Lecture notes on
Probability and Statistics

Ngo Hoang Long

Division of Applied Mathematics
Faculty of Mathematics and Informatics
Hanoi National University of Education
Email: ngolong@hnue.edu.vn

Contents

1 Probability Space **5**

1.1 Introduction 5

1.2 Probability space 7

1.3 Properties of probability 9

1.4 Probabilities on a Finite or Countable Space 13

1.5 Conditional Probability 14

1.6 Independence 16

2 Random Variables and Distributions **29**

2.1 Random variables on a countable space 29

2.1.1 Definitions 29

2.1.2 Examples 31

2.2 Random variables on a general probability space 32

2.2.1 Definition 32

2.2.2 Structure of random variables 33

2.3 Distribution Functions 34

2.3.1 Definition 34

2.3.2 Examples 35

2.4 Expectation 36

2.4.1 Construction of expectation 36

2.4.2 Some limit theorems 39

2.4.3 Some inequalities 40

2.4.4 Expectation of random variable with density 41

2.5 Random elements 42

2.5.1 Definitions 42

2.5.2 Example 43

2.5.3 Density of function of random vectors 43

2.6 Independent random variables 43

2.7 Covariance 45

2.8 Conditional Expectation 46

2.8.1 Definition 46

2.8.2 Examples 47

2.8.3 Properties of conditional expectation 47
2.8.4 Convergence theorem 49
2.8.5 Conditional expectation given a random variable 50
2.9 Exercises 50

3 Fundamental Limit Theorems 62
3.1 Convergence of random variables 62
3.2 Laws of large numbers 67
3.2.1 Weak laws of large numbers 67
3.2.2 Strong laws of large numbers 68
3.3 Central limit theorems 70
3.3.1 Characteristic functions 70
3.3.2 Weak convergence 73
3.3.3 Central limit theorem 75
3.4 Exercises 76
3.4.1 Convergence of random variables 76
3.4.2 Law of large numbers 77
3.4.3 Characteristic function 78
3.4.4 Weak convergence 79
3.4.5 Central limit theorems 80

4 Some useful distributions in statistics 82
4.1 Gamma, chi-square, student and F distributions 82
4.1.1 Gamma distribution 82
4.1.2 Chi-square distribution 82
4.1.3 Student's t distribution 84
4.1.4 F distribution 84
4.2 Sample mean and sample variance 85
4.3 Exercises 86

5 Parameter estimation 88
5.1 Samples and characteristic of sample 88
5.2 Data display 90
5.2.1 Stem-and-leaf diagrams 90
5.2.2 Frequency distribution and histogram 91
5.2.3 Box plots 92
5.2.4 Probability plots 93
5.3 Point estimations 95
5.3.1 Statistics 95
5.3.2 Point estimators 96
5.3.3 Confidence intervals 97
5.4 Method of finding estimation 102

5.4.1 Maximum likelihood estimation 102
5.4.2 Method of Moments 103
5.5 Lower bound for variance 104
5.6 Exercises 106
5.6.1 Confidence interval 106
5.6.2 Point estimator 109
5.6.3 Lower bound for variance 111

6 Hypothesis Testing 113
6.1 Introduction 113
6.2 Method of finding test 116
6.2.1 Likelihood Ratio Tests 116
6.3 Method of evaluating test 117
6.3.1 Most powerful test 117
6.3.2 Uniformly most powerful test 119
6.3.3 Monotone likelihood ratio 121
6.4 Some well-known tests for a single sample 122
6.4.1 Hypothesis test on the mean of a normal distribution, variance σ^2 known . 122
6.4.2 Hypothesis test on the mean of a normal distribution, variance σ^2 unknown 125
6.4.3 Hypothesis test on the variance of a normal distribution 127
6.4.4 Test on a population proportion 128
6.5 Some well-known tests for two samples 129
6.5.1 Inference for a difference in means of two normal distributions, variances known 129
6.5.2 Inference for the difference in means of two normal distributions, variances unknown 131
6.5.3 Paired t-test 133
6.5.4 Inference on the variance of two normal populations 134
6.5.5 Inference on two population proportions 135
6.6 The chi-square test 136
6.6.1 Goodness-of-fit test 136
6.6.2 Tests of independence 142
6.6.3 Test of homogeneity 143
6.7 Exercises 145
6.7.1 Significance level and power function 145
6.7.2 Null distribution 146
6.7.3 Best critical region 146
6.7.4 Some tests for single sample 147
6.7.5 Some tests for two samples 150
6.7.6 Chi-squared tests 154

7 Regression **156**

7.1 Simple linear regression 156

7.1.1 Simple linear regression model 156

7.1.2 Confidence interval for σ^2 157

7.1.3 Confidence interval for β_1 158

7.1.4 Confidence interval for β_0 158

7.1.5 Prediction intervals 158

Chapter 1

Probability Space

[1]

1.1 Introduction

Random experiments are experiments whose output cannot be surely predicted in advance. But when one repeats the same experiment a large number of times one can observe some ”regularity” in the average output. A typical example is the toss of a coin: one cannot predict the result of a single toss, but if we toss the coin many times we get an average of about 50% of ”heads” if the coin is fair. The theory of probability aims towards a mathematical theory which describes such phenomena. This theory contains three main ingredients:

a) The state space: this is the set of all possible outcomes of the experiment, and it is usually denoted by Ω.

Examples

1. A toss of a coin: $\Omega = \{H, T\}$.
2. Two successive tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$.
3. A toss of two dice: $\Omega = \{(i, j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$.
4. The measurement of a length L, with a measurement error: $\Omega = \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the positive real numbers; $\omega \in \Omega$ denotes the result of the measurement, and $w - L$ is the measurement error.
5. The lifetime of a light-bulb: $\Omega = \mathbb{R}_+$.

b) The event: An ”event” is a property which can be observed either to hold or not to hold after the experiment is done. In mathematical terms, an event is a subset of Ω. If A and B are two events, then

- the contrary event is interpreted as the complement set A^c;

[1] This is a draft version which contains many errors. Comments are very welcome

- the event "A or B" is interpreted as the union $A \cup B$;
- the event "A and B" is interpreted as the intersection $A \cap B$;
- the sure event is Ω;
- the impossible event is the empty set $\emptyset$;
- an elementary event is a "singleton*', i.e. a subset $\{w\}$ containing a single outcome w of Ω.

We denote by $\mathcal{A}$ the family of all events. Often (but not always: we will see why later) we have $\mathcal{A} = 2^\Omega$ the set of all subsets of Ω. The family $\mathcal{A}$ should be "stable" by the logical operations described above: if $A, B \in \mathcal{A}$ then we must have $A^c \in \mathcal{A},\ A \cap B \in \mathcal{A},\ A \cup B \in \mathcal{A}$, and also $\Omega \in \mathcal{A}$ and $\emptyset \in \mathcal{A}$.

c) The probability: With each event A one associates a number denoted by $\mathbb{P}(A)$ and called the "probability of A". This number measures the likelihood of the event A to be realized a priori, before performing the experiment. It is chosen between 0 and 1, and the more likely the event is, the closer to 1 this number is.

To get an idea of the properties of these numbers, one can imagine that they are the limits of the "frequency" with which the events are realized: let us repeat the same experiment n times; the n outcomes might of course be different (think of n successive tosses of the same die, for instance). Denote by $f_n(A)$ the frequency with which the event A is realized (i.e. the number of times the event occurs, divided by n). Intuitively we have:

$$\mathbb{P}(A) = \text{ limit of } f_n(A) \text{ as } n \to \infty$$

(we will give a precise meaning to this "limit" later). From the obvious properties of frequencies, we immediately deduce that:

1. $0 \le \mathbb{P}(A) \le 1$;
2. $\mathbb{P}(\Omega) = 1$;
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$.

A mathematical model for our experiment is thus a triple $(\Omega, \mathcal{A}, \mathbb{P})$, consisting of the space Ω, the family $\mathcal{A}$ of all events, and the family of all $\mathbb{P}(A)$ for $A \in \mathcal{A}$; hence we can consider that $\mathbb{P}$ is a map from $\mathcal{A}$ into $[0, 1]$, which satisfies at least the properties (2) and (3) above (plus in fact an additional property, more difficult to understand, and which is given later).

A fourth notion, also important although less basic, is the following one:

d) Random variable: A random variable is a quantity which depends on the outcome of the experiment. In mathematical terms, this is a map from Ω into a space E, where often $E = \mathbb{R}$ or $E = \mathbb{R}^d$. Warning: this terminology, which is rooted in the history of Probability Theory going back 400 years, is quite unfortunate; a random "variable" is not a variable in the analytical sense, but a function!

Let X be such a random variable, mapping Ω into E. One can then ”transport” the probabilistic structure onto the target space E, by setting $\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B))$ for $B \in E$, where $X^{-1}(B) = \{w \in \Omega : X(w) \in B\}$ denotes the pre-image of B by X. This formula defines a new probability, denoted by $\mathbb{P}^X$ but on the space E instead of Ω. The probability $\mathbb{P}^X$ is called the *law of the variable* X.

Example (toss of two dice): One tosses two fair dice and observers the number of dots appearing on each dice. The sample space is $\Omega = \{(i,j) : 1 \leq i \leq 6, 1 \leq i \leq 6\}$, and it is natural to take here $A = 2^\Omega$ and

$$\mathbb{P}(A) = \frac{|A|}{36} \quad \text{if } A \subset \Omega,$$

where $|A|$ denotes the number of points in A. One easily verifies the properties (1), (2), (3) above, and $\mathbb{P}(\{w\}) = 1/36$ for each singleton. The map $X : \Omega \to \mathbb{N}$ defined by $X(i,j) = |j - i|$ is the random variable ”different of the two dice”, and its law is

$$\mathbb{P}_X(B) = \frac{\text{number of pairs } (i,j) \text{ such that } |i-j| \in B}{36}$$

(for example $\mathbb{P}_X(\{1\}) = 5/18, \mathbb{P}_X(\{5\}) = 1/18$, etc ...). We will formalize the concepts of a probability space and random variable in following sections.

1.2 Probability space

Let Ω be a non-empty set without any special structure. Let 2^Ω denote all subsets of Ω, including the empty set denoted by $\emptyset$. With $\mathcal{A}$ being a subset of 2^Ω, we consider the following properties:

1. $\emptyset \in \mathcal{A}$ and $\Omega \in \mathcal{A}$;

2. If $A \in \mathcal{A}$ then $A^c := \Omega \backslash A \in \mathcal{A}$; A^c is called the complement of A;

3. $\mathcal{A}$ is closed under finite unions and finite intersections: that is, if $A_1, \ldots, A_n$ are all in $\mathcal{A}$, then $\cup_{i=1}^n$ and $\cap_{i=1}^n A_i$ are in $\mathcal{A}$ as well (for this it is enough that $\mathcal{A}$ be stable by the union and the intersection of any two sets);

4. $\mathcal{A}$ is closed under countable unions and intersections: that is, if $A_1, A_2 \ldots$ is a countable sequence of events in $\mathcal{A}$ then $\cup_i A_i$ and $\cap_i A_i$ are both also in $\mathcal{A}$.

Definition 1.2.1. $\mathcal{A}$ is an algebra if it satisfies (1), (2) and (3) above. It is a σ-algebra, (or a σ-field) if it satisfies (1), (2), and (4) above.

Definition 1.2.2. If $\mathcal{A}$ is a σ-algebra on Ω then $(\Omega, \mathcal{A})$ is called a measurable space.

Definition 1.2.3. If $\mathcal{C} \subset 2^\Omega$, the σ-algebra generated by $\mathcal{C}$, and written $\sigma(\mathcal{C})$, is the smallest σ-algebra containing $\mathcal{C}$. (It always exists because 2^Ω is a σ-algebra, and the intersection of a family of σ-algebras is again a σ-algebra)

Example: (i) $\mathcal{A} = \{\emptyset, \Omega\}$ (the trivial σ-algebra).
(ii) If A is a subset; then $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$.
(iii) If $\Omega = \mathbb{R}^d$ the Borel σ-algebra, written $\mathcal{B}(\mathbb{R}^d)$, is the σ-algebra generated by all the intervals A of the following type

$$A = (-\infty, x_1] \times (-\infty, x_2] \times \ldots \times (-\infty, x_n],$$

where $x_1, \ldots, x_n \in \mathbb{Q}$.

We can show that $\mathcal{B}(\mathbb{R}^d)$ is also the σ-algebra generated by all open subsets (or by all the closed subsets) of $\mathbb{R}^d$.

Definition 1.2.4. A probability measure defined on a σ-algebra $\mathcal{A}$ of Ω is a function $\mathbb{P} : \mathcal{A} \to [0, 1]$ that satisfies:

1. $\mathbb{P}(\Omega) = 1$;

2. For every countable sequence $(A_n)_{n\geq 1}$ of elements of $\mathcal{A}$, pairwise disjoint (that is, $A_n \cap A_m = \emptyset$ whenever $n \neq m$), one has
$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Axiom (2) above is called countable additivity; the number $\mathbb{P}(A)$ is called the probability of the event A.

In Definition 1.2.4 one might imagine a more elementary condition than (2), namely:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{1.1}$$

for any disjoint sets $A, B \in \mathcal{A}$.

This property is called additivity (or “finite additivity”) and, by an elementary induction, it implies that for every finite $A_1, \ldots, A_m$ of pairwise disjoint events $A_i \in \mathcal{A}$, we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{m} A_i\Big) = \sum_{i=1}^{m} \mathbb{P}(A_i).$$

Theorem 1.2.5. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. The following properties hold:*

(i) $\mathbb{P}(\emptyset) = 0$*;*

(ii) $\mathbb{P}$ *is additive.*

(iii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*;*

(iv) If $A, B \in \mathcal{A}$ *and* $A \subset B$ *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof. If in Axiom (2) we take $A_n = \emptyset$ for all n, we see that the number $a = \mathbb{P}(\emptyset)$ is equal to an infinite sum of itself; since $0 \leq a \leq 1$, this is possible only if $a = 0$, and we have (i).

For (ii) it sufffices to apply Axiom (2) with $A_1 = A$ and $A_2 = B$ and $A_3 = A_4 = ... = \emptyset$, plus the fact that $\mathbb{P}(\emptyset) = 0$, to obtain (1.1).

Applying (1.1) for $A \in \mathcal{A}$ and $B = A^c$ we get (iii).

To show (iv), suppose $A \subset B$ then applying (1.1) for A and $B \backslash A$ we have

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \geq \mathbb{P}(A).$$

□

1.3 Properties of probability

Countable additivity is not implied by additivity. In fact, in spite of its intuitive appeal, additivity is not enough to handle the mathematical problems of the theory. The next theorem shows exactly what is extra when we assume countable additivity instead of just finite additivity.

Theorem 1.3.1. *Let A be a σ-algebra. Suppose that $\mathbb{P} : \mathcal{A} \to [0, 1]$ satisfies $\mathbb{P}(\Omega) = 1$ and is additive. Then the following statements are equivalent:*

(i) Axiom (2) of Definition 1.2.4.

(ii) If $A_n \in \mathcal{A}$ and $A_n \downarrow A$, then $\mathbb{P}(A_n) \downarrow 0$.

(iii) If $A_n \in \mathcal{A}$ and $A_n \downarrow A$, then $\mathbb{P}(A_n) \downarrow \mathbb{P}(A)$.

(iv) If $A_n \in \mathcal{A}$ and $A_n \uparrow \Omega$, then $\mathbb{P}(A_n) \uparrow 1$.

(v) If $A_n \in \mathcal{A}$ and $A_n \uparrow A$, then $\mathbb{P}(A_n) \uparrow \mathbb{P}(A)$.

Proof. The notation $A_n \downarrow A$ means that $A_{n+1} \subset A_n$, each n, and $\cap_{i=1}^{\infty} A_n = A$. The notation $A_n \uparrow A$ means that $A_n \subset A_{n+1}$ and $\cup_{i=1}^{\infty} A_n = A, \forall n \geq 2$.
$(i) \Rightarrow (v)$ Let $A_n \in \mathcal{A}$ with $A_n \uparrow A$. We construct a new sequence as follows: $B_1 = A_1$ and $B_n = A_{n+1} \backslash A_n$.Then $\cup_{i=1}^{\infty} B_n = A; \quad A_n = \cup_{i=1}^{n} B_i$ and the events $(B_n)_{n \geq 1}$ are pairwise disjoint. Therefore

$$\mathbb{P}(A) = \mathbb{P}(\cup_{k \geq 1} B_k) = \sum_{k=1}^{\infty} B_k.$$

Hence

$$\mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(B_k) \uparrow \sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}(A).$$

$(v) \Rightarrow (i)$ Let $A_n \in \mathcal{A}$ be pairwise disjoint. Define $B_n = \cup_{k=1}^{n} A_k$. We have $B_n \uparrow \cup_{k=1}^{\infty} A_k$. Hence

$$\mathbb{P}(\cup_{k=1}^{\infty} A_k) = \lim_{n \to \infty} \mathbb{P}(B_n) = \lim_{n \to \infty} \sum_{k=1}^{n} \mathbb{P}(A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

$(v) \to (iii)$: Suppose that $A_n \downarrow A$. Then $A_n^c \uparrow A^c$. Hence, we have

$$\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) \downarrow 1 - \mathbb{P}(A^c) = \mathbb{P}(A).$$

$(iii) \to (ii)$ is obvious.
$(ii) \to (iv)$: Let $A_n \in \mathcal{A}$ with $A_n \uparrow \Omega$. Thus $\mathbb{P}(A_n^c) \to 0$. Therefore $\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c) \uparrow 1$.
$(iv) \to (v)$: Suppose $A_n \uparrow A$. Denote $B_n = A_n \cup A^c$. One gets $B_n \uparrow \Omega$, it implies $\mathbb{P}(B_n) \uparrow 1$. Hence

$$\mathbb{P}(A_n) = \mathbb{P}(B_n) - \mathbb{P}(A^c) \uparrow 1 - \mathbb{P}(A^c) = \mathbb{P}(A).$$

□

If $A \in 2^{\Omega}$, we define the indicator function by

$$\mathbb{I}_A(w) = \begin{cases} 1 & \text{if } w \in A \\ 0 & \text{if } w \notin A. \end{cases}$$

We can say that $A_n \in \mathcal{A}$ converges to A (we write $A_n \to A$) if $\lim_{n\to\infty} \mathbb{I}_{A_n}(w) = \mathbb{I}_A(w)$ for all $w \in \Omega$. Note that if the sequence A_n increases (resp. decreases) to A, then it also tends to A in the above sense.

Theorem 1.3.2. *Let $\mathbb{P}$ be a probability measure and let A_n be a sequence of events in $\mathcal{A}$ which converges to A. Then $A \in \mathcal{A}$ and $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.*

Proof. Now let $B_n = \cap_{m\geq n} A_m$ and $C_n = \cup_{m\leq n} A_m$. Then B_n increases to A and C_n decreases to A, thus $\lim\limits_{n\to\infty} \mathbb{P}(B_n) = \lim\limits_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}(A)$, by Theorem 1.3.1. However $B_n \subset A_n \subset C_n$, therefore $\mathbb{P}(B_n) \leq \mathbb{P}(A_n) \leq \mathbb{P}(C_n)$, so $\lim\limits_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$ as well. □

Lemma 1.3.3. *Let S be a set. Let $\mathcal{I}$ be a π-system on S, that is, a family of subsets of S stable under finite intersection:*

$$I_1, I_2 \in \mathcal{I} \Rightarrow I_1 \cap I_2 \in \mathcal{I}.$$

Let $\Sigma = \sigma(\mathcal{I})$. Suppose that μ_1 and μ_2 are probability measure on (S, Σ) such that $\mu_1 = \mu_2$ on $\mathcal{I}$. Then $\mu_1 = \mu_2$ on Σ.

Proof. Let

$$\mathcal{D} = \{F \in \Sigma : \mu_1(F) = \mu_2(F)\}.$$

Then $\mathcal{D}$ is a d-system on S, that is, a family of subsets of S satisfied:

$$\begin{aligned} &a) \quad S \in \mathcal{D}, \\ &b) \quad \text{if } A, B \in \mathcal{D} \text{ and } A \subseteq B \text{ then } B \setminus A \in \mathcal{D}, \\ &c) \quad \text{if } A_n \in \mathcal{D} \text{ and } A_n \uparrow A, \text{ then } A \in \mathcal{D}. \end{aligned}$$

Indeed, the fact that $S \in \mathcal{D}$ is given. If $A, B \in \mathcal{D}$, then

$$\mu_1(B \setminus A) = \mu_1(B) - \mu_1(A) = \mu_2(B) - \mu_2(A) = \mu_2(B \setminus A).$$

so that $B \setminus A \in \mathcal{D}$. Finally, if $F_n \in \mathcal{D}$ and $F_n \uparrow F$, then

$$\mu_1(F) = \uparrow \lim \mu_1(F_n) = \uparrow \lim \mu_2(F_n) = \mu_2(F),$$

so that $F \in \mathcal{D}$.
Since $\mathcal{D}$ is a d-system and $\mathcal{D} \supseteq \mathcal{I}$ then $\mathcal{D} \supseteq \sigma(\mathcal{I}) = \Sigma$, and the result follows. □

This lemma implies that if two probability measure agree on a π-system, then they agree on the σ-algebra generated by that π-system.

Theorem 1.3.4 (Carathéodory's Extension Theorem)**.** *Let S be a set and Σ_0 be an algebra on S, and let $\Sigma = \sigma(\Sigma_0)$. If μ_0 is a countably additive map $\mu_0 : \Sigma_0 \to [0,1]$, then there exists a unique measure μ on (S, Σ) such that $\mu = \mu_0$ on Σ_0.*

Proof. Step 1: Let $\mathcal{G}$ be the $\sigma-$algebra of all subsets of S. For $G \in \mathcal{G}$, define

$$\lambda(G) := \inf \sum_n \mu_0(F_n),$$

where the infimum is taken over all sequences (F_n) in Σ_0 with $G \subseteq \cup_n F_n$.
We now prove that
(a) λ *is an outer measure on* $(S, \mathcal{G})$.
The facts that $\lambda() = 0$ and λ is increasing are obvious. Suppose that (G_n) is a sequence in $\mathcal{G}$, such that each $\lambda(G_n)$ is finite. Let $\epsilon > 0$ be given. For each n, choose a sequence $(F_{n,k} : k \in \mathbb{N})$ of elements of Σ_0 such that

$$G_n \subseteq \bigcup_k F_{n,k}, \quad \sum_k \mu_0(F_{n,k}) < \lambda(G_n) + \epsilon 2^{-n}.$$

Then $G := \bigcup_n G_n \subseteq \bigcup_n \bigcup_k F_{n,k}$, so that

$$\lambda(G) \leq \sum_n \sum_k \mu_0(F_{n,k}) < \sum_n \lambda(G_n) + \epsilon.$$

Since ϵ is arbitrary, we have proved result (a).
Step 2: We have λ is a measure on $(S, \mathcal{L})$, where $\mathcal{L}$ is a σ-algebra of $\lambda-$sets in $\mathcal{G}$. All we need show is that
(b) $\Sigma_0 \subseteq \mathcal{L}$, and $\lambda = \mu_0$ on σ_0;
for then $\Sigma := \sigma(\Sigma_0) \subseteq \mathcal{L}$ and we can define μ to be the restriction of λ to (S, Σ).
Step 3: Proof that $\lambda = \mu_0$ on Σ_0.
Let $F \in \Sigma_0$. Then, clearly, $\lambda(F) \leq \mu_0(F)$. Now suppose that $F \subseteq \bigcup_n F_n$, where $F_n \in \Sigma_0$. As usual, we can define a sequence (E_n) of disjoint sets:

$$E_1 := F_1, \quad E_n = F_n \cap \Big(\bigcup_{k<n} F_k\Big)^c$$

such that $E_n \subseteq F_n$ and $\bigcup E_n = \bigcup F_n \supseteq F$. Then

$$\mu_0(F) = \mu_0(\bigcup (F \cap E_n)) = \sum \mu_0(F \cap E_n),$$

by using the countable additivity of μ_0 on Σ_0. Hence

$$\mu_0(F) \leq \sum \mu_0(E_n) \leq \sum \mu_0(F_n),$$

so that $\lambda(F) \geq \mu_0(F)$. Step 3 is complete.
Step 4: *Proof that* $\Sigma_0 \subseteq \mathcal{L}$. Let $E \in \Sigma_0$ and $G \in \mathcal{G}$. Then there exists a sequence (F_n) in Σ_0 such that $G \subseteq \bigcup\limits_n F_n$, and

$$\sum_n \mu_0(F_n) \leq \lambda(G) + \epsilon.$$

Now, by definition of λ,

$$\begin{aligned}\sum_n \mu_0(F_n) &= \sum_n \mu_0(E \cap F_n) + \sum_n \mu_0(E^c \cap F_n) \\ &\geq \lambda(E \cap G) + \lambda(E^c \cap G),\end{aligned}$$

since $E \cap G \subseteq \bigcup(E \cap F_n)$ and $E^c \cap G \subseteq \bigcup(E^c \cap F_n)$. Thus, since ϵ is arbitrary,

$$\lambda(G) \geq \lambda(E \cap G) + \lambda(E^c \cap G).$$

However, since λ is subadditive,

$$\lambda(G) \leq \lambda(E \cap G) + \lambda(E^c \cap G).$$

We see that E is indeed a $\lambda-$set.

□

1.4 Probabilities on a Finite or Countable Space

We suppose that Ω is finite or countable and consider $\mathcal{A} = 2^{\Omega}$. Then a probability on Ω is characterized by its values on the atoms $p_w = \mathbb{P}(\{w\}),\ w \in \Omega$. Indeed, one can easily verify the following theorem.

Theorem 1.4.1. *Let $(p_w)_{w\in\Omega}$ be a family of real numbers indexed by the finite or countable set Ω. Then there exists a unique probability $\mathbb{P}$ such that $\mathbb{P}(\{w\}) = p_w$ if and only if $p_w \geq 0$ and $\sum_{w\in\Omega} p_w = 1$. In this case for any $A \subset \Omega$,*

$$\mathbb{P}(A) = \sum_{w\in A} p_w.$$

Suppose first that Ω is finite. Any family of nonnegative terms summing up to 1 gives an example of a probability on Ω. But among all these examples the following is particularly important:

Definition 1.4.2. A probability $\mathbb{P}$ on the finite set Ω is called uniform if $p_w = \mathbb{P}(\{w\})$ does not depend on w.

In this case, it is immediate that

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of outcomes in favor of } A}{\text{total number of possible outcomes}}.$$

Then computing the probability of any event A amounts to counting the number of points in A. On a given finite set Ω there is one and only one uniform probability.

Example: There are 20 balls in the urn, 10 white 10 red. One draws a set of 5 balls from the urn. Denote X the number of white ball in the set. We want to find the probability that $X = x$, where x is an arbitrary fixed integer.

We label from 1 to 10 for white balls and from 11 to 20 for red balls. Since the balls are drawn at once, it is natural to consider that an outcome is a subset with 5 elements of the set $\{1, \ldots, 20\}$ of all 20 balls. That is, Ω is the family of all subsets with 5 points, and the total number of possible outcomes is $|\Omega| = C_{20}^5$. Next, it is also natural to consider that all possible outcomes are equally likely, that is $\mathbb{P}$ is the uniform probability on Ω. The quantity X is a “random variable” because when the outcome w is known, one also knows the number $X(w)$. The possible values of X is from 0 to 5 and the set $X^{-1}(\{x\}) = \{X = x\}$ contains $C_{10}^x C_{10}^{5-x}$ points for all $0 \leq x \leq 5$. Hence

$$\mathbb{P}(X = x) = \begin{cases} \frac{C_{10}^x C_{10}^{5-x}}{C_{20}^5} & \text{if } 0 \leq x \leq 5 \\ 0 & \text{otherwise.} \end{cases}$$

We thus obtain, when x varies, the distribution or the law, of X. This distribution is called the hypergeometric distribution.

1.5 Conditional Probability

We have known how to answer questions of the following kind: If there are 5 balls in an urn, 2 white and 3 black, what is the probability $\mathbb{P}(A)$ of the event A that a selected ball is white? With the classical approach, $\mathbb{P}(A) = 2/5$.

The concept of conditional probability, which will be introduced below, let us answer questions of the following kind: What is the probability that the second ball is white (event B) under the condition that the first ball was also white (event A)? (We are thinking of sampling without replacement.) It is natural to reason as follows: if the first ball is white, then at the second step we have an urn containing 4 balls, of which 1 is white and 3 black; hence it seems reasonable to suppose that the (conditional) probability in question is $1/4$.

In general, computing the probability of an event A, given that an event B occurs, means finding which fraction of the probability of A is also in the event B.

Definition 1.5.1. Let A, B be events, $\mathbb{P}(B) > 0$. The conditional probability of A given B is

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

Theorem 1.5.2. *Suppose* $\mathbb{P}(B) > 0$*. The operation* $A \mapsto \mathbb{P}(A|B)$ *from* $\mathcal{A} \to [0,1]$ *defines a new probability measure on* $\mathcal{A}$*, called the* conditional probability measure given B.

Proof. We define $Q(A) = \mathbb{P}(A|B)$, with B fixed. We must show Q satisfies (1) and (2) of 1.2.4. But

$$Q(\Omega) = \mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

Therefore, Q satisfies (1). As for (2), note that if $(A_n)_{n\geq 1}$ is a sequence of elements of $\mathcal{A}$ which are pairwise disjoint, then

$$Q(\cup_{n=1}^{\infty} A_n) = \mathbb{P}(\cup_{n=1}^{\infty} A_n|B) = \frac{\mathbb{P}((\cup_{n=1}^{\infty} A_n) \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\cup_{n=1}^{\infty}(A_n \cap B))}{\mathbb{P}(B)}$$

and also the sequence $(A_n \cap B)_{n\geq 1}$ is pairwise disjoint as well; thus

$$Q(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n|B) = \sum_{n=1}^{\infty} Q(A_n).$$

□

Theorem 1.5.3. *If* $A_1, \ldots, A_n \in \mathcal{A}$ *and if* $\mathbb{P}(A_1 \cap \ldots \cap A_{n-1}) > 0$*, then*

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

Proof. We use induction. For $n = 2$, the theorem is simply 1.5.1. Suppose the theorem holds for $n-1$ events. Let $B = A_1 \cap \ldots \cap A_{n-1}$. Then by 1.5.1 $\mathbb{P}(B \cap A_n) = \mathbb{P}(A_n|B)\mathbb{P}(B)$; next we replace $\mathbb{P}(B)$ by its value given in the inductive hypothesis:

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\ldots\mathbb{P}(A_{n-1}|A_1 \cap \ldots \cap A_{n-2}),$$

and we get the result. □

Definition 1.5.4. A collection of events (E_n) is called a *partition* of Ω in $\mathcal{A}$ if

1. $E_n \in \mathcal{A}$ and $\mathbb{P}(E_n) > 0$, each n,
2. they are pairwise disjoint,
3. $\Omega = \cup_n E_n$.

Theorem 1.5.5 (Partition Equation)**.** *Let $(E_n)_{n\geq 1}$ be a finite or countable partition of Ω. Then if $A \in \mathcal{A}$,*

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A|E_n)\mathbb{P}(E_n).$$

Proof. Note that

$$A = A \cap \Omega = \cup_n (A \cap E_n).$$

Since the E_n are pairwise disjoint so also are $(A \cap E_n)_{n\geq 1}$, hence

$$\mathbb{P}(A) = \mathbb{P}\Big(\cup_n (A \cap E_n)\Big) = \sum_n \mathbb{P}(A \cap E_n) = \sum_n \mathbb{P}(A|E_n)\mathbb{P}(E_n).$$

□

Theorem 1.5.6 (Bayes' Theorem)**.** *Let (E_n) be a finite or countable partition of Ω, and suppose $\mathbb{P}(A) > 0$. Then*

$$\mathbb{P}(E_n|A) = \frac{\mathbb{P}(A|E_n)\mathbb{P}(E_n)}{\sum_m \mathbb{P}(A|E_m)\mathbb{P}(E_m)}.$$

Proof. Applying partition equation, we have that the denominator

$$\sum_m \mathbb{P}(A|E_m)\mathbb{P}(E_m) = \mathbb{P}(A).$$

Hence the formula becomes

$$\frac{\mathbb{P}(A|E_n)\mathbb{P}(E_n)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \cap E_n)}{\mathbb{P}(A)} = \mathbb{P}(E_n|A).$$

□

Example 1.5.7. Because a new medical procedure has been shown to be effective in the early detection of an illness, a medical screening of the population is proposed. The probability that the test correctly identifies someone with the illness as positive is 0.99, and the probability that the test correctly identifies someone without the illness as negative is 0.95. The incidence of the illness in the general population is 0.0001. You take the test, and the result is positive. What is the probability that you have the illness? Let D denote the event that you have the illness, and let S denote the event that the test signals positive. The probability requested can be denoted as . The probability that the test correctly signals someone without the illness as negative is 0.95. Consequently, the probability of a positive test without the illness is

$$\mathbb{P}(S|D^c) = 0.05.$$

From Bayes's Theorem,

$$\mathbb{P}(D|S) = \frac{\mathbb{P}(S|D)\mathbb{P}(D)}{\mathbb{P}(S|D)\mathbb{P}(D) + \mathbb{P}(S|D^c)\mathbb{P}(D^c)} = 0.002.$$

Surprisingly, even though the test is effective, in the sense that $\mathbb{P}(S|D)$ is high and $\mathbb{P}(S|D^c)$ is low, because the incidence of the illness in the general population is low, the chances are quite small that you actually have the disease even if the test is positive.

