00%	0.12%
5	[3.88 , 4.60)	4.24	3	7	0.08%	0.20%
6	[4.60 , 5.32)	4.96	8	15	0.22%	0.42%
7	[5.32 , 6.04)	5.68	2	17	0.06%	0.48%
8	[6.04 , 6.76)	6.40	0	17	0.00%	0.48%
9	[6.76 , 7.48)	7.12	4	21	0.11%	0.59%
10	[7.48 , 8.20)	7.84	5	26	0.14%	0.73%
11	[8.20 , 8.92)	8.56	0	26	0.00%	0.73%
12	[8.92 , 10]	9.46	10	36	0.27%	1.00%

Table 7.4: Grouped Frequency Distribution Table With $n = 12$ Classes



(a) $n = 6$

(b) $n = 12$

Fig. 7.2: Histogram and Frequency Polygon

7.2.1 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 7.2.1. *The* ***(arithmetic) mean*** *of the data $x_1, \dots, x_N$ is the value*

$$\overline{x}_a = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{7.3}$$

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 7.2.2. 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

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

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

Definition 7.2.3. *The* ***geometric mean*** *of the data* $x_1,\dots,x_N$ *is the value*

$$\overline{x}_g=\sqrt[N]{x_1\dots x_N}. \tag{7.4}$$

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

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

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 ordinary deviations, emphasized by the arithmetic mean.

Definition 7.2.4. *The* ***harmonic mean*** *of the data* $x_1,\dots,x_N$ *is the value*

$$\overline{x}_h=\frac{N}{\displaystyle\sum_{i=1}^{N}\frac{1}{x_i}}. \tag{7.5}$$

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

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

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

Remark 7.2.5.
1. For any set of data $x_1, \dots, x_N$, the well-known *means inequality* holds:

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

with equality holding if and only if $x_1 = \cdots = x_N$.
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 7.2.6. *The* ***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*

$$P(X < x_{me}) \le \frac{1}{2} \le P(X \le x_{me}). \tag{7.6}$$

Remark 7.2.7.
1. The median may or may not be one of the values in the data. If the sorted primary data is

$$x_1 \le \cdots \le x_N,$$

then

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

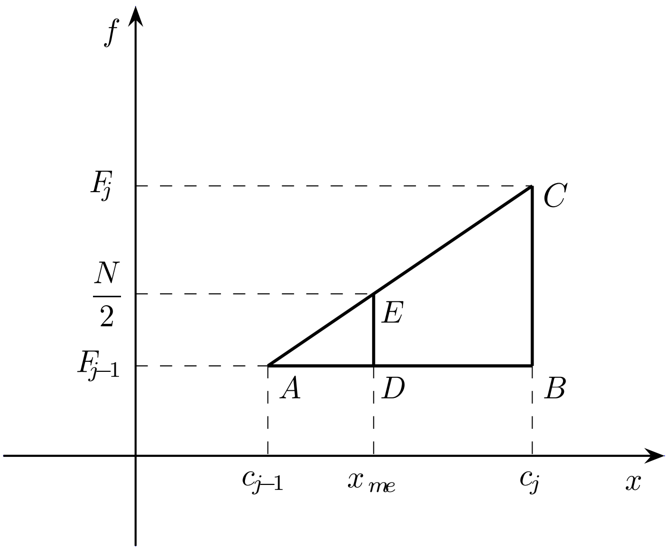


Fig. 7.3: Median Interval

2. For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$, we first determine the *median interval (class)* (the class containing the median), $[c_{j-1}, c_j)$. We use our previous notations: f_j and F_j denote the absolute and cumulative frequency, respectively, of the class $[c_{j-1}, c_j)$. Also, denote by l_j the length of the median interval. Then we have

$$\begin{aligned} f_j &= F_j - F_{j-1}, \\ l_j &= c_j - c_{j-1} \end{aligned}$$

and by (7.6), we must have

$$F_{j-1} < \frac{N}{2} < F_j.$$

The situation is described graphically in Figure 7.3.

Since $\Delta ADE \sim \Delta ABC$, we have $\dfrac{AD}{AB} = \dfrac{DE}{BC}$, so $AD = \dfrac{DE}{BC} \cdot AB$, i.e.

$$x_{me} - c_{j-1} = \frac{\frac{N}{2} - F_{j-1}}{f_j} l_j.$$

Hence

$$x_{me} = c_{j-1} + \frac{N - 2F_{j-1}}{2f_j} l_j.$$

Definition 7.2.8. *A* ***mode****,* x_{mo}*, of a set of data is a most frequent value.*

Remark 7.2.9.
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,

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

3. For grouped data, we first determine the *modal interval* $[c_{k-1}, c_k)$, the interval having a maximum frequency. Then, in a similar manner as for the median, with the usual notations, we find

$$x_{mo} = c_{k-1} + \frac{f_k - f_{k-1}}{2f_k - f_{k-1} - f_{k+1}} l_k.$$

7.2.2 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 much does the data get further away from those middle values. 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 7.2.10. *The **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 7.2.11. *The **mean absolute deviation** 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 7.2.12. *Let X be a set of data sorted increasingly.*

(1) *The **percentiles** 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*

$$P(X < P_k) \le \frac{k}{100}, \ \frac{100-k}{100} \le P(X \le P_k). \tag{7.7}$$

(2) *The **quartiles** of X are the values*

$$Q_1 = P_{25},\ Q_2 = P_{50} = x_{me} \ \textit{and} \ Q_3 = P_{75}, \tag{7.8}$$

that divide the data into 4 equal parts.

Remark 7.2.13.
1. For grouped data, we determine, first, the intervals containing the lower and the upper quartiles (Q_1 and Q_3), i.e the classes $[c_{j-1}, c_j)$, $[c_{k-1}, c_k)$ satisfying the properties

$$F_{j-1} < \frac{N}{4} < F_j,$$

$$F_{k-1} < \frac{3N}{4} < F_k.$$

Then, as before, using linear interpolation, we find

$$\begin{aligned} Q_1 &= c_{j-1} + \frac{N - 4F_{j-1}}{4f_j} l_j, \\ Q_3 &= c_{k-1} + \frac{3N - 4F_{k-1}}{4f_k} l_k. \end{aligned}$$

2. Another important particular case for percentiles are the *deciles*,

$$D_i = P_{10i},\ i = \overline{1,9}.$$

Definition 7.2.14. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range** is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{7.9}$$

(2) *The **interquartile deviation** or the **semi interquartile range** is the value*

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{7.10}$$

(3) *The* ***interquartile deviation coefficient*** *or the* ***relative interquartile deviation*** *is the value*

$$IQDC = \frac{IQD}{x_{me}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{7.11}$$

Remark 7.2.15.
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.

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 a set of data where all values are between 0 and 1, a value of 1000 would surely seem out of place. 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. Once the presence of such outliers is detected, it is suggested that sample statistics be computed both with and without the outliers. 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 Remark 4.5.7 and would work well for unimodal and symmetrical distributions. A more general approach, that works for skewed data, is to consider an outsider any observation that is outside the range

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right] = \left[Q_1 - 3IQD,\ Q_3 + 3IQD\right].$$

Definition 7.2.16.

(1) *The **moment of order k** is the value*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k,\ \ \overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k, \tag{7.12}$$

for primary and for grouped data, respectively.

(2) *The **central moment of order k** is the value*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^k,\ \ \overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^k \tag{7.13}$$

for primary and for grouped data, respectively.

(3) *The **variance** is the value*

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2,\ \ \overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2 \tag{7.14}$$

*for primary and for grouped data, respectively. The quantity $\overline{\sigma} = \sqrt{\overline{\sigma}^2}$ is the **standard deviation**.*

Remark 7.2.17.

1. We will see later that when the data represents a sample (not the entire population), a better formula would be

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2,\ \ s^2 = \frac{1}{N-1}\sum_{i=1}^{n} f_i(x_i - \overline{x})^2, \tag{7.15}$$

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 Chapter 9. For now, we will just agree to use (7.14) to compute the variance of a set of data that represents a population and (7.15) for the variance of a sample.
2. A more efficient computational formula for the variance is

$$\overline{\sigma}^2 = \frac{1}{N}\left(\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right), \tag{7.16}$$

which follows straight from the definition.

Definition 7.2.18. *The **coefficient of variation** is the value*

$$CV = \frac{\overline{\sigma}}{\overline{x}}.$$

Remark 7.2.19.
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 tradeoff.

Definition 7.2.20. *The following are **Pearson's coefficients**:*

(1) *the **absolute and relative asymmetry**, respectively,*

$$AS = \overline{x} - x_{mo}, \quad RAS = \frac{AS}{\overline{\mu}_2^{1/2}} = \frac{\overline{x} - x_{mo}}{\overline{\sigma}},$$

(2) *the **skewness*** $\gamma_1 = \dfrac{\overline{\mu}_3}{\overline{\sigma}^3}$,

(3) *the **kurtosis*** $\gamma_2 = \dfrac{\overline{\mu}_4}{\overline{\sigma}^4}$.

Remark 7.2.21.
1. The skewness is a measure of asymmetry in the sense that negative values indicate data that are skewed to the left (the left tail is long relative to the right tail), while positive values indicate data that are skewed to the right. The normal distribution has a skewness of zero and symmetric data, in general, have a skewness near zero.
2. The kurtosis is a measure of the peakedness or flatness of the data, compared to the normal distribution. The normal distribution has kurtosis 3, values smaller than 3 indicate a "flat" distribution, while values larger than 3 indicate a "peaked" distribution. For this reason, sometimes the kurtosis is defined as $\gamma_2 = \dfrac{\overline{\mu}_4}{\overline{\sigma}^4} - 3$.

7.3 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 and one or more independent variables. 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.

7.3.1 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 $(x_i, y_i)_{i=\overline{1,k}}$, $x_i \in X,\ y_i \in Y$, for $i = \overline{1,k}$.

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 below.

$X \setminus Y$	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 7.5: Correlation Table

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).

Definition 7.3.1. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 7.5 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*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{7.17}$$

(2) *The* ***central moment of order*** $(\boldsymbol{k_1}, \boldsymbol{k_2})$ *of* (X, Y) *is the value*

$$\overline{\mu}_{k_1k_2} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_i - \overline{x})^{k_1}(y_j - \overline{y})^{k_2}, \tag{7.18}$$

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 7.3.2. 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 7.3.3. *Let* (X, Y) *be a two-dimensional characteristic whose distribution is given by Table 7.5.*

(1) *The* ***covariance*** *of* (X, Y) *is the value*

$$\operatorname{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{7.19}$$

(2) *The* ***correlation coefficient*** *of* (X, Y) *is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X\overline{\sigma}_Y}. \tag{7.20}$$

These two notions have been mentioned before, in Chapter 4, 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 $\mathrm{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y},$$

as before, satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{7.21}$$

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. That means we should consider *conditional means*, that were first introduced in Chapter 4.

Definition 7.3.4. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 7.5.*

(1) *The **conditional mean** of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i.}} \sum_{j=1}^{n} f_{ij} y_j, \ i = \overline{1, m}. \tag{7.22}$$

(2) *The **conditional mean** of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{.j}} \sum_{i=1}^{m} f_{ij} x_i, \ j = \overline{1, n}. \tag{7.23}$$

Definition 7.3.5. *Let (X, Y) be a two-dimensional characteristic.*

(1) *The curve $y = f(x)$ consisting of 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 $x = g(x)$ consisting of the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1, n}$, is called the **curve of regression** of X on Y.*

Remark 7.3.6. 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).

7.3.2 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)

$$S = SSE = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - y(x_i)\Big)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)^2$$

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.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \quad (7.24)$$

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 (7.24) 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

$$y - \overline{y} = \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x - \overline{x}) \tag{7.25}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x - \overline{x} = \overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y - \overline{y})\,. \tag{7.26}$$

Remark 7.3.7.

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

$$|\,\overline{\rho}\,| = \overline{a}_{Y|X}\,\overline{a}_{X|Y}.$$

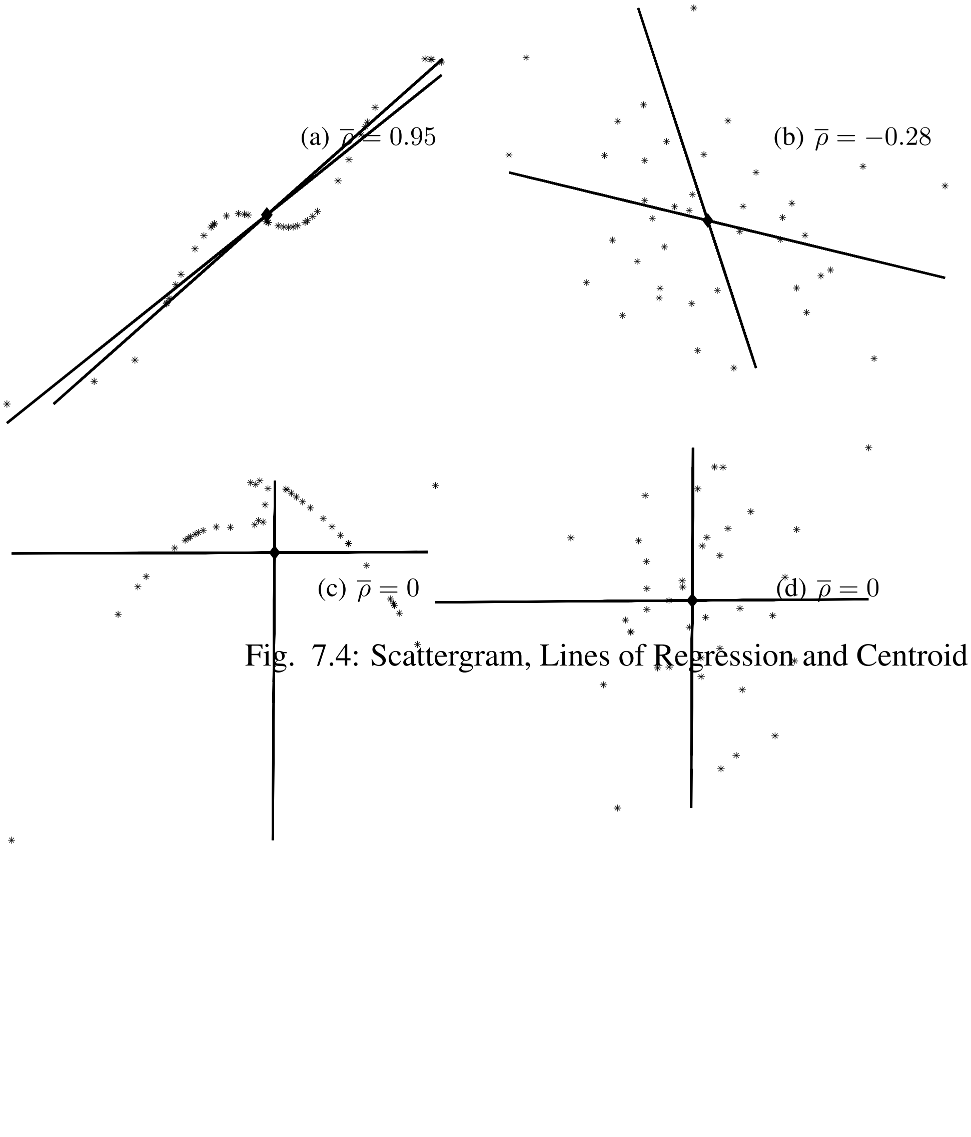
3. For the angle α between the two lines of regression, we have

$$\tan\alpha \;=\; \frac{1 - \overline{\rho}^2}{\overline{\rho}^2} \cdot \frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2 + \overline{\sigma}_Y^2}.$$

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.

Example 7.3.8. Let us examine the situations graphed in Figure 7.4. In Figure 7.4(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 7.4(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 7.4(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 $Y = -X^2 + \sin\left(\dfrac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.
Finally, in Figure 7.4(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 7.3.9. 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 $y = \dfrac{a}{x} + b$.



(a) $\overline{\rho} = 0.95$

(b) $\overline{\rho} = -0.28$

(c) $\overline{\rho} = 0$

(d) $\overline{\rho} = 0$

Fig. 7.4: Scattergram, Lines of Regression and Centroid

Chapter 8

Sample Theory

Suppose we are interested in studying a characteristic (a random variable) X, relative to a population P, of 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.

8.1 Sample Functions

We choose n ($n \leq N$) objects 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 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, $x_1, \dots, x_n$, the observed values of $X_1, \dots, X_n$.

Definition 8.1.1. *A* ***random sample of size*** $\boldsymbol{n}$ *from the distribution of* X*, a characteristic relative to a population* P*, is a collection of* n *independent random variables* $X_1, \dots, X_n$*, having the same distribution as* X*. The variables* $X_1, \dots, X_n$*, are called* ***sample variables*** *and their observed values* $x_1, \dots, x_n$*, are called* ***sample data****.*

Remark 8.1.2. The term *random sample* may refer to the objects selected, to the sample variables, or to the sample data. It is usually clear from the context which meaning is intended. In general, we use capital letters to denote sample variables and corresponding lowercase letters for their values, the sample data.

We are able now to define sample functions, or statistics, in the more precise context of random variables.

Definition 8.1.3. *A* ***sample function*** *or* ***statistic*** *is a random variable*

$$Z_n = h_n(X_1, \dots, X_n),$$

where $h_n : \mathbb{R}^n \to \mathbb{R}$ *is a measurable function. The value of the sample function* Z_n *is* $z_n = h_n(x_1, \dots, x_n)$*.*

We will revisit now some sample numerical characteristics discussed in the previous chapter and define them as sample functions.

8.1.1 Sample Mean

Definition 8.1.4. *The* ***sample mean*** *is the sample function defined by*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8.1}$$

and its value is $\overline{x}_n = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

Now that the sample mean is defined as a random variable, we can discuss its distribution and its numerical characteristics.