Example 1.5.8. Suppose that Bob can decide to go to work by one of three modes of transportation, car, bus, or commuter train. Because of high traffic, if he decides to go by car, there is a 0.5 chance he will be late. If he goes by bus, which has special reserved lanes but is somtimes overcrowded, the probability of being late is only 0.2. The commuter train is almost never late, with a probability of only 0.01, but is more expensive than the bus.
(a) Suppose that Bob is late one day, and his boss wishes to estimate the probability that he drove to work that day by car. Since he does not know which mode of transportation Bod usually uses, he gives a prior probability of $\frac{1}{3}$ to each of the three possibilities. What is the boss' estimate of the probability that Bob drove to work?
(b) Suppose that the coworker of Bob's knows that he almost always takes the commuter train to work, never take the bus, but somtimes, 0.1 of the time, takes the car. What is the coworkers probability that Bob drove to work that day, given that he was late?
We have the following information given in the problem:

$$\begin{aligned}
\mathbb{P}(bus) = \mathbb{P}(car) = \mathbb{P}(train) &= \frac{1}{3} \\
\mathbb{P}(late|car) &= 0.5; \\
\mathbb{P}(late|train) &= 0.01; \\
\mathbb{P}(late|bus) &= 0.2.
\end{aligned}$$

By Bayes Theorem, this is

$$\begin{aligned}
\mathbb{P}(car|late) &= \frac{\mathbb{P}(late|car)\mathbb{P}(car)}{\mathbb{P}(late|car)\mathbb{P}(car) + \mathbb{P}(late|bus)\mathbb{P}(bus) + \mathbb{P}(late|train)\mathbb{P}(train)} \\
&\frac{0.5 \times 1/3}{0.5 \times 1/3 + 0.2 \times 1/3 + 0.01 \times 1/3} \\
&= 0.7042
\end{aligned}$$

Repeat the identical calculations as above, but instead of the prior probabilities being $\frac{1}{3}$, we use $pr(bus) = 0, \mathbb{P}(car) = 0.1$, and $\mathbb{P}(train) = 0.9$. Plugging in to the same equation with these three changes, we get $\mathbb{P}(car|late) = 0.8475$.

1.6 Independence

Definition 1.6.1. 1. Two events A and B are *independent* if

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

2. A (possibly infinite) collection of events $(A_i)_{i\in I}$ is a *pairwise independent* collection if for any distinct elements $i_1, i_2 \in I$,
$$\mathbb{P}(A_{i_1} \cap A_{i_2}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}).$$

3. A (possibly infinite) collection of events $(A_i)_{i\in I}$ is an *independent* collection if for every finite subset J of I, one has
$$\mathbb{P}(\cap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i).$$

If events $(A_i)_{i\in I}$ are independent, they are pairwise independent, but the converse is false.

Proposition 1.6.2. *a) If A and B are independent, so also are A and B^c; A^c and B; and A^c and B^c.*
b) If A and B are independent and $\mathbb{P}(B) > 0$, then
$$\mathbb{P}(A|B) = \mathbb{P}(A|B^c) = \mathbb{P}(A).$$

Proof. a) *A and B^c.* Since A and B are independent, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B^c)) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B^c)$. We have $\mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c)$. Substituting these into the equation $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, we obtain
$$\mathbb{P}(A) - \mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B^c).$$

Hence
$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A)\mathbb{P}(B^c).$$

Therefore, A and B^c are independent.
A^c and B^c. We have $\mathbb{P}(A)\mathbb{P}(B^c) = [1 - \mathbb{P}(A^c)]\mathbb{P}(B^c) = \mathbb{P}(B^c) - \mathbb{P}(A^c)\mathbb{P}(B^c)$ and $\mathbb{P}(A \cap B^c) = \mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c)$. Substituting into the equation $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)\mathbb{P}(B^c)$, we obtain
$$\mathbb{P}(B^c) - \mathbb{P}(A^c \cap B^c) = \mathbb{P}(B^c) - \mathbb{P}(A^c)\mathbb{P}(B^c).$$

So
$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\mathbb{P}(B^c).$$

Therefore, A^c and B^c are independent.
b) We have
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A),$$

and
$$\mathbb{P}(A|B^c) = \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)} = \frac{\mathbb{P}(A)\mathbb{P}(B^c)}{\mathbb{P}(B^c)} = \mathbb{P}(A).$$

□

Examples:

1. Toss a coin 3 times. If A_i is an event depending only on the ith toss, then it is standard to model $(A_i)_{1\leq i\leq 3}$ as being independent.

2. One chooses a card at random from a deck of 52 cards. $A =$ "the card is a heart", and $B =$ "the card is Queen". A natural model for this experiment consists in prescribing the probability $1/52$ for picking any one of the cards. By additivity, $\mathbb{P}(A) = 13/52$ and $\mathbb{P}(B) = 4/52$ and $\mathbb{P}(A \cap B) = 1/52$ hence A and B are independent.

3. Let $n = \{1, 2, 3, 4\}$, and $\mathcal{A} = 2^{\Omega}$. Let $\mathbb{P}(i) = 1/4$, where $i = 1, 2, 3, 4$. Let $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{2, 3\}$. Then A, B, C are pairwise independent but are not independent.

Exercises

Axiom of Probability

1.1. Give a possible sample space for each of the following experiments:

1. A two-sided coin is tossed.

2. A student is asked for the month of the year and the day of the week on which her birthday falls.

3. A student is chosen at random from a class of ten students.

4. You receive a grade in this course.

1.2. Let $\mathcal{A}$ be a σ-algebra of subsets of Ω and let B is a subset of Ω. Show that $\mathcal{F} = \{A \cap B : A \in \mathcal{A}\}$ is a σ-algebra of subsets of B.

1.3. Let f be a function mapping Ω to another space E with a σ-algebra $\mathcal{E}$. Let $\mathcal{A} = \{A \subset \Omega :$ there exists $B \in \mathcal{E}$ with $A = f^{-1}(B)\}$. Show that $\mathcal{A}$ is a σ-algebra on Ω.

1.4. Let $(\mathcal{G}_\alpha)_{\alpha \in I}$ be an arbitrary family of σ-algebras defined on an abstract space Ω. Show that $\mathcal{H} = \cap_{\alpha \in I} \mathcal{G}_\alpha$ is also a σ-algebra.

1.5. Suppose that Ω is an infinite set (countable or not), and let $\mathcal{A}$ be the family of all subsets which are either finite or have a finite complement. Show that $\mathcal{A}$ is an algebra, but not a σ-algebra.

1.6. Give a counterexample that shows that, in general, the union $\mathcal{A} \cup \mathcal{B}$ of two σ-algebras need not be a σ-algebra.

1.7. Let $\Omega = \{a, b, c\}$ be a sample space. Let $\mathbb{P}(\{a\}) = 1/2, \mathbb{P}(\{b\}) = 1/3$, and $\mathbb{P}(\{c\}) = 1/6$. Find the probabilities for all eight subsets of Ω.

1.8. For $A,\ B \in \mathcal{A}$, show

1. $\mathbb{P}(A \cap \overline{B}) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

1.9. Suppose $\mathbb{P}(A) = \frac{3}{4}$ and $\mathbb{P}(B) = \frac{1}{3}$. Show that always $\frac{1}{12} \leq \mathbb{P}(A \cap B) \leq \frac{1}{3}$.

1.10. Let (B_n) be a sequence of events such that $\mathbb{P}(B_n) = 1$ for all $n \geq 1$. Show that

$$\mathbb{P}\Big(\bigcap_n B_n\Big) = 1.$$

1.11. Let $A_1, \ldots, A_n$ be given events. Show the inclusion-exclusion formula:

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j)$$
$$\sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n)$$

where (for example) $\sum_{i<j}$ means to sum over all ordered pairs (i, j) with $i < j$.

1.12. Let $A_i \in \mathcal{A}$ be a sequence of events. Show that

$$P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i),$$

each n, and also

$$P(\cup_{i=1}^\infty A_i) \leq \sum_{i=1}^\infty P(A_i).$$

1.13. *(Bonferroni Inequalities)* Let $A_i \in \mathcal{A}$ be a sequence of events. Show that

1. $P(\cup_{i=1}^n A_i) \geq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j)$,
2. $P(\cup_{i=1}^n A_i \leq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$.

1.14. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Show for events $B_i \subset A_i$ the following inequality

$$\mathbb{P}(\cup_i A_i) - \mathbb{P}(\cup_i B_i) \leq \sum_i \Big(\mathbb{P}(A_i) - \mathbb{P}(B_i)\Big).$$

1.15. If (B_k) are events such that $\sum_{k=1}^n \mathbb{P}(B_k) > n - 1$, then

$$\mathbb{P}\Big(\bigcap_{k=1}^n B_k) > 0.$$

Classical definition of probability

1.16. In the laboratory analysis of samples from a chemical process, five samples from the process are analyzed daily. In addition, a control sample is analyzed two times each day to check the calibration of the laboratory instruments.

1. How many different sequences of process and control samples are possible each day? Assume that the five process samples are considered identical and that the two control samples are considered identical.

2. How many different sequences of process and control samples are possible if we consider the five process samples to be different and the two control samples to be identical.

3. For the same situation as part (b), how many sequences are possible if the first test of each day must be a control sample?

1.17. In the design of an electromechanical product, seven different components are to be stacked into a cylindrical casing that holds 12 components in a manner that minimizes the impact of shocks. One end of the casing is designated as the bottom and the other end is the top.

1. How many different designs are possible?

2. If the seven components are all identical, how many different designs are possible?

3. If the seven components consist of three of one type of component and four of another type, how many different designs are possible? (more difficult)

1.18. The design of a communication system considered the following questions:

1. How many three-digit phone prefixes that are used to represent a particular geographic area (such as an area code) can be created from the digits 0 through 9?

2. As in part (a), how many three-digit phone prefixes are possible that do not start with 0 or 1, but contain 0 or 1 as the middle digit?

3. How many three-digit phone prefixes are possible in which no digit appears more than once in each prefix?

1.19. A byte is a sequence of eight bits and each bit is either 0 or 1.

1. How many different bytes are possible?

2. If the first bit of a byte is a parity check, that is, the first byte is determined from the other seven bits, how many different bytes are possible?

1.20. A bowl contains 16 chips, of which 6 are red, 7 are white, and 3 are blue. If four chips are talmn at random and without replacement, find the probability that: (a) each of the 4 chips is red; (b) none of the 4 chips is red; (c) there is at least 1 chip of each color.

1.21. Three distinct integers are chosen at random from the first 20 positive integers. Compute the probability that: (a) their stun is even; (b) their product is even.

1.22. There are 5 red chips and 3 blue chips in a bowl. The red chips are numbered 1, 2, 3, 4, 5, respectively, and the blue chips are numbered 1, 2, 3, respectively. If 2 chips are to be drawn at random and without replacement, find the probability that these chips have either the same number or the same color.

1.23. In a lot of 50 light bulbs, there are 2 bad bulbs. An inspector examines 5 bulbs, which are selected at random and without replacement. (a) Find the probability of at least 1 defective bulb among the 5. (b) How many bulbs should be examined so that the probability of finding at least 1 bad bulb exceeds 0.2 ?

1.24. Three winning tickets are drawn from urn of 100 tickets. What is the probability of winning for a person who buys:

1. 4 tickets?
2. only one ticket?

1.25. A drawer contains eight different pairs of socks. If six socks are taken at random and without replacement, compute the probability that there is at least one matching pair among these six socks.

1.26. In a classroom there are n students.

1. What is the probability that at least two students have the same birthday?
2. What is the minimum value of n which secures probability 1/2 that at least two have a common birthday?

1.27. Four mice are chosen (without replacement) from a litter containing two white mice. The probability that both white mice are chosen is twice the probability that neither is chosen. How many mice are there in the litter?

1.28. Suppose there are N different types of coupons available when buying cereal; each box contains one coupon and the collector is seeking to collect one of each in order to win a prize. After buying n boxes, what is the probability p_n that the collector has at least one of each type? (Consider sampling with replacement from a population of N distinct elements. The sample size is $n > N$. Use inclusion-exclusion formula)

1.29. An absent-minded person has to put n personal letters in n addressed envelopes, and he does it at random. What is the probability $p_{m,n}$ that exactly m letters will be put correctly in their envelopes?

1.30. N men run out of a men's club after a fire and each takes a coat and a hat. Prove that:

a) the probability that no one will take his own coat and hat is

$$\sum_{k=1}^{N}(-1)^k \frac{(N-k)!}{N!k!};$$

b) the probability that each man takes a wrong coat and a wrong hat is

$$\left[\sum_{k=2}^{N}(-1)^k \frac{1}{k!}\right]^2 .$$

1.31. You throw $6n$ dice at random. Find the probability that each number appears exactly n times.

1.32. * Mary tosses $n+1$ fair coins and John tosses n fair coins. What is the probability that Mary gets more heads than John?

Conditional Probability

1.33. Bowl I contains 6 red chips and 4 blue chips. Five of these 10 chips are selected at random and without replacement and put in bowl II, which was originally empty. One chip is then drawn at random from bowl II. Given that this chip is blue, find the conditional probability that 2 red chips and 3 blue chips are transferred from bowl I to bowl II.

1.34. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

1.35. We roll a fair four-sided die. If the result is 1 or 2, we roll once more but otherwise, we stop. What is the probability that the sum total of our rolls is at least 4?

1.36. There are three coins in a box. One is a two-headed coin, another is a fair coin, and the third is a biased coin that comes up heads 75 percent of the time. When one of the three coins is selected at random and flipped, it shows heads. What is the probability that it was the two-headed coin?

1.37. Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

1.38. At the station there are three payphones which accept 20p pieces. One never works, another always works, while the third works with probability 1/2. On my way to the metropolis for the day, I wish to identify the reliable phone, so that I can use it on my return. The station is empty and I have just three 20p pieces. I try one phone and it does not work. I try another twice in succession and it works both times. What is the probability that this second phone is the reliable one?

1.39. An insurance company insure an equal number of male and female drivers. In any given year the probability that a male driver has an accident involving a claim is α, independently of other years. The analogous probability for females is β. Assume the insurance company selects a driver at random.

a) What is the probability the selected driver will make a claim this year?

b) What is the probability the selected driver makes a claim in two consecutive years?

c) Let A_1, A_2 be the events that a randomly chosen driver makes a claim in each of the first and second years, respectively. Show that $P(A_2|A_1) \geq P(A_1)$.

d) Find the probability that a claimant is female.

1.40. Three newspapers A, B and C are published in a certain city, and a survey shows that for the adult population 20% read A, 16% B, and 14% C, 8% read both A and B, 5% both A and C, 4% both B and C, and 2% read all three. If an adult chosen at random, find the probability that

a) he reads none of these paper;

b) he reads only one of these papers; and

c) he reads at least A and B if is known that he reads at least one paper.

1.41. Customers are used to evaluate preliminary product designs. In the past, 95% of highly successful products received good reviews, 60% of moderately successful products received good reviews, and 10% of poor products received good reviews. In addition, 40% of products have been highly successful, 35% have been moderately successful, and 25% have been poor products.

1. What is the probability that a product attains a good review?
2. If a new design attains a good review, what is the probability that it will be a highly successful product?
3. If a product does not attain a good review, what is the probability that it will be a highly successful product?

1.42. An inspector working for a manufacturing company has a 99% chance of correctly identifying defective items and a 0.5% chance of incorrectly classifying a good item as defective. The company has evidence that its line produces 0.9% of nonconforming items.

1. What is the probability that an item selected for inspection is classified as defective?
2. If an item selected at random is classified as nondefective, what is the probability that it is indeed good?

1.43. A new analytical method to detect pollutants in water is being tested. This new method of chemical analysis is important because, if adopted, it could be used to detect three different contaminantsorganic pollutants, volatile solvents, and chlorinated compoundsinstead of having to use a single test for each pollutant. The makers of the test claim that it can detect high levels of organic pollutants with 99.7% accuracy, volatile solvents with 99.95% accuracy, and chlorinated compounds with 89.7% accuracy. If a pollutant is not present, the test does not signal. Samples are prepared for the calibration of the test and 60% of them are contaminated with organic pollutants, 27% with volatile solvents, and 13% with traces of chlorinated compounds.

A test sample is selected randomly.

1. What is the probability that the test will signal?

2. If the test signals, what is the probability that chlorinated compounds are present?

1.44. Software to detect fraud in consumer phone cards tracks the number of metropolitan areas where calls originate each day. It is found that 1% of the legitimate users originate calls from two or more metropolitan areas in a single day. However, 30% of fraudulent users originate calls from two or more metropolitan areas in a single day. The proportion of fraudulent users is 0.01%. If the same user originates calls from two or more metropolitan areas in a single day, what is the probability that the user is fraudulent?

1.45. The probability of getting through by telephone to buy concert tickets is 0.92. For the same event, the probability of accessing the vendors Web site is 0.95. Assume that these two ways to buy tickets are independent. What is the probability that someone who tries to buy tickets through the Internet and by phone will obtain tickets?

1.46. The British government has stepped up its information campaign regarding foot and mouth disease by mailing brochures to farmers around the country. It is estimated that 99% of Scottish farmers who receive the brochure possess enough information to deal with an outbreak of the disease, but only 90% of those without the brochure can deal with an outbreak. After the first three months of mailing, 95% of the farmers in Scotland received the informative brochure. Compute the probability that a randomly selected farmer will have enough information to deal effectively with an outbreak of the disease.

1.47. In an automated filling operation, the probability of an incorrect fill when the process is operated at a low speed is 0.001. When the process is operated at a high speed, the probability of an incorrect fill is 0.01. Assume that 30% of the containers are filled when the process is operated at a high speed and the remainder are filled when the process is operated at a low speed.

1. What is the probability of an incorrectly filled container?

2. If an incorrectly filled container is found, what is the probability that it was filled during the high-speed operation?

1.48. An encryption-decryption system consists of three elements: encode, transmit, and decode. A faulty encode occurs in 0.5% of the messages processed, transmission errors occur in 1% of the messages, and a decode error occurs in 0.1% of the messages. Assume the errors are independent.

1. What is the probability of a completely defect-free message?

2. What is the probability of a message that has either an encode or a decode error?

1.49. It is known that two defective copies of a commercial software program were erroneously sent to a shipping lot that has now a total of 75 copies of the program. A sample of copies will be selected from the lot without replacement.

1. If three copies of the software are inspected, determine the probability that exactly one of the defective copies will be found.

2. If three copies of the software are inspected, determine the probability that both defective copies will be found.

3. If 73 copies are inspected, determine the probability that both copies will be found. Hint: Work with the copies that remain in the lot.

1.50. A robotic insertion tool contains 10 primary components. The probability that any component fails during the warranty period is 0.01. Assume that the components fail independently and that the tool fails if any component fails. What is the probability that the tool fails during the warranty period?

1.51. A machine tool is idle 15% of the time. You request immediate use of the tool on five different occasions during the year. Assume that your requests represent independent events.

1. What is the probability that the tool is idle at the time of all of your requests?

2. What is the probability that the machine is idle at the time of exactly four of your requests?

3. What is the probability that the tool is idle at the time of at least three of your requests?

1.52. A lot of 50 spacing washers contains 30 washers that are thicker than the target dimension. Suppose that three washers are selected at random, without replacement, from the lot.

1. What is the probability that all three washers are thicker than the target?

2. What is the probability that the third washer selected is thicker than the target if the first two washers selected are thinner than the target?

3. What is the probability that the third washer selected is thicker than the target?

1.53. Continuation of previous exercise. Washers are selected from the lot at random, without replacement.

1. What is the minimum number of washers that need to be selected so that the probability that all the washers are thinner than the target is less than 0.10?

2. What is the minimum number of washers that need to be selected so that the probability that one or more washers are thicker than the target is at least 0.90?

1.54. The alignment between the magnetic tape and head in a magnetic tape storage system affects the performance of the system. Suppose that 10% of the read operations are degraded by skewed alignments, 5% by off-center alignments, 1% by both skewness and offcenter, and the remaining read operations are properly aligned. The probability of a read error is 0.01 from a skewed alignment, 0.02 from an off-center alignment, 0.06 from both conditions, and 0.001 from a proper alignment. What is the probability of a read error.

1.55. Suppose that a lot of washers is large enough that it can be assumed that the sampling is done with replacement. Assume that 60% of the washers exceed the target thickness.

1. What is the minimum number of washers that need to be selected so that the probability that all the washers are thinner than the target is less than 0.10?

2. What is the minimum number of washers that need to be selected so that the probability that one or more washers are thicker than the target is at least 0.90?

1.56. In a chemical plant, 24 holding tanks are used for final product storage. Four tanks are selected at random and without replacement. Suppose that six of the tanks contain material in which the viscosity exceeds the customer requirements.

1. What is the probability that exactly one tank in the sample contains high viscosity material?

2. What is the probability that at least one tank in the sample contains high viscosity material?

3. In addition to the six tanks with high viscosity levels, four different tanks contain material with high impurities. What is the probability that exactly one tank in the sample contains high viscosity material and exactly one tank in the sample contains material with high impurities?

1.57. Plastic parts produced by an injection-molding operation are checked for conformance to specifications. Each tool contains 12 cavities in which parts are produced, and these parts fall into a conveyor when the press opens. An inspector chooses 3 parts from among the 12 at random. Two cavities are affected by a temperature malfunction that results in parts that do not conform to specifications.

1. What is the probability that the inspector finds exactly one nonconforming part?

2. What is the probability that the inspector finds at least one nonconforming part?

1.58. A bin of 50 parts contains five that are defective. A sample of two is selected at random, without replacement.

1. Determine the probability that both parts in the sample are defective by computing a conditional probability.

2. Determine the answer to part (a) by using the subset approach that was described in this section.

1.59. * The Polya urn model is as follows. We start with an urn which contains one white ball and one black ball. At each second we choose a ball at random from the urn and replace it together with one more ball of the same color. Calculate the probability that when n balls are in the urn, i of them are white.

1.60. You have n urns, the rth of which contains $r-1$ red balls and $n-r$ blue balls, $r=1,\ldots,n$. You pick an urn at random and remove two balls from it without replacement. Find the probability that the two balls are of different colors. Find the same probability when you put back a removed ball.

1.61. A coin shows heads with probability p on each toss. Let π_n be the probability that the number of heads after n tosses is even. Show that $\pi_{n+1}=(1-p)\pi_n+p(1-\pi_n)$ and find π_n.

1.62. There are n similarly biased dice such that the probability of obtaining a 6 with each one of them is the same and equal to p $(0<p<1)$. If all the dice are rolled once, show that p_n, the probability that an odd number of 6's is obtained, satisfies the difference equation

$$p_n+(2p-1)p_{n-1}=p,$$

and hence derive an explicit expression for p_n.

1.63. Dubrovsky sits down to a night of gambling with his fellow officers. Each time he stakes u roubles there is a probability r that he will win and receive back $2u$ roubles (including his stake). At the beginning of the night he has 8000 roubles. If ever he has 256000 roubles he will marry the beautiful Natasha and retire to his estate in the country. Otherwise, he will commit suicide. He decides to follow one of two courses of action:

(i) to stake 1000 roubles each time until the issue is decided;

(ii) to stake everything each time until the issue is decided.

Advise him (a) if $r=1/4$ and (b) if $r=3/4$. What are the chances of a happy ending in each case if he follows your advice?

Independence

1.64. Let the events $A_1,\ A_2,\ldots,\ A_n$ be independent and $P(A_i)=p$ $(i=1,\ 2,\ldots,\ n)$. What is the probability that:

a) at least one of the events will occur?

b) at least m of the events will occur?

c) exactly m of the events will occur?

1.65. Each of four persons fires one shot at a target. Let C_k denote the event that the target is hit by person $k, k=1,2,3,4$. If C_1,C_2,C_3,C_4 are independent and if $\mathbb{P}(C_1)=\mathbb{P}(C_2)=0.7,\mathbb{P}(C_3)=0.9$, and $\mathbb{P}(C_4)=0.4$, compute the probability that (a) all of them hit the target; (b) exactly one hits the target; (c) no one hits the target; (d) at least one hits the target.

1.66. The probability of winning on a single toss of the dice is p. A starts, and if he fails, he passes the dice to B, who then attempts to win on her toss. They continue tossing the dice back and forth until one of them wins. What are their respective probabilities of winning?

1.67. Two darts players throw alternately at a board and the first to score a bull wins. On each of their throws player A has probability p_A and player B p_B of success; the results of different throws are independent. If A starts, calculate the probability that he/she wins.

1.68. * A fair coin is tossed until either the sequence HHH occurs in which case I win or the sequence THH occurs, when you win. What is the probability that you win?

1.69. Let $A_1, \ldots, A_n$ be independent events, with $\mathbb{P}(A_i) < 1$. Prove that there exists an event B with $\mathbb{P}(B) > 0$ such that $B \cap A_i = \emptyset$ for $1 \leq i \leq n$.

1.70. n balls are placed at random into n cells. Find the probability p_n that exactly two cells remain empty.

1.71. An urn contains b black balls and r red balls. One of the balls is drawn at random, but when it is put back in the urn c additional balls of the same color are put in with it. Now suppose that we draw another ball. Show that the probability that the first ball drawn was black given that the second ball drawn was red is $b/(b+r+c)$.

1.72. Suppose every packet of the detergent TIDE contains a coupon bearing one of the letters of the word TIDE. A customer who has all the letters of the word gets a free packet. All the letters have the same possibility of appearing in a packet. Find the probability that a housewife who buys 8 packets will get:

a) one free packet,

b) two free packets.

Chapter 2

Random Variables and Distributions

2.1 Random variables on a countable space

2.1.1 Definitions

Throughout this section we suppose that Ω is countable and $\mathcal{A} = 2^{\Omega}$. A *random variable* X in this case is defined as a map from Ω into $\mathbb{R}$. A random variable stands for an observation of the outcome of a random event. Before the random event we may know the range of X but we do not know its exact value until the random event happens. The *distribution* of a random variable X is defined by

$$\mathbb{P}^X(B) = \mathbb{P}(\{w : X(w) \in B\}) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}[X \in B], \quad B \in \mathcal{B}(\mathbb{R}).$$

Since the set Ω is countable, the range of X is also countable. Suppose that $X(\Omega) = \{x_1, x_2, \ldots\}$. Then the distribution of X is completely determined by the following numbers $p_i^X = \mathbb{P}(X = x_i),\ i \geq 1$. Indeed, for any event $A \in \mathcal{A}$,

$$\mathbb{P}^X(A) = \sum_{x_i \in A} \mathbb{P}[X = x_i] = \sum_{x_i \in A} p_i^X.$$

Definition 2.1.1. Let X be a real-valued random variable on a countable space Ω. Suppose that $X(\Omega) = \{x_1, x_2, \ldots\}$. The *expectation of* X, denoted $\mathbb{E}[X]$, is defined to be

$$\mathbb{E}[X] := \sum_i x_i \mathbb{P}[X = x_i] = \sum_i x_i p_i^X$$

provided this sum makes sense: this is the case when at least one of the following conditions is satisfied

1. Ω is finite;

2. Ω is countable and the series $\sum_i x_i p_i^X$ absolutely convergence;

3. $X \geq 0$ always (in this case, the above sum and hence $\mathbb{E}[X]$ as well may take value $+\infty$.

Remark 1. Since Ω is countable, we denote p_w the probability that the elementary event $w \in \Omega$ happens. Then the expectation of X is given by

$$\mathbb{E}[X] = \sum_{w \in \Omega} X(w) p_w.$$

Let $\mathcal{L}^1$ denote the space of all random variables with finite expectation defined on $(\Omega, \mathcal{A}, \mathbb{P})$. The following facts are straightforward from the definition of expectation.

Theorem 2.1.2. *Let $X, Y \in \mathcal{L}^1$. The following statements hold:*

1. *$\mathcal{L}^1$ is a vector space over $\mathbb{R}$ and*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \quad \forall\, a, b \in \mathbb{R};$$

2. *If $X \geq 0$ then $\mathbb{E}X \geq 0$. Moreover, if $X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$;*

3. *If Z is a bounded random variable then $Z \in \mathcal{L}^1$. Furthermore, if $Z' \in \mathcal{L}^1$ and $|Z'| \leq X \in \mathcal{L}^1$ then $Z' \in \mathcal{L}^1$;*

4. *If $X = \mathbb{I}_A$ is the indicator function of an event A, then $\mathbb{E}[X] = \mathbb{P}(A)$;*

5. *Let $\varphi : \mathbb{R} \to \mathbb{R}$. Then*

$$\mathbb{E}[\varphi(X)] = \sum_i \varphi(x_i) p_i^X = \sum_{w \in \Omega} \varphi(X(w)) p_w$$

if the above series is absolutely convergent.

Remark 2. If $\mathbb{E}[X^2] = \sum_i x_i^2 p_i^X < \infty$, then

$$\mathbb{E}[|X|] = \sum_i |x_i| p_i^X \leq \frac{1}{2} \sum_i (|x_i|^2 + 1) p_i^X = \frac{1}{2}(\mathbb{E}(X^2) + 1) < \infty.$$

Definition 2.1.3. *Variation* of a random variable X is defined to be

$$DX = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

It follows from the linearity of expectation operator that

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

Hence

$$DX = \sum_i x_i^2 p_i^X - \Big(\sum_i x_i p_i^X \Big)^2.$$

2.1.2 Examples

Poisson distribution

X has a Poisson distribution with parameter $\lambda > 0$, denoted $X \sim Poi(\lambda)$, if $X(\Omega) = \{0, 1, \ldots\}$ and

$$\mathbb{P}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

The expectation of X is

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{\lambda} e^{-\lambda} = \lambda.$$

A similar calculation gives us the variance of X,

$$DX = \lambda.$$

Bernoulli distribution

X is Bernoulli with parameter $p \in [0, 1]$, denoted $X \sim Ber(p)$, if it takes only two values 0 and 1 and

$$\mathbb{P}[X = 1] = 1 - \mathbb{P}[X = 0] = p.$$

X corresponds to an experiment with only two outcomes, usually called “success” $(X = 1)$ and “failure” $(X = 0)$. The expectation and variance of X are

$$\mathbb{E}[X] = p, \quad DX = p(1 - p).$$

Binomial distribution

X has Binomial distribution with parameters $p \in [0, 1]$ and $n \in \mathbb{N}$, denoted $X \sim B(n, p)$, if X takes on the values $\{0, 1, \ldots, n\}$ and

$$\mathbb{P}[X = k] = C_n^k p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

One has

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} k\mathbb{P}[X = k] = \sum_{k=0}^{n} kC_n^k p^k (1 - p)^{n-k} \\ &= np \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} (1 - p)^{n-k} = np,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{n} k^2\mathbb{P}[X = k] = \sum_{k=0}^{n} k^2 C_n^k p^k (1 - p)^{n-k} \\ &= n(n - 1)p^2 \sum_{k=2}^{n} C_{n-2}^{k-2} p^{k-2} (1 - p)^{n-k} + np \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} (1 - p)^{n-k} \\ &= n(n - 1)p^2 + np,\end{aligned}$$

thus $DX = np(1 - p)$.

Geometric distribution

One repeatedly performs a sequence of independent Bernoulli trials until achieving the first sucesses. Let X denote the number of failures before reaching the first success. X has a Geometric distribution with parameter $q = 1 - p \in [0, 1]$, denoted $X \sim Geo(q)$,

$$\mathbb{P}[X = k] = q^k p, \quad k = 0, 1, \ldots$$

where p is the probability of success. Then we have

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\mathbb{P}[X = k] = \sum_{k=0}^{\infty} kpq^k = \frac{q}{p}.$$

Moreover, one can easily shows that

$$DX = \frac{q}{p^2}.$$

2.2 Random variables on a general probability space

2.2.1 Definition

Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathcal{B}(\mathbb{R})$ the σ-algebra Borel on $\mathbb{R}$.

Definition 2.2.1. A function $X : \Omega \to \mathbb{R}$ is called $\mathcal{A}$-measurable if

$$X^{-1}(B) := \{w : X(w) \in B\} \in \mathcal{A} \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

An $\mathcal{A}$-measurable function X is called a random variable.

Theorem 2.2.2. *Let $X : \Omega \to \mathbb{R}$. The following statements are equivalent*

1. *X is a random variable.*
2. *$\{w : X(w) \leq a\} \in \mathcal{A}$ for all $a \in \mathbb{R}$.*

Proof. Claim $(1) \Rightarrow (2)$ is self-evident, we will prove: $(2) \Rightarrow (1)$. Let

$$\mathcal{C} = \{B \in \mathcal{B}(\mathbb{R}) : X^{-1}(B) \in \mathcal{A}\}.$$

We have $\mathcal{C}$ is a σ-algebra and it contains all sets with the form $(-\infty, a]$ for every $a \in \mathbb{R}$. Thus $\mathcal{C}$ contains $\mathcal{B}(\mathbb{R})$. On the other hand, $\mathcal{C} \subset \mathcal{B}(\mathbb{R})$, so $\mathcal{C} = \mathcal{B}(\mathbb{R})$. This concludes our proof. □

Example 2.2.3. Let $(\Omega, \mathcal{A})$ be a measurable space. For each subset B of Ω one can verifies that $\mathbb{I}_B$ is a random variable iff $B \in \mathcal{A}$. More general, if $x_i \in \mathbb{R}$ and $A_i \in \mathcal{A}$ for all i belongs to some countable index set I, then $X(w) = \sum_{i \in I} x_i \mathbb{I}_{A_i}(w)$ is also a random variable. We call such random variable X *discrete random variable*. When I is finite then X is called *simple random variable*.

Definition 2.2.4. A function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is called *Borel measurable* if $X^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$ for all $B \in \mathcal{B}(\mathbb{R})$.

Remark 3. It implies from the above definition that every continuous function is Borel. Consequently, all the functions $(x, y) \mapsto x + y,\ (x, y) \mapsto xy,\ (x, y) \mapsto x/y,\ (x, y) \mapsto x \vee y,\ (x, y) \mapsto x \wedge y$ are Borel, where $x \vee y = \max(x, y),\ x \wedge y = \min(x, y)$.

Theorem 2.2.5. *Let $X_1, \ldots, X_d$ be random variables defined on a measurable space $(\Omega, \mathcal{A})$ and $\varphi : \mathbb{R}^d \to \mathbb{R}$ a Borel function. Then $Y = \varphi(X_1, \ldots, X_d)$ is also a random variable.*

Proof. Let: $X(w) = (X_1(w), \ldots, X_d(w))$ is the function on $(\Omega, \mathcal{A})$ and takes values in $\mathbb{R}^d$. For every $a_1, \ldots, a_d \in \mathbb{R}$ we have:

$$X^{-1}\Big(\prod_{i=1}^{d}(-\infty, a_i]\Big) = \bigcap_{i=1}^{d}\{w : X_i(w) \le a_i\} \in \mathcal{A}.$$

This implies $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}(\mathbb{R}^d)$. Hence, for every $C \in \mathcal{B}(\mathbb{R}^d)$, $B := \varphi^{-1}(C) \in \mathcal{B}(\mathbb{R}^d)$. Thus,

$$Y^{-1}(C) = X^{-1}(\varphi^{-1}(C)) \in \mathcal{A},$$

i.e. Y is the random variable. □

Corollary 2.2.6. *If X and Y are random variables, so also are $X \pm Y, XY, X \wedge Y, X \vee Y, |X|, X^+ := X \vee 0, X^- = (-X) \vee 0$ and X/Y (if $Y \neq 0$).*

Theorem 2.2.7. *If $X_1, X_2, \ldots$ are random variables then so are $\sup_n X_n, \inf_n X_n, \limsup_n X_n, \liminf_n X_n$*

It follows from Theorem 2.2.7 that if the sequence of random variables $(X_n)_{n \ge 1}$ point-wise converges to X, i.e. $X_n(w) \to X(w)$ for all $w \in \Omega$, then X is a random variable.

2.2.2 Structure of random variables

Theorem 2.2.8. *Let X be a random variable defined on a probability space $(\Omega, \mathcal{A})$.*

1. *There exists a sequence of discrete random variables which uniformly point-wise converges to X.*

2. *If X is non-negative then there exists a sequence of simple random variables Y_n such that $Y_n \uparrow X$.*

Proof.

1. For each $n \ge 1$, denote $X_n(w) = \frac{k}{n}$ if $\frac{k}{n} \le X(w) < \frac{k+1}{n}$ for some $k \in \mathbb{Z}$. X_n is a discrete random variable and $|X_n(w) - X(w)| \le \frac{1}{n}$ for every $w \in \Omega$. Hence, the sequence (X_n) converges uniformly in w to X.

2. Suppose that $X \ge 0$. For each $n \ge 1$, denote $Y_n(w) = \frac{k}{2^n}$ if $\frac{k}{2^n} \le X(w) < \frac{k+1}{2^n}$ for some $k \in \{0, 1, \ldots, n2^n - 1\}$ and $Y_n(w) = 2^n$ if $X(w) \ge 2^n$. We can easily verify that the sequence of simple random variables (Y_n) satisfying $Y_n(w) \uparrow X(w)$ for all $w \in \Omega$.

□

Definition 2.2.9. Let X be a random variable defined on a measurable space $(\Omega, \mathcal{A})$.

$$\sigma(X) := \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$$

is called *σ-algebra generated by X*.

Theorem 2.2.10. *Let X be a random variable defined on a measurable space $(\Omega, \mathcal{A})$ and Y a function $\Omega \to \mathbb{R}$. Then Y is $\sigma(X)$-measurable iff there exists a Borel function $\varphi : \mathbb{R} \to \mathbb{R}$ such that $Y = \varphi(X)$.*

Proof. The sufficient condition is evident. We prove the necessary condition. Firstly, suppose Y is a discrete random variable taking values $y_1, y_2, \ldots$ Since Y is $\sigma(X)$-measurable, sets $A_n = \{w : Y(w) = y_n\} \in \sigma(X)$. By definition of $\sigma(X)$, there exists a sequence $B_n \in \mathcal{B}(\mathbb{R})$ such that $A_n = X^{-1}(B_n)$. Denote

$$C_n = B_n \setminus \cup_{i=1}^{n-1} B_i \in \mathcal{B}(\mathbb{R}),\ n \geq 1.$$

We have sets C_n are pairwise disjoint and $X^{-1}(C_n) = A_n$ for every n. Consider the Borel function φ defined by

$$\varphi(x) = \sum_{n \geq 1} y_n \mathbb{I}_{C_n}(x),$$

we have $Y = \varphi(X)$.

In general case, by Theorem 2.2.8, there exists a sequence of discrete $\sigma(X)$-measurable functions Y_n which uniformly converges to Y. Thus, there exists Borel functions φ_n such that $Y_n = \varphi_n(X)$. Denote

$$B = \{x \in \mathbb{R} : \exists \lim_n \varphi_n(x)\}.$$

Clearly, $B \in \mathcal{B}(\mathbb{R})$ and $B \supset X(\Omega)$. Let: $\varphi(x) = \lim_n \varphi_n(x) \mathbb{I}_B(x)$. We have $Y = \lim_n Y_n = \lim_n \varphi_n(X) = \varphi(X)$. □

2.3 Distribution Functions

2.3.1 Definition

Definition 2.3.1. Let X be a real valued random variable.

$$F_X(x) = \mathbb{P}[X < x], \quad x \in \mathbb{R},$$

is called *distribution funtion* of X.

One can verifies that $F = F_X$ satisfies the following properties

1. F is non-decreasing: if $x \leq y$ then $F(x) \leq F(y)$;

2. F is left continuous and has right limit at any point;

3. $\lim_{x \to -\infty} F(x) = 0,\ \lim_{x \to +\infty} F(x) = 1.$

On the other hand, for any function $F : \mathbb{R} \to [0,1]$ satisfying the these three conditions there exists a (unique) probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F(x) = \mu((-\infty, x))$, for all $x \in \mathbb{R}$ (See [13], section 2.5.2).

If X and Y has the same distribution function we say X and Y are equal in distribution and denote $X \overset{d}{=} Y$.

Definition 2.3.2. If the distribution function F_X has the form

$$F_X(a) = \mathbb{P}[X < a] = \int_{-\infty}^{a} f_X(x)dx, \quad \forall a \in \mathbb{R}$$

we say that X has a *density function* f.

The density function $f = f_X$ has the following properties:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$.
2. $\mathbb{P}[a < X < b] = \int_a^b f(x)dx$ for any $a < b$. Moreover, for any $A \in \mathcal{B}(\mathbb{R})$, it holds

$$\mathbb{P}[X \in A] = \int_A f(x)dx. \tag{2.1}$$

As a consequence we see that if X has a density then $\mathbb{P}[X = a] = 0$ for all $a \in \mathbb{R}$.

2.3.2 Examples

Uniform distribution $U[a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & if\ a \leq x \leq b, \\ 0 & otherwise, \end{cases}$$

is called the *Uniform distribution on* $[a, b]$ and denoted by $U[a, b]$. The distribution function corresponds to f is

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } x > b. \end{cases}$$

Exponential distribution $Exp(\lambda)$

Suppose $\lambda > 0$. X has a *exponential distribution* with rate λ, denoted $X \sim Exp(\lambda)$, if X takes values in $(0, \infty)$ and its density is given by

$$f_X(x) = \lambda^{-1} e^{-x/\lambda} \mathbb{I}_{(0,\infty)}(x).$$

The distribution function of X is

$$F_X(x) = (1 - \lambda^{-1} e^{-x/\lambda}) \mathbb{I}_{(0,\infty)}(x).$$

Normal distribution $\mathcal{N}(a, \sigma^2)$

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

is called the *Normal distribution* with mean a and variance σ^2 and denoted by $\mathcal{N}(a, \sigma^2)$. When $a = 0$ and $\sigma^2 = 1$, $\mathcal{N}(0, 1)$ is called the *Standard normal distribution.*

Gamma distributiton $\mathcal{G}(\alpha, \lambda)$

$$f_X(x) = \frac{x^{\alpha-1} e^{-x/\lambda}}{\Gamma(\alpha)\lambda^\alpha} \mathbb{I}_{(0,\infty)}(x)$$

is called the *Gamma distribution* with parameters $\alpha, \lambda (\alpha, \lambda > 0)$; Γ denotes the gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$. In particular, an $Exp(\lambda)$ distribution is $\mathcal{G}(1, \lambda)$ distribution. The gamma distribution is frequently a probability model for waiting times; for instance, in life testing, the waiting time until ”death” is a random variable which is frequently modeled with a gamma distribution.

2.4 Expectation

2.4.1 Construction of expectation

Definition 2.4.1. Let X be a simple random variable which can be written in the form

$$X = \sum_{i=1}^n a_i \mathbb{I}_{A_i} \tag{2.2}$$

where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{A}$ for all $i = 1, \ldots, n$. *Expectation* of X (or *integration* of X with respect to probability measure $\mathbb{P}$) is defined to be

$$\mathbb{E}[X] := \sum_{i=1}^n a_i \mathbb{P}(A_i).$$

Denote $\mathcal{L}_s = \mathcal{L}_s(\Omega, \mathcal{A}, \mathbb{P})$ the set of simple random variable. It should be noted that a simple random variable has of course many different representations of the form (2.2). However, $\mathbb{E}[X]$ does not depend on the particular representation chosen for X.

Let X and Y be in $\mathcal{L}_s$. We can write

$$X = \sum_{i=1}^n a_i \mathbb{I}_{A_i}, \text{ and } Y = \sum_{i=1}^n b_i \mathbb{I}_{A_i}.$$

for some subsets A_i which form a measurable partition of Ω. Then for any $\alpha, \beta \in \mathbb{R}$, $\alpha X + \beta Y$ is also in $\mathcal{L}_s$ and

$$\alpha X + \beta Y = \sum_{i=1}^n (a_i + b_i) \mathbb{I}_{A_i}.$$

Thus $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$; that is expectation is linear on $\mathcal{L}_s$. Furthermore, expectation is a positive operator, i.e., if $X \le Y$, we have $a_i \le b_i$ for all i, and thus $\mathbb{E}[X] \le \mathbb{E}[Y]$.

Next we define expetation for non-negative random variables. For X non-negative, i.e. $X(\Omega) \subset [0,\infty]$, denote

$$\mathbb{E}[X] = \sup\{\mathbb{E}Y : Y \in \mathcal{L}_s \text{ and } 0 \le Y \le X\}. \tag{2.3}$$

This supremum always exists in $[0,\infty]$. It follows from the positivity of expectation operator that the definition above for $\mathbb{E}[X]$ coincides with Definition 2.4.1 on $\mathcal{L}_s$.

Note that $\mathbb{E}X \ge 0$ but it may happen that $\mathbb{E}X = +\infty$ even when X is never equal to $+\infty$.

Definition 2.4.2. 1. A random variable X is called *integrable* if $\mathbb{E}[|X|] < \infty$. In this case, its expectation is defined to be

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]. \tag{2.4}$$

We also write $\mathbb{E}[X] = \int X(w)d\mathbb{P}(w) = \int Xd\mathbb{P}$.

2. If $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are not both equal to $+\infty$ then the expectation of X is still defined and given by (2.4) where we use the convention that $+\infty + a = +\infty$ and $-\infty + a = -\infty$ for any $a \in \mathbb{R}$.

If $X \ge 0$ then $X = X^+$ and $X^- = 0$. Therefore Definition 2.4.2 again coincides with definition (2.3) on set of non-negative random variables. We denote by $\mathcal{L}^1 = \mathcal{L}^1(\Omega,\mathcal{A},\mathbb{P})$ the set of integrable random variables.

Lemma 2.4.3. *Let X be a non-negative random variable and $(X_n)_{n\ge1}$ a sequence of simple random variables increasing to X. Then $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$ (even if $\mathbb{E}[X] = \infty$).*

Proof. We have $(\mathbb{E}X_n)_{n\ge1}$ is the increasing sequence and upper bounded by $\mathbb{E}X$ by Definition (2.3) so $(\mathbb{E}X_n)_{n\ge1}$ is convergent to a with $a \le \mathbb{E}X$. To prove $a = \mathbb{E}X$, we only show that for every simple random variable Y satisfying $0 \le Y \le X$, we have $\mathbb{E}Y \le a$.

Indeed, suppose Y takes m different values $y_1,\dots,y_m$. Let $A_k = \{w : Y(w) = y_k\}$. For each $\epsilon \in (0,1]$, consider the sequence $Y_{n,\epsilon} = (1-\epsilon)Y\mathbb{I}_{\{(1-\epsilon)Y\le X_n\}}$. We have $Y_{n,\epsilon}$ is simple random variable, $Y_{n,\epsilon} \le X_n$ so

$$\mathbb{E}Y_{n,\epsilon} \le \mathbb{E}X_n \le a \text{ for every } n. \tag{2.5}$$

On the other hand, $Y \le \lim_n X_n$ so for every $w \in \Omega$, there exists $n = n(w)$ such that $(1-\epsilon)Y(w) \le X_n(w)$, i.e. $A_k \cap \{w : (1-\epsilon)Y(w) \le X_n(w)\} \to A_k$ as $n \to \infty$. We have

$$\begin{aligned}\mathbb{E}Y_{n,\epsilon} &= (1-\epsilon)\sum_{k=1}^m y_k\mathbb{P}\Big(A_k \cap [(1-\epsilon)Y \le X_n]\Big)\\ &\to (1-\epsilon)\sum_{k=1}^m y_k\mathbb{P}(A_k) = (1-\epsilon)\mathbb{E}Y, \quad \text{as } n\to\infty.\end{aligned}$$

Asscociate with (2.5), we have $(1-\epsilon)\mathbb{E}Y \le a$ for every $\epsilon \in (0,1]$, i.e. $\mathbb{E}Y \le a$. □

Theorem 2.4.4. *1. $\mathcal{L}^1$ is a vector space on $\mathbb{R}$ and expectation is an linear operator on $\mathcal{L}^1$, i.e., for any $X, Y \in \mathcal{L}^1$ and $x, y \in \mathbb{R}$, one has $\alpha X + \beta Y \in \mathcal{L}^1$ and*

$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}X + \beta \mathbb{E}Y.$$

2. If $0 \leq X \leq Y$ and $Y \in \mathcal{L}^1$ then $X \in \mathcal{L}^1$ and $\mathbb{E}X \leq \mathbb{E}Y$.

Proof. Statement 2 follows exactly from equation 2.3. To prove statement 1, firstly we remark that if X and Y are two non-negative random variables and $\alpha, \beta \geq 0$, by Theorem 2.2.8 there exist two increasing non-negative sequences (X_n) and (Y_n) in $\mathcal{L}_s$ converging to X and Y respectively. Hence, $\alpha X_n + \beta Y_n$ are also simple non-negative random variables, and convege to $\alpha X + \beta Y$. Applying linear and non-negative properties of expectation operator on $\mathcal{L}_s$ and Lemma 2.4.3, we have $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}X + \beta \mathbb{E}Y$.

Now we prove Theorem 2.4.4. Consider two random variables $X, Y \in \mathcal{L}^1$. Since $|\alpha X + \beta Y| \leq |\alpha||X| + |\beta||Y|$, $\alpha X + \beta Y \in \mathcal{L}^1$. We have: if $\alpha > 0$,

$$\mathbb{E}(\alpha X) = \mathbb{E}((\alpha X)^+) - \mathbb{E}((\alpha X)^-) = \mathbb{E}(\alpha(X^+)) - \mathbb{E}(\alpha(X^-)) = \alpha\mathbb{E}(X^+) - \alpha\mathbb{E}(X^-) = \alpha\mathbb{E}X.$$

Similarly to $\alpha < 0$, we also have

$$\mathbb{E}(\alpha X) = \mathbb{E}((\alpha X)^+) - \mathbb{E}((\alpha X)^-) = \mathbb{E}(-\alpha(X^-)) - \mathbb{E}(-\alpha(X^+)) = -\alpha\mathbb{E}(X^-) + \alpha\mathbb{E}(X^+) = \alpha\mathbb{E}X,$$

i.e.

$$\mathbb{E}(\alpha X) = \alpha\mathbb{E}(X) \text{ for every } \alpha \in \mathbb{R}. \tag{2.6}$$

On the other hand, let $Z = X + Y$ we have $Z^+ - Z^- = X + Y = X^+ + Y^+ - (X^- + Y^-)$, so $Z^+ + X^- + Y^- = Z^- + X^+ + Y^+$. Thus $\mathbb{E}(Z^+) + \mathbb{E}(X^-) + \mathbb{E}(Y^-) = \mathbb{E}(Z^-) + \mathbb{E}(X^+) + \mathbb{E}(Y^+)$, then

$$\mathbb{E}Z = \mathbb{E}(Z^+) - \mathbb{E}(Z^-) = \mathbb{E}(X^+) + \mathbb{E}(Y^+) - \mathbb{E}(X^-) - \mathbb{E}(Y^-) = \mathbb{E}X + \mathbb{E}Y.$$

Asscociate with (2.6) we obtain

$$\mathbb{E}(\alpha X + \beta Y) = \mathbb{E}(\alpha X) + \mathbb{E}(\beta Y) = \alpha\mathbb{E}X + \beta\mathbb{E}Y.$$

□

An event A happens *almost surely* if $\mathbb{P}(A) = 1$. Thus we say X equals Y almost surely if $\mathbb{P}[X = Y] = 1$ and denote $X = Y$ a.s.

Corollary 2.4.5. *1. If $Y \in \mathcal{L}^1$ and $|X| \leq Y$, then $X \in \mathcal{L}^1$.*

2. If $X \geq 0$ a.s. and $\mathbb{E}(X) < \infty$, then $X < \infty$ a.s.

3. If $\mathbb{E}(|X|) = 0$ then $X = 0$ a.s.

Proof. 2) Let $A = \{w : X(w) = \infty\}$. For every n, we have $X(w) \geq X(w)\mathbb{I}_A(w) \geq n\mathbb{I}_A(w)$ so $\mathbb{E}(X) \geq n\mathbb{P}(A)$ for every n. Thus $\mathbb{P}(A) \leq \frac{\mathbb{E}(X)}{n} \to 0$ as $n \to \infty$. From this, we have $\mathbb{P}(A) = 0$.

3) Let $A_n = \{w : |X(w)| \geq 1/n\}$. We have $(A_n)_{n\geq 1}$ is the decreasing sequence and $\mathbb{P}(X \neq 0) = \lim_{n\to\infty} \mathbb{P}(A_n)$. Moreover,

$$\frac{1}{n}\mathbb{I}_{A_n}(w) \leq |X(w)|\mathbb{I}_{A_n}(w) \leq |X(w)|$$

so $\mathbb{P}(A_n) \leq n\mathbb{E}|X| = 0$ for every n. Thus $\mathbb{P}(A) = 0$ i.e. $X = 0$ a.s. □

Theorem 2.4.6. *Let X and Y be integrable random variables. If $X = Y$ a.s. then $\mathbb{E}[X] = \mathbb{E}[Y]$.*

Proof. Firstly, we consider the case: X and Y are non-negative. Let $A = \{w : X(w) \neq Y(w)\}$. We have $\mathbb{P}(A) = 0$. Moreover,

$$\mathbb{E}Y = \mathbb{E}(Y\mathbb{I}_A + Y\mathbb{I}_{A^c}) = \mathbb{E}(Y\mathbb{I}_A) + \mathbb{E}(Y\mathbb{I}_{A^c}) = \mathbb{E}(Y\mathbb{I}_A) + \mathbb{E}(X\mathbb{I}_{A^c}).$$

Suppose (Y_n) is a sequence of simple random variables increasing to Y. Hence, $(Y_n\mathbb{I}_A)$ is also a sequence of simple random variables increasing to $(Y\mathbb{I}_A)$. Suppose for each $n \geq 1$, the random variable Y_n is bouned by N_n, so

$$0 \leq \mathbb{E}(Y_n\mathbb{I}_A) \leq \mathbb{E}(N_n\mathbb{I}_A) = N_n\mathbb{P}(A) = 0$$

for each n. Hence $\mathbb{E}(Y\mathbb{I}_A) = 0$. Similarly, $\mathbb{E}(X\mathbb{I}_A) = 0$. Thus $\mathbb{E}Y = \mathbb{E}X$.

In general case, from $X = Y$ a.s. we can easily find that $X^+ = Y^+$ and $X^- = Y^-$ a.s.. Thus, we also have $\mathbb{E}X = \mathbb{E}(X^+) - \mathbb{E}(X^-) = \mathbb{E}(Y^+) - \mathbb{E}(Y^-) = \mathbb{E}Y$. □

2.4.2 Some limit theorems

Theorem 2.4.7 (Monotone convergence theorem)**.** *If the random variables X_n are non-negative and increasing a.s. to X, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$ (even if $\mathbb{E}[X] = \infty$).*

Proof. For each n, let $(Y_{n,k})_{k\geq 1}$ be a sequence of simple random variables increasing to X_n and let $Z_k = \max_{n\leq k} Y_{n,k}$. Then $(Z_k)_{k\geq 1}$ is the sequence of simple non-negative random variables, and thus there exists $Z = \lim_{k\to\infty} Z_k$. Also

$$Y_{n,k} \leq Z_k \leq X \quad \forall n \leq k$$

which implies that

$$X_n \leq Z \leq X \quad a.s.$$

Next let $n \to \infty$ we have $Z = X$ a.s. Since expectation is a positive operator, we have

$$\mathbb{E}Y_{n,k} \leq \mathbb{E}Z_k \leq \mathbb{E}X_k \quad \forall n \leq k.$$

Fix n and let $k \to \infty$, using Lemma 2.4.3 we obtain

$$\mathbb{E}X_n \leq \mathbb{E}Z \leq \lim_{k\to\infty} \mathbb{E}X_k.$$

Now let $n \to \infty$ to obtain

$$\lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}Z \leq \lim_{k\to\infty} \mathbb{E}X_k.$$

Since the left and right sides are the same, $X = Z$ a.s., we deduce the result. □

Theorem 2.4.8 (Fatou's lemma)**.** *If the random variables X_n satisfy $X_n \geq Y$ a.s for all n and some $Y \in L^1$ no . Then*

$$\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

Proof. Firstly we prove Theorem to the case $Y = 0$. Let $Y_n = \inf_{k\geq n} X_k$. We have (Y_n) is the sequence of non-decreasing random variables and

$$\lim_{n\to\infty} Y_n = \liminf_{n\to\infty} X_n.$$

Since $X_n \geq Y_n$, we have $\mathbb{E}X_n \geq \mathbb{E}Y_n$. Asscociate with monotone convergence theorem to the sequence Y_n, we obtain

$$\liminf_{n\to\infty} \mathbb{E}X_n \geq \lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}(\lim_{n\to\infty} Y_n) = \mathbb{E}(\liminf_{n\to\infty} X_n).$$

The general case follows from appling the above result to the sequence of non-negative random variables $\hat{X}_n := X_n - Y$. □

Theorem 2.4.9 (Lebesgue's dominated convergence theorem)**.** *If the random variables X_n converge a.s. to X and $\sup_n |X_n| \leq Y$ a.s. for some $Y \in L^1$. We have $X, X_n \in L^1$ and*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|] = 0.$$

Proof. Since $|X| \leq Y$, $X \in L^1$. Let $Z_n = |X_n - X|$. Since $Z_n \geq 0$ and $-Z_n \geq -2Y$, applying Fatou Lemma to Z_n and $-Z_n$, we obtain

$$0 = \mathbb{E}(\liminf_{n\to\infty} Z_n) \leq \liminf_{n\to\infty} \mathbb{E}Z_n \leq \limsup_{n\to\infty} \mathbb{E}Z_n = -\liminf_{n\to\infty} \mathbb{E}(-Z_n) \leq -\mathbb{E}(\liminf_{n\to\infty}(-Z_n)) = 0.$$

Thus $\lim_{n\to\infty} \mathbb{E}Z_n = 0$ i.e. $\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0$. □

2.4.3 Some inequalities

Theorem 2.4.10. 1. *(Cauchy-Schwarz's inequality) If $X, Y \in L^2$ then $XY \in L^1$ and*

$$|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2). \tag{2.7}$$

2. *$L^2 \subset L^1$ and if $X \in L^2$, then $(\mathbb{E}X)^2 \leq \mathbb{E}(X^2)$.*

3. *L^2 is a vector space on $\mathbb{R}$, i.e., for any $X, Y \in L^2$ and $\alpha, \beta \in \mathbb{R}$, we have $\alpha X + \beta Y \in L^2$.*

Proof. If $\mathbb{E}(X^2)\mathbb{E}(Y^2) = 0$ then $XY = 0$ a.s. Thus $|\mathbb{E}(XY)|^2 = \mathbb{E}(X^2)\mathbb{E}(Y^2) = 0$.

If $\mathbb{E}(X^2)\mathbb{E}(Y^2) \neq 0$, applying the inequality $2|ab| \leq a^2 + b^2$ for $a = X/\sqrt{\mathbb{E}(X^2)}$ and $b = Y/\sqrt{\mathbb{E}(Y^2)}$ and then taking expectation for two sides, we obtain

$$2\mathbb{E}\Big(\frac{XY}{\sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}}\Big) \leq \mathbb{E}\Big(\frac{X^2}{\mathbb{E}(X^2)}\Big) + \mathbb{E}\Big(\frac{Y^2}{\mathbb{E}(Y^2)}\Big) = 2.$$

Hence we have (2.7).

Applying (2.7) for $Y = 1$ we obtain the second claim. The third claim follows from (2.7) and the linearity of expectation. □

If $X \in L^2$, we denote

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

DX is called the *variance* of X. Using the linearity of expectation operator, one can verify that $DX = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

Theorem 2.4.11. *1. (Markov's inequality) Suppose $X \in L^1$, then for any $a > 0$, it holds*

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

2. (Chebyshev's inequality) Suppose $X \in L^2$, then for any $a > 0$, it holds

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{DX}{a^2}.$$

Proof. 1) Since $a\mathbb{I}_{\{|X|\geq a\}}(w) \leq |X(w)|\mathbb{I}_{\{|X|\geq a\}}(w) \leq |X(w)|$ for every $w \in \Omega$. Taking expectation for two sides, we obtain $a\mathbb{P}(|X| \geq a) \leq \mathbb{E}(|X|)$.

2) Applying Markov's inequality', we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) = \mathbb{P}(|X - \mathbb{E}X|^2 \geq a^2) \leq \frac{DX}{a^2}.$$

□

2.4.4 Expectation of random variable with density

Theorem 2.4.12. *Suppose that X has a density function f. Let $h : \mathbb{R} \to \mathbb{R}$ be a Borel function. We have*

$$\mathbb{E}(h(X)) = \int h(x)f(x)dx.$$

provided that either $\mathbb{E}(|h(X)|) < \infty$ or h non-negative.

Proof. Firstly, we consider the case $h \geq 0$. Then there exists a sequence of simple non-negative Borel functions (h_n) increasing to h. Suppose $h_n = \sum_{i=1}^{k_n} a_i^n \mathbb{I}_{A_i^n}$ for $a_i^n \in \mathbb{R}^+$ and $A_i^n \in \mathcal{B}(\mathbb{R})$ for every i. By monotone convergence theorem

$$\mathbb{E}(h(X)) = \mathbb{E}(\lim_n h_n(X)) = \lim_n \mathbb{E}(h_n(X)) = \lim_n \sum_{i=1}^{k_n} h_n(a_i^n)\mathbb{P}[X \in A_i^n].$$

Applying the property (2.1) and monotone convergence theorem, we obtain

$$\mathbb{E}(h(X)) = \lim_n \sum_{i=1}^{k_n} h_n(a_i^n) \int_{A_i^n} f(x)dx = \lim_n \int f(x)h_n(x)dx = \int f(x)h(x)dx.$$

Thus, if h is non-negative, we usually have $\mathbb{E}(h(X)) = \int h(x)f(x)dx$.

In general case, applying above result for h^+ and h^- we deduce this proof. □

Example 2.4.13. Let $X \sim Exp(1)$. Applying Theorem 2.4.12 for $h(x) = x$ and $h(x) = x^2$ respectively, we have

$$\mathbb{E}X = \int_0^\infty xe^{-x}dx = 1,$$

and

$$\mathbb{E}X^2 = \int_0^\infty x^2e^{-x}dx = 2.$$

2.5 Random elements

2.5.1 Definitions

Definition 2.5.1. Let $(E, \mathcal{E})$ be a measure space. A function $X : \Omega \to E$ is called $\mathcal{A}/\mathcal{E}$*-measurable* or *random element* if $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{E}$. The function

$$\mathbb{P}^X(B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{E},$$

is called *probablity distribution* of X on $(E, \mathcal{E})$.

When $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we call X a *random vector.*

Let $X = (X_1, \dots, X_d)$ be a d-dimensional random vector defined on $(\Omega, \mathcal{A}, \mathbb{P})$. The *distribution function* of X is defined by

$$F(x) = \mathbb{P}[X < x] = \mathbb{P}[X_1 < x_1, \dots, X_d < x_d], \quad x \in \mathbb{R}^d.$$

We can easily verify that F satisfying the following properties:

1. $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}^d$.
2. $\lim_{x_k \to -\infty} F(x) = 0$ for all $k = 1, \dots, d$.
3. $\lim_{x_1 \to +\infty, \dots, x_d \to +\infty} F(x) = 1$.
4. F is left continuous.

The random vector X has a density $f : \mathbb{R}^d \to \mathbb{R}^+$ if f is a non-negative Borel measurable function satisfying

$$F(x) = \int_{u<x} f(u) du, \quad \text{for any } x \in \mathbb{R}^d.$$

This equation implies that

$$\mathbb{P}[X \in B] = \int_B f(x) dx, \quad \text{vi mi } B \in \mathcal{B}(\mathbb{R}^d).$$

In particular, we have

$$\mathbb{P}[X_1 \in B_1] = \mathbb{P}[X \in B_1 \times \mathbb{R}^{d-1}] = \int_{B_1} \Big(\int_{\mathbb{R}^{d-1}} f(x_1, \dots, x_d) dx_2 \dots dx_d \Big) dx_1 \text{ for all } B_1 \in \mathcal{B}(\mathbb{R}^d).$$

This implies that if $X = (X_1, \dots, X_d)$ has a density f then X_1 also has a density given by

$$f_{X_1}(x_1) = \int_{\mathbb{R}^{d-1}} f(x_1, x_2, \dots, x_d) dx_2 \dots dx_d, \text{ for all } x_1 \in \mathbb{R}. \tag{2.8}$$

A similar argument as the proof Theorem 2.4.12 yields,

Theorem 2.5.2. *Let $X = (X_1, \dots, X_d)$ be a random vector which has density function f, $\varphi : \mathbb{R}^d \to \mathbb{R}$ a Borel measurable function. We have*

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x) f(x) dx$$

provided that φ is non-negative or $\int_{\mathbb{R}^d} |\varphi(x)| f(x) dx < \infty$.

2.5.2 Example

Multivariate normal distribution

Let $a = (a_1, \dots, a_d)$ be a d-dimensional vector and $M = (m_{i,j})_{i,j=1}^d$ a $d \times d$-square matrix. Suppose that M is symmetric and positive define. Denote $A = M^{-1}$. The random vector $X = (X_1, \dots, X_d)$ has normal distribution $\mathcal{N}(a, M)$ if its density p verifies

$$p(x) = \frac{\sqrt{\det A}}{(2\pi)^{d/2}} \exp\Big\{ -\frac{1}{2}(x-a)A(x-a)^*\Big\},$$

where $(x-a)A(x-a)^* = \sum_i \sum_j a_{ij}(x_i - a_i)(x_j - a_j)$.

Polynomial distribution

The d-dimensional random vector X has a *polynomial distribution* with parameters $n, p_1, \dots, p_d$, denoted by $X \sim MUT(n; p_1, \dots, p_d)$, for $n \in \mathbb{N}^*$ and $p_1, \dots, p_d \geq 0$, if

$$\mathbb{P}[X_1 = k_1, \dots, X_d = k_d] = \frac{n!}{k_1!k_2!\dots k_{d+1}!} p_1^{k_1} p_2^{k_2} \dots p_{d+1}^{k_{d+1}},$$

where $p_{d+1} = 1 - (p_1 + \ldots + p_d),\ 0 \leq k_i \leq n$ and $k_{d+1} = n - (k_1 + \ldots + k_d) \geq 0$.

2.5.3 Density of function of random vectors

Using Theorem 2.5.2 and the change of variables formula we have the following useful result.

Theorem 2.5.3. *Let $X = (X_1, \dots, X_n)$ have a joint density f. Suppose $g : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable and injective, with Jacobian given by*

$$J_g(x) = \Big(\frac{\partial g_i}{\partial x_j}(x)\Big)_{i,j=1,\dots,d}$$

never vanishes. Then $Y = g(X)$ has density

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))|\det J_{g^{-1}}(y)| & \textit{if } y \in g(\mathbb{R}^d) \\ 0 & \textit{otherwise.} \end{cases}$$

2.6 Independent random variables

Definition 2.6.1.

1. The sub-σ-algebras $(\mathcal{A}_i)_{i\in I}$ of $\mathcal{A}$ are *independent* if for all finite subset J of I and for all $A_i \in \mathcal{A}_i$,

$$\mathbb{P}(\cap_{i\in J} A_i) = \prod_{i\in J} \mathbb{P}(A_i).$$

2. The $(E_i, \mathcal{E}_i)$-valued random variables $(X_i)_{i\in I}$ are *independent* if the σ-algebras $(X_i^{-1}(\mathcal{E}_i))_{i\in I}$ are independent.