Proposition 8.1.5. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$. Then*

$$E\left(\overline{X}\right) = \mu \text{ and } V\left(\overline{X}\right) = \frac{\sigma^2}{n}. \tag{8.2}$$

Moreover, if $X \in N(\mu, \sigma)$, then $\overline{X} \in N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Proof. Since $X_1, \dots, X_n$ 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 using the properties of expectation in Theorem 4.1.4, 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 $X_1, \dots, X_n$ are also independent, by Theorem 4.2.2, 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 (see [11]). □

Corollary 8.1.6. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$ and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}.$$

Then the variable Z_n converges in distribution to a standard normal variable, as $n \to \infty$. Moreover, if $X \in N(\mu, \sigma)$, then the statement is true for every $n \in \mathbb{N}$.

Proof. This is a direct consequence of Propositions 8.1.5 and 6.2.4. □

8.1.2 Sample Moments and Sample Variance

Definition 8.1.7. *The statistic*

$$\overline{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{8.3}$$

is called the ***sample moment of order k*** *and its value is* $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.
The statistic

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \tag{8.4}$$

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 8.1.8. Just like for theoretical (population) moments, we have

$$\begin{aligned} \overline{\nu}_1 &= \overline{X}, \\ \overline{\mu}_1 &= 0, \\ \overline{\mu}_2 &= \overline{\nu}_2 - \overline{\nu}_1^2. \end{aligned}$$

Next we discuss the distributions and characteristics of these new sample functions.

Proposition 8.1.9. *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*

$$E\left(\overline{\nu}_k\right) = \nu_k \ \textit{ and } \ V\left(\overline{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \tag{8.5}$$

Proof. First off, the condition that ν_{2k} exists for X ensures the fact that all theoretical moments of X of order up to k also exist, by Remark 4.2.6. The rest follows as before. We have

$$E\left(\overline{\nu}_k\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n}\sum_{i=1}^{n} E(X^k) = \frac{1}{n}\, n\nu_k = \nu_k$$

and

$$\begin{aligned} V\left(\overline{\nu}_k\right) &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i^k) &&= \frac{1}{n^2}\sum_{i=1}^{n} V(X^k) \\ &= \frac{1}{n^2}\, n\left(\nu_{2k} - \nu_k^2\right) &&= \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right). \end{aligned}$$

□

Corollary 8.1.10. *Let X be a characteristic satisfying the hypothesis of Proposition 8.1.9 and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{\nu}_k - \nu_k}{\sqrt{\dfrac{\nu_{2k} - \nu_k^2}{n}}}.$$

Then Z_n converges in distribution to a standard normal variable.

We only discuss the properties of the sample central moment of order 2.

Proposition 8.1.11. *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*

$$\begin{aligned} E\left(\overline{\mu}_2\right) &= \frac{n-1}{n}\,\sigma^2, \qquad (8.6)\\ V\left(\overline{\mu}_2\right) &= \frac{n-1}{n^3}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big], \\ \operatorname{cov}(\overline{X}, \overline{\mu}_2) &= \frac{n-1}{n^2}\,\mu_3. \end{aligned}$$

Proof. We will only prove the first assertion, (8.6), as it is the most important and oftenly used property of $\overline{\mu}_2$. Using Proposition 8.1.9, the properties of expectation and the fact that $X_1, \ldots, X_n$ are independent and identically distributed, we have

$$\begin{aligned}
E\left(\overline{\mu}_2\right) &= E\left(\overline{\nu}_2\right) - E\left(\overline{\nu}_1^2\right) = \nu_2 - E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)^2\right) \\
&= \nu_2 - \frac{1}{n^2} E\Big(\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\Big) \\
&= \nu_2 - \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(X_i^2\right) + 2\sum_{i<j} E(X_i)E(X_j)\right] \\
&= \nu_2 - \frac{1}{n^2}\left[n\nu_2 + 2\,\frac{n(n-1)}{2}\,\nu_1^2\right] = \nu_2 - \frac{1}{n}\nu_2 - \frac{n-1}{n}\nu_1^2 \\
&= \frac{n-1}{n}\left(\nu_2 - \nu_1^2\right) = \frac{n-1}{n}\,\sigma^2.
\end{aligned}$$

□

Remark 8.1.12.
1. For large samples, i.e. when $n \to \infty$, $\overline{X}$ and $\overline{\mu}_2$ are uncorrelated.
2. If X has a symmetric distribution, then $\mu_3 = 0$ and, hence, $\overline{X}$ and $\overline{\mu}_2$ are uncorrelated for every $n \in \mathbb{N}$.
3. As before, one can show that under the assumptions of Proposition 8.1.11, the sequence

$$Z_n = \frac{\overline{\mu}_2 - \sigma^2}{\sqrt{\dfrac{\mu_4 - \sigma^4}{n}}}$$

converges in distribution to a standard normal variable, as $n \to \infty$.
4. 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 8.1.13. *The statistic*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{8.7}$$

is called the ***sample variance*** *and its value is* $\displaystyle\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
The statistic $s = \sqrt{s^2}$ *is called the* ***sample standard deviation****.*

Remark 8.1.14. Notice that the sample central moment of order 2 is no longer equal to the sample variance, as we are used to. In fact, we have

$$s^2 = \frac{n}{n-1}\,\overline{\mu}_2.$$

Then, by Proposition 8.1.11, we have for the sample variance

$$\begin{aligned} E\left(s^2\right) &= \mu_2 \;=\; \sigma^2, \\ V\left(s^2\right) &= \frac{1}{n(n-1)}\Big[(n-1)\mu_4 - (n-3)\sigma^4\Big], \\ \operatorname{cov}(\overline{X}, s^2) &= \frac{1}{n}\,\mu_3. \end{aligned} \tag{8.8}$$

8.1.3 Sample Distribution Function

Thus far, we have been able to define sample functions that mimicked their theoretical correspondents (mean, moments, variance) and, hopefully, will provide good inferential estimates for the entire population. The ultimate goal of Statistics is to derive the probability distribution that generated a sample from the sample itself, i.e. to define a sample function that gives some idea of the cumulative distribution function of a characteristic, relative to the entire population. The idea is suggested by the shape of the cumulative distribution function of discrete random variables.

Definition 8.1.15. *Let X be a characteristic and $X_1, \ldots, X_n$ sample variables for a random sample of size n. The* ***sample distribution function*** *or* ***empirical distribution function*** *is the sample function $\overline{F}_n : \mathbb{R} \to \mathbb{R}$ defined by*

$$\overline{F}_n(x) = \frac{1}{n} I(X_i \leq x) = \frac{\text{card}\{X_i \mid X_i \leq x\}}{n}, \tag{8.9}$$

where $I(A)$ is the indicator of event A, and its value is $\dfrac{\text{card}\{x_i \mid x_i \leq x\}}{n}$.

Remark 8.1.16.
1. So, the sample distribution function at a value x is given by

$$\overline{F}_n(x) = \frac{\text{number of sample elements } x_i \leq x}{n}.$$

Defining it this way makes it an excellent tool for measuring how faithful the sample is to the probability distribution: where observations are densely packed, this function grows rapidly, which is exactly what is expected from the true distribution function (for where the distribution function grows rapidly, the probability density—its derivative, is large, which is propitious to a high concentration of observations), while observations that are few and far between happen in regions of low probability density.
2. Assuming the sample data $x_1, \ldots, x_n$ are sorted in increasing order, a more explicit computational formula for the sample distribution function is

$$\overline{F}_n(x) = \begin{cases} 0, & \text{if } x < x_1 \\ \dfrac{i}{n}, & \text{if } x_i \leq x < x_{i+1},\ i = \overline{1, n-1} \\ 1, & \text{if } x \geq x_n. \end{cases}$$

Thus $\overline{F}_n$ presents similar properties to those of a cumulative distribution function of a discrete random variable:

– it is a step function;
– it monotonically increases from 0 to 1;
– it is constant on semi-open intervals $[x_i, x_{i+1})$;
– its limits at $\pm\infty$ are 1 and 0, respectively.

In addition, here the height of each "step" is $\dfrac{1}{n}$.

3. The sample distribution function can also be viewed as a random variable (since it *is* a sample function). If F denotes the cumulative distribution function of the characteristic X, then for each $x \in \mathbb{R}$, $\overline{F}_n(x)$ is a discrete random variable with probability distribution function

$$\overline{F}_n(x) \begin{pmatrix} \dfrac{i}{n} \\ \\ C_n^i \, (F(x))^i \, (1 - F(x))^{n-i} \end{pmatrix}_{i=\overline{0,n}} .$$

Now that we have seen the similarities between a cumulative distribution function and a sample distribution function, the question that naturally arises is how much does the latter resemble the former, how well and in what sense, does it approximate it. Of course, since these are random variables, convergence of such types of variables should be considered. The Weak Law of Large Numbers can be used to show that

$$\overline{F}_n(x) \xrightarrow{p} F(x),$$

for every fixed $x \in \mathbb{R}$. But an even stronger convergence result holds:

Theorem 8.1.17 (Glivenko-Cantelli)**.** *Let X be a characteristic with cumulative distribution function F and $X_1, \dots, X_n$ sample variables for a random sample of size n, with sample distribution function $\overline{F}_n$. Let*

$$D_n = \sup_{x \in \mathbb{R}} |\overline{F}_n(x) - F(x)|. \tag{8.10}$$

Then

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1,$$

i.e. the sample distribution function converges almost surely to the cumulative distribution function .

Kolmogorov strengthened this result, by effectively providing the rate of this convergence. A random variable is said to follow the **Kolmogorov distribution**, if its cumulative distribution function is given by

$$K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{k=1}^{\infty}e^{-\frac{(2k-1)^2\pi^2}{8x^2}}, \qquad (8.11)$$

for all $x > 0$ and 0, otherwise. The function (8.11) is known as *Kolmogorov's function* and its values can be found in tables.

Theorem 8.1.18 (Kolmogorov)**.** *Assume the hypotheses of Theorem 8.1.17 are satisfied and further, assume that F is continuous. Then*

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \le x\right) = K(x),$$

for all $x > 0$, i.e. the variable $\sqrt{n}D_n$ converges in distribution to the Kolmogorov distribution.

8.2 Properties of Sample Functions

Sample functions are very important in inferential Statistics, since they represent the only "real" information we have about a population. The goal is to be able to make "predictions" on a population characteristic, based on this information that a sample provides and also to measure (in terms of probability) how good those predictions are. Results such as Proposition 8.1.5 and Corollary 8.1.6 can be very useful in that sense. We present next more such results (without proof), that will be used in the next two chapters in making inferences about population characteristics.

Let X be a characteristic of a population from which a random sample of size n is drawn and let $X_1, \ldots, X_n$ be the sample variables.

Proposition 8.2.1. *Assume $X \in N(0,1)$ and let*

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sqrt{n}\, \overline{X} \quad \textit{and} \quad V_n = \sum_{i=1}^{n} (X_i - \overline{X})^2 = (n-1)\, s^2.$$

Then $U_n \in N(0,1)$ and $V_n \in \chi^2(n-1)$.

Proposition 8.2.2. *Assume $X \in N(\mu, \sigma)$ and let*

$$U_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \textit{and} \quad V_n = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{(n-1)\, s^2}{\sigma^2}.$$

Then $U_n \in N(0,1)$ and $V_n \in \chi^2(n-1)$.

Proposition 8.2.3. *Assume $X \in N(\mu, \sigma)$ and let*

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Then $T \in T(n-1)$.

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. We draw from both populations random samples of sizes n_1 and n_2, respectively. Denote the two sets of random variables by

$$X_{11}, \ldots, X_{1n_1} \quad \text{and} \quad X_{21}, \ldots, X_{2n_2}.$$

Denote the sample means and sample variances 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

$$s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(X_{1i}-\overline{X}_1\right)^2, \;\; s_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(X_{2j}-\overline{X}_2\right)^2.$$

In addition, denote by

$$s_p^2 = \frac{\displaystyle\sum_{i=1}^{n_1}\left(X_{1i}-\overline{X}_1\right)^2 + \sum_{j=1}^{n_2}\left(X_{2j}-\overline{X}_2\right)^2}{n_1+n_2-2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

the **pooled variance** of the two samples, i.e. a variance that considers the sample data from both samples.

Proposition 8.2.4. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$ are independent and let*

$$Z = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \quad \text{and} \quad T = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}.$$

Then $Z \in N(0,1)$ and $T \in T(n)$, where

$$\frac{1}{n} = \frac{c^2}{n_1-1}+\frac{(1-c)^2}{n_2-1} \quad \text{and} \quad c = \frac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}.$$

Proposition 8.2.5. *Assume $X_{(1)} \in N(\mu_1, \sigma)$ and $X_{(2)} \in N(\mu_2, \sigma)$ are independent and let*

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$$

Then $T \in T(n_1 + n_2 - 2)$.

Proposition 8.2.6. *Assume $X_{(1)} \in N(\mu_1, \sigma_1)$ and $X_{(2)} \in N(\mu_2, \sigma_2)$ are independent and let*

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

Then $F \in F(n_1 - 1, n_2 - 1)$.

Chapter 9

Estimation

Populations are characterized by parameters. The goal of inferential Statistics is to make inferences about one or more population parameters on the basis of a sample. We will refer to the parameter to be estimated as the *target parameter*.

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 density $f(x;\theta)$ depends on the parameter θ, which is to be estimated. If X is discrete, then f represents the probability distribution function , while if X is continuous, f is the probability density function.

As before, we consider a random sample of size n, represented by the sample variables $X_1, \ldots, X_n$. The notations introduced in the previous chapter for some sample functions still stand. When nothing else is mentioned, μ denotes the population mean and σ^2 the population variance.

9.1 Point Estimation

We are searching for one value that approximates the unknown parameter θ. This approximation is done by using an appropriate statistic. A statistic that estimates the target parameter θ is called a **point estimator** for θ and is denoted by $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ or $\overline{\theta} = \overline{\theta}(X_1, \dots, X_n)$. The value of the point estimator, the **point estimate**, is the actual approximation of the unknown parameter. For instance, to approximate the mean of a population, μ, we may use the sample mean $\overline{X}$. Then the point estimator is $\hat{\mu} = \overline{X}$ (a statistic, a random variable) and the point estimate is $\mu \approx \overline{x}$ (a value, a number).

9.1.1 Unbiased Estimators

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 $\hat{\theta}$ 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.

Definition 9.1.1. *A point estimator $\hat{\theta}$ is called an **unbiased** estimator for θ, if*

$$E(\hat{\theta}) = \theta. \tag{9.1}$$

*The **bias** of $\hat{\theta}$ is the value $B = E(\hat{\theta}) - \theta$.*

Example 9.1.2.
1. Recall from Proposition 8.1.5 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. By Proposition 8.1.11, the sample central moment of order 2 *is not* an

unbiased estimator for the population variance (or it is a *biased* estimator), since

$$E(\overline{\mu}_2) = \frac{n-2}{n}\sigma^2 \neq \sigma^2,$$

while the sample variance *is* one, $E(s^2) = \sigma^2$ (see Remark 8.1.14). That was the main reason for the way the sample variance was defined.

Let $\sigma^2_{\hat{\theta}}$ denote the variance of the distribution of the estimator $\hat{\theta}$. Then the standard deviation of the sampling distribution of $\hat{\theta}$, $\sigma_{\hat{\theta}} = \sqrt{\sigma^2_{\hat{\theta}}}$ is called the **standard error** of the estimator $\hat{\theta}$. The name is justified by the fact that when we have a point estimator, it is highly desirable that its distribution does not vary too much from its mean value. Thus, the measure of this variability, the standard deviation, in a way also measures how "good" the estimate is, it measures the "error" of the approximation.
In Table 9.1, we present some common unbiased estimators, their means and their standard errors. The notation μ stands, as usually, for the mean of a population, p denotes a proportion of individuals in a population having a certain trait (for example, the probability of success in a binomial process), $q = 1 - p$, $\hat{p}$ is the sample proportion (the number of observations in the sample that have that trait, divided by the sample size) and we have the same notations for comparing two population means or two population proportions (when the sample notations refer to two samples, one from each population).

Remark 9.1.3.
1. The expected values and the standard errors in Table 9.1 are valid regardless of the form of the density function of the underlying population. Proposition 8.1.5 states that for the sample mean. Similar results hold for the other three point estimators (see [4]).
2. For large samples (as $n, n_1, n_2 \to \infty$), all four estimators have probability densities that are approximately normal. The Central Limit Theorem

(8.1.6) and similar theorems justify these statements. In practice, it was determined that “large” means $n > 30$ for $\overline{X}$ and $n_1 + n_2 > 40$ for $\overline{X}_1 - \overline{X}_2$.

Target Param. θ	Sample Size	Pt. Estimator $\hat{\theta}$	Mean $E(\hat{\theta})$	St. Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{X}$	μ	$\dfrac{\sigma}{\sqrt{n}}$
p	n	$\hat{p}$	p	$\sqrt{\dfrac{pq}{n}}$
$\mu_1 - \mu_2$	n_1, n_2	$\overline{X}_1 - \overline{X}_2$	$\mu_1 - \mu_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
$p_1 - p_2$	n_1, n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}$

Table 9.1: Common Unbiased Estimators

9.1.2 Minimum Variance Estimators

Our goal is to produce as good estimators as possible. Although unbiasedness is an attractive feature of a point estimator, it is not enough. In order to design procedures for improving our approximations, we need to undertake a more formal and detailed examination of some mathematical properties of point estimators.

Definition 9.1.4. *An estimator $\hat{\theta} = \hat{\theta}_n$, found from a sample of size n, is said to be a* ***consistent*** *estimator for θ, if $\hat{\theta}_n \xrightarrow{p} \theta$, i.e. if for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\Big(|\hat{\theta}_n - \theta| < \varepsilon\Big) = 1.$$

Remark 9.1.5. The property of consistency of a point estimator ensures the fact that the larger the sample size, the better the estimate (the "closer" it gets to the actual value of the target parameter), which is a very reasonable expectation from an estimator.

Definition 9.1.6. *The statistic $S = S(X_1, \dots, X_n)$ is called* ***sufficient*** *for (the estimation of) θ, if the conditional probability distribution of the data $X_1, \dots, X_n$, given the statistic S, does not depend on the parameter θ, i.e.*

$$f(x_1, \dots, x_n; \theta|S) = f(x_1, \dots, x_n|S).$$

Remark 9.1.7.
1. More intuitively, as the name suggests, a statistic is sufficient for the estimation of a parameter if it contains all the information in the given sample about that parameter, if no other statistic which can be computed from the same sample provides any additional information as to the value of the target parameter, than does that sufficient statistic. Then the conditional probability distribution of the data does not depend on the unknown parameter except through the sufficient statistic.
2. If a statistic S is sufficient for a parameter θ, then the conditional distribution of *any* statistic, given S, does not depend on θ, since any statistic is just a function of the sample variables $X_1, \dots, X_n$.

Neither Definition 9.1.6, nor Remark 9.1.7 is easy to verify for a given statistic. An equivalent, more accessible form was given by Fisher, which we present below.

Definition 9.1.8. *The* ***likelihood function*** *of a sample $X_1, \ldots, X_n$ is the joint probability function of the sample, i.e. the sample function*

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta), \tag{9.2}$$

whose value $L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$ represents the joint probability distribution (in the discrete case) or the joint density (in the continuous case) of the random vector $(X_1, \ldots, X_n)$.

Theorem 9.1.9 (Fisher's Factorization Criterion)**.**
A statistic $S = S(X_1, \ldots, X_n)$ is sufficient for θ, if and only if the likelihood function can be factored into two nonnegative functions

$$L(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) g(s; \theta), \tag{9.3}$$

such that one factor, h, does not depend on θ and the other factor, g, which does depend on θ, depends on $(x_1, \ldots, x_n)$ only through the value of $s = S(x_1, \ldots, x_n)$.

Example 9.1.10. Let $X_1, \ldots, X_n$ be sample variables for a random sample drawn from an exponential distribution of parameter $\frac{1}{\theta}$, with $\theta > 0$, unknown. Show that the sample mean $S = \overline{X}$ is a sufficient statistic for the estimation of θ.

Solution 9.1.10: By (3.32), we have that for each $i = \overline{1, n}$,

$$f(x_i; \theta) = \frac{1}{\theta} e^{-\dfrac{x_i}{\theta}},$$

for $x_i > 0$ and 0, otherwise. Then for $x_1, \dots, x_n > 0$, the likelihood function is given by

$$\begin{aligned} L(x_1, \dots, x_n; \theta) &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-\dfrac{x_i}{\theta}} = \frac{1}{\theta^n} e^{-\dfrac{1}{\theta}\sum_{i=1}^{n} x_i} \\ &= \frac{1}{\theta^n} e^{-\dfrac{n\overline{x}}{\theta}} = \left(\frac{1}{\theta} e^{-\dfrac{\overline{x}}{\theta}} \right)^n \\ &= h(x_1, \dots, x_n) g(\overline{x}; \theta), \end{aligned}$$

where $h(x_1, \dots, x_n) = 1$ and $g(\overline{x}; \theta) = \left(\frac{1}{\theta} e^{-\dfrac{\overline{x}}{\theta}} \right)^n$.

Thus, by Theorem 9.1.9, $\overline{X}$ is a sufficient statistic for the estimation of θ.

■

Sufficient statistics play an important role in finding good estimators. We have already mentioned that for an unbiased estimator, we want the standard error (i.e., the variance of the sampling distribution) to be small. Having an unbiased estimator, sufficient statistics can be used to find an unbiased estimator with a smaller variance.

Theorem 9.1.11 (Rao-Blackwell)**.** *Let $\hat{\theta}$ be an unbiased estimator and S be a sufficient statistic for θ. Then*

$$\overline{\theta} = E(\hat{\theta}|S) \tag{9.4}$$

is also an unbiased estimator for θ and $V(\overline{\theta}) \leq V(\hat{\theta})$.

Proof. First, we have to show that the estimator given by (9.4) is well-defined for our purposes, i.e. that it only depends on the sample variables (so it is a statistic) and not on θ. Since $\hat{\theta}$ is a statistic and S is sufficient for

θ, by Remark 9.1.7(2), the conditional distribution of $\hat{\theta}$, given S, does not depend on θ and thus, neither does its expectation. So $\overline{\theta}$ is a statistic. Recall from Proposition 4.4.4 the properties

$$\begin{aligned} E(E(X|Y)) &= E(X), \\ V(X) &= E(V(X|Y)) + V(E(X|Y)). \end{aligned}$$

Then, since $\hat{\theta}$ is unbiased, we have

$$\begin{aligned} E(\overline{\theta}) &= E(E(\hat{\theta}|S)) = E(\hat{\theta}) = \theta, \\ V(\hat{\theta}) &= V(E(\hat{\theta}|S)) + E(V(\hat{\theta}|S)) = V(\overline{\theta}) + E(V(\hat{\theta}|S)). \end{aligned}$$

So $\overline{\theta}$ is unbiased and since $V(\hat{\theta}|S=s) \geq 0, \forall s$, $E(V(\hat{\theta}|S)) \geq 0$ and thus

$$V(\overline{\theta}) \leq V(\hat{\theta}).$$

□

So, using sufficient statistics we can improve an unbiased estimator by lowering its variance. Ideally, we would like to find an unbiased estimator with the smallest variance possible.

Definition 9.1.12. *An unbiased estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ for θ is called a **minimum-variance unbiased estimator (MVUE)**, if it has lower variance than any other unbiased estimator for θ.*

Remark 9.1.13. It can be shown that if an unbiased estimator exists for a parameter, then an MVUE also exists and it is unique.

Unfortunately, repeated applications of Theorem 9.1.11 to an unbiased estimator $\hat{\theta}$ using *the same* sufficient statistic S, will not lead to the MVUE, since

$$E\left(E(\hat{\theta}|S)|S\right) = E\left(E(\hat{\theta}|S)\right) = E(\hat{\theta}|S).$$

Stronger conditions on the statistic S need to be imposed.

Definition 9.1.14. *The statistic $S = S(X_1, \dots, X_n)$ is called* ***complete*** *for the family of probability distributions $f(x;\theta)$, $\theta \in A$, if*

$$E(\varphi(S)) = 0, \forall \theta \in A \Longrightarrow \varphi \overset{a.s.}{=} 0. \tag{9.5}$$

Theorem 9.1.15 (Lehmann-Scheffé)**.** *Let $\hat{\theta}$ be an unbiased estimator and S be a sufficient and complete statistic for θ. Then $\overline{\theta} = E(\hat{\theta}|S)$ is the MVUE.*

Proof. Let $\widetilde{\theta} = \widetilde{\theta}(X_1, \dots, X_n)$ be any unbiased estimator for θ. By Theorem 9.1.11, $\theta_1 = E(\widetilde{\theta}|S)$ is unbiased ($E(\theta_1) = \theta$), and $V(\theta_1) \leq V(\widetilde{\theta})$. The same things are true for $\overline{\theta}$. Then $E(\theta_1) = E(\overline{\theta}) = \theta$, i.e.

$$E(E(\hat{\theta}|S)) = E(E(\widetilde{\theta}|S)) = \theta,$$

which can be rewritten as

$$E\left(E(\hat{\theta}|S) - E(\widetilde{\theta}|S)\right) = 0$$

and this is true for any unbiased estimator $\widetilde{\theta}$.
Since S is complete,

$$E(\hat{\theta}|S) \overset{a.s.}{=} E(\widetilde{\theta}|S),$$

so

$$\overline{\theta} \overset{a.s.}{=} \theta_1$$

and $V(\overline{\theta}) = V(\theta_1)$.
Finally, we have $V(\overline{\theta}) = V(\theta_1) \leq V(\widetilde{\theta})$ i.e.

$$V(\overline{\theta}) \leq V(\widetilde{\theta}),$$

for any unbiased estimator $\widetilde{\theta}$. Thus $\overline{\theta}$ is an MVUE. □

Example 9.1.16. Let $X_1, \dots, X_n$ be sample variables for a random sample drawn from a Bernoulli distribution with parameter $p \in (0, 1)$, unknown. Find the MVUE for p.

Solution 9.1.16: Recall from (3.7) that for each $i = \overline{1, n}$,

$$f(x_i; p) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \;\; x_i \in \{0, 1\}.$$

Then

$$\begin{aligned} L(x_1, \ldots, x_n; p) &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= p^s(1-p)^{n-s}, \end{aligned}$$

where S is the statistic

$$S = S(X_1, \ldots, X_n) = \sum_{1=1}^{n} X_i = n\overline{X}.$$

By Theorem 9.1.9, with $h(x_1, \ldots, x_N) = 1$ and $g(s; p) = p^s(1-p)^{n-s}$, S is sufficient.

Now assume that $E(\varphi(S)) = 0$, for all $p \in (0, 1)$. Recall from Remark 4.1.6 that since $X_1, \ldots, X_n$ are independent and identically distributed with a Bernoulli distribution, S follows a binomial distribution with parameters n and p. Then

$$E(\varphi(S)) = \sum_{s=0}^{n} \varphi(s) C_n^s p^s (1-p)^{n-s} = (1-p)^n \sum_{s=0}^{n} \varphi(s) C_n^s \left(\frac{p}{1-p}\right)^s.$$

If $E(\varphi(S)) = 0$, for all $p \in (0, 1)$, then

$$\sum_{s=0}^{n} \varphi(s) C_n^s \left(\frac{p}{1-p}\right)^s = 0,$$

for all $p \in (0, 1)$, which is possible only if $\varphi(s) = 0$, for all $s = \overline{0, n}$, i.e. $\varphi \overset{a.s.}{=} 0$. Thus S is also complete.

Now let us consider the estimator $\hat{p} = \overline{X} = \dfrac{1}{n} S$. Since $S \in B(n, p)$, we know that $E(S) = np$ and, hence $E(\hat{p}) = p$, so $\hat{p}$ is an unbiased estimator for p. Then by Theorem 9.1.15, the MVUE is given by

$$\overline{p} = E(\hat{p}|S) = E\left(\frac{1}{n}S|S\right) = E\left(\frac{1}{n}S\right) = \frac{1}{n}S = \overline{X}.$$

■

9.1.3 Efficient Estimators

Definition 9.1.17. *For a sample of size n, the* ***Fisher information*** *relative to θ, is the quantity*

$$I_n(\theta) = E\left[\left(\frac{\partial \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta}\right)^2\right], \tag{9.6}$$

if the likelihood function L is differentiable with respect to θ.

Remark 9.1.18. The Fisher information is a way of measuring the amount of information that a random sample $X_1, \ldots, X_n$ carries about an unknown parameter θ, upon which the likelihood function depends.

An easier computational formula than (9.6) is given below.

Proposition 9.1.19. *If the range of X does not depend on θ and the likelihood function L is twice differentiable with respect to θ, then*

$$I_n(\theta) = -E\left[\frac{\partial^2 \ln L(X_1, \ldots, X_n; \theta)}{\partial \theta^2}\right], \tag{9.7}$$

Corollary 9.1.20. *If the range of X does not depend on θ, then*

$$I_n(\theta) = nI_1(\theta). \tag{9.8}$$

Proof. Since the sample is repeated,

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

Then

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \ln f(X_i; \theta), \\ \frac{\partial^2 \ln L}{\partial \theta^2} &= \sum_{i=1}^{n} \frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2} . \end{aligned}$$

By Proposition 9.1.19,

$$I_n(\theta) = -\sum_{i=1}^{n} E\left[\frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2}\right] = \sum_{i=1}^{n} I_1(\theta) = nI_1(\theta).$$

□

Definition 9.1.21. *An estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is called an* ***absolutely correct*** *estimator for θ, if it satisfies the conditions*

(i) $E(\hat{\theta}) = \theta$,

(ii) $\lim_{n\to\infty} V(\hat{\theta}) = 0$.

Proposition 9.1.22. *An absolutely correct estimator is consistent.*

Proof. Let $\hat{\theta}$ be an absolutely correct estimator. By Chebyshev's inequality (4.19), for every $\varepsilon > 0$,

$$P(|\hat{\theta} - E(\hat{\theta})| > \varepsilon) \leq \frac{V(\hat{\theta})}{\varepsilon^2}.$$

Since $\hat{\theta}$ is unbiased, $E(\hat{\theta}) = \theta$, so we have

$$0 \leq P(|\hat{\theta} - \theta| > \varepsilon) \leq \frac{V(\hat{\theta})}{\varepsilon^2}.$$

Let $n \to \infty$ to get

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

Thus $\hat{\theta}$ is a consistent estimator. □

From Proposition 8.1.9 we have now the following:

Proposition 9.1.23. *Let X be a characteristic with the property that for some $k \in \mathbb{N}$, the theoretical moment $\nu_{2k} = E\left(X^{2k}\right)$ exists. Then the sample moment of order k, $\overline{\nu}_k = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^k$ is an absolutely correct estimator for ν_k.*

Remark 9.1.24. In particular, the sample mean $\overline{X}$ is an absolutely correct estimator for the theoretical mean $\mu = E(X)$.

In our quest for finding unbiased estimators with small variance, a natural question arises: how small can that variance be? A lower bound for the variance of an estimator, under certain regularity conditions, is given below.

Theorem 9.1.25 (Cramér-Rao Inequality). *Let X be a characteristic whose probability function $f(x;\theta)$ is differentiable with respect to θ, for $\theta \in (a,b)$ and let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an absolutely correct estimator for θ. Then*

$$V(\hat{\theta}) \geq \frac{1}{I_n(\theta)}. \tag{9.9}$$

Definition 9.1.26. *Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an absolutely correct estimator for θ. The* ***efficiency*** *of $\hat{\theta}$ is the quantity*

$$e(\hat{\theta}) = \frac{I_n^{-1}(\theta)}{V(\hat{\theta})} = \frac{1}{I_n(\theta)V(\hat{\theta})}. \tag{9.10}$$

The estimator $\hat{\theta}$ is said to be ***efficient*** *for θ, if $e(\hat{\theta}) = 1$.*

Remark 9.1.27.
1. So, by Theorem 9.1.25, the efficiency $e(\hat{\theta})$ is the minimum possible variance for an unbiased estimator $\hat{\theta}$ divided by its actual variance. Its value is always $e(\hat{\theta}) \le 1$.
2. An efficient estimator may not exist, but if it does, it is also the MVUE. This is because an efficient estimator maintains equality on the Cramér-Rao inequality for all parameter values, which means it attains the minimum variance for all parameters. The MVUE, even if it exists, is not necessarily efficient.

Example 9.1.28. Let X be a characteristic with probability density function

$$f(x;\theta) = \frac{1}{\theta^2} x e^{-\dfrac{x}{\theta}},$$

for $x > 0$ and 0, otherwise, where $\theta > 0$ is unknown. For a random sample $X_1, \ldots, X_n$, consider the estimator $\hat{\theta} = \frac{1}{2}\overline{X}$. Show that it is absolutely correct and find its efficiency.

Solution 9.1.16: First, let us check that $f(x;\theta)$ is indeed a density function.

$$\int\limits_{\mathbb{R}} f(x;\theta)\, dx = \frac{1}{\theta^2} \int\limits_0^\infty x e^{-\dfrac{x}{\theta}}\, dx,$$

which, with the change of variables $u = \dfrac{x}{\theta}$, is equal to

$$\int_0^\infty u e^{-u}\,du = \Gamma(2) = 1$$

(see Appendix A.1 for Euler's Gamma function, Γ).
With the same change of variables, we compute

$$\begin{aligned}
E(X) &= \frac{1}{\theta^2}\int_0^\infty x^2 e^{-\dfrac{x}{\theta}}\,dx = \theta\int_0^\infty u^2 e^{-u}\,du = \theta\,\Gamma(3) = 2\theta,\\
E(X^2) &= \frac{1}{\theta^2}\int_0^\infty x^3 e^{-\dfrac{x}{\theta}}\,dx = \theta^2\int_0^\infty u^3 e^{-u}\,du = \theta^2\,\Gamma(4) = 6\theta^2,\\
V(X) &= E(X^2) - (E(X))^2 = 6\theta^2 - 4\theta^2 = 2\theta^2.
\end{aligned}$$

Then for $\hat{\theta}$ we have

$$E(\hat{\theta}) = \frac{1}{2}E(\overline{X}) = \frac{1}{2}E(X) = \theta,$$

which means $\hat{\theta}$ is unbiased and

$$V(\hat{\theta}) = \frac{1}{4}V(\overline{X}) = \frac{1}{4}\frac{V(X)}{n} = \frac{\theta^2}{2n} \to 0, \text{ as } n \to \infty,$$

so $\hat{\theta}$ is absolutely correct.
To compute the Fisher information, since the range of X does not depend on θ, we use (9.8). We have

$$L(X_1;\theta) = L(X;\theta) = \frac{1}{\theta^2}X e^{-\dfrac{1}{\theta}X}, \;\; \ln L = -2\ln\theta + \ln X - \frac{1}{\theta}X,$$

so
$$\frac{\partial \ln L}{\partial \theta} = -\frac{2}{\theta} + \frac{1}{\theta^2}X, \quad \frac{\partial^2 \ln L}{\partial \theta^2} = \frac{2}{\theta^2} - \frac{2}{\theta^3}X.$$

Then
$$I_1(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) = -\frac{2}{\theta^2} + \frac{2}{\theta^3}E(X) = -\frac{2}{\theta^2} + \frac{4}{\theta^2} = \frac{2}{\theta^2}.$$

Thus
$$I_n(\theta) = \frac{2n}{\theta^2} \text{ and } e(\hat{\theta}) = 1,$$

so $\hat{\theta} = \frac{1}{2}\overline{X}$ is an efficient estimator and, by Remark 9.1.27, also the MVUE for θ.

■

9.1.4 The Method of Moments and The Method of Maximum Likelihood

We present two of the most popular methods of finding point estimators.

Method of Moments

This is one of the oldest and easiest methods for obtaining point estimators, first formalized by K. Pearson in the late 1800's. It is based on the assumption that the sample moments, $\overline{\nu}_k$, should provide good estimates for the corresponding population moments, ν_k, an assumption supported theoretically by the fact that the former are absolutely correct estimators for the latter (Proposition 9.1.23). So it simply involves equating the two and solving the resulting system

$$\nu_k = \overline{\nu}_k, \tag{9.11}$$

for the unknown parameters. The values of k in (9.11) start at 1 and go as far up as needed, depending on the number of parameters to be estimated.