A class $\mathcal{C}$ of subsets of Ω is called a π-system if is closes under finite intersections, so that $A \cap B \in \mathcal{C}$ implies $A, B \in \mathcal{C}$. Furthermore, a class $\mathcal{D}$ is a λ-system if contains Ω and is closed under proper differences and increasing limits, i.e.,

- $A_1, A_2, \ldots \in \mathcal{D}$ with $A_n \uparrow A$ implies $A \in \mathcal{D}$;
- $A, B \in \mathcal{D}$ with $A \subset B$ implies $B \backslash A \in \mathcal{D}$.

Lemma 2.6.2 (Monotone classes)**.** *Let $\mathcal{C}, \mathcal{D}$ be classes of subsets of Ω where $\mathcal{C}$ is a π-system and $\mathcal{D}$ is a λ-system such that $\mathcal{C} \subset \mathcal{D}$. Then $\sigma(\mathcal{C}) \subset \mathcal{D}$.*

Lemma 2.6.3. *Let $\mathcal{G}$ and $\mathcal{F}$ be sub-σ-algebras of $\mathcal{A}$. Let $\mathcal{G}_1$ and $\mathcal{F}_1$ be π-systems such that $\sigma(\mathcal{G}_1) = \mathcal{G}$ and $\sigma(\mathcal{F}_1) = \mathcal{F}$. Then $\mathcal{G}$ is independent of $\mathcal{F}$ if $\mathcal{F}_1$ and $\mathcal{G}$ are independent, i.e.,*

$$\mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G), \quad F \in \mathcal{F}_1,\ G \in \mathcal{G}_1.$$

Proof. Suppose that $\mathcal{F}_1$ and $\mathcal{G}_1$ are independent. We fix any $F \in \mathcal{F}_1$ and define

$$\sigma_F = \{G \in \mathcal{G} : \mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G)\}.$$

Then σ_F is a λ-system containing π-system $\mathcal{G}_1$. Applying monotone classes theorem, we have $\sigma_F = \mathcal{G}$, it means that

$$\mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G), \quad F \in \mathcal{F}_1,\ G \in \mathcal{G}.$$

Next, for any $G \in \mathcal{G}$ we define

$$\sigma_G = \{F \in \mathcal{F} : \mathbb{P}(F \cap G) = \mathbb{P}(F)\mathbb{P}(G)\}.$$

We also have that σ_G is a λ-system containing π-system $\mathcal{F}_1$ so that $\sigma_G = \mathcal{F}$, which yields the desired property. □

Theorem 2.6.4. *Let X and Y be two random variables. The following statements are equivalent:*

(i) *X is independent of Y;*

(ii) *$F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$;*

(iii) *$f(X)$ and $g(Y)$ are independent for any Borel functions $f, g : \mathbb{R} \to \mathbb{R}$;*

(iv) *$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ for any Borel function $f, g : \mathbb{R} \to \mathbb{R}$ which are either positive or bounded.*

Proof. $(i) \Rightarrow (ii)$: Suppose X be independent of Y, then two events $\{w : X(w) < x\}$ v $\{w : Y(w) < y\}$ are also independent for every $x, y \in \mathbb{R}$. We have (ii).

$(ii) \Rightarrow (i)$: Since the set of events $\{w : X(w) < x\},\ x \in \mathbb{R}$, is a π-system generating $\sigma(X)$ and $\{w : X(w) < y\},\ y \in \mathbb{R}$, is a π-system generating $\sigma(Y)$, so applying Lemma 2.6.3 we have X is independent of Y.

$(i) \Rightarrow (iii)$: For every $A, B \in \mathcal{B}(\mathbb{R})$, we have $f^{-1}(A), g^{-1}(B) \in \mathcal{B}(\mathbb{R})$ then

$$\begin{aligned}&\mathbb{P}(f(X) \in A, g(Y) \in B) = \mathbb{P}(X \in f^{-1}(A), Y \in g^{-1}(B))\\&= \mathbb{P}(X \in f^{-1}(A))\mathbb{P}(Y \in g^{-1}(B)) = \mathbb{P}(f(X) \in A)\mathbb{P}(g(Y) \in B).\end{aligned}$$

Thus, $f(X)$ is independent of $g(Y)$.

$(iii) \Rightarrow (i)$: We choose $f(x) = g(x) = x$.

$(i) \Rightarrow (iv)$: Since (i) is equivalent of (iii), we only prove

$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ for every random variable which is integrable or non-negative X and Y.

Firstly, we suppose that: X and Y are non-negative. By Theorem 2.2.8 there exists a sequence of simple random variables $X_n = \sum_{i=1}^{k_n} a_i\mathbb{I}_{A_i}$ increasing to X and $Y_n = \sum_{j=1}^{l_n} b_j\mathbb{I}_{B_j}$ increasing to Y for $A_i \in \sigma(X)$ v $B_i \in \sigma(Y)$. Applying monotone convergence theorem, we have

$$\begin{aligned}\mathbb{E}(XY) &= \lim_{n\to\infty} \mathbb{E}(X_nY_n) = \lim_{n\to\infty} \sum_{i=1}^{k_n}\sum_{j=1}^{l_n} a_ib_j\mathbb{P}(A_iB_j) = \lim_{n\to\infty} \sum_{i=1}^{k_n}\sum_{j=1}^{l_n} a_ib_j\mathbb{P}(A_i)\mathbb{P}(B_j)\\&= \lim_{n\to\infty} \Big(\sum_{i=1}^{k_n} a_i\mathbb{P}(A_i)\Big)\Big(\sum_{j=1}^{l_n} b_j\mathbb{P}(B_j)\Big) = \lim_{n\to\infty} \mathbb{E}(X_n)\mathbb{E}(Y_n) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

In general case, we write $X = X^+ - X^-$ v $Y = Y^+ - Y^-$. Since (iii), we have $X^\pm$ are independent of $Y^\pm$, then

$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}(X^+Y^+) + \mathbb{E}(X^-Y^-) - \mathbb{E}(X^+Y^-) - \mathbb{E}(X^-Y^+)\\&= \mathbb{E}(X^+)\mathbb{E}(Y^+) + \mathbb{E}(X^-)\mathbb{E}(Y^-) - \mathbb{E}(X^+)\mathbb{E}(Y^-) - \mathbb{E}(X^-)\mathbb{E}(Y^+) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

$(iv) \Rightarrow (i)$: Choose $f = \mathbb{I}_{(-\infty,x)}$ and $g = \mathbb{I}_{(-\infty,y)}$.

$(iv) \Rightarrow (v)$ is evident. □

2.7 Covariance

Definition 2.7.1. The *covariance* of random variables $X, Y \in L^2$ is defined by

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

The *correlation coefficient* of $X, Y \in L^2$ is

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{DXDY}}.$$

Using the linearity of expectation operator, we have

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

Furthermore, it follows from Cauchy-Schwarz's inequality that $|cov(X, Y)| \leq \sqrt{DXDY}$, it means

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

Example 2.7.2. Let X and Y be independent random variables whose distributions are $N(0,1)$. Denote $Z = XY$ and $T = X - Y$. We have

$$cov(Z,T) = \mathbb{E}(XY(X-Y)) - \mathbb{E}(XY)\mathbb{E}(X-Y) = 0,$$

and

$$cov(Z,T^2) = \mathbb{E}(XY(X-Y)^2) - \mathbb{E}(XY)\mathbb{E}((X-Y)^2) = -2,$$

since $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y = 0$, $\mathbb{E}(X^3Y) = \mathbb{E}(X^3)\mathbb{E}Y = 0$, $\mathbb{E}(XY^3) = \mathbb{E}X\mathbb{E}(Y^3) = 0$ and $\mathbb{E}(X^2Y^2) = \mathbb{E}(X^2)\mathbb{E}(Y^2) = 1$. Thus Z and T are uncorrelated random variables but not independent.

Proposition 2.7.3. *Let $(X_n)_{n\geq 1}$ be a sequence of pair-wise uncorrelated random variables. Then*

$$D(X_1 + \ldots + X_n) = D(X_1) + \ldots + D(X_n).$$

Proof. We have

$$\begin{aligned} D(X_1 + \ldots + X_n) &= \mathbb{E}\Big[\Big((X_1 - \mathbb{E}X_1) + \ldots (X_n - \mathbb{E}X_n)\Big)^2\Big] \\ &= \sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}X_i)^2] + 2\sum_{1\leq i<j\leq n} \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] \\ &= \sum_{i=1}^n \mathbb{E}[(X_i - \mathbb{E}X_i)^2] = \sum_{i=1}^n D(X_i), \end{aligned}$$

since $\mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0$ for any $1 \leq i < j \leq n$. □

2.8 Conditional Expectation

2.8.1 Definition

Definition 2.8.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X an integrable random variable. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{A}$. Then there exists a random variable Y such that

1. Y is $\mathcal{G}$-measurable,
2. $\mathbb{E}[|Y|] < \infty$,
3. for every set $G \in \mathcal{G}$, we have
$$\int_G Y d\mathbb{P} = \int_G X d\mathbb{P}.$$

Moreover, if Z is another random variable with these properties then $\mathbb{P}[Z = Y] = 1$. Y is called a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ of X given $\mathcal{G}$, and we write $Y = \mathbb{E}[X|\mathcal{G}]$, a.s.

We often write $\mathbb{E}[X|Z_1, Z_2, \ldots]$ for $\mathbb{E}[X|\sigma(Z_1, Z_2, \ldots)]$.

2.8.2 Examples

Example 2.8.2. Let X be an integrable random variable and $\mathcal{G} = \sigma(A_1, \ldots, A_m)$ where $(A_i)_{1\le i\le m}$ is a measurable partition of Ω. Suppose that $\mathbb{P}(A_i) > 0$ for all $i = 1, \ldots, m$. Then

$$\mathbb{E}(X|\mathcal{G}) = \sum_{i=1}^{n} \Big(\frac{1}{\mathbb{P}(A_i)} \int_{A_i} X d\mathbb{P}\Big) I_{A_i}.$$

Example 2.8.3. Let X and Z be random variables whose joint density is $f_{X,Z}(x,z)$. We know that $f_Z(z) = \int_{\mathbb{R}} f_{X,Z}(x,z)dx$ is density of Z. Define the elementary conditional density $f_{X|Z}$ of X given Z by

$$f_{X|Z}(x|z) := \begin{cases} \frac{f_{X,Z}(x,z)}{f_Z(z)} & \text{if } f_Z(z) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let h be a Borel function on $\mathbb{R}$ such that $\mathbb{E}[|h(X)|] < \infty$. Set

$$g(z) = \int_{\mathbb{R}} h(x) f_{X|Z}(x|z)dx.$$

Then $Y = g(Z)$ is a version of the conditional expectation $\mathbb{E}[h(X)|Z]$.

Indeed, for a typical element of $\sigma(Z)$ which has the form $\{w : Z(w) \in B\}$, where $B \in \mathcal{B}$, we have

$$\mathbb{E}[h(X)\mathbb{I}_B(Z)] = \int\int h(x)\mathbb{I}_B(z) f_{X,Z} dxdz = \int g(z)\mathbb{I}_B(z) f_Z(z)dz = \mathbb{E}[g(Z)\mathbb{I}_B(Z)].$$

2.8.3 Properties of conditional expectation

Theorem 2.8.4. *Let ξ, η be integrable random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G}$ be a sub-σ-algebras of $\mathcal{F}$.*

1. *If c is a constant, then $\mathbb{E}[c|\mathcal{G}] = c$ a.s.*
2. *If $\xi \ge \eta$ a.s. then $\mathbb{E}(\xi|\mathcal{G}) \ge \mathbb{E}(\eta|\mathcal{G})$ a.s.*
3. *If a, b are constants, then*

$$\mathbb{E}(a\xi + b\eta|\mathcal{G}) = a\mathbb{E}(\xi|\mathcal{G}) + b\mathbb{E}(\eta|\mathcal{G}).$$

4. *If $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}(\xi)$ a.s.*
5. *$\mathbb{E}(\xi|\mathcal{F}) = \xi$ a.s.*
6. *$\mathbb{E}(\mathbb{E}(\xi|\mathcal{G})) = \mathbb{E}(\xi)$.*
7. (Tower property) *Let $\mathcal{G}_1 \subset \mathcal{G}_2$ be sub-σ-algebras of $\mathcal{F}$ then*

$$\mathbb{E}(\mathbb{E}(\xi|\mathcal{G}_1)|\mathcal{G}_2) = \mathbb{E}(\mathbb{E}(\xi|\mathcal{G}_2)|\mathcal{G}_1) = \mathbb{E}(\xi|\mathcal{G}_1) \quad a.s.$$

8. *If ξ is independent of $\mathcal{G}$, then $\mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}(\xi)$ a.s.*

9. *If η is $\mathcal{G}$-measurable and $\mathbb{E}(|\xi\eta|) < \infty$, then*

$$\mathbb{E}(\xi\eta|\mathcal{G}) = \eta\mathbb{E}(\xi|\mathcal{G}) \quad a.s.$$

10. *Let $\mathcal{H}$ be a sub-σ-algebras of $\mathcal{F}$ which is independent of $\sigma(\mathcal{G}, \xi)$, then*

$$\mathbb{E}\big(\xi|\sigma(\mathcal{G}, \mathcal{H})\big) = \mathbb{E}(\xi|\mathcal{G}) \quad a.s.$$

Proof. 1. Statement 1 is evident.
2. Since $\xi \geq \eta$ a.s. so $\int_A \xi d\mathbb{P} \geq \int_A \eta d\mathbb{P}$ for every $A \in \mathcal{G}$. Hence, $\int_A \mathbb{E}(\xi|\mathcal{G})d\mathbb{P} \geq \int_A \mathbb{E}(\eta|\mathcal{G})d\mathbb{P}$ for every $A \in \mathcal{G}$. Thus, $\mathbb{E}(\xi|\mathcal{G}) \geq \mathbb{E}(\eta|\mathcal{G})$ a.s.
3. If $A \in \mathcal{G}$,

$$\int_A (a\xi + b\eta)d\mathbb{P} = a\int_A \xi d\mathbb{P} + b\int_A \eta d\mathbb{P}$$
$$= a\int_A \mathbb{E}(\xi|\mathcal{G})d\mathbb{P} + b\int_A \mathbb{E}(\eta|\mathcal{G})d\mathbb{P} = \int_A (a\mathbb{E}(\xi|\mathcal{G}) + b\mathbb{E}(\eta|\mathcal{G}))d\mathbb{P}$$

From this, we have proof.
4. Since $\mathbb{E}\xi$ is measurable with respect to σ-algebra $\mathcal{G} = \{\emptyset, \Omega\}$ and if $A = \emptyset$ or $A = \Omega$, we have

$$\int_A \xi d\mathbb{P} = \int_A \mathbb{E}\xi d\mathbb{P} \Rightarrow \mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}(\xi) \quad a.s.$$

5. Statement 5 is evident.
6. Using Definition 2.8.1 for $G = \Omega$, we have:

$$\int_\Omega \mathbb{E}(\xi|\mathcal{G})d\mathbb{P} = \int_\Omega \xi d\mathbb{P} \Rightarrow \mathbb{E}(\mathbb{E}(\xi|\mathcal{G})) = \mathbb{E}\xi \quad a.s.$$

7. If $A \in \mathcal{G}$, we have:

$$\int_A \mathbb{E}[\mathbb{E}(\xi|\mathcal{G}_2)|\mathcal{G}_1]d\mathbb{P} = \int_A \mathbb{E}(\xi|\mathcal{G}_2)d\mathbb{P} = \int_A \xi d\mathbb{P}.$$

From this and Definition 2.8.1, the first equation is proven. The second one follows from Statement 5 and remark that $\mathbb{E}(\xi|\mathcal{G}_1)$ is $\mathcal{G}_2$ -measurable.
8. If $A \in \mathcal{G}$, X and $\mathbb{I}_A$ are independent. Hence, we have:

$$\int_A \xi d\mathbb{P} = \mathbb{E}(\xi\mathbb{I}_A) = \mathbb{E}\xi.\mathbb{P}(A) = \int_A (\mathbb{E}\xi)d\mathbb{P} \Rightarrow \mathbb{E}(\xi|\mathcal{G}) = \mathbb{E}(\xi) \quad a.s.$$

9. First suppose that ξ and η are non-negative. For $\eta = \mathbb{I}_A$, where A is $\mathcal{G}$-measurable, $B \cap A \in \mathcal{G}$, so that, by the defining relation,

$$\int_A \eta\mathbb{E}(\xi|\mathcal{G})d\mathbb{P} = \int_{B\cap A} \mathbb{E}(\xi|\mathcal{G})d\mathbb{P} = \int_{B\cap A} \xi d\mathbb{P} = \int_B \xi\eta d\mathbb{P},$$

which proves the desired relation for indicators, and hence for simple random variables. Next, if $\{\eta_n, n \geq 1\}$ are simple random variables, such that $\eta_n \uparrow \eta$ almost surely as $n \to \infty$, it follows that $\eta_n\xi \uparrow \eta\xi$ and $\eta_n\mathbb{E}(\xi|\mathcal{G}) \uparrow \eta\mathbb{E}(\xi\mathcal{G}|)$ almost surely as $n \to \infty$, from which the conclusion follows by monotone convergence. The general case follows by the decomposition $\xi = \xi^+ - \xi^-$ and $\eta = \eta^+ - \eta^-$. □

2.8.4 Convergence theorem

Let $(\xi_n), \xi$ and η be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G}$ be a sub-σ-algebras of $\mathcal{F}$.

Theorem 2.8.5 (Monotone convergence theorem). *a) Suppose that $\xi_n \uparrow \xi$ a.s. and there exists a positive integer m such that $\mathbb{E}(\xi_m^-) < \infty$. Then, $\mathbb{E}(\xi_n|\mathcal{G}) \uparrow \mathbb{E}(\xi|\mathcal{G})$ a.s.*
b) Suppose that $\xi_n \downarrow \xi$ a.s. and there exists a positive integer m such that $\mathbb{E}(\xi_m^+) < \infty$, then $\mathbb{E}(\xi_n|\mathcal{G}) \downarrow \mathbb{E}(\xi|\mathcal{G})$ a.s.

Proof. Suppose $\mathbb{E}\xi_{n_0}^- < \infty$. Hence $0 \leq \xi_n + \xi_{n_0}^- \uparrow \xi + \xi_{n_0}^-$, by Theorem **??**

$$\int_A \lim_n \mathbb{E}(\xi_n + \xi_{n_0}^-|\mathcal{G})d\mathbb{P} = \lim_n \int_A \mathbb{E}(\xi_n + \xi_{n_0}^-|\mathcal{G})d\mathbb{P}$$

$$= \lim_n \int_A (\xi_n + \xi_{n_0}^-)d\mathbb{P} = \int_A \lim_n (\xi_n + \xi_{n_0}^-)d\mathbb{P} = \int_A (\xi + \xi_{n_0}^-)d\mathbb{P}$$

By linearity we have

$$\int_A \lim_n \mathbb{E}(\xi_n|\mathcal{G})d\mathbb{P} = \int_A \xi d\mathbb{P} = \int_A \mathbb{E}(\xi|\mathcal{G})d\mathbb{P}, \quad \forall A \in \mathcal{G}$$

and then

$$\lim_n \mathbb{E}(\xi_n|\mathcal{G}) = \mathbb{E}(\xi|\mathcal{G}) \quad a.s.$$

Similarly to claim (b). □

Theorem 2.8.6 (Fatou's lemma). *a) If $\xi_n \leq \eta, \forall n \geq 1$ a.s., and $\mathbb{E}(\eta) < \infty$ then*

$$\limsup_n \mathbb{E}(\xi_n|\mathcal{G}) \leq \mathbb{E}(\limsup_n \xi_n|\mathcal{G}), \quad a.s.$$

b) If $\xi_n \geq \eta, \forall n \geq 1$ a.s., and $\mathbb{E}(\eta) > -\infty$, then

$$\liminf_n \mathbb{E}(\xi_n|\mathcal{G}) \geq \mathbb{E}(\liminf_n \xi_n|\mathcal{G}), \quad a.s.$$

c) If $|\xi_n| \leq \eta, \forall n \geq 1$ a.s., and $\mathbb{E}(\eta) < \infty$, then

$$\mathbb{E}(\liminf_n \xi_n|\mathcal{G}) \leq \liminf_n \mathbb{E}(\xi_n|\mathcal{G}) \leq \limsup_n \mathbb{E}(\xi_n|\mathcal{G}) \leq \mathbb{E}(\limsup_n \xi_n|\mathcal{G}), \ a.s.$$

Theorem 2.8.7 (Lebesgue's dominated convergence theorem). *Suppose that $\mathbb{E}(\eta) < \infty$, $|\xi_n| \leq \eta$ a.s., and $\xi_n \xrightarrow{a.s} \xi$. Then,*

$$\lim_n \mathbb{E}(\xi_n|\mathcal{G}) = \mathbb{E}(\xi|\mathcal{G}) \quad \textit{and} \quad \lim_n \mathbb{E}(|\xi_n - \xi||\mathcal{G}) = 0, \quad a.s.$$

The proofs of Fatou's lemma and Lebesgue's dominated convergence theorem are analogous in a similar vein to the proofs of Fatou's lemma and the Dominated convergence theorem without conditioning.

Theorem 2.8.8 (Jensen's inequality). *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\varphi(\xi)$ is integrable. Then*

$$\mathbb{E}(\varphi(\xi)|\mathcal{G}) \geq \varphi(\mathbb{E}(\xi|\mathcal{G})), \quad a.s.$$

Proof. A result in real analysis is that if $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then $\varphi(x) = \sup_n(a_n x + b_n)$ for a countable collection of real numbers (a_n, b_n). Then

$$\mathbb{E}(a_n\xi + b_n|\mathcal{G}) = a_n\mathbb{E}(\xi|\mathcal{G}) + b_n.$$

But $\mathbb{E}(a_n\xi + b_n|\mathcal{G}) \le \mathbb{E}(\varphi(\xi)|\mathcal{G})$, hence $a_n\mathbb{E}(\xi|\mathcal{G}) + b_n \le \mathbb{E}(\xi|\mathcal{G})$, for every n. Taking the supremum in n, we get the result. □

In particular, if $\varphi(x) = x^2$ then $\mathbb{E}(\xi^2|\mathcal{G}) \ge \big(\mathbb{E}(\xi|\mathcal{G})\big)^2$.

2.8.5 Conditional expectation given a random variable

Since $\mathbb{E}(\xi|\eta)$ is $\sigma(\eta)$-measurable, there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(\xi|\eta) = f(\eta)$. We denote $f(x) = \mathbb{E}(\xi|\eta = x)$.
a) Since $\mathbb{E}(\xi) = \mathbb{E}(f(\eta)) = \int_{\mathbb{R}} f(x)dF_\eta(x)$,

$$\mathbb{E}(\xi) = \int_{\mathbb{R}} \mathbb{E}(\xi|\eta = x)dF_\eta(x). \tag{2.9}$$

b) Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a Borel function such that both ξ and $\xi\varphi(\eta)$ are integrable. Then, the equation

$$\mathbb{E}(\xi\varphi(\eta)|\eta = y) = \varphi(y)\mathbb{E}(\xi|\eta = y)$$

holds $\mathbb{P}_\eta$-a.s.
c) If ξ and η are independent, then

$$\mathbb{E}(\xi|\eta = y) = \mathbb{E}(\xi).$$

Moreover, let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ satisfy $\mathbb{E}|\varphi(\xi, \eta)| < \infty$, then

$$\mathbb{E}(\varphi(\xi, \eta)|\eta = y) = \mathbb{E}(\varphi(\xi, y)) \quad (\mathbb{P}_\eta - a.s.). \tag{2.10}$$

2.9 Exercises

Discrete random variables

2.1. An urn contains five red, three orange, and two blue balls. Two balls are randomly selected. What is the sample space of this experiment? Let X represent the number of orange balls selected. What are the possible values of X? Calculate expectation and variance of X.

2.2. An urn contains 7 white balls numbered 1,2,...,7 and 3 black ball numbered 8,9,10. Five balls are randomly selected, (a) with replacement, (b) without replacement. For each of cases (a) and (b) give the distribution:

1. of the number of white balls in the sample;
2. of the minimum number in the sample;

3. of the maximum number in the sample;

4. of the minimum number of balls needed for selecting a white ball.

2.3. A machine normally makes items of which 4% are defective. Every hour the producer draws a sample of size 10 for inspection. If the sample contains no defective items he does not stop the machine. What is the probability that the machine will not be stopped when it has started producing items of which 10% are defective.

2.4. Let X represent the difference between the number of heads and the number of tails obtained when a fair coin is tossed n times. What are the possible values of X? Calculate expectation and variance of X.

2.5. An urn contains N_1 white balls and N_2 black balls; n balls are drawn at random, (a) with replacement, (b) without replacement. What is the expected number of white balls in the sample?

2.6. A student takes a multiple choice test consisting of two problems. The first one has 3 possible answers and the second one has 5. The student chooses, at random, one answer as the right one from each of the two problems. Find:

a) the expected number, $E(X)$ of the right answers X of the student;

b) the $Var(X)$.

2.7. In a lottery that sells 3,000 tickets the first lot wins $1,000, the second $500, and five other lots that come next win $100 each. What is the expected gain of a man who pays 1 dollar to buy a ticket?

2.8. A pays 1 dollar for each participation in the following game: three dice are thrown; if one ace appears he gets 1 dollar, if two aces appear he gets 2 dollars and if three aces appear he gets 8 dollars; otherwise he gets nothing. Is the game fair, i.e., is the expected gain of the player zero? If not, how much should the player receive when three aces appear to make the game fair?

2.9. Suppose a die is rolled twice. What are the possible values that the following random variables can take on?

1. The maximum value to appear in the two rolls.

2. The minimum value to appear in the two rolls.

3. The sum of the two rolls.

4. The value of the first roll minus the value of the second roll.

2.10. Suppose X has a binomial distribution with parameters n and $p \in (0, 1)$. What is the most likely outcome of X?

2.11. An airline knows that 5 percent of the people making reservations on a certain flight will not show up. Consequently, their policy is to sell 52 tickets for a flight that can hold only 50 passengers. What is the probability that there will be a seat available for every passenger who shows up?

2.12. Suppose that an experiment can result in one of r possible outcomes, the ith outcome having probability $p_i, i = 1, \ldots, r, \sum_{i=1}^r p_i = 1$. If n of these experiments are performed, and if the outcome of any one of the n does not affect the outcome of the other $n1$ experiments, then show that the probability that the first outcome appears x_1 times, the second x_2 times, and the rth x_r times is

$$\frac{n!}{x_1!x_2!\cdots x_r!}p_1^{x_1}p_2^{x_2}\cdots p_r^{x_r}$$

when $x_1 + x_2 + \ldots + x_r = n$. This is known as the *multinomial* distribution.

2.13. A television store owner figures that 50 percent of the customers entering his store will purchase an ordinary television set, 20 percent will purchase a color television set, and 30 percent will just be browsing. If five customers enter his store on a certain day, what is the probability that two customers purchase color sets, one customer purchases an ordinary set, and two customers purchase nothing?

2.14. Let X be Geometric. Show that for $i, j > 0$,

$$\mathbb{P}[X > i + j|X > i] = \mathbb{P}[X > j].$$

2.15. If a fair coin is successively flipped, find the probability that a head first appears on the fifth trial.

2.16. A coin having probability p of coming up heads is successively flipped until the rth head appears. Argue that X, the number of flips required, will be $n, n \geq r$, with probability

$$\mathbb{P}[X = n] = C_{n-1}^{r-1}p^r(1-p)^{n-r}, \quad n \geq r.$$

This is known as the *negative binomial* distribution. Find the expectation and variance of X.

2.17. A fair coin is independently flipped n times, k times by A and $n - k$ times by B. Show that the probability that A and B flip the same number of heads is equal to the probability that there are a total of k heads.

2.18. Suppose that we want to generate a random variable X that is equally likely to be either 0 or 1, and that all we have at our disposal is a biased coin that, when flipped, lands on heads with some (unknown) probability p. Consider the following procedure:

1. Flip the coin, and let 0_1, either heads or tails, be the result.

2. Flip the coin again, and let 0_2 be the result.

3. If 0_1 and 0_2 are the same, return to step 1.

4. If 0_2 is heads, set $X = 0$, otherwise set $X = 1$.

(a) Show that the random variable X generated by this procedure is equally likely to be either 0 or 1.

(b) Could we use a simpler procedure that continues to flip the coin until the last two flips are different, and then sets $X = 0$ if the final flip is a head, and sets $X = 1$ if it is a tail?

2.19. Consider n independent flips of a coin having probability p of landing heads. Say a changeover occurs whenever an outcome differs from the one preceding it. For instance, if the results of the flips are $HHTHTHHT$, then there are a total of five changeovers. If $p = 1/2$, what is the probability there are k changeovers?

2.20. Let X be a Poisson random variable with parameter λ. What is the most likely outcome of X?

2.21. * *Poisson Approximation to the Binomial* Let P be a Binomial probability with probability of success p and number of trial n. Let $\lambda = np$. Show that

$$P(k \text{ successes}) = \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)\left\{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)\right\}\left(1-\frac{\lambda}{n}\right)^{-k}.$$

Let $n \to \infty$ and let p change so that λ remains constant. Conclude that for small p and large n,

$$P(k \text{ successes}) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad \text{where } \lambda = pn.$$

2.22. * Let X be the Binomial $B(n,p)$.

a) Show that for $\lambda > 0$ and $\varepsilon > 0$ then

$$P(X - np > n\varepsilon) \le E[\exp(\lambda(X - np - n\varepsilon))].$$

b) With $a > 0$ show that

$$P\left(\mid \frac{X}{n} - p \mid > a\right) \le \frac{\sqrt{p(1-p)}}{a^2 n}\min\{\sqrt{p(1-p}, a\sqrt{n}\}.$$

2.23. Let X be Poisson (λ).

a) With λ a positive integer. Show $E\{|X-\lambda|\} = \frac{2\lambda^\lambda e^{-\lambda}}{(\lambda-1)!}$,

b) Show for $r = 2, 3, 4, \ldots$,

$$E\{X(X-1)\ldots(X-r+1)\} = \lambda^r.$$

2.24. Let X be Geometric (p).

a) Show $E\left\{\frac{1}{1+X}\right\} = \log\left((1-p)^{\frac{p}{p-1}}\right)$.

b) Show for $r = 2, 3, 4, \ldots$, $E\{X(X-1)\ldots(X-r+1)\} = \frac{r!p^r}{(1-p)^r}$.

2.25. Suppose X takes all its values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. Show that

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}[X > n].$$

The following exercises use the additivity of expectation

2.26. Liam's bowl of spaghetti contains n strands. He selects two ends at random and joins them together. He does this until there are no ends left. What is the expected number of spaghetti hoops in the bowl?

2.27. Sarah collects figures from cornflakes packets. Each packet contains one figure, and n distinct figures make a complete set. Find the expected number of packets Sarah needs to collect a complete set.

2.28. Each packet of the breakfast cereal Soggies contains exactly one token, and tokens are available in each of the three colours blue, white and red. You may assume that each token obtained is equally likely to be of the three available colours, and that the (random) colours of different tokens are independent. Find the probability that, having searched the contents of k packets of Soggies, you have not yet obtained tokens of every colour.

Let N be the number of packets required until you have obtained tokens of every colour. Show that $\mathbb{E}[N] = \frac{11}{2}$.

2.29. Each box of cereal contains one of $2n$ different coupons. The coupons are organized into n pairs, so that coupons 1 and 2 are a pair, coupons 3 and 4 are a pair, and so on.

Once you obtain one coupon from every pair, you can obtain a prize. Assuming that the coupon in each box is chosen independently and uniformly at random from the $2n$ possibilities, what is the expected number of boxes you must buy before you can claim the prize?

Continuous random variables

2.30. The amount of bread (in hundreds of kilos) that a bakery sells in a day is a random variable with density

$$f(x) = \begin{cases} cx & \text{for } 0 \le x < 3, \\ c(6-x) & \text{for } 3 \le x < 6, \\ 0 & \text{otherwise.} \end{cases}$$

a) Find the value of c which makes f a probability density function.

b) What is the probability that the number of kilos of bread that will be sold in a day is, (i) more than 300 kilos? (ii) between 150 and 450 kilos?

c) Denote by A and B the events in (i) and (ii), respectively. Are A and B independent events?

2.31. Suppose that the duration in minutes of long-distance telephone conversations follows an exponential density function:

$$f(x) = \frac{1}{5}e^{-x/5} \text{ for } x > 0.$$

Find the probability that the duration of a conversation:

a) will exceed 5 minutes;

b) will be between 5 and 6 minutes;

c) will be less than 3 minutes;

d) will be less than 6 minutes given that it was greater than 3 minutes.

2.32. A number is randomly chosen from the interval (0;1). What is the probability that:

a) its first decimal digit will be a 1;

b) its second decimal digit will be a 5;

c) the first decimal digit of its square root will be a 3?

2.33. The height of men is normally distributed with mean μ=167 cm and standard deviation σ=3 cm.

a) What is the percentage of the population of men that have height, (i) greater than 167 cm, (ii) greater than 170 cm, (iii) between 161 cm and 173 cm?

b) In a random sample of four men what is the probability that:

i) all will have height greater than 170 cm;

ii) two will have height smaller than the mean (and two bigger than the mean)?

2.34. Find the constant k and the mean and variance of the population defined by the probability density function

$$f(x) = k(1+x)^{-3} \text{ for } 0 \le x < \infty$$

and zero otherwise.