Example 9.1.29. Let X be a characteristic following the binomial model with parameters $m \in \mathbb{N}$ and $p \in (0, 1)$, both unknown. Based on a random sample $X_1, \dots, X_n$, find the method of moments estimators for m and p.

Solution 9.1.29: We have two parameters to estimate, so we will consider two equations in (9.11), for $k = 1, 2$. Recall that for $X \in B(m, p)$ (see [11]),

$$\begin{aligned} \nu_1 &= E(X) &&= mp \\ \nu_2 &= E(X^2) &&= mp(mp - p + 1). \end{aligned}$$

Then system (9.11) becomes

$$\begin{cases} mp &= \overline{\nu}_1 \\ mp(mp - p + 1) &= \overline{\nu}_2, \end{cases}$$

with solution

$$\begin{aligned} \hat{m} &= \frac{\overline{X}^2}{\overline{X}^2 + \overline{X} - \overline{\nu}_2} \\ \hat{p} &= \frac{\overline{X}^2 + \overline{X} - \overline{\nu}_2}{\overline{X}}. \end{aligned}$$

■

Method of Maximum Likelihood

Maximum-likelihood estimation was first recommended, analyzed and then vastly popularized by R. A. Fisher in the 1920's, although it had been used earlier by Gauss and Laplace. For a fixed random sample from an underlying probability distribution, the maximum likelihood method picks the values of the population parameters that make the data "more likely" than any other values of the parameters would make them.

Let us illustrate it, first, with a simple example. Suppose there are 5 balls in a box, black or white, the number of each being unknown. Suppose

further, that we randomly select 3 of them, without replacement, and we get all three white. What would be a good estimate, $\hat{w}$, for the number of white balls in the box, w? Obviously, $w \in \{3, 4, 5\}$.
If the true value was $w = 3$, then the probability of us randomly selecting 3 white balls would be (see the hypergeometric model (2.3))

$$p = \frac{C_3^3 C_2^0}{C_5^3} = \frac{1}{10}.$$

If the true value was $w = 4$, then the probability of us randomly selecting 3 white balls would be

$$p = \frac{C_4^3 C_1^0}{C_5^3} = \frac{4}{10}.$$

And, finally, if the true value was $w = 5$, then the probability of us randomly selecting 3 white balls would be

$$p = \frac{C_5^3 C_0^0}{C_5^3} = 1.$$

So, it would seem reasonable to choose $\hat{w} = 5$ as our estimate for w, since this would *maximize* the probability of obtaining our observed sample.
This, in essence, describes the method of maximum likelihood estimation. Since the probability of obtaining an observed sample is measured by the likelihood function of a sample (9.2), this method chooses the values of an estimator $\hat{\Theta} = (\theta_1, \ldots, \theta_l) = \hat{\Theta}(X_1, \ldots, X_n)$ that maximize the function $L(X_1, \ldots, X_n; \Theta)$. i.e. if L is twice differentiable with respect to Θ, the solutions of the maximum-likelihood system

$$\frac{\partial L(X_1, \ldots, X_n; \Theta)}{\partial \theta_j} = 0, \;\; j = \overline{1, l}, \tag{9.12}$$

or, equivalently, but easier to compute, the maximum-likelihood equations

$$\frac{\partial \ln L(X_1, \ldots, X_n; \Theta)}{\partial \theta_j} = 0, \;\; j = \overline{1, l}. \tag{9.13}$$

If the system (9.13) has a solution, then it is unique ant it is called the **maximum likelihood (MLE)** estimator.

Example 9.1.30. Consider Example 9.1.28. Find the method of moments estimator and the MLE for θ.

Solution 9.1.30: The method of moments estimator will be the solution of the equation

$$2\theta = \overline{X},$$

i.e. $\hat{\theta} = \dfrac{1}{2}\overline{X}$.

The likelihood function is given by

$$\begin{aligned} L(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^{n} \frac{1}{\theta^2} x_i e^{-\dfrac{x_i}{\theta}} \\ &= \left(\prod_{i=1}^{n} x_i\right) \frac{1}{\theta^{2n}} e^{-\dfrac{1}{\theta}\sum\limits_{i=1}^{n} x_i} \\ &= K \frac{1}{\theta^{2n}} e^{-\dfrac{n\overline{x}}{\theta}}, \end{aligned}$$

where $K = \prod\limits_{i=1}^{n} x_i$ is a constant with respect to θ.
Then system (9.13) becomes

$$-\frac{2n}{\theta} + \frac{n\overline{X}}{\theta^2} = 0,$$

whose solution is the MLE

$$\overline{\theta} = \frac{1}{2}\overline{X},$$

the same as the method of moments estimator $\hat{\theta}$. ∎

Remark 9.1.31. In our example, the two methods yielded the same point estimator. That is not always the case. If they differ, the natural question is: which one is better? In some respects, when estimating parameters of a known family of probability distributions, the method of moments is superseded by Fisher's method of maximum likelihood, because maximum likelihood estimators have higher probability of being close to the quantities to be estimated. However, in some cases, the likelihood equations (9.13) may be intractable without computers, whereas the method of moments estimators can be quickly and easily calculated by hand as shown above. Estimates by the method of moments may be used as the first approximation to the solutions of the likelihood equations (9.13) , and successive improved approximations may then be found by the Newton-Raphson method. In this way, the method of moments and the method of maximum likelihood are symbiotic. In some cases, infrequent with large samples but not so infrequent with small samples, the estimates given by the method of moments are outside of the parameter space and it does not make sense to rely on them then. That problem never arises in the method of maximum likelihood. Also, estimates by the method of moments are not necessarily sufficient statistics, i.e., they sometimes fail to take into account all relevant information in the sample.

9.2 Confidence Interval Estimation

So far, point estimators provided one single value, $\hat{\theta}$, to estimate the value of an unknown parameter θ, but no measure of the accuracy of the estimate. In contrast, an **interval estimator** specifies instead a range within which the parameter is estimated to lie. More specifically, the sample will be used to produce two sample functions, $\hat{\theta}_L(X_1, \ldots, X_n) < \hat{\theta}_U(X_1, \ldots, X_n)$, with values $\hat{\theta}_L = \hat{\theta}_L(x_1, \ldots, x_n)$, $\hat{\theta}_U = \hat{\theta}_U(x_1, \ldots, x_n)$, respectively, such that for a given $\alpha \in (0, 1)$,

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha. \tag{9.14}$$

Then the range $(\hat{\theta}_L, \hat{\theta}_U)$ is called **confidence interval (CI)**, more specifically, a $100(1-\alpha)\%$ confidence interval, the values $\hat{\theta}_L, \hat{\theta}_U$ are called (lower and upper) **confidence limits**, $1 - \alpha$ is called **confidence level** or **confidence coefficient** and α is called **significance level** .

Remark 9.2.1.
1. It may seem a little peculiar that we use $1 - \alpha$ instead of simply α in (9.14), since both values are in $(0, 1)$, but the reasons are in close connection with hypothesis testing and will be revealed in the next chapter.
2. The condition (9.14) does not uniquely determine a $100(1 - \alpha)\%$ CI.
3. Evidently, the smaller α and the length of the interval $\hat{\theta}_U - \hat{\theta}_L$ are, the better the estimate for θ.

To produce a CI estimate for θ, we need a *pivotal quantity*, i.e. a statistic S that satisfies two conditions:

(i) $S = S(X_1, \dots, X_n; \theta)$ is a function of the sample measurements and the unknown parameter θ, *only*,

(ii) the distribution of S is known and does not depend on θ.

9.2.1 Large Sample Confidence Intervals

If the target parameter θ is one of the four in Table 9.1, then for large samples ($n \geq 30$, $n_1 + n_2 > 40$), the statistic

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \tag{9.15}$$

has an approximately standard normal distribution (see Remark 9.1.3) and, consequently, can be used as a pivotal quantity to construct a CI for estimating θ.
We will choose two values, Z_L, Z_U such that for a given $\alpha \in (0, 1)$,

$$P(Z_L < Z < Z_U) = 1 - \alpha.$$

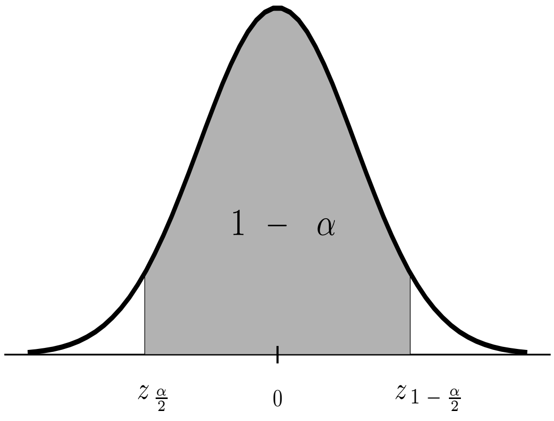


Fig. 9.1: Confidence Interval

Since $Z \in N(0,1)$, the relation above can be interpreted as: the area under the standard normal density function, between the values Z_L and Z_U is equal to $1-\alpha$. We mentioned earlier that the interval is not unique. We will choose the two values such that the area $1-\alpha$ is *in the middle*, i.e. (since the total area under the graph is 1) the two portions left on the two sides, both have an area of $\dfrac{\alpha}{2}$, as seen in Figure 9.1. That means we want

$$\begin{aligned} P(Z < Z_L) &= \frac{\alpha}{2} \\ P(Z < Z_U) &= 1 - \frac{\alpha}{2}, \end{aligned}$$

so the two values will be the quantiles $Z_L = z_{\frac{\alpha}{2}}$ and $Z_U = z_{1-\frac{\alpha}{2}}$ of order $\dfrac{\alpha}{2}$ and $1-\dfrac{\alpha}{2}$, respectively, for the standard normal distribution (see Definition 4.3.7 and Remark 4.3.8(2)).

So, now we have

$$P\left(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

i.e.

$$P\left(z_{\frac{\alpha}{2}} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

and by rewriting the double inequality inside,

$$P\left(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}} < \theta < \hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right) = 1 - \alpha,$$

so the $100(1-\alpha)\%$ CI for θ is given by

$$\left(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}},\ \hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right). \tag{9.16}$$

Remark 9.2.2. 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

$$\left(\hat{\theta} - z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}},\ \hat{\theta} + z_{1-\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right) \quad \text{or} \quad \left(\hat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}},\ \hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right).$$

Remark 9.2.3. 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 $(-\infty, \hat{\theta}_U)$ and $(\hat{\theta}_L, \infty)$, 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. Then, we can find a $100(1-\alpha)\%$ *upper confidence interval* for θ,

$$\left(-\infty, \hat{\theta} - z_{\alpha}\sigma_{\hat{\theta}}\right)$$

and a $100(1-\alpha)\%$ *lower confidence interval* for θ,

$$\left(\hat{\theta} - z_{1-\alpha}\sigma_{\hat{\theta}}, \infty\right).$$

9.2.2 Confidence Intervals for the Mean

Assume the characteristic X has mean μ and variance σ^2 and $X_1, \ldots, X_n$ is a random sample.

CI for the Mean, Known Variance

If either X is approximately normally distributed or $n > 30$, and σ is known, then the statistic

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution, by Corollary 8.1.6 and can be used to derive as above, the $100(1-\alpha)\%$ CI

$$\left(\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right). \tag{9.17}$$

CI for the Mean, Unknown Variance

In practice, it is unreasonable to expect to know the value of σ, if the value of μ is unknown. If X is approximately normally distributed, then the statistic

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

follows the Student $T(n-1)$ distribution, by Proposition 8.2.3. Then we find the $100(1-\alpha)\%$ CI for the mean to be

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}},\ \overline{X} + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right), \tag{9.18}$$

where the quantile $t_{1-\frac{\alpha}{2}}$ refers to the $T(n-1)$ distribution.

Example 9.2.4. 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 9.2.4: For our sample, $n = 64$, $\overline{x} = 33$ and $s^2 = 256$. For $\alpha = 0.05$,

$$t_{1-\frac{\alpha}{2}} = t_{0.975} = 1.9983,$$

so the limits of the 95% CI for the population mean μ are, by (9.18)

$$\begin{aligned} \overline{x} - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= 15 - 1.9983 \cdot \frac{16}{8} &= 11.0034, \\ \overline{x} + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= 15 + 1.9983 \cdot \frac{16}{8} &= 18.9966. \end{aligned}$$

So

$$\mu \in (29.0034,\ 36.9966)\,,$$

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.

■

9.2.3 Confidence Intervals for the Difference of Two Means

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, approximately normally distributed with $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ distributions, respectively. We draw two random samples

$$X_{11}, \ldots, X_{1n_1} \text{ and } X_{21}, \ldots, X_{2n_2},$$

with sample means

$$\overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, \quad \overline{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

and sample variances

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1 \right)^2, \; s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(X_{2j} - \overline{X}_2 \right)^2.$$

CI for the Difference of Means, Known Variances

Assume μ_1 and μ_2 are unknown, while σ_1, σ_2 are both known. We want a CI for the difference of the means $\mu_1 - \mu_2$. By Proposition 8.2.4, the statistic

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is standard normally distributed. Then the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left(\overline{X}_1 - \overline{X}_2 - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \; \overline{X}_1 - \overline{X}_2 + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right). \quad (9.19)$$

CI for the Difference of Means, Unknown, but Equal Variances

If the variances are not known, but known to be equal, then we consider the pooled variance of the two samples, introduced in Chapter 8,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Then by Proposition 8.2.5, the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

follows a $T(n_1 + n_2 - 2)$ distribution. We use it to find the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ to be

$$\left(\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{X}_1 - \overline{X}_2 + t_{1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right), \tag{9.20}$$

where the quantile $t_{1-\frac{\alpha}{2}}$ refers to the $T(n_1 + n_2 - 2)$ distribution.

CI for the Difference of Means, Unknown Variances

If no information is known on the two population variances, then we use the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

which, by Proposition 8.2.4, follows a $T(n)$ distribution, where

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \quad \text{and} \quad c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{9.21}$$

Then the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left(\overline{X}_1 - \overline{X}_2 - t_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 + t_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right). \tag{9.22}$$

where the quantile $t_{1-\frac{\alpha}{2}}$ refers to the $T(n)$ distribution, with n given by (9.21).

CI for the Difference of Means, Paired Data

In many applications, we want to compare the means of two populations, when two random samples are available, but they are *not* independent, but rather, each observation in one sample is naturally or by design *paired* with an observation in the other. For instance, we want to compare the average values of the same measurement made under two different conditions, like the response of the same group of patients to two different treatments, or their health status ”before” and “after” some treatment. When such pairing occurs, the methods described above for finding a CI for the difference of population means, no longer apply. For one thing, both samples would have the same length, n. Then it is a matter of finding a CI for the mean of *one* population, which is the *difference* of the corresponding values. So, for the two paired random samples

$$X_{11}, \dots, X_{1n} \quad \text{and} \quad X_{21}, \dots, X_{2n},$$

we consider the sample

$$D_1, \dots, D_n,$$

where $D_i = X_{1i} - X_{2i}$, $i = \overline{1, n}$. Then the sample mean and sample variance are computed by

$$\overline{X}_d = \frac{1}{n} \sum_{i=1}^{n} D_i \quad \text{and} \quad s_d^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(D_i - \overline{X}_d\right)^2 .$$

Since the statistic

$$T = \frac{\overline{X}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}}$$

follows the Student $T(n-1)$ distribution, the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left(\overline{X}_d - t_{1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}},\ \overline{X}_d + t_{1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}\right), \tag{9.23}$$

where the quantile $t_{1-\frac{\alpha}{2}}$ refers to the $T(n-1)$ distribution.

Example 9.2.5. Employees in a manufacturing plant are trained to do a certain assembling operation. A new method of training is in a trial period and the management wants to compare the new method with the standard procedure. Two groups were observed, one using the standard method, the other the new method, their assembling times (in minutes) were recorded and the results displayed in the table below. Assuming the assembling times are approximately normally distributed and that the variances using the two methods are equal, find a 99% confidence interval for the average difference in assembling time using the two methods.

St. Method	New Method
$n_1 = 12$	$n_2 = 15$
$\overline{x}_1 = 35.3$	$\overline{x}_2 = 31.2$
$s_1^2 = 17.7$	$s_2^2 = 17.5$

Solution 9.2.5: Since the two samples are independent and the population variances are assumed equal, we use formula (9.20). For $\alpha = 0.01$, with $12 + 15 - 2 = 25$ degrees of freedom,

$$t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.7874.$$

The pooled variance is

$$\begin{aligned} s_p^2 &= \frac{11 \cdot 17.7 + 14 \cdot 17.5}{25} = 17.588, \\ s_p &= 4.1938. \end{aligned}$$

The limits of the 99% CI for $\mu_1 - \mu_2$ are

$$4.1 \pm 2.7874 \cdot 4.1938 \cdot \sqrt{\frac{1}{12} + \frac{1}{15}},$$

so $\mu_1 - \mu_2 \in (-0.4274,\ 8.6274)$ with probability 0.99. ■

9.2.4 Confidence Intervals for the Variance and the Ratio of Two Variances

Confidence Intervals for the Variance

Assume the characteristic X is approximately normally distributed, with mean μ and variance σ^2 , with σ unknown and $X_1, \ldots, X_n$ is a random sample. We know by Proposition 8.2.2 that the statistic

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{(n-1)\,s^2}{\sigma^2}$$

follows a $\chi^2(n-1)$ distribution. We use it to find the $100(1-\alpha)\%$ CI for σ^2

$$\left(\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}},\ \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}} \right), \tag{9.24}$$

where the quantiles $\chi^2_{\frac{\alpha}{2}}, \chi^2_{1-\frac{\alpha}{2}}$ refer to the $\chi^2(n-1)$ distribution. Notice that in this case, we used both quantiles, since the χ^2 distribution is no longer symmetric.

Remark 9.2.6. A CI for the standard deviation σ can be found from (9.23), by simply taking the square root of the two limits.

Example 9.2.7. A media firm wants to check the variability of equipment designed to measure the volume of an audio source. Thirty independent measurements recorded by this equipment for the same sound yielded a variance of 10.57. Assuming the volume values recorded by the equipment for the audio source are approximately normally distributed, find a 90% CI for the standard deviation of all such volume values.

Solution 9.2.7: We have $n = 30$, $s^2 = 10.57$, $\alpha = 0.1$ and the quantiles

$$\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.05} = 17.7084,\ \ \chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.95} = 42.557,$$

with 29 degrees of freedom. Then the limits of the 90% CI for σ^2 are

$$\begin{aligned}\frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2}}} &= \frac{29\cdot 10.57}{42.557} &= 7.2028,\\ \frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2}}} &= \frac{29\cdot 10.57}{17.7084} &= 17.3099.\end{aligned}$$

So $\sigma^2 \in (7.2028,\ 17.3099)$ and $\sigma \in (2.6838,\ 4.1605)$, with probability 0.9.

■

Confidence Intervals for the Ratio of Variances

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, approximately normally distributed with $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ distributions, respectively, and $\sigma_1,\ \sigma_2$ are not known. We draw two random samples

$$X_{11}, \ldots, X_{1n_1} \text{ and } X_{21}, \ldots, X_{2n_2},$$

having sample means $\overline{X}_1$, $\overline{X}_2$ and sample variances s_1^2, s_2^2, respectively. We can find a CI for the ratio of population variances $\frac{\sigma_1^2}{\sigma_2^2}$, using the statistic

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2},$$

following a $F \in F(n_1-1, n_2-1)$ distribution, by Proposition 8.2.6. Then the $100(1-\alpha)\%$ CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\left(\frac{1}{f_{1-\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2},\ \frac{1}{f_{\frac{\alpha}{2}}}\cdot\frac{s_1^2}{s_2^2}\right), \tag{9.25}$$

where the quantiles $f_{\frac{\alpha}{2}}, f_{1-\frac{\alpha}{2}}$ refer to the $F(n_1-1, n_2-1)$ distribution. As before, (9.24) can be used to find a CI for the ratio of standard deviations.

Chapter 10

Hypothesis Testing

In the last chapter 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.

10.1 Introduction, Basic Concepts

As such, we will work with **statistical hypotheses**, i.e. assumptions about some characteristic X (relative to a population), whose density $f(x; \theta)$ depends on the parameter θ, which is to be estimated. As before, if X is discrete, then f represents the probability distribution function, while if X is continuous, f is the probability density function.
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. Any hypothesis test will involve two theories, two hypotheses, one being proposed by the researcher, called the **alternative (research)**, denoted by H_1 (or H_a) and the **null hypothesis**, denoted by H_0, the negation (in some way) of the research hypothesis. To determine the truth value of a hypothesis, we use a sample function called the **test statistic (TT)**. The set of values of the test statistic for which we 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 **non-parametric** 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

$$\begin{array}{ll} H_0: & \theta \in A_0 \\ H_1: & \theta \in A_1. \end{array}$$

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

$$\begin{array}{lll} H_1: & \theta < \theta_0 & \text{(left-tailed test),} \\ H_1: & \theta > \theta_0 & \text{(right-tailed test),} \\ H_1: & \theta \neq \theta_0 & \text{(two-tailed test).} \end{array}$$

Remark 10.1.1. 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. Assume that a supermarket receives complaints about a certain brand of 1 liter bottles of juice. Customers claim there is less than 1 liter in the bottles. Random bottles are selected and their weights recorded. To test if the customers are right, the appropriate hypotheses on the average weight would be

$$\begin{aligned} H_0 :& \quad \mu = 1 \\ H_1 :& \quad \mu < 1, \end{aligned}$$

since, in this case, an average greater than or equal to 1 fall under the same category, the bottles are "good", while an average of less than 1 would mean there is a problem with the bottles.

Designing a hypothesis test means constructing the rejection region RR, or equivalently, a region $U \in \mathbb{R}^n$, such that for a given $\alpha \in (0,1)$, the conditional probability, conditioned by H_0 being true,

$$P(TT \in RR \mid H_0) = P((X_1, \ldots, X_n) \in U \mid H_0) = \alpha, \tag{10.1}$$

where $X_1, \ldots, X_n$ are sample variables. The value α is called **significance level** or **risk probability**.

For any given hypothesis testing problem, we have the following possibilities:

Actual situation Decision	H_0 true	H_1 true
Reject H_0	Type I error (prob α)	Right decision
Do not reject H_0	Right decision	Type II error (prob β)

In two of the cases, we make the right decision, in the other two, we make an error. An error of type I occurs when we reject a true null hypothesis and by (10.1), the probability of making such an error is the significance level

$$P(\text{ type I error}) = P(\text{ reject } H_0 \,|\, H_0) = P(TT \in RR \,|\, H_0) = \alpha, \quad (10.2)$$

while an error of type II happens when we fail to reject a false null hypothesis, and its probability is denoted by β,

$$P(\text{ type II error}) = P(\text{ accept } H_0 \,|\, H_1) = P(TT \notin RR \,|\, H_1) = \beta. \quad (10.3)$$

Remark 10.1.2.

1. The rejection region and hence, the hypothesis test, are not uniquely determined by (10.1), 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.

10.2 Common Large Sample Tests (Z-Tests)

We want to test a set of hypotheses concerning a parameter θ, based on a sample $X_1, \dots, X_n$ (or two samples $X_{11}, \dots, X_{1n_1}$, $X_{21}, \dots, X_{2n_2}$, if the parameter θ involves comparing two populations). We design a hypothesis testing procedure based on an estimator $\hat{\theta}$ that has an approximately normal distribution with mean θ and standard error $\sigma_{\hat{\theta}}$. The large sample estimators given in Table 9.1 satisfy these requirements.
So for a given level of significance $\alpha \in (0, 1)$, consider the hypotheses

$$H_0 : \quad \theta = \theta_0,$$

with one of the alternatives

$$H_1 : \quad \begin{cases} \theta < \theta_0 \\ \theta > \theta_0 \\ \theta \neq \theta_0. \end{cases} \tag{10.4}$$

We will use the test statistic

$$Z_0 = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}, \tag{10.5}$$

Recall that for n large ($n > 30$ or $n_1 + n_2 > 40$),

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

follows an $N(0, 1)$ distribution. So, if the true value is $\theta = \theta_0$, then $Z_0 \in N(0, 1)$.

Let us start with the left-tailed case. We need to determine RR such that (10.1) 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 $RR = \{z \mid z < k_1\}$. We have

$$\begin{aligned} \alpha &= P(z_0 \in RR \mid H_0) \\ &= P(z_0 < k_1 \mid \theta = \theta_0) \\ &= P(z_0 < k_1 \mid Z_0 \in N(0,1)). \end{aligned}$$

Now, we know that if $Z_0 \in N(0,1)$, $P(z_0 < z_\alpha) = \alpha$, where z_α is the quantile of order α for the $N(0,1)$ distribution. Thus, we choose $k_1 = z_\alpha$ and

$$RR_{\text{left}} = \{z_0 < z_\alpha\}. \tag{10.6}$$

Similarly, for a right-tailed test, we want to find a rejection region of the form $RR = \{z \mid z > k_2\}$, so that

$$\begin{aligned} \alpha &= P(z_0 \in RR \mid H_0) \\ &= P(z_0 > k_2 \mid \theta = \theta_0) \\ &= P(z_0 > k_2 \mid Z_0 \in N(0,1)). \end{aligned}$$

Since $P(z_0 < z_{1-\alpha}) = 1 - \alpha$, then $P(z_0 > z_{1-\alpha}) = \alpha$ and so we choose $k_2 = z_{1-\alpha}$, the quantile of order $1 - \alpha$ for the $N(0,1)$ distribution and

$$RR_{\text{right}} = \{z_0 > z_{1-\alpha}\}. \tag{10.7}$$

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 $RR = \{z \mid z < k_1 \text{ or } z > k_2\}$. The rejection region should be chosen such that

$$P(z_0 < k_1 \text{ or } z_0 > k_2 \mid \theta = \theta_0) = \alpha,$$

or, equivalently,

$$P(k_1 < z_0 < k_2 \mid Z_0 \in N(0,1)) = 1 - \alpha.$$

We encountered such problems before in Chapter 9, 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

$$RR_{\text{two}} = \{z_0 < z_{\frac{\alpha}{2}} \text{ or } z_0 > z_{1-\frac{\alpha}{2}}\}, \tag{10.8}$$

or, since the distribution of Z is symmetric and $z_{1-\frac{\alpha}{2}} > 0$,

$$\begin{aligned} RR_{\text{two}} &= \{z_0 < -z_{1-\frac{\alpha}{2}} \text{ or } z_0 > z_{1-\frac{\alpha}{2}}\} \\ &= \{z_0 > |z_{1-\frac{\alpha}{2}}|\}. \end{aligned}$$

To summarize, the rejection regions for the three alternatives (10.4) are given by

$$RR : \begin{cases} \{z_0 < z_\alpha\} \\ \{z_0 > z_{1-\alpha}\} \\ \{z_0 < z_{\frac{\alpha}{2}} \text{ or } z_0 > z_{1-\frac{\alpha}{2}}\} = \{z_0 > |z_{1-\frac{\alpha}{2}}|\}. \end{cases} \tag{10.9}$$

Remark 10.2.1.
1. Since for large samples a variable $Z \in N(0,1)$ is 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.

Example 10.2.2. The number of sales per month at a large car dealership is known to have a mean of 20 and a standard deviation of 3.2 and all salary, tax and bonus figures are based on those 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 19.15 sales. At the 5% significance level, does the data confirm the manager's suspicion?

Solution 10.2.2: Since the sample size $n > 30$, 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 mean is less than 20, which is supposed to be, so the two hypotheses 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, $\sigma = 3.2$ and the sample mean is $\overline{x} = 19.15$.
The value of the test statistic is

$$z_0 = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{19.15 - 20}{\frac{3.2}{6}} = -1.5938.$$

The rejection region is, by (10.6),

$$RR = \{z \mid z < z_\alpha\} = (-\infty, -1.645).$$

Since $z_0 \notin RR$, we do not reject H_0, i.e. at the 5% significance level, the data *does not* confirm the manager's suspicion.

■

10.3 Significance Testing, P-Values

There is a problem that might occur in hypothesis testing: we may find that having preset α, the probability of a type I error, and henceforth having determined 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, yet the probability of getting that value of the test statistic under the assumption that H_0 is true, is still very small, comparable with our preset α. 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 we should reject the null hypothesis H_0.

To avoid this situation, we perform a **significance test**: for a given random sample $(X_1, \ldots, X_n)$, we still set up H_0 and H_1 as before, we determine a test statistic and then we compute the probability of observing a value *at least as extreme* (in the sense of the test conducted) of the test statistic as the value observed from the sample, under the assumption that H_0 is true. This probability is called the ***P*-value** of the test. If it is small, we reject H_0, otherwise we do not reject it. 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, i.e. the smallest significance level of rejection. 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 (10.4), if TT_0 is the value of the test statistic TT under the assumption that H_0 is true and F is the cumulative distribution function of TT, the P-value is given by

$$P = \begin{cases} P(TT < TT_0 \mid H_0) & = & F(TT_0) \\ P(TT > TT_0 \mid H_0) & = & 1 - F(TT_0) \\ P(TT > |TT_0| \mid H_0) & = & 2(1 - F(|TT_0|)). \end{cases} \tag{10.10}$$

Then

$$\begin{array}{l} \text{if } P \le \alpha, \text{ reject } H_0, \\ \text{if } P > \alpha, \text{ do not reject } H_0. \end{array} \tag{10.11}$$

Example 10.3.1. For the problem in Example 10.2.2, the P-value of the test is

$$P = P(Z < z_0) = P(Z < -1.5938) = 0.0555.$$

Since $\alpha = 0.05 < 0.0555 = P$, we do not reject H_0.

Remark 10.3.2. Whether we perform a hypothesis or a significance test, the conclusion we arrive at, the decision of rejecting or not the null hypothesis, is the same, for a given significance level α. However, significance testing is preferable to hypothesis testing, especially from the computer implementation point of view, since it avoids the inversion of a cumulative distribution function, which is, oftenly, a complicated improper integral.

10.4 Hypothesis Tests for the Mean

Assume the characteristic X has mean μ and variance σ^2 and $X_1, \dots, X_n$ are sample variables for a random sample.
Consider the hypotheses

$$H_0 : \quad \mu = \mu_0,$$

with one of the alternatives

$$H_1 : \quad \begin{cases} \mu < \mu_0 \\ \mu > \mu_0 \\ \mu \ne \mu_0. \end{cases}$$

Test for the Mean, Known Variance (Z-Tests)

If either X is approximately normally distributed or $n > 30$, and σ is known, then the statistic

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution, by Corollary 8.1.6 and then the statistic $Z_0 = Z(\mu = \mu_0)$ can be used as a test statistic as above. The rejection regions for the three alternatives are given by

$$RR: \begin{cases} \{z_0 < z_\alpha\} \\ \{z_0 > z_{1-\alpha}\} \\ \{z_0 < z_{\frac{\alpha}{2}} \text{ or } z_0 > z_{1-\frac{\alpha}{2}}\} = \{z_0 > |z_{1-\frac{\alpha}{2}}|\}. \end{cases} \tag{10.12}$$

Test for the Mean, Unknown Variance (T-Test)

If X is approximately normally distributed or n is large ($n > 30$) and the value of σ is not known, then the statistic

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

follows the Student $T(n-1)$ distribution, by Proposition 8.2.3. We use

$$T_0 = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{10.13}$$

as a test statistic. For $\alpha \in (0, 1)$, the rejection regions for the three alternatives are

$$RR: \begin{cases} \{t_0 < t_\alpha\} \\ \{t_0 > t_{1-\alpha}\} \\ \{t_0 < t_{\frac{\alpha}{2}} \text{ or } t_0 > t_{1-\frac{\alpha}{2}}\} = \{t_0 > |t_{1-\frac{\alpha}{2}}|\}, \end{cases} \tag{10.14}$$

where the quantiles t_γ refer to the $T(n-1)$ distribution.

10.5 Hypothesis Tests for Comparing Two Population Means

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, approximately normally distributed with $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ distributions, respectively and two random samples $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2n}$, with means $\overline{X}_1, \overline{X}_2$ and sample variances s_1^2, s_2^2, respectively.
We have the hypotheses

$$H_0: \ \ \mu_1 = \mu_2,$$

with one of the alternatives

$$H_1: \ \begin{cases} \mu_1 < \mu_2 \\ \mu_1 > \mu_2 \\ \mu_1 \neq \mu_2. \end{cases}$$

Test for the Difference of Means, Known Variances

If μ_1 and μ_2 are unknown, but σ_1, σ_2 are both known, then by Proposition 8.2.4, the statistic

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is standard normally distributed. Then

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{10.15}$$

is the test statistic we use. Given $\alpha \in (0,1)$, the rejection regions for the three alternatives are given by

$$RR: \ \begin{cases} \{z_0 < z_\alpha\} \\ \{z_0 > z_{1-\alpha}\} \\ \{z_0 < z_{\frac{\alpha}{2}} \text{ or } \ z_0 > z_{1-\frac{\alpha}{2}}\} = \{z_0 > |z_{1-\frac{\alpha}{2}}|\}. \end{cases} \tag{10.16}$$

Test for the Difference of Means, Unknown, but Equal Variances

If the variances are not known, but known to be equal, then we consider the pooled variance of the two samples, introduced in Chapter 8,

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

Then by Proposition 8.2.5, the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

follows a $T(n_1+n_2-2)$ distribution. Then we use the test statistic

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{10.17}$$

and for $\alpha \in (0,1)$, we find the rejection regions for the three alternatives to be

$$RR: \begin{cases} \{t_0 < t_\alpha\} \\ \{t_0 > t_{1-\alpha}\} \\ \{t_0 < t_{\frac{\alpha}{2}} \text{ or } t_0 > t_{1-\frac{\alpha}{2}}\} = \{t_0 > |t_{1-\frac{\alpha}{2}}|\}, \end{cases} \tag{10.18}$$

where the quantiles t_γ refer to the $T(n_1+n_2-2)$ distribution.

Test for the Difference of Means, Unknown Variances

If no information is known on the two population variances, then by Proposition 8.2.4, the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}},$$

follows a $T(n)$ distribution, where

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \quad \text{and} \quad c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{10.19}$$

Then the test statistic

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10.20}$$

will be used to find the rejection regions

$$RR: \begin{cases} \{t_0 < t_\alpha\} \\ \{t_0 > t_{1-\alpha}\} \\ \{t_0 < t_{\frac{\alpha}{2}} \text{ or } t_0 > t_{1-\frac{\alpha}{2}}\} = \{t_0 > |t_{1-\frac{\alpha}{2}}|\}, \end{cases} \tag{10.21}$$

for $\alpha \in (0,1)$, where the quantiles t_γ refer to the $T(n)$ distribution, with n given by (10.19).

Test for the Difference of Means, Paired Data (Paired T-Test)

If we have paired data, i.e. two samples

$$X_{11}, \dots, X_{1n} \text{ and } X_{21}, \dots, X_{2n},$$

that are not independent, then we consider the sample

$$D_1, \dots, D_n,$$

where $D_i = X_{1i} - X_{2i},\ i = \overline{1,n}$, with sample mean and sample variance

$$\overline{X}_d = \frac{1}{n}\sum_{i=1}^{n} D_i \quad \text{and} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \overline{X}_d\right)^2.$$

Since the statistic

$$T = \frac{\overline{X}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}}$$

follows the Student $T(n-1)$ distribution, we use

$$T_0 = \frac{\overline{X}_d}{\frac{s_d}{\sqrt{n}}} \tag{10.22}$$

as our test statistic and for a given $\alpha \in (0,1)$, we determine the rejection regions corresponding to the three alternatives to be

$$RR: \begin{cases} \{t_0 < t_\alpha\} \\ \{t_0 > t_{1-\alpha}\} \\ \{t_0 < t_{\frac{\alpha}{2}} \text{ or } t_0 > t_{1-\frac{\alpha}{2}}\} = \{t_0 > |t_{1-\frac{\alpha}{2}}|\}, \end{cases} \tag{10.23}$$

where the quantiles t_γ refer to the $T(n-1)$ distribution.