2.35. A mode of a distribution of one random variable X is a value of x that maximizes the pdf or pmf. For X of the continuous type, $f(x)$ must be continuous. If there is only one such x, it is called the mode of the distribution. Find the mode of each of the following distributions

1. $f(x) = 12x^2(1-x),\ 0 < x < 1$, zero elsewhere.
2. $f(x) = \frac{1}{2}x^2e^{-x},\ 0 < x < \infty$, zero elsewhere.

2.36. A median of a distribution of a random variable X is a value x such that $\mathbb{P}[X < x] \le \frac{1}{2}$ and $\mathbb{P}[X \le x] \ge \frac{1}{2}$. Find the median of each of the following distribution:

1. $f(x) = 3x^2,\ 0 < x < 1$, zero elsewhere.

2. $f(x) = \frac{1}{\pi(1+x^2)}$.

2.37. Let $0 < p < 1$. A $(100p)$th percentile (quantile of order p) of the distribution of a random variable X is a value ζ_p such that

$$\mathbb{P}[X < \zeta_p] \leq p, \quad \text{and} \quad \mathbb{P}[X \leq \zeta_p] \geq p.$$

Find the pdf $f(x)$, the $25th$ percentile and the 60th percentile for each of the the followin cdfs.

1. $F(x) = (1 + e^x)^{-1},\ -\infty < x < \infty$.

2. $F(x) = e^{-e^{-x}},\ -\infty < x < \infty$.

3. $F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x),\ -\infty < x < \infty$.

2.38. If X is a random variable with the probability density function f, find the probability density function of $Y = X^2$ if

(a) $f(x) = 2xe^{-x^2}$, for $0 \leq X < \infty$

(b) $f(x) = (1 + x)/2$, for $-1 \leq X \leq 1$

(c) $f(x) = \frac{1}{2}$, for $-\frac{1}{2} \leq X \leq \frac{3}{2}$.

2.39. Let X be a standard normal random variable. Denote $Y = e^X$.

1. Find the density of Y. This is known as *log-normal distribution.*

2. Find the expectation and variance of Y.

2.40. Let X be a uniform distribution $U(0, 1)$. Find the density of the following random variable.

1. $Y = -\frac{1}{\lambda}\ln(1 - X)$.

2. $Z = \ln\frac{X}{1-X}$. This is known as *Logistic* distribution.

3. $T = \sqrt{2\ln\frac{1}{1-X}}$. This is known as *Rayleigh* distribution.

2.41. Let X have the uniform distribution $U(-\frac{\pi}{2}, -\frac{\pi}{2})$.

1. Find the pdf of $Y = \tan X$. This is the pdf of a Cauchy distribution.

2. Show that Y is not integrable.

3. Denote $Z = (X \vee (-a)) \wedge a$ for some $a > 0$. Find $\mathbb{E}[Z]$.

2.42. Let X be a random variable with distribution function F that is continuous. Show that $Y = F(X)$ is uniform.

2.43. Let F be a distribution function that is continuous and is such that the inverse function F^{-1} exists. Let U be uniform on $[0,1]$. Show that $X = F^{-1}(U)$ has distribution function F.

2.44.

1. Let X be a non-negative random variable satisfying $\mathbb{E}[X^\alpha] < \infty$ for some $\alpha > 0$. Show that
$$\mathbb{E}[X^\alpha] = \alpha \int_0^\infty x^{\alpha-1}(1-F(x))dx.$$
2. Let Y be a continuous random variable. Show that
$$\mathbb{E}[Y] = \int_0^{+\infty} (\mathbb{P}[Y > t] - \mathbb{P}[Y < -t])dt.$$

2.45. Suppose that the density function of X satisfies $\int_a^b f(x)dx = 1$ for some real constants $a < b$. Show that $a < \mathbb{E}[X] < b$ and $DX \leq \frac{(b-a)^2}{4}$.

2.46. Let X be a nonnegative random variable with mean μ and variance σ^2, both finite. Show that for any $b > 0$,
$$\mathbb{P}[X \geq \mu + b\sigma] \leq \frac{1}{1+b^2}.$$
Hint: consider the function $g(x) = \frac{[(x-\mu)b+\sigma]^2}{\sigma^2(1+b^2)^2}$.

2.47. Let X be a random variable with mean μ and variance σ^2, both finite. Show that for any $d > 1$,
$$\mathbb{P}[\mu - d\sigma < X < \mu + d\sigma] \geq 1 - \frac{1}{d^2}.$$

2.48. Divide a line segment into two parts by selecting a point at random. Find the probability that the larger segment is at least three times the shorter. Assume a uniform distribution.

2.49. Let X be an integrable random variable.

1. Let (A_n) be a sequence of events such that $\lim_n \mathbb{P}(A_n) = 0$. Show that $\lim_{n\to\infty} \mathbb{E}[XI_{A_n}] = 0$.
2. Show that for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any event A satisfying $\mathbb{P}(A) < \delta$, $\mathbb{E}[XI_A] < \epsilon$.

2.50. Let (X_n) be a sequence of non-negative random variable. Show that
$$\mathbb{E}[\sum_{n=1}^\infty X_n] = \sum_{n=1}^\infty \mathbb{E}[X_n].$$

2.51. Given the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, suppose X is a non-negative random variable and $\mathbb{E}[X] = 1$. Define $Q : \mathcal{A} \to \mathbb{R}$ by $Q(A) = \mathbb{E}[XI_A]$.

1. Show that Q defines a probability measure on $(\Omega, \mathcal{A})$.
2. Show that if $\mathbb{P}(A) = 0$, then $Q(A) = 0$.
3. Suppose $\mathbb{P}(X > 0) = 1$. Let $\mathbb{E}_Q$ denote expectation with respect to Q. Show that $\mathbb{E}_Q[Y] = \mathbb{E}_{\mathbb{P}}[YX]$.

Random elements

2.52. An urn contains 3 red balls, 4 blue balls and 2 yellow balls. Pick up randomly 2 ball from that urn and denote X and Y the number of red and yellow balls in the 2 balls, respectively.

1. Make the joint distribution table of X and Y.
2. Are X and Y independent?
3. Find the distribution of $Z = XY$.

2.53. Suppose that the joint pmf of X and Y is

$$\mathbb{P}[X = i, Y = j] = C_j^i e^{-2\lambda}\lambda^j/j!, \quad 0 \le i \le j.$$

1. Find the probability mass function of Y.
2. Find the probability mass function of X.
3. Find the probability mass function of $Y - X$.

2.54. Let X and Y be independent random variables taking values in $\mathbb{N}$ with

$$\mathbb{P}[X = i] = \mathbb{P}[Y = i] = \frac{1}{2^i}, \quad i = 1, 2, \dots$$

Find the following probability

1. $\mathbb{P}[X \wedge Y \le i]$.
2. $\mathbb{P}[X = Y]$.
3. $\mathbb{P}[X > Y]$.
4. $\mathbb{P}[X \text{ divides } Y]$.
5. $\mathbb{P}[X \ge kY]$ for a given positive integer k.

2.55. Let X and Y be independent geometric random variables with parameters λ and μ.

1. Let $Z = X \wedge Y$. Show that Z is geometric and find its parameter.
2. Find the probability that $X = Y$.

2.56. Let X and Y be independent random variables with uniform distribution on the set $\{-1, 1\}$. Let $Z = XY$. Show that X, Y, Z are pairwise independent but that they are not mutually independent.

2.57. * Let n be a prime number greater than 2; and X, Y be independent and uniformly distributed on $\{0, 1, \dots, n-1\}$. For each $r, 0 \le r \le n-1$, define $Z_r = X + rY(\mod n)$. Show that the random variable $Z_r, r = 0, \dots, n-1$, are pairwise independent.

2.58. Let (X_n) be a sequence of independent random variables with $\mathbb{P}[X_n = 1] = \mathbb{P}[X_n = -1] = \frac{1}{2}$ for all n. Let $Z_n = X_0 X_1 \dots X_n$. Show that $Z_1, Z_2, \dots$ are independent.

2.59. Let $(a_1, \dots, a_n)$ be a random permutation of $(1, \dots, n)$, equally likely to be any of the $n!$ possible permutations. Find the expectation of

$$L = \sum_{i=1}^{n} |a_i - i|.$$

2.60. A blood test is being performed on n individuals. Each person can be tested separately. but this is expensive. Pooling can decrease the cost. The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the group of k individuals. If the test is positive, then each of the k persons must be tested separately and thus $k + 1$ total tests are required for the k people. Suppose that we create n/k disjoint groups of k people (where k divides n) and use the pooling method. Assume that each person has a positive result on the test independently with probability p.

(a) What is the probability that the test for a pooled sample of k people will be positive?

(b) What is the expected number of tests necessary?

(c) Describe how to find the best value of k.

(d) Give an inequality that shows for what values of p pooling is better than just testing every individual.

2.61. You need a new staff assistant, and you have n people to interview. You want to hire the best candidate for the position. When you interview a candidate, you can give them a score, with the highest score being the best and no ties being possible. You interview the candidates one by one. Because of your company's hiring practices, after you interview the kth candidate, you either offer the candidate the job before the next interview or you forever lose the chance to hire that candidate. We suppose the candidates are interviewed in a random order, chosen uniformly at random from all $n!$ possible orderings.

We consider the following strategy. First, interview m candidates but reject them allL these candidates give you an idea of how strong the field is. After the mth candidate. hire the first candidate you interview who is better than all of the previous candidates you have interviewed.

1. Let E be the event that we hire the best assistant, and let E_i be the event that ith candidate is the best and we hire him. Determine $\mathbb{P}(E_i)$, and show that

$$\mathbb{P}(E) = \frac{m}{n} \sum_{j=m+1}^{n} \frac{1}{j-1}.$$

2. Show that

$$\frac{m}{n}(\ln n - \ln m) \leq \mathbb{P}(E) \leq \frac{m}{n}(\ln(n-1) - \ln(m-1)).$$

3. Show that $m(\ln n - \ln m)/n$ is maximized when $m = n/e$, and explain why this means $\mathbb{P}(E) \geq 1/e$ for this choice of m.

2.62. Let X and Y have the joint pdf

$$f(x, y) = \begin{cases} 6(1-x-y) & \text{if } x+y<1, x>0, y>0, \\ 0 & \text{otherwise.} \end{cases}$$

Compute $\mathbb{P}[2X + 3Y < 1]$ and $\mathbb{E}[XY + 2X^2]$.

2.63. Let X and Y have the joint pdf

$$f_{X,Y}(x, y) = \begin{cases} 10xy & \text{if } 0<x<y<1 \\ 0 & \text{otherwise} \end{cases}.$$

Find the joint pdf of X/Y and Y.

2.64. Let X be a normal with $\mu = 0$ and $\sigma^2 < \infty$, and let Θ be uniform on $[0, \pi]$. Assume that X and θ are independent. Find the distribution of $Z = X + a \cos \Theta$.

2.65. Let X and Y be independent random variable with the same distribution $N(0, \sigma^2)$.

1. Let $U = X + Y$ and $V = X - Y$. Show that U and V are independent.
2. Let $Z = \sqrt{X^2 + Y^2}$ and $W = \arctan \frac{X}{Y} \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Show that X has a Rayleight distribution, that W is uniform, and that Z and W are independent.

2.66. (Simulation of Normal Random Variables) Let U and V be two independent uniform random variable on $[0, 1]$. Let $\theta = 2\pi U$ and $S = -\ln(V)$.

1. Show that S has an exponential distribution, and that R

2.67. Let $(X_1, \ldots, X_n)$ be random variables. Define

$$\begin{aligned} Y_1 &= \min\{X_i, 1 \leq i \leq n\}, \\ Y_2 &= \text{ second smallest of } X_1, \ldots, X_n, \\ &\vdots \\ Y_n &= \max\{X_i, 1 \leq i \leq n\}. \end{aligned}$$

Then $Y_1, \ldots, Y_n$ are also random variables, and $Y_1 \leq Y_2 \leq \ldots \leq Y_n$. They are called the order statistics of $(X_1, \ldots, X_n)$ and are usually denoted $Y_k = X_{(k)}$. Assume that X_i are i.i.d. with common density f.

1. Show that the joint density of the order statistics is given by

$$f_{(X_{(1)}, X_{(2)}, \ldots, X_{(n)})}(y_1, \ldots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i) & \text{for } y_1 < y_2 < \ldots < y_n \\ 0 & \text{otherwise.} \end{cases}$$

2. Show that $X_{(k)}$ has density

$$f_{(k)}(y) = kC_n^k f(y)(1-F(y))^{n-k}F(y)^{k-1}$$

where F is distribution function of X_k.

2.68. Show that the function

$$F(x,y) = \begin{cases} 0 \text{ for } x+y<1, \\ 1 \text{ for } x+y \geq 1, \end{cases}$$

is not a joint distribution function.

2.69. Let X and Y be independent and suppose $\mathbb{P}[X+Y=\alpha]=1$ for some constant α. Show that both X and Y are constant random variables.

2.70. Let $(X_n)_{n\geq 1}$ be iid with common continuous distribution function $F(x)$. Denote $R_n = \sum_{j=1}^n \mathbb{I}_{\{X_j \geq X_n\}}$, and $A_n = \{R_n = 1\}$.

1. Show that the sequence of random variables $(R_n)_{n\geq 1}$ is independent and

$$\mathbb{P}[R_n = k] = \frac{1}{n}, \quad \text{for } k = 1, \dots, n.$$

2. The sequence of events $(A_n)_{n\geq 1}$ is independent and

$$\mathbb{P}(A_n) = \frac{1}{n}.$$

Chapter 3

Fundamental Limit Theorems

3.1 Convergence of random variables

In this section, we study about *convergence of random variables*. This is an important concept in probability theory and it has many applications in statistics. Here we study a sequence of random events or variables and we consider whether it obeys some behavior. Such a behavior can be characterized in two cases: the limit is a constant value or the limit is still random but we can describe its law.

When discussing the convergence of random variables, we need to define the metric between two random variables or the manner that the random variables are close to each other. Then in the following, we give some ”manners” or some modes of convergence.

Definition 3.1.1. Let $(X_n)_{n\geq 1}$ be a sequence of random variables defined on$(\Omega, \mathcal{A}, \mathbb{P})$. We say that X_n

- *converges almost surely* to X and denoted by $X_n \xrightarrow{a.s.} X$ or $\lim_n X_n = X$ a.s., if
$$\mathbb{P}\Big(w : \lim_{n\to\infty} X_n(w) = X(w)\Big) = 1;$$
- *converges in probability* to X and denoted by $X_n \xrightarrow{\mathbb{P}} X$, if for any $\epsilon > 0$,
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0;$$
- *converges in L^p* $(p > 0)$ to X and denoted by $X_n \xrightarrow{L^p} X$ if $\mathbb{E}(|X_n|^p) < \infty$ for any n, $\mathbb{E}(|X|^p) < \infty$ and
$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0.$$

Note that, the value of a random variable is a number, so the most natural way to consider the convergence of random variables is via the convergence of a sequence of numbers; and here comes the convergence almost surely. But sometimes this mode of convergence can fail, then the convergence in probability is defined in the meaning that the larger n is, the smaller and smaller the probability that X_n is far away from X becomes; and the convergence in L^p is considered in the sense that the average distance between X_n and X must tends to 0.

We have the following example.

Example 3.1.2. Let $\{X_n\}$ be a sequence of random variables such that

$$\mathbb{P}(X_n = 0) = 1 - \frac{1}{n^2} \quad \text{and } \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

Then X_n converges to 0 in probability, in L^p for $0 < p < 2$ and almost surely.

• At first, we consider the convergence in probability. For any $\epsilon > 0$, observe that the event $\{|X_n - 0| > \epsilon\}$ is included in the event $\{X_n \neq 0\} = \{X_n = n\}$. Then

$$0 \leq \mathbb{P}(|X_n - 0| > \epsilon) \leq \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

Therefore from the sandwich theorem,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \epsilon) = 0.$$

It implies that $X_n \xrightarrow{\mathbb{P}} 0$.

• In order to prove the convergence in L^p for $p \in (0, 2)$, we must check that

$$\lim_{n\to\infty} \mathbb{E}(|X_n - 0|^p) = 0.$$

This limit can be deduced from the computation that $\mathbb{E}\left(|X_n|^p\right) = n^{p-2}$.

• Usually, in order to prove or disprove the convergence almost surely, we use the *Borel-Cantelli lemma* that can be stated as follows.

Lemma 3.1.3 (Borel-Cantelli). *Let A_n be a sequence of events in a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$. Denote* $\limsup A_n = \cap_{n=1}^{\infty}\left(\cup_{m\geq n} A_m\right)$.

1. *If* $\Sigma_{n=1}^{\infty}\mathbb{P}(A_n) < \infty$, *then* $\mathbb{P}(\limsup A_n) = 0$.

2. *If* $\Sigma_{n=1}^{\infty}\mathbb{P}(A_n) = \infty$ *and A_n's are independent, then* $\mathbb{P}(\limsup A_n) = 1$.

Proof. 1. From the definition of $\limsup A_n$, it is clear that for every i,

$$\limsup A_n \subset \cup_{m\geq i} A_m.$$

So $\mathbb{P}(\limsup A_n) \leq \mathbb{P}(\cup_{m\geq i} A_m) \leq \Sigma_{m=i}^{\infty}\mathbb{P}(A_m)$. Since $\Sigma_{n=1}^{\infty}\mathbb{P}(A_n) < \infty$, the right hand side can be arbitrary small for suitable i. Then $\mathbb{P}(\limsup A_n) = 0$.

2. We have

$$\begin{aligned}1 - \mathbb{P}(\limsup A_n) &= \mathbb{P}\left(\overline{\cap_{n=1}^{\infty} \cup_{m\geq n} A_m}\right)\\ &= \mathbb{P}\left(\cup_{n=1}^{\infty}\overline{\cup_{m\geq n} A_m}\right)\\ &= \mathbb{P}\left(\cup_{n=1}^{\infty} \cap_{m\geq n} \overline{A_m}\right).\end{aligned}$$

In order to prove that $\mathbb{P}(\limsup A_n) = 1$, i.e $1 - \mathbb{P}(\limsup A_n) = 0$, we will show that $\mathbb{P}\left(\cap_{m\geq n}\overline{A_m}\right) = 0$ for every n. Indeed, since A_n's are independent,

$$\begin{aligned}\mathbb{P}\left(\cap_{m\geq n}\overline{A_m}\right) &= \Pi_{m\geq n}\mathbb{P}(\overline{A_m}) \\ &= \Pi_{m\geq n}[1-\mathbb{P}(A_m)] \\ &\leq \Pi_{m\geq n}e^{-\mathbb{P}(A_m)} = e^{-\Sigma_{m\geq n}\mathbb{P}(A_m)} = e^{-\infty} = 0.\end{aligned}$$

Then the result follows.

□

The meaning of the event $\limsup A_n$ is that A_n occurs for an infinite number of n. Therefore $\mathbb{P}(\limsup A_n) = 0$ means that almost surely there exists just a finite number of n that we can see A_n.

Now, let's see the application of the Borel-Cantelli Lemma in our example. We denote the event $A_n = \{X_n \neq 0\} = \{X_n = n\}$. Then,

$$\Sigma_{n=1}^{\infty}\mathbb{P}(A_n) = \Sigma_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

It implies that almost surely A_n occurs a finite number of n, i.e the number of n such that X_n differs from zero is finite. Hence, almost surely the limit of X_n exists and it must be zero. So $X_n \xrightarrow{a.s.} X$.

For any random variables X v Y, we denote

$$d_{\mathbb{P}}(X,Y) = \mathbb{E}\Big[\frac{|X-Y|}{|X-Y|+1}\Big].$$

The following proposition characterizes the convergence in probability via metric $d_{\mathbb{P}}$.[1]

Proposition 3.1.4. *X_n converges in probability to X iff*

$$\lim_{n\to\infty} d_{\mathbb{P}}(X_n, X) = 0. \tag{3.1}$$

Proof. $\Rightarrow$) Suppose that $X_n \xrightarrow{\mathbb{P}} X$. For any $\epsilon > 0$ and $w \in \Omega$, because of the increasing property of the function $x \mapsto \frac{x}{x+1}$ on the interval $[0,\infty)$, we have

$$\frac{|X_n(w)-X(w)|}{|X_n(w)-X(w)|+1} \leq \frac{\epsilon}{\epsilon+1}\mathbb{I}_{\{|X_n-X|\leq\epsilon\}}(w) + \mathbb{I}_{\{|X_n-X|<\epsilon\}}(w).$$

Taking expectation of both sides, we have

$$d_{\mathbb{P}}(X_n,X) \leq \frac{\epsilon}{\epsilon+1}\mathbb{P}(|X_n-X|\leq\epsilon) + \mathbb{P}(|X_n-X|>\epsilon).$$

Hence

$$\limsup_{n\to\infty} d_{\mathbb{P}}(X_n,X) \leq \epsilon + \limsup_{n\to\infty}\mathbb{P}(|X_n-X|>\epsilon) = \epsilon \text{ for all } \epsilon > 0.$$

[1] $d_{\mathbb{P}}$ is indeed a metric on L^0.

This implies (3.1).

$\Leftarrow$) On the other hand, we suppose that condition (3.1) holds. Then for any $\epsilon > 0$, it follows from Markov's inequality that

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{P}\Big(\frac{|X_n - X|}{|X_n - X| + 1} \geq \frac{\epsilon}{\epsilon + 1}\Big) \leq \frac{\epsilon + 1}{\epsilon}\mathbb{E}\Big[\frac{|X_n - X|}{|X_n - X| + 1}\Big] \to 0 \text{ as } n \to \infty.$$

□

The following proposition shows that among the three modes of convergence, the convergence in probability is the weakest form.

Proposition 3.1.5. *Let $(X_n)_{n\geq 1}$ be a sequence of random variables.*

1. *If $X_n \xrightarrow{L^p} X$ for some $p > 0$ then $X_n \xrightarrow{\mathbb{P}} X$.*
2. *If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. 1. Suppose that $X_n \xrightarrow{L^p} X$. Then by Markov inequality, for each $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^p > \epsilon^p) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\epsilon^p}.$$

Since $\mathbb{E}(|X_n - X|^p) \to 0$, by the sandwich theorem, $\mathbb{P}(|X_n - X| > \epsilon)$ converges also to 0. Therefore $X_n \xrightarrow{\mathbb{P}} X$.

2. Suppose that $X_n \xrightarrow{a.s.} X$. It is clear that

$$\frac{|X_n - X|}{1 + |X_n - X|} \leq 1,$$

then by Lebesgue's Dominated Convergence Theorem (see **??**);

$$\lim_{n\to} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = \mathbb{E}\left(\lim_{n\to} \frac{|X_n - X|}{1 + |X_n - X|}\right) = \mathbb{E}(0) = 0.$$

From the Proposition 3.1.4, we have $X_n \xrightarrow{\mathbb{P}} X$.

□

In the above example, we can see that convergence in probability is not sufficient for convergence almost surely. However, we have the following result.

Proposition 3.1.6. 1. *Suppose $X_n \xrightarrow{\mathbb{P}} X$. Then there exists a subsequence $(n_k)_{k\geq 1}$ such that $X_{n_k} \xrightarrow{a.s.} X$.*

2. *On the contrary, if for all subsequence $(n_k)_{k\geq 1}$, there exists a further subsequence $(m_k)_{k\geq 1}$ such that $X_{m_k} \xrightarrow{a.s.} X$ then $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. 1. Suppose $X_n \xrightarrow{\mathbb{P}} X$. Then from Proposition 3.1.4,

$$\lim_{n\to}\mathbb{E}\left(\frac{|X_n - X|}{1+|X_n - X|}\right) = 0.$$

So there exists a subsequence $\{n_k\}$ such that

$$\mathbb{E}\left(\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|}\right) < \frac{1}{2^k}.$$

It is clear that

$$\Sigma_{k=1}^{\infty}\mathbb{E}\left(\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|}\right) < \infty.$$

Therefore,

$$\Sigma_{k=1}^{\infty}\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|} < \infty \quad \text{a.s.}$$

Then, almost surely

$$\lim_{k\to\infty}\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|} = 0,$$

it implies that

$$\lim_{k\to\infty}|X_{n_k} - X| = 0,$$

i.e, $\lim\limits_{k\to\infty} X_{n_k} = X$.

2. Indeed, if we assume that X_n does not converge in probability to X, then from Proposition 3.1.4, the sequence $\mathbb{E}\left(\frac{|X_n-X|}{1+|X_n-X|}\right)$ does not converge to 0, i.e. there exists a positive constant $\epsilon > 0$ such that we can find a subsequence n_k satisfying

$$\mathbb{E}\left(\frac{|X_{n_k} - X|}{1+|X_{n_k} - X|}\right) > \epsilon, \quad \forall k.$$

It implies that for all subsequence $\{m_k\}$ of $\{n_k\}$, X_{n_k} can not converge almost surely to X. This is in contradiction with the hypothesis.

So we must have that $X_n \xrightarrow{\mathbb{P}} X$.

□

We have the following elementary but useful proposition.

Proposition 3.1.7. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function.*

1. *If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$ then $f(X_n, Y_n) \xrightarrow{a.s.} f(X, Y)$.*
2. *If $X_n \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$ then $f(X_n, Y_n) \xrightarrow{\mathbb{P}} f(X, Y)$.*

Proof. 1. Denote by $A = \{w \in \Omega : \lim_{n\to\infty} X_n(w) = X(w)\} \cap \{w \in \Omega : \lim_{n\to\infty} Y_n(w) = Y(w)\}$. It is clear that $\mathbb{P}(A) = 1$ and for all $w \in A$, we have $\lim_{n\to\infty} f(X_n(w), Y_n(w)) = f(X(w), Y(w))$ since f is continuous. Then $f(X_n) \xrightarrow{a.s} f(X)$.

2. From the second part of Proposition 3.1.6, in order to prove that $f(X_n, Y_n) \xrightarrow{\mathbb{P}} f(X, Y)$, we can check that for all subsequence $\{n_k\}$, there exists a subsequence $\{m_k\}$ such that $f(X_{m_k}, Y_{m_k}) \xrightarrow{a.s} f(X, Y)$. Indeed, since $X_{n_k} \xrightarrow{\mathbb{P}} X$ and $Y_{n_k} \xrightarrow{\mathbb{P}} Y$, then from the first part of Proposition 3.1.6, we can extract a subsequence $\{m_k\}$ satisfying $X_{m_k} \xrightarrow{a.s} X$ and $Y_{m_k} \xrightarrow{a.s} Y$. Then from the first part of this theorem, the result follows.

□

3.2 Laws of large numbers

In this section, we study the first special and classical limit theorem named ”Law of large number”. It was first stated but without proof by Cardano. Later, the first proof was given by Bernoulli when he considered the binary random variables. This theorem shows that in some cases, the limit behaviour of the average of some random variables is a constant.

More precise, throughout this section, we consider $(X_n)_{n\geq 1}$ a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and denote

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

We have the ”weak law” and the ”strong law”.

3.2.1 Weak laws of large numbers

In the weak law, we have the *convergence in probability* as follows.

Theorem 3.2.1. *Suppose that*

$$\lim_{n\to\infty} \frac{D(S_n)}{n^2} = 0.$$

Then

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0, \text{ as} \, n \to \infty.$$

Proof. For every $\epsilon > 0$, by the Chebyshev inequality,

$$\mathbb{P}\Big(\Big|\frac{S_n - \mathbb{E}S_n}{n}\Big| \geq \epsilon\Big) \leq \frac{D(S_n)}{n^2\epsilon^2} \to 0 \text{ khi } n \to \infty.$$

It implies the result. □

We have a recent corollary.

Corollary 3.2.2. *Let $(X_n)_{n\geq 1}$ be a sequence of pairwise uncorrelated random variables satisfying*

$$\lim_{n\to\infty} \frac{D(X_1) + \ldots + D(X_n)}{n^2} = 0.$$

Then

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0, \text{ as} n \to \infty.$$

Proof. Observe that $D(S_n) = D(X_1) + \ldots + D(X_n)$ and apply Theorem 3.2.1. □

In a special case, when X_n's are i.i.d with finite variance, we have

$$\lim_{n\to\infty} \frac{D(X_1)+\ldots+D(X_n)}{n^2} = \lim_{n\to\infty} \frac{n.D(X_1)}{n^2} = \lim_{n\to\infty} \frac{D(X_1))}{n} = 0.$$

So the condition in Theorem 3.2.1 is met. Moreover, by linearity,

$$\mathbb{E}S_n = \mathbb{E}X_1 + \mathbb{E}X_2 + \ldots + \mathbb{E}X_n = n\mathbb{E}X_1.$$

So we have proved the following corollary

Lemma 3.2.3. *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d random variables with finite variance. Then*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mathbb{E}X_1, \text{ as } n \to \infty.$$

Note that when X_n has the Bernoulli law, then S_n is the number of successful trials and Bernoulli showed that S_n/n converges in probability to the probability of success of a trial. However, his proof is much more complicated than the one given here.

3.2.2 Strong laws of large numbers

We first claim a simple version of the strong laws.

Theorem 3.2.4. *Let $(X_n)_{n\geq 1}$ be a sequence of pair-wise uncorrelated random variables satisfying $\sup_n D(X_n^2) \leq \sigma^2 < \infty$. Then*

$$\lim_{n\to\infty} \frac{S_n - \mathbb{E}S_n}{n} = 0 \quad \text{a.s and in } L^2.$$

Proof. At first, we assume that $\mathbb{E}(X_n) = 0$. Denote $Y_n = S_n/n$. Then $\mathbb{E}(Y_n) = 0$, and from Proposition 2.7.3,

$$\mathbb{E}(Y_n^2) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i \leq \frac{\sigma^2}{n}.$$

Hence $Y_n \overset{L^2}{\to} 0$. We also have

$$\sum_{n=1}^{\infty} \mathbb{E}(Y_{n^2}^2) \leq \sum_{n=1}^{\infty} \frac{\sigma^2}{n^2} < \infty.$$

From the Monotone Convergence Theorem,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} Y_{n^2}^2\right] = \sum_{n=1}^{\infty} \mathbb{E}(Y_{n^2}^2) < \infty,$$

so $\sum_{n=1}^{\infty} Y_{n^2}^2 < \infty$ almost surely. It implies that

$$Y_{n^2} \overset{a.s}{\to} 0. \tag{3.2}$$

For each $n \in \mathbb{N}$, denote by $p(n)$ the integer part of $\sqrt{n}$. Since

$$Y_n - \frac{p(n)^2}{n} Y_{p(n)^2} = \frac{1}{n} \sum_{j=p(n)^2+1}^{n} X_j,$$

we have

$$\mathbb{E}\Big[\Big(Y_n - \frac{p(n)^2}{n}Y_{p(n)^2}\Big)^2\Big] \le \frac{n-p(n)^2}{n^2}\sigma^2 \le \frac{2p(n)+1}{n^2}\sigma^2 \le \frac{2\sqrt{n}+1}{n^2}\sigma^2 \le \frac{3}{n^{3/2}}\sigma^2,$$

with the observations $n \le (p(n)+1)^2$ and $p(n) \le \sqrt{n}$. By the same argument, since

$$\sum_{n=1}^{\infty}\mathbb{E}\Big[\Big(Y_n - \frac{p(n)^2}{n}Y_{p(n)^2}\Big)^2\Big] \le \sum_{n=1}^{\infty}\frac{3}{n^{3/2}}\sigma^2 < \infty,$$

then

$$Y_n - \frac{p(n)^2}{n}Y_{p(n)^2} \overset{h.c.c}{\to} 0.$$

From (3.2) and the observation $\frac{p(n)^2}{n} \to 1$, we deduce that $Y_n \overset{a.s}{\to} 0$.

In general, if $\mathbb{E}(X_n) \neq 0$, we denote $Z_n = X_n - \mathbb{E}(X_n)$. Then $\{Z_n\}$ is a sequence of pair-wise uncorrelated random variables with mean zero satisfying the condition of the theorem. Therefore

$$\frac{S_n - \mathbb{E}S_n}{n} = \frac{Z_1 + \ldots + Z_n}{n} \overset{a.s}{\to} 0.$$

□

In the following, we state without proof two general versions of strong law of large numbers.

Theorem 3.2.5. *Let $(X_n)_{n\ge1}$ be a sequence of independent random variable and, $(b_n)_{n\ge1}$ a sequence of positive numbers satisfying $b_n \uparrow \infty$. If*

$$\sum_{n=1}^{\infty}\frac{DX_n}{b_n^2} < \infty \textit{ then } \frac{S_n - \mathbb{E}(S_n)}{b_n} \xrightarrow{a.s.} 0.$$

Theorem 3.2.6. *Let $(X_n)_{n\ge1}$ be a sequence of iid random variables. Then*

$$\lim_{n\to\infty}\frac{S_n}{n} = \mathbb{E}(X_1) \textit{ iff} \mathbb{E}(|X_1|) < \infty.$$

Example 3.2.7. Consider $(X_n)_{n\ge1} \overset{i.i.d}{\sim} B(1,p)$. From Theorem 3.2.6,

$$\frac{S_n}{n} \xrightarrow{h.c.c} \mathbb{E}(X_1) = p.$$

Then, to approximate the probability of success of each trial, we can use the approximation S_n/n for n large enough.

An application of Strong law of large numbers that is quite simple but very useful is the Monte Carlo method.

Example 3.2.8. Let f be an integrable function over $[0,1]$, i.e.

$$\int_0^1 |f(x)|dx < \infty. \tag{3.3}$$

In most of the practical applications, the quantity $I = \int_0^1 f(x)dx$ can not be calculated exactly by the analytical method. Therefore one usually approximate it by the numerical method. When f

is smooth enough, I can be approximated well by taking the average (with some weight) of the values of f at some fixed points. For example, if f is twice differentiable, we have

$$I \approx \frac{f(t_0^n) + 2f(t_1^n) + \ldots + 2f(t_{n-1}^n) + f(t_n^n)}{2n},$$

where $t_i^n = \frac{i}{n}$, $i = 0, 1, \ldots, n$.