10.6 Hypothesis Tests Concerning Variances

Test for the Variance (χ^2-Test)

Assume the characteristic X is approximately normally distributed, with mean μ and variance σ^2 , with σ unknown. Let $X_1, \ldots, X_n$ be a random sample with sample mean $\overline{X}$ and sample variance s^2.
We have the hypotheses

$$H_0: \ \ \sigma^2 = \sigma_0^2,$$

with one of the alternatives

$$H_1: \begin{cases} \sigma^2 < \sigma_0^2 \\ \sigma^2 > \sigma_0^2 \\ \sigma^2 \neq \sigma_0^2. \end{cases}$$

By Proposition 8.2.2 the statistic

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma^2}$$

follows a $\chi^2(n-1)$ distribution. Hence we use

$$\chi_0^2 = \frac{(n-1)\, s^2}{\sigma_0^2} \tag{10.24}$$

as a test statistic. For $\alpha \in (0,1)$, the rejection regions for the three alternatives are

$$RR: \begin{cases} \{\chi_0^2 < \chi_\alpha^2\} \\ \\ \{\chi_0^2 > \chi_{1-\alpha}^2\} \\ \\ \{\chi_0^2 < \chi_{\frac{\alpha}{2}}^2 \text{ or } \chi_0^2 > \chi_{1-\frac{\alpha}{2}}^2\}, \end{cases} \tag{10.25}$$

where the quantiles χ_γ^2 refer to the $\chi^2(n-1)$ distribution.

Test for the Ratio of Variances (*F*-Test)

Assume we have two characteristics $X_{(1)}$ and $X_{(2)}$, approximately normally distributed with $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ distributions, respectively, and $\sigma_1,\ \sigma_2$ are not known. We have two random samples

$$X_{11}, \ldots, X_{1n_1} \quad \text{and} \quad X_{21}, \ldots, X_{2n_2},$$

with sample means $\overline{X}_1,\ \overline{X}_2$ and sample variances s_1^2, s_2^2, respectively.
We have the hypotheses

$$H_0: \quad \sigma_1^2 = \sigma_2^2,$$

with one of the alternatives

$$H_1: \begin{cases} \sigma_1^2 < \sigma_2^2 \\ \\ \sigma_1^2 > \sigma_2^2 \\ \\ \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

We know by Proposition 8.2.6 that the statistic

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

follows an $F(n_1 - 1, n_2 - 1)$ distribution. Then we use

$$F_0 = \frac{s_1^2/\sigma_0^2}{s_2^2/\sigma_0^2} = \frac{s_1^2}{s_2^2} \tag{10.26}$$

as a test statistic. For a significance level $\alpha \in (0,1)$, we determine the rejection regions for the three alternatives to be

$$RR: \begin{cases} \{f_0 < f_\alpha\} \\ \\ \{f_0 > f_{1-\alpha}\} \\ \\ \{f_0 < f_{\frac{\alpha}{2}} \text{ or } f_0 > f_{1-\frac{\alpha}{2}}\}, \end{cases} \tag{10.27}$$

where the quantiles f_γ refer to the $F(n_1 - 1, n_2 - 1)$ distribution.

Example 10.6.1. Suppose the strengths to a certain load of two types of material, $M1$ and $M2$, are studied, knowing that they are approximately normally distributed. 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

1. At the 5% significance level, do the variances of the two populations seem to be equal or not?
2. At the same significance level, does the data suggest that on average, $M1$ is stronger than $M2$?

Solution 10.6.1:
1. First, we compare the variances of the two populations. 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 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

$$\begin{aligned} f_{\frac{\alpha}{2}} &= f_{0.025} = 0.4103 \\ f_{1-\frac{\alpha}{2}} &= f_{0.975} = 2.7006. \end{aligned}$$

Thus, by (10.27), 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, by (10.10)

$$P = 2\left(1 - F_{F(24,15)}(|f_0|)\right) = 2\left(1 - F_{F(24,15)}(2.7398)\right) = 0.0469.$$

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.

2. Next we want to compare the population means. If $M1$ is to be stronger than $M2$ on average, than our test is

$$\begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2, \end{array}$$

a right-tailed test. Which one of the tests for the difference of means should we use? Our answer is in part 1. At this significance level, the variances are unknown and different. Then the value of the test statistic is, by (10.20)

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{380 - 370}{\sqrt{\frac{537}{25} + \frac{196}{16}}} = 1.7218.$$

By (10.19), we have

$$c = 0.6368, \quad n = 38.9244 \approx 39$$

and the quantile for a $T(39)$ distribution

$$t_{1-\alpha} = t_{0.95} = 1.6849.$$

By (10.21), 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. That means that yes, the data suggests that material $M1$ is, on average, stronger than material $M2$.
The P-value of this test is

$$P = 1 - F_{T(39)}(t_0) = 1 - F_{T(39)}(1.7218) = 0.0465,$$

again lower than $\alpha = 0.05$, which indicates the rejection of H_0. ■

Remark 10.6.2. Most of the tests we presented are based on 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 violated. In modern Statistics there is an ongoing search for finding robust methods of estimation for variances. For more information, see [12].

10.7 Power of a Test and the Neyman-Pearson Lemma

The "goodness" of a test is measured by the two probabilities of risk

$$\begin{array}{rclcl} \alpha & = & P(\text{type I error}) & = & P(\text{reject } H_0 \mid H_0) \\ \beta & = & P(\text{type II error}) & = & P(\text{accept } H_0 \mid H_1). \end{array}$$

The smaller both of them are, the more reliable the test is. For some problems, a type I error is more dangerous, while for others, a type II error is unacceptable. In general, α is preset, at most 0.05 and the test is designed so that β is also small enough to be acceptable.

10.7.1 Type II Errors and Power of a Test

The computation of β can be more difficult, because the condition that H_1 is true does not specify an actual value for the unknown parameter and thus, does not identify a distribution, for which the probability can be computed. The simple condition that a parameter θ is less than, greater than or not equal to a value is not enough to help us compute the probability. However, if the alternate H_1 is also a simple hypothesis, then β can be computed. Thus β, unlike α, depends on the value specified in the alternative hypothesis, $\beta = \beta(\theta_1)$.

Example 10.7.1. Let us consider again the problem in Example 10.2.2 and find β for the test

$$\begin{aligned} H_0: &\quad \mu = \mu_0 = 20 \\ H_1: &\quad \mu = \mu_1 = 18.5 < 20, \end{aligned}$$

i.e. find $\beta(\mu_1)$.
For $\alpha = 0.05$, we have determined the rejection region

$$\begin{aligned} RR &= \left\{ Z_0 = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{0.05} \right\} = \left\{ \frac{\overline{X} - 20}{\frac{3.2}{6}} < -1.645 \right\} \\ &= \left\{ \overline{X} < -1.645 \cdot \frac{3.2}{6} + 20 \right\} = \left\{ \overline{X} < 19.1227 \right\}. \end{aligned}$$

Then

$$\beta(\mu_1) = P\left(Z_0 \notin RR \mid H_1\right) = P\left(\overline{X} \geq 19.1227 \mid \mu = \mu_1\right).$$

If the true value of μ is μ_1, then the statistic

$$Z_1 = \frac{\overline{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - 18.5}{\frac{3.2}{6}}$$

follows a standard normal distribution. Hence

$$\begin{aligned} \beta(\mu_1) &= P\left(\frac{\overline{X} - 18.5}{\frac{3.2}{6}} \geq \frac{19.1227 - 18.5}{\frac{3.2}{6}} \mid \mu = 18.5 \right) \\ &= P(Z_1 \geq 1.1676 \mid Z_1 \in N(0,1)) = 0.1215. \end{aligned}$$

Notice that the value of β in Example 10.7.1 is far too large to be acceptable. In order to have a better control over β, we introduce a new notion.

Definition 10.7.2. *The **power of a test** involving a parameter θ, unknown, is the probability of rejecting the null hypothesis, when the true value of the parameter is $\widetilde{\theta}$, i.e.*

$$\pi(\widetilde{\theta}) = P(TT \in RR \mid \theta = \widetilde{\theta}). \tag{10.28}$$

If the null hypothesis is true, i.e. $\theta = \theta_0$, then

$$\pi(\theta_0) = P(TT \in RR \mid \theta = \theta_0) = P(\text{reject } H_0 \mid H_0) = \alpha.$$

For any other value $\theta = \theta_1 \neq \theta_0$,

$$\begin{aligned} \pi(\theta_1) &= P(\text{reject } H_0 \mid \theta = \theta_1) &= P(\text{reject } H_0 \mid H_1) \\ &= 1 - P(\text{accept } H_0 \mid H_1) &= 1 - \beta(\theta_1). \end{aligned}$$

So there is a strong relationship between the power of a test and the probability of a type II error. The larger the power is, the smaller β is, which is what we want in a test. Then we can state a hypothesis testing problem the following way:
For a parametric test where both hypotheses are simple

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta = \theta_1, \end{aligned}$$

we preset $\alpha = \pi(\theta_0)$ and we determine a rejection region for which

$$\pi(\theta_1) = 1 - \beta(\theta_1)$$

is *the largest possible*. Such a test is called a **most powerful test**.

10.7.2 The Neyman-Pearson Lemma

The following result gives a procedure for finding a most powerful test.

Lemma 10.7.3 (Neyman-Pearson (NPL))**.** *Let X be a characteristic with density $f(x;\theta)$, with $\theta \in A \subset \mathbb{R}$, unknown. Suppose we test on θ the simple hypotheses*

$$\begin{array}{rl} H_0: & \theta = \theta_0 \\ H_1: & \theta = \theta_1, \end{array}$$

based on a random sample $X_1, \ldots, X_n$. Let $L(\theta) = L(X_1, \ldots, X_n; \theta)$ denote the likelihood function for this sample. Then for a fixed $\alpha \in (0,1)$, a most powerful test is the test with rejection region given by

$$RR = \left\{ \frac{L(\theta_1)}{L(\theta_0)} \geq k_\alpha \right\}, \tag{10.29}$$

where the constant $k_\alpha > 0$ depends on α.

Example 10.7.4. Suppose Y represents a single observation from a probability density given by

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & \text{if } x \in (0,1) \\ 0, & \text{otherwise.} \end{cases}$$

Find the NPL most powerful test that at the significance level 5% tests

$$\begin{array}{rl} H_0: & \theta = 1 \\ H_1: & \theta = 50. \end{array}$$

Find β for that test.

Solution 10.7.4: Since our sample has size 1, we have

$$\frac{L(\theta_1)}{L(\theta_0)} = \frac{f(Y;\theta_1)}{f(Y;\theta_0)} = \frac{50Y^{49}}{1} = 50Y^{49}.$$

So the rejection region given by the NPL is

$$RR = \{50Y^{49} \geq k_\alpha\} = \{Y \geq k'\},$$

where $k' = \left(\frac{1}{50}k_\alpha\right)^{1/49}$.
We find the value of k' from

$$\alpha = P(Y \in RR \mid H_0) = P(Y \geq k' \mid \theta = 1) = \int_{k'}^{1} dx = 1 - k',$$

so $k' = 1 - \alpha = 0.95$.
So of all tests for testing H_0 versus H_1, based on a sample of size 1, the observation Y, at the significance level $\alpha = 0.05$, the most powerful test has rejection region $RR = \{Y \geq 0.95\}$.
For this test,

$$\begin{aligned} \beta(\theta_1) &= P(Y < k' \mid \theta = 50) = \int_0^{k'} 50x^{49}\,dx \\ &= x^{50}\Big|_0^{k'} = (k')^{50} = (1-\alpha)^{50} = 0.0769. \end{aligned}$$

■

Notice that the rejection region and, hence, the most powerful test we found in Example 10.7.4, depend on the value stated in H_1. For a different value of θ_1, we would have found a different rejection region. That is usually the case. However, sometimes, a test obtained with the NPL actually maximizes the power for *every* value in H_1, i.e. even if H_1 is not a simple hypothesis. Such a test is called a **uniformly most powerful test**.

Example 10.7.5. Let $X_1, \dots, X_n$ be a random sample drawn from a normal $N(\mu, \sigma)$ distribution, with μ unknown and σ known. At the significance level $\alpha \in (0, 1)$, find the most powerful right-tailed test for testing

$$\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0. \end{array}$$

Solution 10.7.5: First we use the NPL to find a most powerful test for

$$\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 > \mu_0. \end{array}$$

We have the normal density

$$f(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2}\Big), \ \forall x \in \mathbb{R}.$$

The likelihood function is

$$L(\mu) = \prod_{i=1}^{n} f(x_i; \mu) = \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big).$$

Then by the NPL, we want

$$\frac{L(\mu_1)}{L(\mu_0)} = \exp\Big(\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \mu_1)^2\Big]\Big) \geq k_\alpha,$$

or, taking the logarithm ln (which is an increasing function) on both sides,

$$\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \mu_1)^2\Big] \geq \ln k_\alpha.$$

After cancellations and using $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we have

$$2n\overline{x}(\mu_1 - \mu_0) \geq 2\sigma^2 \ln k_\alpha + n(\mu_1^2 - \mu_0^2).$$

Since $\mu_1 > \mu_0,$ we get

$$\overline{x} \geq \frac{\sigma^2 \ln k_\alpha}{n(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2} = K_\alpha.$$

Then we use the test statistic $\overline{x}$, for which we found the rejection region

$$RR = \{\overline{x} \geq K_\alpha\}.$$

But

$$\begin{aligned}\alpha = \ & P\Big(\overline{x} \geq K_\alpha \mid \mu = \mu_0\Big) = P\Big(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{K_\alpha - \mu_0}{\sigma/\sqrt{n}}\Big) \\ = \ & P\Big(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}\Big),\end{aligned}$$

since $Z_0 = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \in N(0,1)$. Then we must have

$$\frac{K_\alpha - \mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha}, \quad K_\alpha = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}},$$

so K_α is independent of μ_1. Then the test with $RR = \{\overline{x} \geq K_\alpha\}$ is a uniformly most powerful test for testing

$$\begin{aligned} H_0 : &\quad \mu = \mu_0 \\ H_1 : &\quad \mu > \mu_0, \end{aligned}$$

at the significance level α. ■

Remark 10.7.6. In a similar manner, we can find a uniformly most powerful test for the left-tailed case

$$\begin{aligned} H_0 : &\quad \mu = \mu_0 \\ H_1 : &\quad \mu < \mu_0, \end{aligned}$$

but we *cannot* use the NPL to find a uniformly most powerful test for the two-tailed alternative.

10.8 Nonparametric Tests

The concepts of statistical hypothesis or significance testing presented so far are based on the sampling distribution of a particular statistic. In short, if we have a basic knowledge of the underlying distribution of a variable, then we can make predictions about how, in repeated samples of equal size, this particular statistic will "behave" that is, how it is distributed. Such basic knowledge includes, for example, the assumption that the samples are drawn from an approximately normal population, or that two samples are independent, or we require large enough samples. For many variables of interest, we simply do not know for sure that all the assumptions we need are met. Also, these tests use arithmetic operations in the manipulation of scores on which the inference is based, and therefore they are useful only with numerical observations.

Hence, the need is evident for statistical procedures to be used in cases when nothing is known about the parameters of the variable of interest in the population, the so-called *nonparametric* or *distribution-free* statistical tests. These tests do not make numerous or stringent assumptions about the underlying population and most of them can be used with non-numerical data, as well. Some nonparametric methods have long been among the standard tools of statisticians, others are relatively new and have just started to gain prominence. We present here two of the most popular nonparametric tests.

10.8.1 The χ^2-Test for Contingency Tables

This test assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other. Let us consider two characteristics X and Y, relative to the same population and assume they are grouped into r (for rows) and c (for columns) cells, respectively.

When an individual is randomly selected, let A_i denote the event that the individual belongs to the i^{th} class with respect to X and by B_j the event that it belongs to the j^{th} class with respect to Y, $i = \overline{1, r}$, $j = \overline{1, c}$. Denote by

$$\begin{aligned} p_{ij} &= P(A_i \cap B_j),\ i = \overline{1, r},\ j = \overline{1, c}, \\ p_{i.} &= \sum_{j=1}^{c} p_{ij},\ i = \overline{1, r}, \\ p_{.j} &= \sum_{i=1}^{r} p_{ij},\ j = \overline{1, c}. \end{aligned}$$

Then the hypotheses

$$\begin{aligned} H_0 &: \ X \text{ and } Y \text{ are independent} \\ H_1 &: \ X \text{ and } Y \text{ are not independent,} \end{aligned}$$

can be rewritten as

$$\begin{aligned} &H_0 : \ p_{ij} = p_{i.} p_{.j},\ \forall\, i = \overline{1, r},\ \forall\, j = \overline{1, c} \\ &H_1 : \ p_{i_0 j_0} \neq p_{i_0 .} p_{. j_0},\ \text{for some } i_0 \in \{1, \dots, r\},\ j_0 \in \{1, \dots, c\}. \end{aligned} \tag{10.30}$$

To tests these hypotheses, we represent the n sample data in a contingency table (as in section 7.3),

$X \setminus Y$	B_1	$\dots$	B_j	$\dots$	B_c	
A_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1c}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{ic}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	$\dots$	n_{rj}	$\dots$	n_{rc}	$n_{r.}$
	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.c}$	$n_{..} = n$

where n_{ij} represents the absolute frequency of the class (i, j), for $i = \overline{1, r},\ j = \overline{1, c}$ and

$$n_{i.} = \sum_{j=1}^{c} n_{ij},\ i = \overline{1, r}, \quad n_{.j} = \sum_{i=1}^{r} n_{ij},\ j = \overline{1, c}, \quad n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}\ .$$

The maximum likelihood estimators for the marginal probabilities are

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, i = \overline{1, r}, \quad \hat{p}_{.j} = \frac{n_{.j}}{n},\ j = \overline{1, c}. \tag{10.31}$$

Then for an $r \times c$ contingency table, the "theoretical frequency" or "expected frequency" for a cell, given the hypothesis of independence, is

$$\hat{E}_{ij} = \frac{n_{i.} n_{.j}}{n},\ i = \overline{1, r},\ j = \overline{1, c}. \tag{10.32}$$

The test statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{10.33}$$

and for n large enough, it follows a χ^2 distribution. We determine the number of degrees of freedom the following way: there are rc cells in the table, but the marginal probabilities are fixed by (10.31), so the number of degrees of freedom left is

$$df = rc - (r + c - 1) = (r - 1)(c - 1).$$

Then the test statistic X^2 given by (10.33) follows a $\chi^2((r - 1)(c - 1))$ distribution. For a given significance level $\alpha \in (0, 1)$, the rejection region for the test (10.30) is

$$RR = \{X^2 > \chi^2_{1-\alpha}\}, \tag{10.34}$$

where the quantile $\chi^2_{1-\alpha}$ refers to the $\chi^2((r-1)(c-1))$ distribution. If the observed value of the test statistic from the sample is X^2_{obs}, then the P-value of the test is computed, as the probability of the test statistic taking values at least as extreme as the observed value, i.e.

$$P = P(X^2 > X^2_{obs}). \tag{10.35}$$

Remark 10.8.1. The only assumption that this test makes is related to the sample size. For the test to work well, a common rule is to have 5 or more observations in all cells of a 2×2 table and 5 or more in 80% of the cells in larger tables, but no cells with zero count.

Example 10.8.2. A study is conducted to reveal if there is a relationship between gender and the public or private sector of employment of citizens of Cluj-Napoca. A public opinion poll surveyed a simple random sample of 1000 persons. Respondents were classified by gender, male or female, and by place of employment, state employees (SE), people who work in private firms (PE), or unemployed (U). The results are shown in the contingency table below. At the 5% significance level, do the two characteristics seem to be related in some way, or not?

	SE	PE	U
Male	200	150	50
Female	250	300	50

Solution 10.8.2: We test the hypotheses

$$\begin{aligned} H_0 &: \quad \text{gender and place of employment are independent} \\ H_1 &: \quad \text{gender and place of employment are not independent.} \end{aligned}$$

Here $n = 1000$, $r = 2$ and $c = 3$. We add to the table the marginal frequencies

$$n_{1.} = 400, \;\; n_{2.} = 600,$$

$$n_{.1} = 450, \;\; n_{.2} = 450, \;\; n_{.3} = 100$$

and the expected frequencies (10.32) (in parentheses, in each cell)

$$\begin{aligned}
\hat{E}_{11} &= \frac{n_{1.}n_{.1}}{n} = \frac{400 \cdot 450}{1000} = 180, \\
\hat{E}_{12} &= \frac{n_{1.}n_{.2}}{n} = \frac{400 \cdot 450}{1000} = 180, \\
\hat{E}_{13} &= \frac{n_{1.}n_{.3}}{n} = \frac{400 \cdot 100}{1000} = 40, \\
\hat{E}_{21} &= \frac{n_{2.}n_{.1}}{n} = \frac{600 \cdot 450}{1000} = 270, \\
\hat{E}_{22} &= \frac{n_{2.}n_{.2}}{n} = \frac{600 \cdot 450}{1000} = 270, \\
\hat{E}_{23} &= \frac{n_{2.}n_{.3}}{n} = \frac{600 \cdot 100}{1000} = 60.
\end{aligned}$$

The complete contingency table is shown below.

	SE	PE	U	
Male	200	150	50	400
	(180)	(180)	(40)	
Female	250	300	50	600
	(270)	(270)	(60)	
	450	450	100	1000

The value of the test statistic (10.33) is

$$\begin{aligned}
X^2_{obs} &= \sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(n_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \\
&= \frac{(200-180)^2}{180}+\frac{(150-180)^2}{180}+\frac{(50-40)^2}{40} \\
&+ \frac{(250-270)^2}{270}+\frac{(300-270)^2}{270}+\frac{(50-60)^2}{60} \\
&= 16.2037.
\end{aligned}$$

For $df = (2-1)(3-1) = 2$ degrees of freedom, the quantile of order 0.95 is

$$\chi^2_{0.95} = 5.9915,$$

so the value $X^2_{obs} \in RR = (5.9915, \infty)$ and, hence, we reject the null hypothesis. For significance testing, the P-value of the test

$$P = P(X^2 > 16.2037) = 0.0003$$

is very small, much smaller than α, so, of course, we reject H_0.
Thus, we conclude that there is a relationship between gender and place of employment.

■

10.8.2 The Kolmogorov-Smirnov (K-S) Test

The Kolmogorov-Smirnov (K-S) test for one sample is used to decide if the sample comes from a population with a specific distribution, i.e. if the distribution of the sample data “ fits” into some known distribution, which is why it is also known as the K-S “goodness-of-fit” test. We only present the two-tailed test.

Let X be a continuous characteristic with theoretical cumulative distribution function F, unknown. We want to test the hypotheses

$$\begin{aligned} H_0 :& \quad F = F_0 \\ H_1 :& \quad F \neq F_0, \end{aligned}$$

where F_0 is a completely specified cumulative distribution function . Let $X_1, \ldots, X_n$ be a random sample with sample distribution function F_n, as described in (8.9). We use the test statistic

$$D_n = \sup_{x \in \mathbb{R}} |\overline{F}_n(x) - F_0(x)|, \tag{10.36}$$

defined in (8.10). If the sample indeed comes from the distribution F_0, then by Theorem 8.1.17, D_n converges to 0 almost surely, and by Theorem 8.1.18, $\sqrt{n}D_n$ converges in distribution to the Kolmogorov distribution, K, defined in (8.11).

Given the way the sample distribution function is defined and the fact that both F_0 and $\overline{F}_n$ are monotonely increasing, if the sample data is sorted in increasing order $x_1 < \cdots < x_n$, the value of the test statistic (10.36) is computed by

$$d_n = \max_{1 \leq i \leq n} \left\{ F_0(x_i) - \frac{i-1}{n}, \frac{i}{n} - F_0(x_i) \right\}. \tag{10.37}$$

If the data is grouped into m classes, with y_i the class representative and n_i the class frequency for $i = \overline{1, m}$, then

$$d_n = \max_{1 \leq i \leq m} |\overline{F}_n(y_i) - F_0(y_i)|. \tag{10.38}$$

For $\alpha \in (0, 1)$, the rejection region for the test is

$$RR = \{d_n > \frac{1}{\sqrt{n}} k_{1-\alpha} = d_{n,1-\alpha}\}, \tag{10.39}$$

where $k_{1-\alpha}$ is the quantile for the Kolmogorov distribution (8.11). The critical values $d_{n,\gamma}$ are given in Appendix B.5.
The P-value of the test is computed by

$$P = P(D_n > d_n) = P(\sqrt{n}D_n > \sqrt{n}d_n) = 1 - K(\sqrt{n}d_n). \qquad (10.40)$$

Remark 10.8.3. An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it does not depend on an adequate sample size for the approximations to be valid. However, the K-S test has several limitations, such as: it only applies to continuous distributions, it tends to be more sensitive near the center of the distribution than at the tails and the distribution F_0 must be fully specified (i.e. if location, scale, and shape parameters are estimated from the data, then the rejection region of the K-S test is no longer valid and must be determined by simulation).

Example 10.8.4. The values of a characteristic X are believed to be normally distributed. A sample of 100 measurements is taken and the grouped frequency distribution is given in the table below.

Data	Frequency
$1.25 - 2.50$	1
$2.50 - 3.75$	3
$3.75 - 5.00$	5
$5.00 - 6.25$	42
$6.25 - 7.50$	23
$7.50 - 8.75$	11
$8.75 - 10.00$	10
$10.00 - 11.25$	5

Table 10.1: Frequency Distribution Table for Example 10.8.4

At the 1% significance level, does the characteristic X seem to have a $N(6.25, 1)$ distribution? What about at 5%?

Solution 10.8.4: We test

$$\begin{aligned} H_0: &\ \ F = F_0 \\ H_1: &\ \ F \neq F_0, \end{aligned}$$

where F_0 is the cumulative distribution function of the $N(6.25, 1)$ distribution

$$F_0(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{(t-6.25)^2}{2}}\, dt.$$

The sample data are $y_i = 1.25 + 1.25i$, $i = \overline{0,8}$. The sample distribution function is $\overline{F}_{100}$ and its values are

$$\overline{F}_{100}(y_i) = \frac{1}{100}\sum_{j=1}^{i} n_j, \ \ i = \overline{0,8}.$$

The computations are given in the table below.

y_i	n_i	$F_0(y_i)$	$\overline{F}_{100}(y_i)$	$\lvert\overline{F}_{100}(y_i) - F_0(y_i)\rvert$
1.25	0	0.0000	0.0000	0.0000
2.50	1	0.0001	0.0100	0.0099
3.75	3	0.0062	0.0400	0.0338
5.00	5	0.1056	0.0900	0.0156
6.25	42	0.5000	0.5100	0.0100
7.50	23	0.8944	0.7400	0.1544
8.75	11	0.9938	0.8500	0.1438
10.00	10	0.9999	0.9500	0.0499
11.25	5	1.0000	1.0000	0.0000

The value of the test statistic is

$$d_n = 0.1544.$$

For $\alpha = 0.01$, $d_{n,1-\alpha} = d_{100,0.99} = 0.163$, so $d_n \notin RR$, thus we accept the null hypothesis that the data is drawn from a $N(6.25, 1)$ distribution, while if $\alpha = 0.05$, $d_{n,1-\alpha} = d_{100,0.95} = 0.136$, $d_n \in RR$, so we reject H_0.
The P-value of the test is

$$P = 1 - K(1.544) = 0.0175.$$

■

Appendix A

Euler's Functions

A.1 Euler's Gamma Function

Definition A.1.1. *Euler's Gamma function* $\Gamma : 0, \infty) \to (0, \infty)$ *is defined by*

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Proposition A.1.2. *The Gamma function has the following properties:*

(1) $\Gamma(1) = 1;$

(2) $\Gamma(a+1) = a\Gamma(a),$ *for all* $a > 0;$

(3) $\Gamma(n+1) = n!\,,$ *for all* $n \in \mathbb{N};$

(4) $\Gamma\left(\frac{1}{2}\right) = \sqrt{2} \int_0^\infty e^{-\frac{t^2}{2}} dt = \int_{\mathbb{R}} e^{-t^2} dt = \sqrt{\pi}.$

Proof.
(1) By definition,

$$\Gamma(1) = \int_0^\infty e^{-x} dx = -e^{-x}\Big|_0^\infty = 1.$$

(2) Using integration by parts, we have

$$\begin{aligned}\Gamma(a) &= \int_0^\infty x^{a-1} e^{-x} dx \\ &= -x^{a-1}e^{-x}\Big|_0^\infty + (a-1)\int_0^\infty x^{a-2}e^{-x}dx \\ &= (a-1)\Gamma(a-1)\,.\end{aligned}$$

(3) Using successively the property proved in (2), we get

$$\Gamma(n) = (n-1)\Gamma(n-1) = \cdots = (n-1)!\Gamma(1) = (n-1)!\,.$$

(4) With the change of variables $x = \frac{t^2}{2}$, we obtain

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \sqrt{2}\int_0^\infty e^{-\frac{t^2}{2}} dt.$$

With the substitution $x = t^2$, we have

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = 2\int_0^\infty e^{-u^2} du = \int_{\mathbb{R}} e^{-u^2} du\,.$$

Then we write

$$\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \int_{\mathbb{R}} e^{-u^2} du \int_{\mathbb{R}} e^{-v^2} dv = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-(u^2+v^2)} dudv\,.$$

Switching to polar coordinates $u = r\cos t, v = r\sin t$, the domain $\mathbb{R}^2$ becomes $\{(r,t) \in \mathbb{R}^2 \mid r > 0, t \in [0, 2\pi)\}$, the determinant of the Jacobian matrix is equal to r and we obtain

$$\left(\Gamma\left(\frac{1}{2}\right)\right)^2 = \int_0^\infty \int_0^{2\pi} re^{-r^2} dr dt = \pi\,.$$

□

A.2 Euler's Beta Function

Definition A.2.1. *Euler's Beta function $\beta : (0,\infty) \times (0,\infty) \to (0,\infty)$ is defined by*

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx.$$

Proposition A.2.2. *The Beta function has the following properties:*

(1) $\beta(a,1) = \dfrac{1}{a}$, *for all* $a > 0$;

(2) $\beta(a,b) = \beta(b,a)$, *for all* $a > 0, b > 0$;

(3) $\beta(a,b) = \dfrac{a-1}{b}\beta(a-1,b+1)$, *for all* $a > 1, b > 0$;

(4) $\beta(a,b) = \dfrac{b-1}{a+b-1}\beta(a,b-1) = \dfrac{a-1}{a+b-1}\beta(a-1,b)$, *for all* $a > 1$, $b > 1$;

(5) $\beta(a,b) = \displaystyle\int_0^\infty t^{a-1}(1+t)^{-(a+b)} dt$, *for all* $a > 0, b > 0$;

(6) $\beta(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, *for all* $a > 0, b > 0.$

Proof.
(1) By definition,

$$B(a,1) = \int\limits_0^1 x^{a-1}dx = \frac{1}{a}\,.$$

(2) We substitute $x = 1 - t$ to get

$$\beta(a,b) = \int\limits_0^1 x^{a-1}(1-x)^{b-1}dx = -\int\limits_1^0 (1-t)^{a-1}t^{b-1}dy = \beta(b,a).$$

(3) Integrating by parts, we get

$$\begin{aligned}
&\beta(a,b) = \int\limits_0^1 x^{a-1}\left(-\frac{(1-x)^b}{b}\right)' dx \\
&= -x^{a-1}\frac{(1-x)^b}{b}\bigg|_0^1 + \frac{a-1}{b}\int\limits_0^1 x^{a-2}(1-x)^b dx \\
&= \frac{a-1}{b}\beta(a-1,b+1)\,.
\end{aligned}$$

(4) Integrating by parts, we have

$$\begin{aligned}
\beta(a,b) &= \int\limits_0^1 \left(\frac{x^a}{a}\right)'(1-x)^{b-1}dx \\
&= \frac{x^a}{a}(1-x)^{b-1}\Big|_0^1 + \frac{b-1}{a}\int\limits_0^1 x^a(1-x)^{b-2}dx \\
&= \frac{b-1}{a}\int\limits_0^1 \Big(x^{a-1}(1-x)^{b-2} - x^{a-1}(1-x)^{b-1}\Big)dx \\
&= \frac{b-1}{a}\beta(a,b-1) - \frac{b-1}{a}\beta(a,b)\,,
\end{aligned}$$

i.e.

$$\beta(a,b) = \frac{b-1}{a+b-1}\beta(a,b-1).$$

Then by symmetry, we have

$$\beta(a,b) = \frac{a-1}{a+b-1}\beta(a-1,b)\,.$$

(5) With the change of variables $1+t = \dfrac{1}{1-x}$, we have

$$\begin{aligned}
x &= \frac{t}{1+t} \\
dx &= \frac{1}{(1+t)^2}\,dt
\end{aligned}$$

and the domain of integration $(0,1)$ becomes $(0,\infty)$. Thus

$$\beta(a,b) = \int\limits_0^\infty t^{a-1}(1+t)^{-(a+b)}dt.$$

(6) Using the substitution $x = u^2$ for the Gamma function, we have

$$\Gamma(a) = \int\limits_0^\infty x^{a-1}e^{-x}dx = 2\int\limits_0^\infty u^{2a-1}e^{-u^2}du.$$

Then

$$\Gamma(a)\Gamma(b) = 4\int\limits_0^\infty\int\limits_0^\infty u^{2a-1}v^{2b-1}e^{-(u^2+v^2)}dudv.$$

Using polar coordinates $u = r\cos t, v = r\sin t$, the domain $(0,\infty) \times (0,\infty)$ becomes $\left\{(r,t) \in \mathbb{R}^2 \mid r > 0, t \in \left[0, \frac{\pi}{2}\right)\right\}$, the determinant of the Jacobian matrix is again r and we obtain

$$\begin{aligned}\Gamma(a)\Gamma(b) &= 4\int\limits_0^\infty\int\limits_0^{\frac{\pi}{2}} r^{2(a+b)-1}e^{-r^2}(\cos t)^{2a-1}(\sin t)^{2b-1}drdt \\ &= 2\int\limits_0^\infty r^{2(a+b)-1}e^{-r^2}dr\, 2\int\limits_0^{\frac{\pi}{2}}(\cos t)^{2a-1}(\sin t)^{2b-1}dt \\ &= \Gamma(a+b)B(a,b).\end{aligned}$$

We use the substitution $x = \cos^2 t$, to compute

$$2\int\limits_0^{\frac{\pi}{2}}(\cos t)^{2a-1}(\sin t)^{2b-1}dt = \int\limits_0^1 x^{a-1}(1-x)^{b-1}dx = B(a,b).$$

Hence,

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

□

Appendix B

Tables

B.1 Standard Normal Cumulative Distribution Function

$$F_Z(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt, \;\; x \in \mathbb{R}; \;\; F_Z(-x) = 1 - F_Z(x), \;\; x > 0.$$

	$F_Z(x)$									
x	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

(Continued)

	$F_Z(x)$									
x	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

B.2 Values for the T Distribution

$$\gamma = F_T(t_{n,\gamma}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{t_{n,\gamma}} \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx.$$

	$t_{n,\gamma}$					
$n = df$	γ= 0.90	0.95	0.975	0.99	0.995	0.999
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
35	1.306	1.690	2.030	2.438	2.724	3.340
40	1.303	1.684	2.021	2.423	2.704	3.307
45	1.301	1.679	2.014	2.412	2.690	3.281
50	1.299	1.676	2.009	2.403	2.678	3.261
55	1.297	1.673	2.004	2.396	2.668	3.245
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