However, the above method is not good in the sense that we must take too many points to have a good approximation when f is not smooth enough. In this case, we can use the Monte Carlo method that can be stated in the simplest version as follows. Let $(U_j)_{j\geq 1}$ be a sequence of i.i.d random variables of the uniform distribution over $[0, 1]$ and denote

$$I_n = \frac{1}{n}\sum_{j=1}^{n} f(U_j).$$

Since $\mathbb{E}[|f(U_j)|] = \int_0^1 |f(x)|dx < \infty$, then from Theorem 3.2.6, I_n converges almost surely to $\mathbb{E}[f(U_1)] = I$ as $n \to \infty$. To evaluate the error of the approximation, we assume more that

$$\int_0^1 |f(x)|^2 dx < \infty. \tag{3.4}$$

Then, the square of the error is

$$\mathbb{E}[(I_n - I)^2] = \mathbb{E}[(I_n - \mathbb{E}[I_n])^2] = \frac{1}{n} Df(U_1) \leq \frac{1}{n}\int_0^1 |f(x)|^2 dx.$$

In practical, we use the computer to generate the sequence $(U_j)_{j\geq 1}$ and obtain an approximation of I for any function f satisfying the condition (3.3). Under the condition (3.4), the error of the approximation only depends on the size n and not on the smoothness of f. The Monte Carlo method also seems to be more useful than the other deterministic ones in approximating the multiple integral. The only thing we must care about is the square of the error. If we can reduce it, then the calculation will be more accurate and we can also reduce the time on computer (see [**?**]). That is the way one wants to improve the Monte Carlo method.

The error of the Monte Carlo method will be analysed in more detail based on the Central limit theorems that will be explained in the following.

3.3 Central limit theorems

In this section, we will state and prove the second classical limit theorem in probability theory named ”Central limit theorem”. It is the most beautiful pearl of probability and has a lot of applications in statistics. However, to understand this theorem, we need to define a new mode of convergence and the tools to study it.

3.3.1 Characteristic functions

Sometimes to analyse a quantity or a function, it is better to transform it in another form. Since a random variable can be seen as a special function, we can do the same. In this section, we study the Fourier transformation of a random variable. We have the following definition.

Definition 3.3.1. 1. Let X be a random variable. We define its *characteristics function* by

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_{\mathbb{R}} e^{itx} dF_X(x).$$

2. The *characteristic function* of random vector $X = (X_1, \dots, X_n)$ is defined by

$$\varphi_X(t_1, \dots, t_n) = \mathbb{E}\Big[\exp\Big(i \sum_{j=1}^{n} t_j X_j\Big)\Big].$$

Theorem 3.3.2. *For every random variable X, the characteristic function φ_X has the following properties;*

1. $\varphi_X(0) = 1$*;*
2. $\varphi_X(-t) = \overline{\varphi_X(t)}$*;*
3. $|\varphi_X(t)| \le 1$*;*
4. $|\varphi_X(t+h) - \varphi_X(t)| \le \mathbb{E}[|e^{ihX} - 1|]$*, so φ_X is uniformly continuous on $(-\infty, +\infty)$;*
5. $\mathbb{E}[e^{it(aX+b)}] = e^{itb}\varphi_X(at)$*.*

Proof. It is easy to see that $\varphi_X(0) = 1$. Applying the inequality $(\mathbb{E}X)^2 \le \mathbb{E}(X^2)$,

$$|\varphi_X(t)| = \sqrt{(\mathbb{E}\cos tX)^2 + (\mathbb{E}\sin tX)^2} \le \sqrt{\mathbb{E}(\cos^2 tX) + \mathbb{E}(\sin^2 tX)} = 1,$$

then φ_X is bounded. And the continuity of can be deduced by Lebesgue dominated convergence theorem. □

The following theorem shows the connection between the characteristic function of a random variable and its moments.

Theorem 3.3.3. *If $\mathbb{E}[|X|^m] < \infty$ for some positive integer m. Then φ_X has continuous derivatives up to order m, and*

$$\varphi^{(k)}(t) = i^k \mathbb{E}[X^k e^{itX}], \tag{3.5}$$

$$\mathbb{E}[X^k] = \frac{\varphi^{(k)}(0)}{i^k}, \tag{3.6}$$

$$\varphi_X(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!}\mathbb{E}[X^k] + \frac{(it)^n}{n!}\alpha_n(t), \tag{3.7}$$

where $|\alpha_n(t)| \le 2\mathbb{E}(|X^n|)$ and $\alpha_n(t) \to 0$ as $t \to 0$.
On the other hand, if $\varphi^{(2m)}(0)$ exists and is finite for some positive integer m, then $\mathbb{E}[X^{2m}] < \infty$.

Proof. Since $\mathbb{E}(|X|^m) < \infty$, we have $\mathbb{E}(|X|^k) < \infty$ for all $k = 1, \dots, m$. Then

$$\sup_t \int |(ix)^k e^{itx}| dF_X(x) \le \int |x|^k dF_X(x) < \infty.$$

From Lebesgue theorem, we can take the differentation under the integral sign and obtain (3.5). In (3.5), let $t = 0$ then we have (3.6).

Consider the Taylor expansion of function $\exp(x)$ at $x = 0$,

$$\begin{aligned}\mathbb{E}(e^{itX}) &= \mathbb{E}\Big(\sum_{k=0}^{n-1}\frac{(itX)^k}{k!} + \frac{(itX)^n}{n!}e^{i\theta X}\Big)\\ &= \sum_{k=0}^{n-1}\frac{(it)^k}{k!}\mathbb{E}(X^k) + \frac{(it)^n}{n!}\Big(\mathbb{E}(X^n) + \alpha_n(t)\Big),\end{aligned}$$

where $|\theta| \le |t|,\ \alpha_n(t) = \mathbb{E}\big(X^n(e^{i\theta X} - 1)\big)$. Therefore $|\alpha_n(t)| \le 2\mathbb{E}(|X|^n)$, i.e it is bounded. So from the Dominated convergence theorem, we have $\alpha_n(t) \to 0$ as $t \to 0$.

The inverse statement can be proved by concurrence, see [13, page 190-193]. □

We consider the characteristic function of some usual distributions.

Example 3.3.4. Let $X \sim Poi(\lambda)$. We have

$$\varphi_X(t) = \sum_{k=0}^{\infty}\frac{e^{-\lambda}\lambda^k}{k!}e^{itk} = e^{-ld}\frac{(\lambda e^{it})^k}{k!} = e^{\lambda(e^{it}-1)}.$$

Example 3.3.5. Let $X \sim N(0,1)$. We have

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = \int_{-\infty}^{\infty}\frac{\cos tx}{\sqrt{2\pi}}e^{-x^2/2}dx + \int_{-\infty}^{\infty}\frac{\sin tx}{\sqrt{2\pi}}e^{-x^2/2}dx.$$

Since the funtion $x \mapsto e^{-x^2/2}\sin tx$ is an odd and integrable function, the second integral equals to 0. Thanks to Theorem 3.3.3, we can take the derivative of both side with respect to t and get

$$\varphi_X'(t) = -\int_{-\infty}^{\infty}\frac{\sin tx}{\sqrt{2\pi}}xe^{-x^2/2}dx.$$

By integration by parts,

$$\varphi_X'(t) = -\int_{-\infty}^{\infty}\frac{\cos tx}{\sqrt{2\pi}}te^{-x^2/2}dx = -t\varphi_X(t).$$

The differential equation $\frac{\varphi_X'}{\varphi_X} = -t$ with initial condition $\varphi_X(0) = 1$ has a solution

$$\varphi_X(t) = e^{-t^2/2}.$$

If $X \sim N(a,\sigma^2)$ then

$$\varphi_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\int e^{itx}e^{-\frac{(x-a)^2}{2\sigma^2}}dx.$$

Using the change of variable: $y = (x-a)/\sigma$, we get

$$\varphi_X(t) = \frac{e^{ita}}{\sqrt{2\pi}}\int e^{it\sigma y}e^{-y^2/2}dy = e^{ita - t^2\sigma^2/2}.$$

The following theorem shows the meaning of the name ”characteristic function”.

Theorem 3.3.6. *Two random vectors have the same distribution if their characteristic functions coincide. Moreover, if* $\int |\varphi_X(t)|dt < \infty$ *then* X *has bounded continuous density given by*

$$f_X(y) = \frac{1}{2\pi} e^{-ity}\varphi(t)dt.$$

Example 3.3.7. Let X and Y have Poisson distribution with the corresponding parameters μ and λ. Assume more that X and Y are independent. Let us consider the distribution of the random variable $X+Y$. We can compute its characteristic function as

$$\varphi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) = e^{(\lambda+\mu)(e^{it}-1)}.$$

Then this characteristic function agrees with the one of $Poi(\mu+\lambda)$. So the random variable $X+Y$ has the Poisson distribution with the parameter $\mu+\lambda$.

We can also use the characteristic function to check whether the random variables are independent.

Theorem 3.3.8. $X_1, \ldots, X_n$ *are independent random variables iff*

$$\varphi_{(X_1,\ldots,X_n)}(t_1,\ldots,t_n) = \varphi_{X_1}(t_1)\ldots\varphi_{X_n}(t_n) \text{ for all } (t_1,\ldots,t_n) \in \mathbb{R}^n.$$

Example 3.3.9. Let X and Y be independent random variables which have standard normal distribution $N(0,1)$. According to Example 3.3.5, we have

$$\varphi_{(X+Y,X-Y)}(t,s) = \mathbb{E}e^{it(X+Y)+is(X-Y)} = \mathbb{E}e^{i(t+s)X}\mathbb{E}e^{i(t-s)Y} = e^{-t^2-s^2}.$$

Put $s=0$ and $t=0$, we have $\varphi_{X+Y}(t) = e^{-t^2}$ and $\varphi_{X-Y}(s) = e^{-s^2}$. Hence both $X+Y$ and $X-Y$ have normal distribution $N(0,2)$. Furthermore, they are independent since

$$\varphi_{(X+Y,X-Y)}(t,s) = \varphi_{X+Y}(t)\varphi_{X-Y}(s) \text{ for all } t,s \in \mathbb{R}.$$

3.3.2 Weak convergence

In this section, we consider another mode of convergence of a sequence of random variables that is the *weak convergence*. In the first three modes of convergence, we can see the trace of analysis such as the limit of a sequence of numbers or a Cauchy sequence. Here the weak convergence totally comes from the probability theory.

Definition 3.3.10. X_n *converges weakly* to X and denoted by $X_n \xrightarrow{w} X$, if $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for each f which is real-valued, continuous and bounded.
When $X_n \xrightarrow{w} X$ we also say that F_{X_n} converges weakly to F_X and denote $F_{X_n} \xrightarrow{w} F_X$.

Note that we do not require or suppose that the random variables $(X_n)_{n\geq 1}$ are defined on the same probability space in the above definition. We just care about the expectation or the distribution. Therefore sometimes we call weakly convergence by convergence in distribution (See Exercise 3.27).

If we suppose that X_n's and X are defined on the same probability space, we have the following propositions.

Proposition 3.3.11. *Let $(X_n)_{n\geq 1}$ and X be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \xrightarrow{w} X$.*

Proof. We prove by contradiction. Assume that $X_n \xrightarrow{\mathbb{P}} X$ but $X_n \overset{w}{\not\longrightarrow} X$. Then there exist a bounded continuous function f, a constant $\epsilon > 0$ and a subsequence $(n_k)_{k\geq 1}$ such that

$$|\mathbb{E}(f(X_{n_k})) - \mathbb{E}(f(X))| > \epsilon \text{ for all } k \geq 1. \tag{3.8}$$

From Proposition 3.1.6, there exists a subsequence $(m_k)_{k\geq 1}$ of the sequence $(n_k)_{k\geq 1}$ such that $X_{m_k} \xrightarrow{a.s} X$. Since f is continuous, $f(X_{m_k}) \xrightarrow{h.c.c} f(X)$. By Dominated Convergence Theorem, $\mathbb{E}(f(X_{m_k})) \to \mathbb{E}(f(X))$. It is in contradiction with (3.8). Then the result follows. □

Proposition 3.3.12. *Let $(X_n)_{n\geq 1}$ and X be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X_n \xrightarrow{w} X$ and $X = const$ a.s then $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. Let $X \equiv a$ a.s. Consider the bounded continuous function $f(x) = \frac{|x-a|}{|x-a|+1}$. Since $X_n \xrightarrow{w} a$, $\mathbb{E}(f(X_n)) \to f(a) = 0$. From Proposition 3.1.4, $X_n \overset{\mathbb{P}}{\to} a$. □

The following theorem gives us a very useful criterion to verify the weak convergence of random variables by using the characteristic function. Its proof is provided in [13, page 196-199].

Theorem 3.3.13. *Let $(F_n)_{n\geq 1}$ be a sequence of distribution function whose characteristic functions are $(\varphi_n)_{n\geq 1}$ respectively,*

$$\varphi_n(t) = \int_{\mathbb{R}} e^{itx} dF_n(x).$$

1. *If $F_n \overset{w}{\to} F$ for some distribution function F then (φ_n) converges point-wise to the characteristic function φ of F.*
2. *If $\varphi_n(t) \to \varphi(t),\ t \in \mathbb{R}$. Then the following statements are equivalent.*
   (a) *$\varphi(t)$ is a characteristic function and $F_n \overset{w}{\to} F$ where F is a distribution function whose characteristic function is φ;*
   (b) *φ is continuous at $t = 0$.*

Example 3.3.14. Let X_n be normal $N(a_n, \sigma_n^2)$. Suppose that $a_n \to 0$ and $\sigma_n^2 \to 1$ as $n \to \infty$. Then the sequence (X_n) converges weakly to $N(0, 1)$ since

$$\varphi_{X_n}(t) = e^{ita_n - \sigma_n^2 t^2/2} \to e^{-t^2/2}.$$

Example 3.3.15 (Weak laws of large numbers)**.** Let $(X_k)_{k\geq 1}$ be a iid sequence of random variables whose mean is finite. Then

$$\frac{1}{n}(X_1 + \ldots + X_n) \xrightarrow{\mathbb{P}} a.$$

Indeed, denote $S_n = X_1 + \ldots + X_n$ and φ is characteristic function of X_k. Then,

$$\varphi_{S_n/n}(t) = \varphi_{S_n}(t/n) = [\varphi(t/n)]^n.$$

Thanks to Theorem 3.3.3, we have

$$\varphi(t/n) = 1 + \frac{ita}{n} + \frac{t}{n}\alpha(t/n),$$

with $\alpha(t) \to 0$ as $t \to 0$. Thus,

$$\varphi_{S_n/n}(t) = \Big(1 + \frac{ita}{n} + \frac{t}{n}\alpha(t/n)\Big)^n \to e^{ita}$$

as $n \to \infty$. Note that if $X \equiv a$ then $\varphi_X(t) = e^{ita}$. Hence, $S/n \overset{w}{\to} a$. Applying Proposition 3.3.12 we obtain the desired result.

The following theorem is very useful in statistics.

Theorem 3.3.16 (Slutsky's theorem)**.** *Suppose that $X_n \overset{w}{\to} X$, $A_n \overset{\mathbb{P}}{\to} a$ and $B_n \overset{\mathbb{P}}{\to} b$ where a and b are some constants. Then*

$$A_n + B_n X_n \overset{w}{\to} a + bX.$$

3.3.3 Central limit theorem

The central limit theorem is stated as follows.

Theorem 3.3.17. *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d random variables and $\mathbb{E}(X_n) = \mu$ and $DX_n = \sigma^2 \in (0,\infty)$. Denote $S_n = X_1 + \ldots + X_n$. Then $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \overset{w}{\to} N(0,1)$.*

Proof. Denote φ by the characteristic function of the random variable $X_n - \mu$. Since X_n's have the same law, φ does not depend on n. Moreover, since X_n's are independent,

$$\varphi_{Y_n}(t) = \mathbb{E}\exp\Big(it\sum_{j=1}^n \frac{X_j - \mu}{\sigma\sqrt{n}}\Big) = \prod_{j=1}^n \mathbb{E}\exp\Big(it\frac{X_j - \mu}{\sigma\sqrt{n}}\Big) = \Big(\varphi\big(\frac{t}{\sigma\sqrt{n}}\big)\Big)^n.$$

It is clear that $\mathbb{E}(X_j - \mu) = 0$ and $\mathbb{E}((X_j - \mu)^2) = \sigma^2$. Then from Theorem 3.3.3, φ has the continuous second derivative and

$$\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + t^2\alpha(t),$$

where $\alpha(t) \to 0$ as $t \to 0$. Using the expansion $\ln(1+x) = x + o(x)$ as $x \to 0$,

$$\ln\varphi_{Y_n}(t) = n\ln\Big(1 - \frac{t^2}{2n} + \frac{t^2}{n\sigma^2}\alpha\big(\frac{t}{\sigma\sqrt{n}}\big)\Big) \to -\frac{t^2}{2}.$$

Therefore $\varphi_{Y_n}(t) \to e^{-t^2/2}$ as $n \to \infty$. Applying Theorem 3.3.13, we have the desired result. □

In the following, we give an example of the central limit theorem. More detail, we will approximate the binomial probability by the normal probability.

Example 3.3.18. We know that a binomial random variable $S_n \sim B(n,p)$ can be written as the sum of n i.i.d random variables $\sim B(1,p)$. Then as n large enough, from the central limit theorem, we can approximate the random variable $(S_n - np)/\sqrt{np(1-p)}$ by the standard normal variable $\mathcal{N}(0,1)$.

Usually, the probability that $a \leq S_n \leq b$ can be formulated as

$$\Sigma_{i=a}^{b} C_n^i p^i (1-p)^{n-i}.$$

However, when n is too large, calculating C_n^i for some i is impossible since it exceeds the capacity of the calculator or the computer (please, consider $1000!$ or $5000!$). Then in practical, we can estimate this probability by

$$\begin{aligned}\mathbb{P}(a \leq S_n \leq b) &= \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \in \left[\frac{a-np}{\sqrt{np(1-p)}}, \frac{b-np}{\sqrt{np(1-p)}}\right]\right) \\ &\cong \mathbb{P}\left(\mathcal{N}(0,1) \in \left[\frac{a-np}{\sqrt{np(1-p)}}, \frac{b-np}{\sqrt{np(1-p)}}\right]\right).\end{aligned}$$

Note that to compute the last probability, we can write down it as an integral from the density function of the normal variable. It can be computed or approximated easily.

In order to define the rate that the distribution of F_{Y_n} converges to normal distribution, we use the Berry-Esseen's inequality: Suppose $\mathbb{E}(|X_1|^3) < \infty$ then

$$\sup_{-\infty < x < \infty} \left| F_{Y_n}(x) - \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \right| \leq K_{BE} \frac{\mathbb{E}(|X_1 - \mathbb{E}X_1|^3)}{\sigma^3 \sqrt{n}}, \tag{3.9}$$

where K_{BE} is some constant in $(\frac{\sqrt{10}+3}{6\sqrt{2\pi}}, 0.4748)$ (see[12]).

The condition that X_n's are iid is too restrictive. Many authors manage to weaken this condition. In the following, we state the Lindeberg's central limit theorem. Its proof can be found in [13, page 221-225].

Theorem 3.3.19. *Let $(X_n)_{n\geq 1}$ be a sequence of independent random variables with finite variance. Denote $S_n = X_1 + \ldots + X_n,\ B_n = DX_1 + \ldots + DX_n$. Suppose that*

$$L_n(\epsilon) := \frac{1}{B_n^2} \sum_{k=1}^{n} \mathbb{E}\Big((X_k - \mathbb{E}X_k)^2 \mathbb{I}_{\{|X_k - \mathbb{E}X_k| > \epsilon B_n\}}\Big) \to 0,\ \forall \epsilon > 0. \tag{3.10}$$

Then $S_n^* = \frac{S_n - \mathbb{E}S_n}{B_n} \xrightarrow{w} N(0,1)$.

3.4 Exercises

3.4.1 Convergence of random variables

3.1. Prove that if $X_n \xrightarrow{\mathbb{P}} X$ and, at the same time, $X_n \xrightarrow{\mathbb{P}} Y$, then X and Y are equivalent, in the sense that $\mathbb{P}[X \neq Y] = 0$.

3.2. Show that $d_{\mathbb{P}}$ is a metric on L^0, it means that

1. $d(X,Y) \geq 0$ and $d(X,Y) = 0$ iff $X = Y$ a.s.;

2. $d(X,Y) = d(Y,X)$;

3. $d(X,Y) \leq d(X,Z) + d(Z,Y)$;

for any random variables X, Y, Z.

3.3. Show that $(X_n)_{n\geq 1}$ converges in probability to X iff

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X| \wedge 1) = 0.$$

3.4. Consider the probability space $([0;1], \mathcal{B}([0;1]), P)$. Let $X = 0$ and $X_1, X_2, \ldots$ be random variables

$$X_n(\omega) = \begin{cases} 0 & \text{if } \frac{1}{n} \leq \omega \leq 1 \\ e^n & \text{if } 0 \leq \omega < 1/n. \end{cases}$$

Show that $X \xrightarrow{P} X$, but $E|X_n - X|^p$ does not converge for any $p > 0$.

3.5. Consider the probability space $([0;1], \mathcal{B}([0;1]), P)$. Let $X = 0$. For each $n = 2^m + k$ where $0 \leq k < 2^m$, we define

$$X_n(\omega) = \begin{cases} 1 & \text{if } \frac{k}{2^m} \leq \omega \leq \frac{k+1}{2^m} \\ 0 & \text{otherwise}. \end{cases}$$

Show that $X \xrightarrow{P} X$, but $\{X_n\}$ does not converge to X a.s.

3.6. Let $(X_n)_{n\geq 1}$ be a sequence of exponential random variables with parameter $\lambda = 1$. Show that

$$\mathbb{P}\Big[\limsup_{n\to\infty} \frac{X_n}{\ln n} = 1\Big] = 1.$$

3.7. Let $X_1, X_2, \ldots$ be a sequence of identically distributed random variables with $E|X_1| < \infty$ and let $Y_n = n^{-1} \max_{1\leq i\leq n} |X_i|$. Show that $\lim_n E(Y_n) = 0$ and $\lim_n Y_n = 0$ a.s.

3.8. [5] Let $(X_n)_{n\geq 1}$ be random variables with $X_n \xrightarrow{P} X$. Suppose $|X_n(\omega)| \leq C$ for a constant $C > 0$ and all ω. Show that $\lim_{n\to\infty} E|X_n - X| = 0$.

3.4.2 Law of large numbers

3.9. [10] Let $X_1, \ldots, X_n$ be independent and identically distributed random variables such that for $x = 3, 4, \ldots, P(X_1 = \pm x) = (2cx^2 \log x)^{-1}$, where $c = \sum_{x=3}^{\infty} x^{-2}/\log x$. Show that $E|X_1| = \infty$ but $n^{-1} \sum_{i=1}^{n} X_i \xrightarrow{P} 0$.

3.10. [10] Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $Var(S_1) < \infty$. Show that

$$\frac{1}{n(n+1)} \sum_{j=1}^{n} jX_j \xrightarrow{P} EX_1.$$

3.11. [2] If for every n, $Var(X_i) \leq c < \infty$ and $Cov(X_i, X_j) < 0$ $(i, j = 1, 2, \ldots)$, then the WLLN holds.

3.12. [2] *(Theorem of Bernstein)* Let $\{X_n\}$ be a sequence of random variables so that $Var(X_i) \leq c < \infty$ $(i = 1, 2, \ldots)$ and $Cov(X_i, X_j) \to 0$ when $|i - j| \to \infty$ then the WLLN holds.

3.13. [5] Let $(Y_j)_{j\geq 1}$ be a sequence of independent Binomial random variables, all defined on the same probability space, and with law $B(p, 1)$. Let $X_n = \sum_{j=1}^{n} Y_j$. Show that X_j is $B(p, j)$ and that $\frac{X_j}{j}$ converges a.s to p.

3.14. [5] Let $\{X_j\}_{j\geq 1}$ be i.i.d with X_j in L^1. Let $Y_j = e^{X_j}$. Show that $\left(\prod_{j=1}^{n} Y_j\right)^{\frac{1}{n}}$ converges to a constant α a.s.

3.15. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with X_j in L^1 and $EX_j = \mu$. Let $(Y_j)_{j\geq 1}$ be also i.i.d with Y_j in L^1 and $EY_j = \nu \neq 0$. Show that

$$\lim_{n\to\infty} \frac{1}{\sum_{j=1}^{n} Y_j} \sum_{j=1}^{n} X_j = \frac{\mu}{\nu} \quad \text{a.s.}$$

3.16. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with X_j in L^1 and suppose $\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (X_j - \nu)$ converges in distribution to a random variable Z. Show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j = \nu \quad \text{a.s.}$$

3.17. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with X_j in L^p. Show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j^p = EX^p \quad \text{a.s.}$$

3.18. [5] Let $(X_j)_{j\geq 1}$ be i.i.d. $N(1; 3)$ random variables. Show that

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \ldots X_n}{X_1^2 + X_2^2 + \ldots + X_n^2} = \frac{1}{4} \quad \text{a.s.}$$

3.19. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with mean μ and variance σ^2. Show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 = \sigma^2 \quad \text{a.s.}$$

3.4.3 Characteristic function

3.20. Find the characteristic function of X,

1. $\mathbb{P}[X = 1] = \mathbb{P}[X = -1] = 1/2$;

2. $\mathbb{P}[X = 1] = \mathbb{P}[X = 0] = 1/2$;

3. $X \sim U(a, b)$;

4. the density of X is $f(x) = (1-|x|)I_{|x|<1}$;

5. $X \sim Exp(\lambda)$;

3.21. Show that if $X_1, \ldots, X_n$ are independent and uniformly distribution on $(-1,1)$, then for $n \geq 2$, $X_1 + \ldots + X_n$ has density

$$f(x) = \frac{1}{\pi}\int_0^\infty \Big(\frac{\sin t}{t}\Big)^n \cos tx dt.$$

3.22. Suppose that X has density

$$f(x) = \frac{1-\cos x}{\pi x^2}.$$

Show that

$$\varphi_X(t) = (1-|t|)^+.$$

3.23. 1. Suppose that X has Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Show that

$$\varphi_X(t) = e^{-|t|}.$$

2. Let $X_1, \ldots, X_n$ be a sequence of independent Cauchy random variables. Find the distribution of $(X_1 + \ldots + X_n)/n$.

3.24. Let $X_1, X_2, \ldots$ be independent taking values 0 and 1 with probability $1/2$ each.

1. Find the distribution of $\xi = \sum_{i=1}^\infty \frac{X_i}{2^i}$.

2. Find the characteristic function of $\zeta = 2\sum_{i=1}^\infty \frac{X_i}{3^i}$. We say that ζ has the Cantor distribution.

3.4.4 Weak convergence

3.25. Show that if X_n and Y_n are independent for $1 \leq n$, $X_n \overset{w}{\to} X$ and $Y_n \overset{w}{\to} Y$, then $X_n + Y_n \overset{w}{\to} X + Y$.

3.26. Consider the probability space $([0;1], \mathcal{B}([0;1]), P)$. Let X and $X_1, X_2, \ldots$ be random variables

$$X_{2n}(\omega) = \begin{cases} 1 & \text{if } 0 \leq \omega \leq 1/2 \\ 0 & \text{if } 1/2 < \omega \leq 1. \end{cases}$$

and

$$X_{2n+1}(\omega) = \begin{cases} 0 & \text{if } 0 \leq \omega \leq 1/2 \\ 1 & \text{if } 1/2 < \omega \leq 1. \end{cases}$$

Show that the sequence (X_n) converges in distribution? Does it converge in probability?

3.27. Let $(X_n)_{n\geq 1}$ and X are random variables whose distribution functions are $(F_n)_{n\geq 1}$ and F, respectively.

1. If $X_n \xrightarrow{w} X$ then $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in D$ where D is a dense subset of $\mathbb{R}$ given by
$$D = \{x \in \mathbb{R} : F(x+) = F(x)\}.$$

2. If $\lim_{n\to\infty} F_n(x) = F(x)$ for any x in some dense subset of $\mathbb{R}$ then $X_n \xrightarrow{w} X$.

3.28. If $X_n \xrightarrow{w} X, Y_n \xrightarrow{P} c$, then

a) $X_n + Y_n \xrightarrow{w} X + c$

b) $X_n Y_n \xrightarrow{w} cX$

c) $X_n/Y_n \xrightarrow{w} X/c$ if $Y_n \neq 0$ a.s for all n and $c \neq 0$.

3.29. [10] Show that if $X_n \xrightarrow{d} X$ and $X = c$ a.s for a real number c, then $X_n \xrightarrow{P} X$.

3.30. [10] A family of random variable $(X_i)_{i\in I}$ is called *uniformly integrable* if
$$\lim_{N\to\infty} \sup_{i\in I} \mathbb{E}[|X_i| I_{\{|X_i|\geq N\}} = 0.$$

Let $X_1, X_2, \dots$ be random variables. Show that $\{|X_n|\}$ is uniformly integrable if one of the following condition holds:

a) $\sup_n E|X_n|^{1+\delta} < \infty$ for a $\delta > 0$.

b) $P(|X_n| \geq c) \leq P(|X| \geq c)$ for all n and $c > 0$, where X is an integrable random variable.

3.31. Let X_n be random variable distributed as $N(\mu_n, \sigma_n^2),\ n = 1, 2, \dots$ and X be a random variable distributed as $N(\mu, \sigma^2)$. Show that $X_n \xrightarrow{d} X$ if and only if $\lim_n \mu_n = \mu$ and $\lim_n \sigma_n^2 = \sigma^2$.

3.32. If Y_n are random variables with characteristic function φ_n, then $Y_n \overset{w}{\to} 0$ iff there is a $\delta > 0$ so that $\varphi_n(t) \to 1$ for $|t| \leq \delta$.

3.4.5 Central limit theorems

3.33. [10] Let $U_1, U_2, \dots$ be independent random variables having the uniform distribution on [0;1] and $Y_n = (\prod_{i=1}^n U_i)^{-1/n}$. Show that $\sqrt{n}(Y_n - e) \xrightarrow{d} N(0, e^2)$.

3.34. [10] Suppose that X_n is a random variable having the binomial distribution with size n and probability $\theta \in (0,1),\ n = 1, 2, \dots$ Define $Y_n = \log(X_n/n)$ when $X_n \geq 1$ and $Y_n = 1$ when $X_n = 0$. Show that $\lim_n Y_n = \log\theta$ a.s and $\sqrt{n}(Y_n - \log\theta) \xrightarrow{d} N(0, \frac{1-\theta}{\theta})$.

3.35. [2] Show that for the sequence $\{X_n\}$ of independent random variables with

a) $P[X_n = \pm 1] = \frac{1-2^{-n}}{2}, \quad P[X_n = \pm 2^n] = \frac{1}{2^{n+1}}, \quad n = 1, 2, \dots,$

b) $P[X_n = \pm n^2] = \frac{1}{2},$

the CLT holds.

3.36. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with $EX_1 = 1$ and $\sigma^2_{X_1} = \sigma^2 \in (0;\infty)$. Show that

$$\frac{2}{\sigma}(\sqrt{S_n} - \sqrt{n}) \xrightarrow{d} N(0,1).$$

3.37. [5] Show that

$$\lim_{n\to\infty} e^{-n}\left(\sum_{k=0}^{n} \frac{n^k}{k!}\right) = \frac{1}{2}.$$

3.38. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with $EX_j = 0$ and $\sigma^2_{X_j} = \sigma^2 < \infty$. Let $S_n = \sum_{j=1}^{n} X_j$. Show that

$$\lim_{n\to\infty} E\left\{\frac{|S_n|}{\sqrt{n}}\right\} = \sqrt{\frac{2}{n}}\sigma.$$

3.39. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with the uniform distribution on (-1;1). Let

$$Y_n = \frac{\sum_{j=1}^{n} X_j}{\sum_{j=1}^{n} X_j^2 + \sum_{j=1}^{n} X_j^3}.$$

Show that $\sqrt{n}Y_n$ converges in distribution.

3.40. [5] Let $(X_j)_{j\geq 1}$ be i.i.d with the uniform distribution on $(-j;j)$.

a) Show that

$$\frac{S_n}{n^{\frac{3}{2}}} \xrightarrow{d} N(0;\frac{1}{9}).$$

b) Show that

$$\frac{S_n}{\sqrt{\sum_{j=1}^{n} \sigma_j^2}} \xrightarrow{d} N(0,1).$$

Chapter 4

Some useful distributions in statistics

4.1 Gamma, chi-square, student and F distributions

4.1.1 Gamma distribution

Recall that a random variable X has a Gamma distribution $\mathcal{G}(\alpha, \lambda)$ if its density is given by

$$f_X(x) = \frac{x^{\alpha-1}e^{-x/\lambda}}{\Gamma(\alpha)\lambda^{\alpha}} I_{\{x>0\}}.$$

Note that $\mathcal{G}(1, \lambda) = Exp(\lambda)$.

Proposition 4.1.1. *If X is $\mathcal{G}(\alpha, \lambda)$ distributed, then*

$$\mathbb{E}[X] = \alpha\lambda, \quad DX = \alpha\lambda^2.$$

Moreover, the characteristic function of X is given by

$$\varphi_X(t) = \int_0^{\infty} e^{itx} \frac{x^{\alpha-1}e^{-x/\lambda}}{\Gamma(\alpha)\lambda^{\alpha}} dx = \Big(\frac{1}{1-i\lambda t}\Big)^{\alpha}.$$

Corollary 4.1.2. *Let $(X_i)_{1\le i\le n}$ be a sequence of independent random variables. Suppose that X_i is $\mathcal{G}(\alpha_i, \lambda)$ distributed. Then $S = X_1 + \cdots + X_n$ is $\mathcal{G}(\alpha_1 + \cdots + \alpha_i, \lambda)$ distributed.*

4.1.2 Chi-square distribution

Definition 4.1.3. Let $(Z_i)_{1\le i\le n}$ be a sequence of independent, standard normal distributed random variables. The distribution of $V = Z_1^2 + \ldots + Z_n^2$ is called *chi-square distribution with n degrees of freedom* and is denoted by χ^2_n.

Note that since Z_i^2 is $\mathcal{G}(\frac{1}{2}, 2)$ distributed, χ^2_n is $\mathcal{G}(\frac{n}{2}, 2)$ distributed. Moreover,

$$\mathbb{E}[\chi^2_n] = n, \quad D\chi^2_n = 2n.$$

A notable consequence of the definition of the chi-square distribution is that if U and V are independent and $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then $U + V \sim \chi^2_{m+n}$.

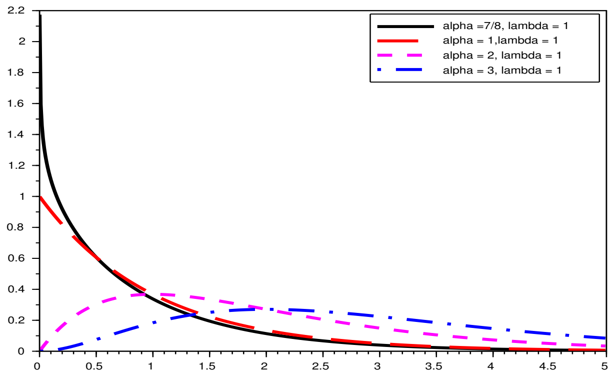


Figure 4.1: Density of gamma distribution

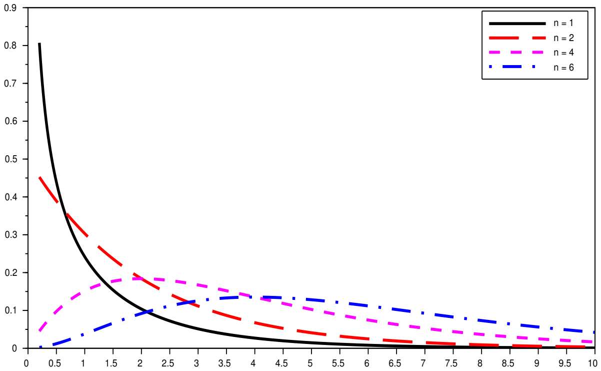


Figure 4.2: Density of χ^2 distribution

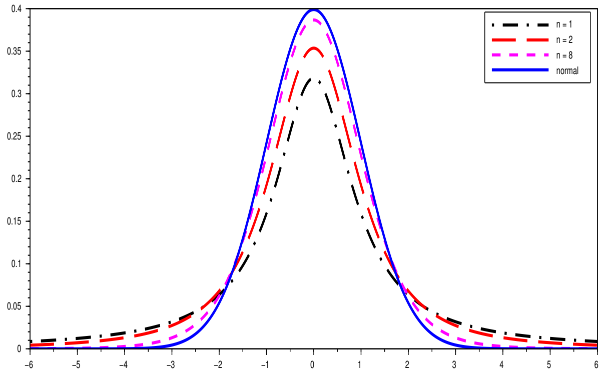


Figure 4.3: Density of student distribution

4.1.3 Student's t distribution

Definition 4.1.4. If $Z \sim N(0;1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $\dfrac{Z}{\sqrt{U/n}}$ is called student's t distribution with n degrees of freedom.

Student's t distribution is also call t distribution.

A direct computation with the density gives the following result.

Proposition 4.1.5. *The density function of the student's t distribution with n degrees of freedom is*

$$f_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}.$$

In addition,

$$f_n(t) \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

4.1.4 F distribution

Definition 4.1.6. Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F *distribution with m and n degrees of freedom* and is denoted by $F_{m,n}$.

The density of W is given by

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{m/2}x^{m/2-1}\left(1+\frac{m}{n}x\right)^{-(m+n)/2}, \quad x \geq 0.$$

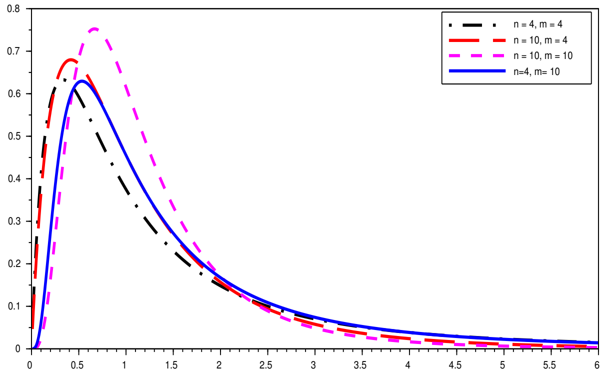


Figure 4.4: Density of F distribution

4.2 Sample mean and sample variance

Let (X_n) be independent $\mathcal{N}(\mu, \sigma^2)$ random variables, and

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(X_i - \overline{X}_n)^2$$

Proposition 4.2.1. *The random variable* $\overline{X}_n$ *and the vector of random variables* $(X_1 - \overline{X}_n, X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n)$ *are independent.*

Proof. We write

$$s\overline{X}_n + \sum_{i=1}^{n} t_i(X_i - \overline{X}_n) = \sum_{i=1}^{n} a_i X_i$$

where $a_i = \frac{s}{n} + (t_i - \overline{t})$. Note that

$$\sum_{i=1}^{n} a_i = s \quad \text{and} \quad \sum_{i=1}^{n} a_i^2 = \frac{s^2}{n} + \sum_{i=1}^{n}(t_i - \overline{t})^2.$$

Therefore, the characteristic function of $(\overline{X}_n, X_1 - \overline{X}_n, X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n)$ is

$$\begin{aligned}
&= \mathbb{E}[\exp(is\overline{X}_n + i\sum_{j=1}^{n} t_j(X_j - \overline{X}_n)] = \prod_{j=1}^{n} \exp\Big(i\mu a_j - \frac{\sigma^2}{2}a_j^2\Big)\\
&= \exp\Big(i\mu s - \frac{\sigma^2}{2n}s^2\Big)\exp\Big(-\frac{\sigma^2}{2}\sum_{i=1}^{n}(t_i - \overline{t})^2\Big).
\end{aligned}$$

The first factor is the cf of $\overline{X}_n$ while the second factor is the cf of $(X_1 - \overline{X}_n, X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n)$ (this is obtained by let $s = 0$ in the formula). This implies the desired result. □

Corollary 4.2.2. $\overline{X}_n$ *and* s_n^2 *are independently distributed.*

Theorem 4.2.3. *The distribution of $(n-1)s_n^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom.*

Proof. We first note that

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 = \sum_{i=1}^n \Big(\frac{X_i-\mu}{\sigma}\Big)^2 \sim \chi_n^2.$$

Also,

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\overline{X}_n)^2 + \Big(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\Big)^2 =: U+V.$$

Since U and V are independent, $\varphi_W(t) = \varphi_U(t)\varphi_V(t)$. Since W and V both follow chi-square distribution, so

$$\varphi_U(t) = \frac{\varphi_W(t)}{\varphi_V(t)} = \frac{(1-i2t)^{-n/2}}{(1-i2t)^{-1/2}} = (1-i2t)^{-(n-1)/2}.$$

The last expression is the c.f. of a random variable with a χ_{n-1}^2 distribution. □

We end up with the following result.

Corollary 4.2.4.

$$\frac{\overline{X}_n-\mu}{s_n/\sqrt{n}} \sim t_{n-1}.$$

4.3 Exercises

4.1. Show that

1. if $X \sim F_{n,m}$ then $X^{-1} \sim F_{m,n}$.
2. if $X \sim t_n$ then $X^2 \sim F_{1,n}$.
3. the Cauchy distribution and the t distribution with 1 degree of freedom are the same.
4. Iif X and Y are independent exponential random variable with $\lambda = 1$, then X/Y follows an F distribution.

4.2. Show how to use the chi-square distribution to calculate $\mathbb{P}(a < s_n^2/\sigma^2 < b)$.

4.3. Let $X_1,\dots,X_n$ be a sequence of independent and $\mathcal{N}(\mu_X,\sigma^2)$ distributed random variables. and $Y_1,\dots,Y_m$ be a sequence of independent and $\mathcal{N}(\mu_Y,\sigma^2)$ distributed random variables. Show how to use F distribution to find $\mathbb{P}(s(X))_n^2/s(Y)_n^2 > c)$, for some positive constant c.

4.4. Let $W \sim F_{n,m}$ and denote $Y = \frac{m}{m+nW}$. Show that Y has a beta distribution.

4.5. Let X_1, X_2 and X_3 be three independent chi-square variables with r_1, r_2 and r_3 degrees of freedom, respectively.

1. Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1+X_2$ are independent.

2. Deduce that
$$\frac{X_1/r_1}{X_2/r_2} \quad \text{and} \quad \frac{X_2/r_3}{(X_1+X_2)/(r_1+r_2)}$$
are independent F-variables.

Chapter 5

Parameter estimation

5.1 Samples and characteristic of sample

Definition 5.1.1 (Random samples)**.** A sequence of random variable $X_1, \ldots, X_n$ is called a *random sample* observing from a random variable X if

- $(X_i)_{1 \leq i \leq n}$ are independent;
- X_i has the same distribution as X for all $i = 1, \ldots, n$.

We call n the *sample size.*

Example 5.1.2. An urn contains m balls, labeled from $1, 2, \ldots, m$ and are identical except for the number. The experiment is to choose a ball at random and record the number. Let X denote the number. Then the distribution of X is given by

$$\mathbb{P}[X = k] = \frac{1}{m}, \ for \ x = 1, \ldots, m.$$

In case m is unknown, to obtain information on m we take a sample of n balls, which we will denote as $\mathbf{X} = (X_1, \ldots, X_n)$ where X_i is the number on the ith ball.

The sample can be drawn in several ways.

1. *Sampling with replacement:* We randomly select a ball, record its number and put it back to the urn. All the ball are then remixed, and the next ball is chosen. We can see that $X_1, \ldots, X_n$ are mutually independent random variables and each has the same distribution as X. Hence $(X_1, \ldots, X_n)$ is a random sample.
2. *Sampling without replacement:* Here n balls are selected at random. After a ball is selected, we do not return it to the urn. The $X_1, \ldots, X_n$ are not independent, but each X_i has the same distribution as X.

If m is much greater than n, the sampling schemes are practically the same.

Definition 5.1.3. The *empirical distribution function* is defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i < x}, \quad x \in \mathbb{R}.$$

1. $F_n(x)$ is non-dercreasing with respect to x;

2. $F_n(x)$ is left continuous and has right limit at any points;

3. $\lim_{x\to-\infty} F_n(x) = 0,\ \lim_{x\to+\infty} F_n(x) = 1.$

4. $F_n(x) \xrightarrow{a.s.} F(x)$ for any $x \in \mathbb{R}$. Indeed, applying law of large numbers for the iid sequence $\mathbb{I}_{X_i<x}$, we have
$$F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{X_i<x} \xrightarrow{a.s.} \mathbb{E}[\mathbb{I}_{X_1<x}] = F(x).$$

Definition 5.1.4. Let $(X_1, \ldots, X_n)$ be a random sample.

1. *Sample mean*
$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

2. *Population variance*
$$S_n^2(X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2,$$
and *sample variance*
$$s_n^2(X) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

3. k*th sample moment*
$$m_k = \frac{1}{n}\sum_{i=1}^n X_i^k,$$
and *centralized* k*th sample moment*
$$v_k = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^k.$$

4. *Sample covariance* of 2-dimensional sample $(X_1, Y_1), \ldots, (X_n, Y_n)$
$$r = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_n(X)S_n(Y)}.$$

5. The *sample mode* is the most frequently occurring data value.

6. The *sample median* is a measure of central tendency that divides the data into two equal parts, half below the median and half above. If the number of observations is even, the median is halfway between the two central values. If the number of observations is odd, the median is the central value.

7. When an ordered set of data is divided into four equal parts, the division points are called *quartiles.* The first or lower quartile, q_1, is a value that has approximately 25% of the observations below it and approximately 75% of the observations above. The second quartile, q_2, has approximately 50% of the observations below its value. The second quartile is exactly equal to the median. The third or upper quartile, q_3, has approximately 75% of the observations below its value.

8. The *interquartile range* is defined as $IQR = q_3 - q_1$. The IQR is also used as a measure of variability.

5.2 Data display

Well-constructed data display is essential to good statistical thinking, because it helps us exploring important features of the data and providing insight about the type of model that should be used in solving the problem. In this section we will briefly introduce some methods to display data.

5.2.1 Stem-and-leaf diagrams

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

1. Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits and a leaf, consisting of the remaining digit.

2. List the stem values in a vertical column.

3. Record the leaf for each observation beside its stem.

4. Write the units for stems and leaves on the display.

It is usually best to choose between 5 and 20 stems.

Example 5.2.1. The weights of 80 students are given in the following table.

59.0	59.5	52.7	47.9	55.7	48.3	52.1	53.1	55.2	45.3
46.5	54.8	48.4	53.1	56.9	47.4	50.2	52.1	49.6	46.4
52.9	41.1	51.0	50.0	56.8	45.9	59.5	52.8	46.7	55.7
48.6	51.6	53.2	54.1	45.8	50.4	54.1	52.0	56.2	62.7
62.0	46.8	54.6	54.7	50.2	45.9	49.1	42.6	49.8	52.1
56.5	53.5	46.5	51.9	46.5	53.5	45.5	50.2	55.1	49.6
47.6	44.8	55.0	56.2	49.4	57.0	52.4	48.4	55.0	47.1
52.4	56.8	53.2	50.5	56.6	49.5	53.1	51.2	55.5	53.7

Construct a stem-and-leaf diagram for their weight.

Stem	Leaf	Frequency
41	1	1
42	6	1
44	8	1
45	3 5 8 9 9	5
46	4 5 5 5 7 8	6
47	1 4 6 9	4
48	3 4 4 6	4
49	1 4 5 6 6 8	6
50	0 2 2 2 4 5	6
51	0 2 6 9	4
52	0 1 1 1 4 4 7 8 9	9
53	1 1 1 2 2 5 5 7	8
54	1 1 6 7 8	5
55	0 0 1 2 5 7 7	7
56	2 2 5 6 8 8 9	7
57	1	1
59	0 5 5	3
62	0 7	2

5.2.2 Frequency distribution and histogram

A frequency distribution is a more compact summary of data than a stem-and-leaf diagram. To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called class intervals, cells, or bins. If possible, the bins should be of equal width in order to enhance the visual information in the frequency distribution. Some judgment must be used in selecting the number of bins so that a reasonable display can be developed. The number of bins depends on the number of observations and the amount of scatter or dispersion in the data. A frequency distribution that uses either too few or too many bins will not be informative. We usually find that between 5 and 20 bins is satisfactory in most cases and that the number of bins should increase with n. Choosing the number of bins approximately equal to the square root of the number of observations often works well in practice.

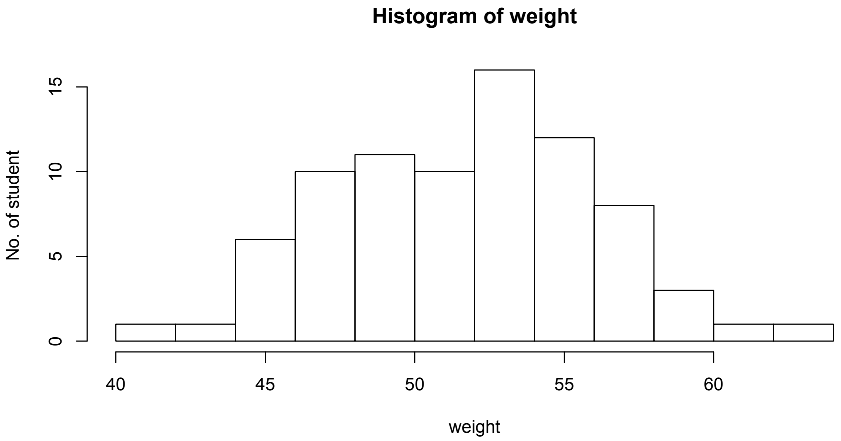
The histogram is a visual display of the frequency distribution. The stages for constructing a histogram follow.

1. Label the bin (class interval) boundaries on a horizontal scale.

2. Mark and label the vertical scale with the frequencies or the relative frequencies.

3. Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

Example 5.2.2. Histogram of the students' weight given in Example 5.2.1.

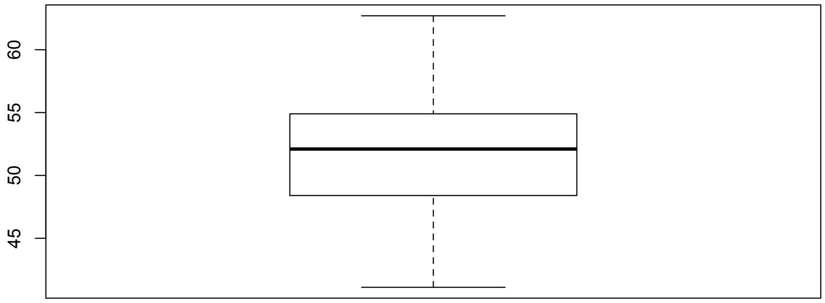


5.2.3 Box plots

The *box plot* is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.

A box plot displays the three quartiles, the minimum, and the maximum of the data on a rectangular box, aligned either horizontally or vertically. The box encloses the interquartile range with the left (or lower) edge at the first quartile, q_1, and the right (or upper) edge at the third quartile, q_3. A line is drawn through the box at the second quartile (which is the 50th percentile or the median). A line, or *whisker,* extends from each end of the box. The lower whisker is a line from the first quartile to the smallest data point within 1.5 interquartile ranges from the first quartile. The upper whisker is a line from the third quartile to the largest data point within 1.5 interquartile ranges from the third quartile. Data farther from the box than the whiskers are plotted as individual points. A point beyond a whisker, but less than 3 interquartile ranges from the box edge, is called an *outlier.* A point more than 3 interquartile ranges from the box edge is called an *extreme outlier.*

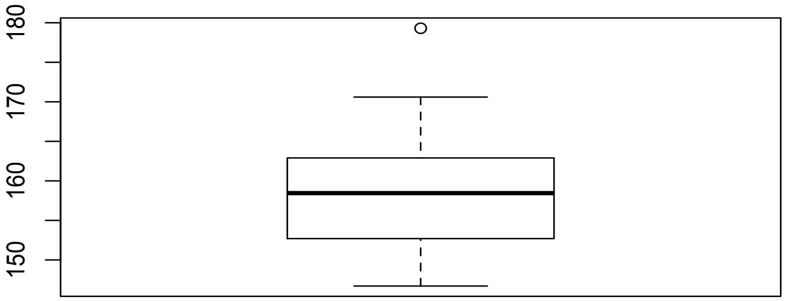
Example 5.2.3. Consider the sample in Example 5.2.1. The quantiles of the sample are $q_1 = 48.40, q_2 = 52.10, q_3 = 54.85$. Bellow is the box plot of the students' weight.



Example 5.2.4. Construct a box plot of the following data.

158.7	167.6	164.0	153.1	179.3	153.0	170.6	152.4	161.5	146.7
147.2	158.2	157.7	161.8	168.4	151.2	158.7	161.0	147.9	155.5

The quantiles of this sample are $q_1 = 152.85, q_2 = 158.45, q_3 = 162.35$



5.2.4 Probability plots

How do we know if a particular probability distribution is a reasonable model for data? Some of the visual displays we have used earlier, such as the histogram, can provide insight about the form of the underlying distribution. However, histograms are usually not really reliable indicators of the distribution form unless the sample size is very large. Probability plotting is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. The general procedure is very simple and can be performed quickly. It is also more reliable than the histogram for small to moderate size samples.

To construct a probability plot, the observations in the sample are first ranked from smallest to largest. That is, the sample $x_1, x_2, \ldots, x_n$ is arranged as $x_{(1)} \leq x_{(2)} < \ldots \leq x_{(n)}$. The ordered observations $x_{(j)}$ are then plotted against their observed cumulative frequency $(j - 0.5)/n$. If the hypothesized distribution adequately describes the data, the plotted points will fall approximately along a straight line which is approximately between the 25th and 75th percentile points; if the plotted points deviate significantly from a straight line, the hypothesized model is not appropriate. Usually, the determination of whether or not the data plot as a straight line is subjective.

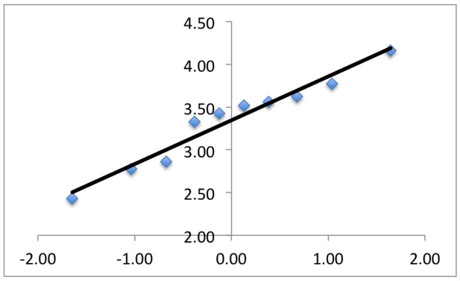
In particular, a normal probability plot can be constructed by plotting the standardized normal scores $z_j = \Phi^{-1}\left(\frac{j-0.5}{n}\right)$ against $x_{(j)}$.

Example 5.2.5. Consider the following sample:

$$2.86, 3.33, 3.43, 3.77, 4.16, 3.52, 3.56, 3.63, 2.43, 2.78.$$

We construct a normal probability plot for this sample as follows.

j	$x_{(j)}$	$(j-0.5)/10$	$\Phi^{-1}\left(\frac{j-0.5}{n}\right)$
1	2.43	0.05	-1.64
2	2.78	0.15	-1.04
3	2.86	0.25	-0.67
4	3.33	0.35	-0.39
5	3.43	0.45	-0.13
6	3.52	0.55	0.13
7	3.56	0.65	0.39
8	3.63	0.75	0.67
9	3.77	0.85	1.04
10	4.16	0.95	1.64



Since all the points are very close to the straight line, one may conclude that a normal distribution adequately describes the data.

Remark 4. This is very surjective method. Please use it at your own risk! Later we will introduce the Shapiro and Wilcoxon tests for the normal distribution hypothesis.

5.3 Point estimations

5.3.1 Statistics

Example 5.3.1. We continue Example 5.1.2. Recall that we do not know the number of balls m and have to use the sample $(X_1, \dots, X_n)$ to obtain information about m.

Since $\mathbb{E}(X) = \frac{m+1}{2}$, using laws of large numbers, we have

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{a.s.} \frac{m+1}{2}.$$

Therefore, we get the first estimator for m given by

$$\hat{m}_n := 2\frac{X_1 + \dots + X_n}{n} - 1 \xrightarrow{a.s.} m.$$

Another estimation for m is defined by

$$\tilde{m}_n := \max\{X_1, \dots, X_n\}.$$

Since

$$\mathbb{P}[\tilde{m}_n \neq m] = \mathbb{P}[X_1 < m, \dots, X_n < m] = \prod_{i=1}^{n} \mathbb{P}[X_i < m] = \Big(\frac{m-1}{m}\Big)^n \to 0$$

as $n \to \infty$, we have $\tilde{m}_n \xrightarrow{a.s.} m$.

The estimator $\hat{m}_n$ and $\tilde{m}_n$ are called statistics which depend only on the observations $X_1, \dots, X_n$ not m.

Definition 5.3.2. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample observed from X and $(T, \mathcal{B}_T)$ a measurable space. Then any function $\varphi(\mathbf{X}) = \varphi(X_1, \dots, X_n)$, where $\varphi : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \to (T, \mathcal{B}_T)$ is a measurable function, of the sample is called a *statistic.*

In the following we only consider the case that $(T, \mathcal{B}_T)$ is a subset of $(\mathbb{R}^d, \mathbb{B}(\mathbb{R}^d))$.

Definition 5.3.3. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a sample observed from a distribution with density $f(x;\theta)$, $\theta \in \Theta$. Let $Y = \varphi(\mathbf{X})$ be a statistic with density $f_Y(y;\theta)$. Then Y is called a ***sufficient statistic*** for θ if

$$\frac{f(\mathbf{x};\theta)}{f_Y(\varphi(\mathbf{x});\theta)} = H(\mathbf{x}),$$

where $\mathbf{x} = (x_1, \dots, x_n)$, $f(\mathbf{x};\theta)$ is density of $\mathbf{X}$ at $\mathbf{x}$, and $H(\mathbf{x})$ does not depend on $\theta \in \Theta$.

Example 5.3.4. Let $(X_1, \dots, X_n)$ be a sample observed from a Poisson distribution with parameter $\lambda > 0$. Then $Y_n = \varphi(X) = X_1 + \dots + X_n$ has Poisson distribution with parameter $n\lambda$. Hence

$$\frac{f(\mathbf{X};\theta)}{f_Y(\varphi(\mathbf{X});\theta)} = \frac{\prod_{i=1}^n f(X_i;\theta)}{f_{Y_n}(Y_n;n\theta)} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!}\frac{Y_n!}{e^{-n\lambda}(n\lambda)^Y} = \frac{Y_n!}{n^{Y_n}\prod_{i=1}^n X_i!}.$$

Therefore Y_n is a sufficient statistic for λ.

In order to directly verify the sufficiency of statistic $\varphi(\mathbf{X})$ we need to know the density of $\varphi(\mathbf{X})$ which is not always the case in practice. We next introduce the following criterion of Neyman to overcome this difficulty.

Theorem 5.3.5. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution that has density $f(x;\theta)$, $\theta \in \Theta$. The statistic $Y_1 = \varphi(\mathbf{X})$ is a sufficient statistic for θ iff we can find two nonnegative functions k_1 and k_2 such that*

$$f(\mathbf{x};\theta) = k_1(\varphi(\mathbf{x});\theta)k_2(\mathbf{x}) \tag{5.1}$$

where k_2 does not depend upon θ.

Example 5.3.6. Let $(X_1, \ldots, X_n)$ be a sample from normal distribution $N(\theta, 1)$ with $\theta \in \Theta = \mathbb{R}$. Denote $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. The joint density of $X_1, \ldots, X_n$ at $(x_1, \ldots, x_n)$ is given by

$$\frac{1}{(2\pi)^{n/2}} \exp\Big[-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2}\Big] = \exp\Big[-\frac{n(\bar{x}-\theta)^2}{2}\Big] \frac{\exp\Big[-\sum_{i=1}^n \frac{(x_i-\bar{x})^2}{2}\Big]}{(2\pi)^{n/2}}.$$

We see that the first factor on the right hand side depends upon $x_1, \ldots, x_n$ only through $\bar{x}$ and the second factor does not depend upon θ, the factorization theorem implies that the mean $\bar{X}$ of the sample is a sufficient statistic for θ, the mean of the normal distribution.

5.3.2 Point estimators

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample from distribution $F(x;\theta)$ which depends on a unknown parameter $\theta \in \Theta$. Even thought function φ does not depend on the unknown parameter θ, the statistic $\varphi(\mathbf{X})$ may convey information about θ. In such cases, we may call the statistic a *point estimator* of θ.

Definition 5.3.7. A statistic $\varphi(X_1, \ldots, X_n)$ is called

1. an *unbiased estimator* of θ if $\mathbb{E}_\theta[\varphi(X_1, \ldots, X_n)] = \theta$;
2. an *asymptotic unbiased estimator* of θ if $\lim_{n\to\infty} \mathbb{E}_\theta[\varphi(X_1, \ldots, X_n)] = \theta$;
3. a *best unbiased estimator* of θ if
   (a) $\mathbb{E}_\theta[\varphi(X_1, \ldots, X_n)] = \theta$;
   (b) $D_\theta\varphi(X_1, \ldots, X_n) \le D_\theta\bar{\varphi}(X_1, \ldots, X_n)$ for any unbiased estimator $\bar{\varphi}(X_1, \ldots, X_n)$ of θ.
4. a *consistent estimator* of θ if

$$\varphi(X_1, \ldots, X_n) \xrightarrow{\mathbb{P}_\theta} \theta \quad \text{khi } n \to \infty.$$

Here we denote $\mathbb{E}_\theta, D_\theta, \mathbb{P}_\theta$ the expectation, variance and probability under the condition that the distribution of X_i is $F(x;\theta)$.

Example 5.3.8. Let $(X_1, \ldots, X_n)$ be a sample from normal distribution $N(a, \sigma^2)$. Using the linearity of expectation and laws of large numbers, we have

- $\bar{X}_n$ is an unbiased estimator of a;
- $s_n^2(X)$ is an unbiased and consistent estimator of σ^2.
- $S_n^2(X)$ is an asymptotic unbiased and consistent estimator of σ^2.

Example 5.3.9. In Example 5.3.1, both $\hat{m}_n$ and $\tilde{m}_n$ are consistent estimators of m. Moreover, $\hat{m}_n$ is unbiased and $\tilde{m}_n$ is asymptotic unbiased.

5.3.3 Confidence intervals

Let X be a random variable whose density is $f(x,\theta),\ \theta\in\Theta$, where θ is unknown. In the last section, we discussed estimating θ by a statistic $\varphi(X_1,\dots,X_n)$ where $X_1,\dots,X_n$ is a sample from the distribution of X. When the sample is drawn, it is unlikely that the value of φ is the true value of the parameter. In fact, if φ has a continuous distribution then $\mathbb{P}_\theta[\varphi=\theta]=0$. What is needed is an estimate of the error of the estimation.

Example 5.3.10. Let $(X_1,\dots,X_n)$ be a sample from normal distribution $\mathcal{N}(a;\sigma^2)$ where σ^2 is known. We know that $\bar{X}_n$ is an unbiased, consistent estimator of a. But how close is $\bar{X}_n$ to a? Since $\bar{X}_n\sim\mathcal{N}(a;\sigma^2/n)$, we have $(\bar{X}_n-a)/(\sigma/\sqrt{n})$ has a standard normal $\mathcal{N}(0;1)$ distribution. Therefore,

$$0.954=\mathbb{P}\Big[-2<\frac{\bar{X}_n-a}{\sigma/\sqrt{n}}<2\Big]=\mathbb{P}\Big[\bar{X}_n-2\frac{\sigma}{\sqrt{n}}<a<\bar{X}_n+2\frac{\sigma}{\sqrt{n}}\Big]. \tag{5.2}$$

Expression (5.2) says that before the sample is drawn the probability that a belongs to the random interval $\Big(\bar{X}_n-2\frac{\sigma}{\sqrt{n}}<a<\bar{X}_n+2\frac{\sigma}{\sqrt{n}}\Big)$ is 0.954. After the sample is drawn the realized interval $\Big(\bar{x}_n-2\frac{\sigma}{\sqrt{n}}<a<\bar{x}_n+2\frac{\sigma}{\sqrt{n}}\Big)$ has either trapped a or it has not. But because of the high probability of success before the sample is drawn, we call the interval $\Big(\bar{X}_n-2\frac{\sigma}{\sqrt{n}}<a<\bar{X}_n+2\frac{\sigma}{\sqrt{n}}\Big)$ a 95.4% *confidence interval* for a. We can say, with some confidence, that $\bar{x}$ is within $2\frac{\sigma}{\sqrt{n}}$ from a. The number $0.954=95.4\%$ is called a *confidence coefficient.* Instead of using 2, we could use, say, 1.645, 1.96 or 2.576 to obtain $90\%, 95\%$ or 99% confidence intervals for a. Note that the lengths of these confidence intervals increase as the confidence increases; i.e., the increase in confidence implies a loss in precision. On the other hand, for any confidence coefficient, an increase in sample size leads to shorter confidence intervals.

In the following, thanks to the central limit theorems, we will present a general method to find the confident interval for parameters of a large class of distribution. To avoid confusion, let θ_0 denote the true, unknown value of the parameter θ. Suppose φ is an estimator of θ_0 such that

$$\sqrt{n}(\varphi-\theta_0)\xrightarrow{w}\mathcal{N}(0,\sigma_\varphi^2). \tag{5.3}$$

The parameter σ_φ^2 is the asymptotic variance of $\sqrt{n}\varphi$ and, in practice, it is usually unknown. For the present, though, we suppose that σ_φ^2 is known.

Let $Z = \sqrt{n}(\varphi - \theta_0)/\sigma_\varphi$ be the standardized random variable. Then Z is asymptotically $\mathcal{N}(0,1)$. Hence, $\mathbb{P}[-1.96 < Z < 1.96] = 0.95$. This implies

$$0.95 = \mathbb{P}\Big[\varphi - 1.96\frac{\sigma_\varphi}{\sqrt{n}} < \theta_0 < \varphi + 1.96\frac{\sigma_\varphi}{\sqrt{n}}\Big] \tag{5.4}$$

Because the interval $\left(\varphi - 1.96\frac{\sigma_\varphi}{\sqrt{n}} < \theta_0 < \varphi + 1.96\frac{\sigma_\varphi}{\sqrt{n}}\right)$ is a function of the random variable φ, we will call it a *random interval.* The probability that the random interval contains θ is approximately 0.95.

Since in practice, we often do not know σ_φ. Suppose that there exists a consistent estimator of σ_φ, say S_φ. It then follows from Slutsky's theorem that

$$\frac{\sqrt{n}(\varphi - \theta_0)}{S_\varphi} \xrightarrow{w} N(0,1).$$

Hence the interval $\left(\varphi - 1.96S_\varphi/\sqrt{n}, \varphi - 1.96S_\varphi/\sqrt{n}\right)$ would be a random interval with approximate probability 0.95% of covering θ_0.

In general, we have the following definition.