B.3 Values for the χ^2 Distribution

$$\gamma = F_{\chi^2}(\chi^2_{n,\gamma}) = \frac{1}{\Gamma(\frac{n}{2})\,2^{\frac{n}{2}}} \int_0^{\chi^2_{n,\gamma}} x^{\frac{n}{2}-1} e^{-\dfrac{x}{2}}\,dx.$$

	$\chi^2_{n,\gamma}$									
$n = df$	γ =0.001	0.005	0.010	0.025	0.050	0.010	0.125	0.200	0.250	0.333	0.500
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335
55	28.173	31.735	33.570	36.398	38.958	42.060	43.220	46.036	47.610	49.972	54.335
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335

(Continued)

	$\chi^2_{n,\gamma}$										
$n = df$	0.600	0.667	0.750	0.800	0.875	0.900	0.950	0.975	0.990	0.995	0.999
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
55	57.016	58.945	61.665	63.577	67.211	68.796	73.311	77.380	82.292	85.749	93.168
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

B.4 Values for the F Distribution

$$\gamma = F_F(f_{m,n,\gamma}) = \frac{1}{\beta(\frac{m}{2},\frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} \int_0^{f_{m,n,\gamma}} x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}} dx.$$

	γ	$f_{m,n,\gamma}$							
$n\backslash m$		1	2	3	4	5	6	7	8
1	0.950	161	199	216	225	230	234	237	239
	0.975	648	799	864	900	922	937	948	957
	0.990	4052	4999	5404	5624	5764	5859	5928	5981
	0.995	16212	19997	21614	22501	23056	23440	23715	23924
2	0.950	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4
	0.975	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4
	0.990	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4
	0.995	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4
3	0.950	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	0.975	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54
	0.990	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	0.995	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13
4	0.950	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	0.975	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98
	0.990	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
	0.995	31.33	26.28	24.26	23.15	22.46	21.98	21.62	21.35
5	0.950	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818
	0.975	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757
	0.990	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289
	0.995	22.785	18.314	16.530	15.556	14.939	14.513	14.200	13.961
6	0.950	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
	0.975	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600
	0.990	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102
	0.995	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566
7	0.950	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726
	0.975	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899
	0.990	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840
	0.995	16.235	12.404	10.883	10.050	9.522	9.155	8.885	8.678
8	0.950	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438
	0.975	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433
	0.990	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029
	0.995	14.688	11.043	9.597	8.805	8.302	7.952	7.694	7.496

(Continued)

	γ	$f_{m,n,\gamma}$								
$n\backslash m$		9	10	15	20	30	60	120	240	∞
1	0.950	241	242	246	248	250	252	253	254	254
	0.975	963	969	985	993	1001	1010	1014	1016	1018
	0.990	6022	6056	6157	6209	6260	6313	6340	6353	6366
	0.995	24091	24222	24632	24837	25041	25254	25358	25414	25466
2	0.950	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
	0.975	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5
	0.990	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
	0.995	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5
3	0.950	8.81	8.79	8.70	8.66	8.62	8.57	8.55	8.54	8.53
	0.975	14.47	14.42	14.25	14.17	14.08	13.99	13.95	13.92	13.90
	0.990	27.34	27.23	26.87	26.69	26.50	26.32	26.22	26.17	26.13
	0.995	43.88	43.68	43.08	42.78	42.47	42.15	41.99	41.91	41.83
4	0.950	6.00	5.96	5.86	5.80	5.75	5.69	5.66	5.64	5.63
	0.975	8.90	8.84	8.66	8.56	8.46	8.36	8.31	8.28	8.26
	0.990	14.66	14.55	14.20	14.02	13.84	13.65	13.56	13.51	13.46
	0.995	21.14	20.97	20.44	20.17	19.89	19.61	19.47	19.40	19.32
5	0.950	4.772	4.735	4.619	4.558	4.496	4.431	4.398	4.382	4.365
	0.975	6.681	6.619	6.428	6.329	6.227	6.123	6.069	6.042	6.015
	0.990	10.158	10.051	9.722	9.553	9.379	9.202	9.112	9.066	9.020
	0.995	13.772	13.618	13.146	12.903	12.656	12.402	12.274	12.209	12.144
6	0.950	4.099	4.060	3.938	3.874	3.808	3.740	3.705	3.687	3.669
	0.975	5.523	5.461	5.269	5.168	5.065	4.959	4.904	4.877	4.849
	0.990	7.976	7.874	7.559	7.396	7.229	7.057	6.969	6.925	6.880
	0.995	10.391	10.250	9.814	9.589	9.358	9.122	9.001	8.941	8.879
7	0.950	3.677	3.637	3.511	3.445	3.376	3.304	3.267	3.249	3.230
	0.975	4.823	4.761	4.568	4.467	4.362	4.254	4.199	4.171	4.142
	0.990	6.719	6.620	6.314	6.155	5.992	5.824	5.737	5.694	5.650
	0.995	8.514	8.380	7.968	7.754	7.534	7.309	7.193	7.135	7.076
8	0.950	3.388	3.347	3.218	3.150	3.079	3.005	2.967	2.947	2.928
	0.975	4.357	4.295	4.101	3.999	3.894	3.784	3.728	3.699	3.670
	0.990	5.911	5.814	5.515	5.359	5.198	5.032	4.946	4.903	4.859
	0.995	7.339	7.211	6.814	6.608	6.396	6.177	6.065	6.008	5.951

(Continued)

	γ	$f_{m,n,\gamma}$							
$n\backslash m$		1	2	3	4	5	6	7	8
9	0.950	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230
	0.975	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102
	0.990	10.562	8.022	6.992	6.422	6.057	5.802	5.613	5.467
	0.995	13.614	10.107	8.717	7.956	7.471	7.134	6.885	6.693
10	0.950	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
	0.975	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855
	0.990	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057
	0.995	12.827	9.427	8.081	7.343	6.872	6.545	6.303	6.116
11	0.950	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948
	0.975	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664
	0.990	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744
	0.995	12.226	8.912	7.600	6.881	6.422	6.102	5.865	5.682
12	0.950	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849
	0.975	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512
	0.990	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499
	0.995	11.754	8.510	7.226	6.521	6.071	5.757	5.524	5.345
13	0.950	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767
	0.975	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388
	0.990	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302
	0.995	11.374	8.186	6.926	6.233	5.791	5.482	5.253	5.076
14	0.950	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699
	0.975	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285
	0.990	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140
	0.995	11.060	7.922	6.680	5.998	5.562	5.257	5.031	4.857
15	0.950	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641
	0.975	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199
	0.990	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004
	0.995	10.798	7.701	6.476	5.803	5.372	5.071	4.847	4.674
16	0.950	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
	0.975	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125
	0.990	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890
	0.995	10.576	7.514	6.303	5.638	5.212	4.913	4.692	4.521

(Continued)

	γ	$f_{m,n,\gamma}$								
$n\backslash m$		9	10	15	20	30	60	120	240	∞
9	0.950	3.179	3.137	3.006	2.936	2.864	2.787	2.748	2.727	2.707
	0.975	4.026	3.964	3.769	3.667	3.560	3.449	3.392	3.363	3.333
	0.990	5.351	5.257	4.962	4.808	4.649	4.483	4.398	4.354	4.311
	0.995	6.541	6.417	6.032	5.832	5.625	5.410	5.300	5.244	5.188
10	0.950	3.020	2.978	2.845	2.774	2.700	2.621	2.580	2.559	2.538
	0.975	3.779	3.717	3.522	3.419	3.311	3.198	3.140	3.110	3.080
	0.990	4.942	4.849	4.558	4.405	4.247	4.082	3.996	3.953	3.909
	0.995	5.968	5.847	5.471	5.274	5.071	4.859	4.750	4.695	4.639
11	0.950	2.896	2.854	2.719	2.646	2.570	2.490	2.448	2.426	2.404
	0.975	3.588	3.526	3.330	3.226	3.118	3.004	2.944	2.914	2.883
	0.990	4.632	4.539	4.251	4.099	3.941	3.776	3.690	3.647	3.602
	0.995	5.537	5.418	5.049	4.855	4.654	4.445	4.337	4.281	4.226
12	0.950	2.796	2.753	2.617	2.544	2.466	2.384	2.341	2.319	2.296
	0.975	3.436	3.374	3.177	3.073	2.963	2.848	2.787	2.756	2.725
	0.990	4.388	4.296	4.010	3.858	3.701	3.535	3.449	3.405	3.361
	0.995	5.202	5.085	4.721	4.530	4.331	4.123	4.015	3.960	3.904
13	0.950	2.714	2.671	2.533	2.459	2.380	2.297	2.252	2.230	2.206
	0.975	3.312	3.250	3.053	2.948	2.837	2.720	2.659	2.628	2.595
	0.990	4.191	4.100	3.815	3.665	3.507	3.341	3.255	3.210	3.165
	0.995	4.935	4.820	4.460	4.270	4.073	3.866	3.758	3.703	3.647
14	0.950	2.646	2.602	2.463	2.388	2.308	2.223	2.178	2.155	2.131
	0.975	3.209	3.147	2.949	2.844	2.732	2.614	2.552	2.520	2.487
	0.990	4.030	3.939	3.656	3.505	3.348	3.181	3.094	3.050	3.004
	0.995	4.717	4.603	4.247	4.059	3.862	3.655	3.547	3.492	3.436
15	0.950	2.588	2.544	2.403	2.328	2.247	2.160	2.114	2.090	2.066
	0.975	3.123	3.060	2.862	2.756	2.644	2.524	2.461	2.429	2.395
	0.990	3.895	3.805	3.522	3.372	3.214	3.047	2.959	2.914	2.868
	0.995	4.536	4.424	4.070	3.883	3.687	3.480	3.372	3.317	3.260
16	0.950	2.538	2.494	2.352	2.276	2.194	2.106	2.059	2.035	2.010
	0.975	3.049	2.986	2.788	2.681	2.568	2.447	2.383	2.350	2.316
	0.990	3.780	3.691	3.409	3.259	3.101	2.933	2.845	2.799	2.753
	0.995	4.384	4.272	3.920	3.734	3.539	3.332	3.224	3.168	3.111

(Continued)

	γ	$f_{m,n,\gamma}$							
$n\backslash m$		1	2	3	4	5	6	7	8
17	0.950	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548
	0.975	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061
	0.990	8.400	6.112	5.185	4.669	4.336	4.101	3.927	3.791
	0.995	10.384	7.354	6.156	5.497	5.075	4.779	4.559	4.389
18	0.950	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510
	0.975	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005
	0.990	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705
	0.995	10.218	7.215	6.028	5.375	4.956	4.663	4.445	4.276
20	0.950	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447
	0.975	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913
	0.990	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564
	0.995	9.944	6.987	5.818	5.174	4.762	4.472	4.257	4.090
25	0.950	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337
	0.975	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753
	0.990	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324
	0.995	9.475	6.598	5.462	4.835	4.433	4.150	3.939	3.776
30	0.950	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266
	0.975	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651
	0.990	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173
	0.995	9.180	6.355	5.239	4.623	4.228	3.949	3.742	3.580
50	0.950	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130
	0.975	5.340	3.975	3.390	3.054	2.833	2.674	2.553	2.458
	0.990	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890
	0.995	8.626	5.902	4.826	4.232	3.849	3.579	3.376	3.219
100	0.950	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032
	0.975	5.179	3.828	3.250	2.917	2.696	2.537	2.417	2.321
	0.990	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694
	0.995	8.241	5.589	4.542	3.963	3.589	3.325	3.127	2.972
∞	0.950	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938
	0.975	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192
	0.990	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511
	0.995	7.879	5.298	4.279	3.715	3.350	3.091	2.897	2.744

(Continued)

	γ	$f_{m,n,\gamma}$								
$n\backslash m$		9	10	15	20	30	60	120	240	∞
17	0.950	2.494	2.450	2.308	2.230	2.148	2.058	2.011	1.986	1.960
	0.975	2.985	2.922	2.723	2.616	2.502	2.380	2.315	2.282	2.247
	0.990	3.682	3.593	3.312	3.162	3.003	2.835	2.746	2.700	2.653
	0.995	4.254	4.142	3.793	3.607	3.412	3.206	3.097	3.041	2.984
18	0.950	2.456	2.412	2.269	2.191	2.107	2.017	1.968	1.943	1.917
	0.975	2.929	2.866	2.667	2.559	2.445	2.321	2.256	2.222	2.187
	0.990	3.597	3.508	3.227	3.077	2.919	2.749	2.660	2.613	2.566
	0.995	4.141	4.030	3.683	3.498	3.303	3.096	2.987	2.931	2.873
20	0.950	2.393	2.348	2.203	2.124	2.039	1.946	1.896	1.870	1.843
	0.975	2.837	2.774	2.573	2.464	2.349	2.223	2.156	2.121	2.085
	0.990	3.457	3.368	3.088	2.938	2.778	2.608	2.517	2.470	2.421
	0.995	3.956	3.847	3.502	3.318	3.123	2.916	2.806	2.749	2.690
25	0.950	2.282	2.236	2.089	2.007	1.919	1.822	1.768	1.740	1.711
	0.975	2.677	2.613	2.411	2.300	2.182	2.052	1.981	1.944	1.906
	0.990	3.217	3.129	2.850	2.699	2.538	2.364	2.270	2.220	2.169
	0.995	3.645	3.537	3.196	3.013	2.819	2.609	2.496	2.437	2.377
30	0.950	2.211	2.165	2.015	1.932	1.841	1.740	1.683	1.654	1.622
	0.975	2.575	2.511	2.307	2.195	2.074	1.940	1.866	1.827	1.787
	0.990	3.067	2.979	2.700	2.549	2.386	2.208	2.111	2.060	2.006
	0.995	3.451	3.344	3.006	2.823	2.628	2.415	2.300	2.239	2.176
50	0.950	2.073	2.026	1.871	1.784	1.687	1.576	1.511	1.476	1.438
	0.975	2.381	2.317	2.109	1.993	1.866	1.721	1.639	1.594	1.545
	0.990	2.785	2.698	2.419	2.265	2.098	1.909	1.803	1.745	1.683
	0.995	3.092	2.988	2.653	2.470	2.272	2.050	1.925	1.858	1.786
100	0.950	1.975	1.927	1.768	1.676	1.573	1.450	1.376	1.333	1.283
	0.975	2.244	2.179	1.968	1.849	1.715	1.558	1.463	1.409	1.347
	0.990	2.590	2.503	2.223	2.067	1.893	1.692	1.572	1.504	1.427
	0.995	2.847	2.744	2.411	2.227	2.024	1.790	1.652	1.573	1.485
∞	0.950	1.880	1.831	1.666	1.571	1.459	1.318	1.221	1.155	1
	0.975	2.114	2.048	1.833	1.708	1.566	1.388	1.268	1.187	1
	0.990	2.407	2.321	2.039	1.878	1.696	1.473	1.325	1.225	1
	0.995	2.621	2.519	2.187	2.000	1.789	1.533	1.364	1.251	1

B.5 Critical Values for the K-S Test $d_{n,\gamma}$

	$d_{n,\gamma}$				
$n\backslash\gamma$	0.2	0.1	0.05	0.02	0.01
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

References

[1] Agratini O., *Probabilităţi - Culegere de Probleme*, Lito Universitatea Babeş-Bolyai, Cluj-Napoca, 1992.

[2] Blaga P., *Calculul probabilităţilor - Culegere de Probleme*, Lito Universitatea Babeş-Bolyai, Cluj-Napoca, 1984.

[3] ———, *Calculul probabilităţilor şi statistică matematică*, Vol. II, Curs şi culegere de probleme , Universitatea Babeş-Bolyai, Cluj-Napoca, 1994.

[4] ———, *Statistică ... prin MATLAB*, Presa Universitară Clujeană, Cluj-Napoca, 2002.

[5] Blaga P., Rădulescu M., *Calculul probabilităţilor*, Lito Universitatea Babeş-Bolyai, Cluj-Napoca, 1987.

[6] Chow Y. S., Teicher H., *Probability Theory: Independence, Interchangeability, Martingales*, Springer Verlag, New York, 1997.

[7] Forthofer R. N., Lee Eun Sul, Hernandez M., *Biostatistics: A Guide to Design, Analysis, and Discovery*, Second Edition, Elsevier Academic Press, 2007.

[8] DeGroot M. H., *Probability and Statistics*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

[9] Grinstead Ch. M., Snell J. L., *Introduction in Probability*, Second edition, American Mathematical Society, 1997.

[10] Lisei H., *Probability Theory*, Casa Cărţii de Ştiinţă, Cluj-Napoca, 2004.

[11] Lisei H., Micula S., Soós A., *Probability Theory through Problems and Applications*, Cluj University Press, 2006.

[12] Maronna R. A., Martin R. D., Yohai V. J., *Robust Statistics, Theory and Methods*, John Wiley & Sons, Ltd. Chichester, 2006.

[13] Mihoc I., Fătu C. I., *Calculul probabilităţilor şi statistică matematică*, Partea a III-a, Casa de Editură Transilvania Press, Cluj-Napoca, 2003.

[14] Mihoc Gh., Ciucu G., Craiu V., *Teoria probabilităţilor şi statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1970.

[15] Milton J. S., Arnold J. C., *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*, 3rd Edition, McGraw-Hill, New York, 1995.

[16] Rao M. M., *Probability Theory with Applications*, Academic Press , New York, 1984.

[17] Reischer C., Sâmboan A., *Culegere de probleme de teoria probabilităţilor şi statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1972.

[18] Shiryaev A. N., *Probability*, Springer Verlag, New York, 1995.

[19] Stapleton J. H., *Linear statistical models*, John Wiley & Sons , New York -Chichester -Brisbane, 1995.

[20] J. Stoyanov, I. Mirazchiiski, Z. Ignatov, M. Tanushev, *Exercise Manual in Probability Theory*, Kluwer Academic Publishers, Dordrecht, 1989.

[21] Trîmbiţaş R. T., *Metode statistice*, Presa Universitară Clujeană, Cluj-Napoca, 2000.

[22] Wentzel E., Ovcharov L., *Applied Problems in Probability Theory* , Mir Publishers Moscow, 1986.

Index