Definition 5.3.11. Let $(X_1, \ldots, X_n)$ be a sample from a distribution $F(x,\theta),\ \theta \in \Theta$. A random interval (φ_1, φ_2), where φ_1 and φ_2 are some estimator of θ, is called a $(1-\alpha)$-*confidence interval* for θ if

$$\mathbb{P}(\varphi_1 < \theta < \varphi_2) = 1 - \alpha,$$

for some $\alpha \in [0,1]$.

Confidence interval for the Mean a

Let $X_1, \ldots, X_n$ be a random sample from the distribution of a random variable X which has unknown mean a and unknown variance σ^2. Let $\bar{X}$ and s^2 be sample mean and sample variance, respectively. By the Central limit theorem, the distribution of $\sqrt{n}(\bar{X} - a)/s$ approximates $\mathcal{N}(0;1)$. Hence, an approximated $(1-\alpha)$ confidence interval for a is

$$\Big(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\Big), \tag{5.5}$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

1. Because $\alpha < \alpha'$ implies that $x_{\alpha/2} > x_{\alpha'/2}$, selection of higher values for confidence coefficients leads to larger error terms and hence, longer confidence intervals, assuming all else remains the same.

2. Choosing a larger sample size decreases the error part and hence, leads to shorter confidence intervals, assuming all else stays the same.

3. Usually the parameter σ is some type of scale parameter of the underlying distribution. In these situations, assuming all else remains the same, an increase in scale (noise level), generally results in larger error terms and, hence, longer confidence intervals.

Confidence interval for p

Let $X_1, \dots, X_n$ be a random sample from the Bernoulli distribution with probability of success p. Let $\hat{p} = \bar{X}$ be the sample proportion of successes. It follows from the Central limit theorem that $\hat{p}$ has an approximate $\mathcal{N}(p; \frac{p(1-p)}{n})$ distribution. Since $\hat{p}$ and $\hat{p}(1-\hat{p})$ are consistent estimators for p and $p(1-p)$, respectively, an approximate $(1-\alpha)$ confidence interval for p is given by

$$\Big(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}\Big).$$

Confidence interval for mean of normal distribution

In general, the confidence intervals developed so far in this section are approximate. They are based on the Central Limit Theorem and also, often require a consistent estimate of the variance. In our next example, we develop an exact confidence interval for the mean when sampling from a normal distribution

Theorem 5.3.12. *Let $X_1, \dots, X_n$ be a random sample from a $\mathcal{N}(a, \sigma^2)$ distribution. Recall that $\bar{X}$ and s^2 are sample mean and sample variance, respectively. The random variable $T = (\bar{X} - a)/(s/\sqrt{n})$ has a t-distribution with $n-1$ degrees of freedom.* [a]

[a] In statistics, the t-distribution was first derived as a posterior distribution in 1876 by Helmert and Lroth. The t-distribution also appeared in a more general form as Pearson Type IV distribution in Karl Pearson's 1895 paper.

In the English-language literature the distribution takes its name from William Sealy Gosset's 1908 paper in Biometrika under the pseudonym "Student".

For each $\alpha \in (0,1)$, denote $t_{\alpha/2,n-1}$ satisfying

$$\frac{\alpha}{2} = \mathbb{P}\Big(T > t_{\alpha/2,n-1}\Big).$$

Thanks to the symmetry of t-distribution, we have

$$\begin{aligned} 1-\alpha &= \mathbb{P}\Big(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\Big) = \mathbb{P}\Big(-t_{\alpha/2,n-1} < \frac{\bar{X}-a}{S/\sqrt{n}} < t_{\alpha/2,n-1}\Big) \\ &= \mathbb{P}\Big(\bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < a < \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\Big). \end{aligned}$$

Thus, a $(1-\alpha)$ confidence interval for a is given by

$$\Big(\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\Big). \tag{5.6}$$

Note that the only difference between this confidence interval and the large sample confidence interval (5.5) is that $t_{\alpha/2,n-1}$ replaces $z_{\alpha/2}$. This one is exact while (5.5) is approximate. Of course, we have to assume we are sampling a normal population to get the exactness. In practice, we often do not know if the population is normal. Which confidence interval should we use? Generally, for the same α, the intervals based on $t_{\alpha/2,n-1}$ are larger than those based on $z_{\alpha/2}$. Hence, the interval (5.6) is generally more conservative than the interval (5.5). So in practice, statisticians generally prefer the interval (5.6).

Confidence interval on the variance and standard deviation of a normal population

Sometimes confidence intervals on the population variance or standard deviation are needed. When the population is modelled by a normal distribution, the tests and intervals described in this section are applicable. The following result provides the basis of constructing these confidence intervals.

Theorem 5.3.13. *Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a normal distribution with mean μ and variance σ^2, and let s^2 be the sample variance, i.e,*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_i)^2.$$

Then the random variable

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

has a χ^2-distribution with $n-1$ degrees of freedom.

We recall that the pdf of a χ^2 random variable with k degree of freedom is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \quad x > 0.$$

Theorem 5.3.13 leads to the following construction of the CI for σ^2.

Theorem 5.3.14. *If s^2 is the sample variance from a random sample of n observation from a normal distribution with unknown variance σ^2, then a $100(1-\alpha)\%$ CI on σ^2 is*

$$\frac{(n-1)s^2}{c^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{c^2_{1-\alpha/2,n-1}},$$

where $c^2_{a,n-1}$ satisfies $\mathbb{P}(\chi^2_{n-1} > c^2_{a,n-1}) = a$ and the random variable χ^2_{n-1} has a chi-square distribution with $n-1$ degrees of freedom.

Confidence interval for differences in means

A practical problem of interest is the comparison of two distributions; that is, comparing the distributions of two random variables, say X and Y. In this section, we will compare the means of X and Y. Denote the means of X and Y by a_X and a_Y, respectively. In particular, we shall obtain confidence intervals for the difference $\Delta = a_X - a_Y$. Assume that the variances of X and Y are finite and denote them as $\sigma_X^2 = Var(X)$ and let $\sigma_Y^2 = Var(Y)$. Let $X_1 ...., X_n$ be a random sample from the distribution of X and let $Y_1, ..., Y_m$ be a random sample from the distribution of Y. Assume that the sample were gathered independently of one another. Let $\bar{X}$ and $\bar{Y}$ the sample means of X and Y, respectively. Let $\hat{\Delta} = \bar{X} - \bar{Y}$. Next we obtain a large sample confidence interval for Δ based on the asymptotic distribution of $\hat{\Delta}$.

Proposition 5.3.15. *Let $N = n + m$ denote the total sample size. We suppose that*

$$\frac{n}{N} \to \lambda_X, \text{ and } \frac{m}{N} \to \lambda_Y \text{ where } \lambda_X + \lambda_Y = 1.$$

Then a $(1-\alpha)$ confidence interval for Δ is

1. *(if σ_X^2 and σ_Y^2 are known)*

$$\Big((\bar{X} - \bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{X} - \bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\Big); \tag{5.7}$$

2. *(if σ_X^2 and σ_Y^2 are unknown)*

$$\Big((\bar{X} - \bar{Y}) - z_{\alpha/2}\sqrt{\frac{s^2(X)}{n} + \frac{s^2(Y)}{m}}, (\bar{X} - \bar{Y}) + z_{\alpha/2}\sqrt{\frac{s^2(X)}{n} + \frac{s^2(Y)}{m}}\Big), \tag{5.8}$$

where $s^2(X)$ and $s^2(Y)$ are sample variances of (X_n) and (Y_m), respectively.

Proof. It follows from the Central limit theorem that $\sqrt{n}(\bar{X} - a_X) \xrightarrow{w} \mathcal{N}(0; \sigma_X^2)$. Thus,

$$\sqrt{N}(\bar{X} - a_X) \xrightarrow{w} \mathcal{N}(0; \frac{\sigma_X^2}{\lambda_X}).$$

Likewise,

$$\sqrt{N}(\bar{Y} - a_Y) \xrightarrow{w} \mathcal{N}(0; \frac{\sigma_Y^2}{\lambda_Y}).$$

Since the samples are independent of one another, we have

$$\sqrt{N}\Big((\bar{X} - \bar{Y}) - (a_X - a_Y)\Big) \xrightarrow{w} \mathcal{N}(0; \frac{\sigma_X^2}{\lambda_X} + \frac{\sigma_Y^2}{\lambda_Y}).$$

This implies (5.7). Since $S^{*2}(X)$ and $S^{*2}(Y)$ are consistent estimators of σ_X^2 and σ_Y^2, applying Slutsky's theorem, we obtain (5.8). □

Confidence interval for difference in proportions

Let X and Y be two independent random variables with Bernoulli distributions $B(1, p_1)$ and $B(1, p_2)$, respectively. Let $X_1, \dots, X_n$ be a random sample from the distribution of X and let $Y_1, \dots, Y_m$ be a random sample from the distribution of Y.

Proposition 5.3.16. *A $(1-\alpha)$ confidence interval for $p_1 - p_2$ is*

$$\Big(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}},\ \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}\Big),$$

where $\hat{p}_1 = \bar{X}$ and $\hat{p}_2 = \bar{Y}$.

5.4 Method of finding estimation

5.4.1 Maximum likelihood estimation

The method of maximum likelihood is one of the most popular technique for deriving estimators. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution with pdf/pdm $f(x;\theta)$. The likelihood function is defined by

$$L(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Definition 5.4.1. For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\mathbf{x};\theta)$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

Example 5.4.2. Let $(X_1, \ldots, X_n)$ be a random sample from the distribution $N(\theta, 1)$, where θ is unknown. We have

$$L(\mathbf{x};\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2},$$

A simple calculus shows that the MLE of θ is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. One can easily verify that $\hat{\theta}$ is an unbiased and consistent estimator of θ.

Example 5.4.3. Let $(X_1, \ldots, X_n)$ be a random sample from the Bernoulli distribution with a unknown parameter p. The likelihood function is

$$L(\mathbf{x};p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

A simple calculus shows that the MLE of p is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$. One can easily verify that $\hat{\theta}$ is an unbiased and consistent estimator of θ.

Let $X_1, \ldots, X_n$ denote a random sample from the distribution with pdf $f(x;\theta),\ \theta \in \Theta$. Let θ_0 denote the true value of θ. The following theorem gives a theoretical reason for maximizing the likelihood function. It says that the maximum of $L(\theta)$ asymptotically separates the true model at θ_0 from models at $\theta \neq \theta_0$.

Theorem 5.4.4. *Suppose that*

(R0) $f(.;\theta) \neq f(.;\theta')$ *for all* $\theta \neq \theta'$*;*

(R1) all $f(.;\theta), \theta \in \Theta$ *have common support for all* θ*.*

Then

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}[L(\mathbf{X};\theta_0) > L(\mathbf{X};\theta)] = 1, \quad \textit{for all } \theta \neq \theta_0.$$

Proof. By taking logs, we have

$$\mathbb{P}_{\theta_0}[L(\mathbf{X};\theta_0) > L(\mathbf{X};\theta)] = \mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^{n} \ln\Big(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\Big) < 0\Big].$$

Since the function $\phi(x) = -\ln x$ is strictly convex, it follows from the Law of Large Numbers and Jensen's inequality that

$$\frac{1}{n}\sum_{i=1}^{n} \ln\Big(\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\Big) \xrightarrow{\mathbb{P}} \mathbb{E}_{\theta_0}\Big[\ln\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\Big] < \ln \mathbb{E}_{\theta_0}\Big[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\Big] = 0.$$

Note that condition $f(.;\theta) \neq f(.;\theta')$ for all $\theta \neq \theta'$ is needed to obtain the last strict inequality while the common support is needed to obtain the last equality. □

Theorem 5.4.4 says that asymptotically the likelihood function is maximized at the true value θ_0. So in considering estimates of θ_0, it seems natural to consider the value of θ which maximizes the likelihood.

We close this section by showing that maximum likelihood estimators, under regularity conditions, are consistent estimators.

Theorem 5.4.5. *Suppose that the pdfs $f(x,\theta)$ satisfying (R0), (R1) and*

(R2) The point θ_0 is an interior point in Θ.

(R3) $f(x;\theta)$ is differentiable with respect to θ in Θ.

Then the likelihood equation,

$$\frac{\partial}{\partial\theta}L(\theta) = 0 \Leftrightarrow \frac{\partial}{\partial\theta}\ln L(\theta) = 0,$$

has a solution $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

5.4.2 Method of Moments

Let $(X_1,\dots,X_n)$ be a random sample from a distribution with density $f(x;\theta)$ where $\theta = (\theta_1,\dots,\theta_k) \in \Theta \subset \mathbb{R}^k$. Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$\mu_j = \mathbb{E}[X^j] = g_j(\theta_1,\dots,\theta_k), \quad j = 1,\dots,k.$$

and

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

The moments estimator $(\hat{\theta}_1,\dots,\hat{\theta}_k)$ is obtained by solving the system of equations

$$m_j = g_j(\theta_1,\dots,\theta_k),\ j = 1,\dots,k.$$

Example 5.4.6 (Binomial distribution)**.** Let $(X_1, \dots, X_n)$ be a random sample from the Binomial distribution $B(k,p)$, that is,

$$\mathbb{P}_{k,p}[X_i = x] = C_k^x p^x (1-p)^{k-x}, \quad x = 0, 1, \dots, k.$$

Here we assume that p and k are unknown parameters. Equating the first two sample moments to those of the population yields

$$\begin{cases} \bar{X}_n = kp \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = kp(1-p) + k^2p^2 \end{cases} \Leftrightarrow \begin{cases} k = \hat{k} = \frac{\bar{X}_n^2}{\bar{X}_n - \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2} \\ p = \hat{p} = \frac{\bar{X}_n}{\hat{k}}. \end{cases}$$

5.5 Lower bound for variance

In this section we establish a remarkable inequality called the Rao-Cramer lower bound which gives a lower bound on the variance of any unbiased estimate. We then show that, under regularity conditions, the variances of the maximum likelihood estimates achieve this lower bound asymptotically.

Theorem 5.5.1 (Rao-Cramer Lower Bound)**.** *Let $X_1, \dots, X_n$ be iid with common pdf $f(x;\theta)$ for $\theta \in \Theta$. Assume that the regularity conditions (R0)-(R2) hold. Moreover, suppose that*

(R4) The pdf $f(x;\theta)$ is twice differentiable as a function of θ.

(R5) The integral $\int f(x;\theta)dx$ can be differentiated twice under integral sign as a function of θ.

Let $Y = u(X_1, \dots, X_n)$ be a statistic with mean $\mathbb{E}[Y] = \mathbb{E}[u(X_1, \dots, X_n)] = k(\theta)$. Then

$$DY \geq \frac{[k'(\theta)]^2}{nI(\theta)},$$

where $I(\theta)$ is called Fisher information *and given by*

$$I(\theta) = -\int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} f(x;\theta)dx = D\Big[\frac{\partial \ln f(X;\theta)}{\partial\theta}\Big].$$

Proof. Since

$$k(\theta) = \int_{\mathbb{R}^n} u(x_1, \dots, x_n) f(x_1;\theta) \dots f(x_n;\theta) dx_1 \dots dx_n,$$

we have

$$k'(\theta) = \int_{\mathbb{R}^n} u(x_1, \dots, x_n) \Big(\sum_{i=1}^n \frac{\partial \ln f(x_i;\theta)}{\partial\theta}\Big) f(x_1;\theta) \dots f(x_n;\theta) dx_1 \dots dx_n.$$

Denote $Z = \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta}$. It is easy to verify that $\mathbb{E}[Z] = 0$ and $DZ = nI(\theta)$. Moreover, $k'(\theta) = \mathbb{E}[YZ]$. Hence, we have

$$k'(\theta) = \mathbb{E}[YZ] = \mathbb{E}[Y]\mathbb{E}[Z] + \rho\sqrt{nI(\theta)DY},$$

where ρ is the correlation coefficient between Y and Z. Since $\mathbb{E}[Z] = 0$ and $\rho^2 \leq 1$, we get

$$\frac{|k'(\theta)|^2}{nI(\theta)DY} \leq 1,$$

which implies the desired result. □

Definition 5.5.2. Let Y be an unbiased estimator of a parameter θ in the case of point estimation. The statistic Y is called an *efficient estimator* of θ if and only if the variance of Y attains the Rao-Cramer lower bound.

Example 5.5.3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from a exponential distribution that has the mean $\lambda > 0$. Show that $\bar{X}$ is an efficient estimator of λ.

Example 5.5.4 (Poisson distribution)**.** Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Poisson distribution that has the mean $\theta > 0$. Show that $\bar{X}$ is an efficient estimator of θ.

In the above examples, we were able to obtain the MLEs in closed form along with their distributions and, hence, moments. This is often not the case. Maximum likelihood estimators, however, have an asymptotic normal distribution. In fact, MLEs are asymptotically efficient.

Theorem 5.5.5. *Assume $X_1, \ldots, X_n$ are iid with pdf $f(x;\theta_0)$ for $\theta_0 \in \Theta$ such that the regularity condition (R0)-(R5) are satisfied. Suppose further that $0 < I(\theta_0) < \infty$, and*

(R6) The pdf $f(x;\theta)$ is three times differentiable as a function of θ. Moreover, for all $\theta \in \Theta$, there exists a constant c and a function $M(x)$ such that

$$\Big|\frac{\partial^2 \ln f(x;\theta)}{\partial \theta^3}\Big| \leq M(x),$$

with $\mathbb{E}_{\theta_0}[M(X)] < \infty$, for all $\theta_0 - c < \theta < \theta_0 + c$ and all x in the support of X.

Then any consistent sequence of solutions of the mle equations satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{w}{\to} N(0, I(\theta_0)^{-1}).$$

Proof. Expanding the function $l'(\theta)$ into a Taylor series of order two about θ_0 and evaluating it at $\hat{\theta}_n$, we get

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0)l''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l'''(\theta_n^*),$$

where θ_n^* is between θ_0 and $\hat{\theta}_n$. But $l'(\hat{\theta}_n) = 0$. Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2}l'(\theta_0)}{-n^{-1}l''(\theta_0) - (2n)^{-1}(\hat{\theta}_n - \theta_0)l'''(\theta_n^*)}.$$

By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}} l'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ln f(X_i;\theta_0)}{\partial \theta} \xrightarrow{w} N(0, I(\theta_0)).$$

Also, by the Law of Large Numbers,

$$-\frac{1}{n} l''(\theta_0) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f(X_i;\theta_0)}{\partial \theta^2} \xrightarrow{\mathbb{P}} I(\theta_0).$$

Note that $|\hat{\theta}_n - \theta_0| < c_0$ implies that $|\theta_n^* - \theta_0| < c_0$, thanks to Condition (R6), we have

$$\Big| -\frac{1}{n} l'''(\theta_n^*) \Big| \leq \frac{1}{n} \sum_{i=1}^{n} \Big| \frac{\partial^2 \ln f(X_i;\theta)}{\partial \theta^3} \Big| \leq \frac{1}{n} \sum_{i=1}^{n} M(X_i).$$

Since $\mathbb{E}_{\theta_0}|M(X)| < \infty$, applying Law of Large Numbers, we have $\frac{1}{n}\sum_{i=1}^{n} M(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}_{\theta_0}[M(X)]$. Moreover, since $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, for any $\epsilon > 0$, there exists $N > 0$ so that $\mathbb{P}[|\hat{\theta}_n - \theta_0| < c_0] \geq 1 - \frac{\epsilon}{2}$ and

$$\mathbb{P}\Big[\Big| \frac{1}{n} \sum_{i=1}^{n} M(X_i) - \mathbb{E}_{\mu_0}[M(X)] \Big| < 1\Big] \geq 1 - \frac{\epsilon}{2},$$

for all $n \geq N$. Therefore,

$$\mathbb{P}\Big[\Big| -\frac{1}{n} l'''(\theta_n^*) \Big| \leq 1 + \mathbb{E}_{\theta_0}[M(X)]\Big] \geq 1 - \frac{\epsilon}{2},$$

hence $n^{-1} l'''(\theta_n^*)$ is bounded in probability. This implies the desired result. □

5.6 Exercises

5.6.1 Confidence interval

5.1. For a normal population with known variance σ^2, answer the following questions:

1. What is the confidence level for the interval $\bar{x} - 2.14\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 2.14\sigma/\sqrt{n}$.
2. What is the confidence level for the interval $\bar{x} - 2.49\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 2.49\sigma/\sqrt{n}$.
3. What is the confidence level for the interval $\bar{x} - 1.85\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.84\sigma/\sqrt{n}$.

5.2. A confidence interval estimate is desired for the gain in a circuit on a semiconductor device. Assume that gain is normally distributed with standard deviation $\sigma = 20$.

1. Find a 95% CI for μ when $n = 10$ and $\bar{x} = 1000$.
2. Find a 95% CI for μ when $n = 25$ and $\bar{x} = 1000$.
3. Find a 99% CI for μ when $n = 10$ and $\bar{x} = 1000$.
4. Find a 99% CI for μ when $n = 25$ and $\bar{x} = 1000$.

5.3. Following are two confidence interval estimates of the mean μ of the cycles to failure of an automotive door latch mechanism (the test was conducted at an elevated stress level to accelerate the failure).

$$3124.9 \leq \mu \leq 3215.7 \qquad 3110.5 \leq \mu \leq 3230.1.$$

1. What is the value of the sample mean cycles to failure?

2. The confidence level for one of these CIs is 95% and the confidence level for the other is 99%. Both CIs are calculated from the same sample data. Which is the 95% CI? Explain why.

5.4. $n = 100$ random samples of water from a fresh water lake were taken and the calcium concentration (milligrams per liter) measured. A 95% CI on the mean calcium concentration is $0.49 \leq \mu \leq 0.82$.

1. Would a 99% CI calculated from the same sample data been longer or shorter?

2. Consider the following statement: There is a 95% chance that μ is between 0.49 and 0.82. Is this statement correct? Explain your answer.

3. Consider the following statement: If $n = 100$ random samples of water from the lake were taken and the 95% CI on μ computed, and this process was repeated 1000 times, 950 of the CIs will contain the true value of μ. Is this statement correct? Explain your answer.

5.5. A research engineer for a tire manufacturer is investigating tire life for a new rubber compound and has built 16 tires and tested them to end-of-life in a road test. The sample mean and standard deviation are 60,139.7 and 3645.94 kilometers. Find a 95% confidence interval on mean tire life.

5.6. An Izod impact test was performed on 20 specimens of PVC pipe. The sample mean is $\bar{X} = 1.25$ and the sample standard deviation is $S = 0.25$. Find a 99% lower confidence bound on Izod impact strength.

5.7. The compressive strength of concrete is being tested by a civil engineer. He tests 12 specimens and obtains the following data.

2216	2237	2225	2301	2318	2255
2249	2204	2281	2263	2275	2295

1. Is there evidence to support the assumption that compressive strength is normally distributed? Does this data set support your point of view? Include a graphical display in your answer.

2. Construct a 95% confidence interval on the mean strength.

5.8. A machine produces metal rods. A random sample of 15 rods is selected, and the diameter is measured. The resulting date (in millimetres) are as follows

8.24	8.25	8.2	8.23	8.24
8.21	8.26	8.26	8.2	8.25
8.23	8.23	8.19	8.28	8.24

1. Check the assumption of normality for rod diameter.

2. Find a 95% CI on mean rod diameter.

5.9. A rivet is to be inserted into a hole. A random sample of $n = 15$ parts is selected, and the hole diameter is measured. The sample standard deviation of the hole diameter measurements is $s = 0.008$ millimeters. Construct a 99% CI for σ^2.

5.10. The sugar content of the syrup in canned peaches is normally distributed with standard deviation σ. A random sample of $n = 10$ cans yields a sample standard deviation of $s = 4.8$ milligrams. Find a 95% CI for σ.

5.11. Of 1000 randomly selected cases of lung cancer, 823 resulted in death within 10 years.

1. Construct a 95% CI on the death rate from lung cancer.

2. How large a sample would be required to be at least 95% confident that the error in estimating the 10-year death rate from lung cancer is less than 0.03?

5.12. A random sample of 50 suspension helmets used by motorcycle riders and automobile race-car drivers was subjected to an impact test, and on 18 of these helmets some damage was observed.

1. Find a 95% CI on the true proportion of helmets of this type that would show damage from this test.

2. Using the point estimate of p obtained from the preliminary sample of 50 helmets, how many helmets must be tested to be 95% confident that the error in estimating the true value of p is less than 0.02?

3. How large must the sample be if we wish to be at least 95% confident that the error in estimating p is less than 0.02, regardless of the true value of p?

5.13. Consider a CI for the mean μ when σ is known,

$$\bar{x} - z_{\alpha_1}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha_2}\sigma/\sqrt{n}$$

where $\alpha_1 + \alpha_2 = \alpha$. If $\alpha_1 = \alpha_2 = \alpha/2$, we have the usual $100(1-\alpha)\%$ CI for μ. In the above, when $\alpha_1 \neq \alpha_2$, the CI is not symmetric about μ. The length of the interval is $L = \sigma(z_{\alpha_1} + z_{\alpha_2})/\sqrt{n}$. Prove that the length of the interval L is minimized when $\alpha_1 = \alpha_2 = \alpha/2$.

5.14. Let the observed value of the mean $\bar{X}$ of a random sample of size 20 from a distribution that is $N(\mu, 80)$ be 81.2. Find a 95 percent confidence interval for μ.

5.15. Let $\bar{X}$ be the mean of a random sample of size n from a distribution that is $N(\mu, 9)$. Find n such that $\mathbb{P}[\bar{X} - 1 < \mu < \bar{X} + 1] = 0.90$, approximately.

5.16. Let a random sample of size 17 from the normal distribution $N(\mu, \sigma^2)$ yield $\bar{x} = 4.7$ and $s^2 = 5.76$. Determine a 90 percent confidence interval for μ.

5.17. Let $\bar{X}$ denote the mean of a random sample of size n from a distribution that has mean μ and variance $\sigma^2 = 10$. Find n so that the probability is approximately 0.954 that the random interval $(\bar{X} - \frac{1}{2}, \bar{X} + \frac{1}{2})$ includes μ.

5.18. Find a $1 - \alpha$ confidence interval for θ, given $X_1, \dots, X_n$ iid with pdf

1. $f(x;\theta) = 1$ if $\theta - \frac{1}{2} < x < \theta + \frac{1}{2}$.
2. $f(x;\theta) = 2x/\theta^2$ if $0 < x < \theta, \quad \theta > 0$.

5.19. Let $(X_1, \dots, X_n)$ be a random sample from a $\mathcal{N}(0, \sigma_X^2)$, and let $(Y_1, \dots, Y_m)$ be a random sample from a $\mathcal{N}(0, \sigma_Y^2)$, independent of the Xs. Define $\lambda = \sigma_Y^2/\sigma_X^2$. Find a $(1-\alpha)$ CI for λ.

5.20. Suppose that $X_1, \dots, X_n$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ population.

1. If σ^2 is known, find a minimum value for n to guarantee that a 0.95 CI for μ will have length no more than $\sigma/4$.
2. If σ^2 is unknown, find a minimum value for n to guarantee, with probability 0.90, that a 0.95 CI for μ will have length no more than $\sigma/4$.

5.21. Let $(X_1, \dots, X_n)$ be iid uniform $\mathcal{U}(0;\theta)$. Let Y be the largest order statistics. Show that the distribution of Y/θ does not depend on θ, and find the shortest $(1-\alpha)$ CI for θ.

5.6.2 Point estimator

5.22. Let X_1, X_2, X_3 be a random sample of size three from a uniform $(\theta, 2\theta)$ distribution, where $\theta > 0$.

1. Find the method of moments estimator of θ.
2. Find the MLE of θ.

5.23. Let $X_1, \dots, X_n$ be a random sample from the pdf

$$f(x;\theta) = \theta x^{-2}, \quad 0 < \theta < x < \infty.$$

1. What is a sufficient statistics for θ.
2. Find the mle of θ.
3. Find the method of moments estimator of θ.

5.24. Let $X_1, \dots, X_n$ be iid with density

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

Show that the mle of θ exists and is unique.

5.25. Let $X_1, \dots, X_n$ represent a random sample from each of the distributions having the following pdfs or pmfs:

1. $f(x;\theta) = \theta^x e^{-\theta}/x!,\ x = 0, 1, 2, \dots,\ 0 \leq \theta < \infty$, zero elsewhere.
2. $f(x;\theta) = \frac{1}{\theta}\mathbb{I}_{\{0<x<\theta\}},\ \theta > 0.$
3. $f(x;\theta) = \theta x^{\theta-1}\mathbb{I}_{\{0<x<1\}},\ 0 < \theta < \infty.$
4. $f(x;\theta) = \frac{e^{-x/\theta}}{\theta}\mathbb{I}_{\{x>0\}},\ 0 < \theta < \infty.$
5. $f(x;\theta) = e^{\theta-x}\mathbb{I}_{\{x>\theta\}},\ -\infty < \theta < \infty.$
6. $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}, \quad -\infty < x < \infty, -\infty < \theta < \infty$. Find the mle of θ.

In each case find the mle $\hat{\theta}$ of θ.

5.26. Let $X_1, \dots, X_n$ be a sample from the inverse Gaussian pdf

$$f(x;\mu,\lambda) = \Big(\frac{\lambda}{2\pi x^3}\Big)^{1/2} \exp\big(-\lambda(x-\mu)^2/(2\mu^2 x)\big), \quad x > 0.$$

Find the mles of μ and λ.

5.27. Suppose $X_1, \dots, X_n$ are iid with pdf $f(x;\theta) = \frac{2x}{\theta^2}I_{\{0<x\leq\theta\}}$. Find

1. the mle $\hat{\theta}$ for θ;
2. the constant c so that $\mathbb{E}[c\hat{\theta}] = \theta$;
3. the mle for the median of the distribution.

5.28. Suppose $X_1, \dots, X_n$ are iid with pdf $f(x;\theta) = e^{-x/\theta}\mathbb{I}_{\{0<x<\infty\}}$. Find the mle of $\mathbb{P}[X \geq 3]$.

5.29. The independent random variables $X_1, \dots, X_n$ have the common distribution

$$\mathbb{P}(X_i \leq x|\alpha,\beta) = \begin{cases} 0 & \text{if } x < 0 \\ (x/\beta)^\alpha & \text{if } 0 \leq x \leq \beta \\ 1 & \text{if } x > \beta, \end{cases}$$

where the parameter α and β are positive.

1. Find a two dimensional sufficient statistics for (α,β).
2. Find the mles of α and β.
3. The length (in millimeters) of cuckoos' eggs found in hedge sparrow nests can be modelled with this distribution. Fot the data

   $$22,0,\ 23,9,\ 20,9,\ 23,8,\ 25,0,\ 24,0\ 21,7,\ 23,8,\ 22,8,\ 23,1,\ 23,1,\ 23,5,\ 23,0,\ 23,0,$$

   find the mles of α and β.

5.30. Suppose that the random variables $Y_1, \dots, Y_n$ satisfy

$$Y_i = \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where $x_1, \dots, x_n$ are fixed constants, and $\epsilon_!, \dots, \epsilon_n$ are iid $\mathcal{N}(0, \sigma^2)$, σ^2 unknown.

1. Show that $\hat{\beta}_n := \sum Y_i / \sum x_i$ is an unbiased estimator of β. Find the variance of $\hat{\beta}$.
2. Find a two-dimensional sufficient statistics for (β, σ^2).
3. Find the mle $\bar{\beta}_n$ of β, and show that it is an unbiased estimator of β. Compare the variances of $\bar{\beta}_n$ and $\hat{\beta}_n$.
4. Find the distribution of the mle of β.

5.6.3 Lower bound for variance

5.31. Let $(X_1, \dots, X_n)$ be a random sample from a population with eman μ and variance σ^2.

1. Show that the estimator $\sum_{i=1}^n a_i X_i$ is an unbiased estimator of μ if $\sum_{i=1}^n a_i = 1$.
2. Among all such unbiased estimator, find the one with minimum variance, and calculate the variance.

5.32. Given the pdf

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

show that the Rao-Cramér lower bound is $\frac{2}{n}$, where n is the sample size. What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ if $\hat{\theta}$ is the mle of θ?

5.33. Let X have a gamma distribution with $\alpha = 4$ and $\beta = \theta > 0$.

1. Find the Fisher information $I(\theta)$.
2. If $(X_1, \dots, X_n)$ is a random sample from this distribution, show that the mle of θ is an efficient estimator of θ.
3. What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$.

5.34. Let X be $\mathcal{N}(0;\theta)$, $0 < \theta < \infty$.

1. Find the Fisher information $I(\theta)$.
2. If $(X_1, \dots, X_n)$ is a random sample from this distribution, show that the mle of θ is an efficient estimator of θ.
3. What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$.

5.35. Let $(X_1, \dots, X_n)$ be a random sample from a $\mathcal{N}(0;\theta)$ distribution. We want to estimate the standard deviation $\sqrt{\theta}$. Find the constant c so that $Y = c\sum_{i=1}^n |X_i|$ is an unbiased estimator of $\sqrt{\theta}$ and determine its efficiency.

5.36. If $(X_1, \dots, X_n)$ is a random sample from a distribution with pdf

$$f(x;\theta) = \begin{cases} \frac{3\theta^3}{(x+\theta)^4} & 0 < x < \infty,\ 0 < \theta < \infty \\ 0 & \text{otherwise,} \end{cases}$$

show that $Y = 2\bar{X}_n$ is an unbiased estimator of θ and determine its efficiency.

5.37 (Beta $(\theta, 1)$ distribution)**.** Let $X_1, X_2, \dots, X_n$ denote a random sample of size $n > 2$ from a distribution with pdf

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

where the parameter space $\Omega = (0, \infty)$.

1. Show that $\hat{\theta} = -\frac{n}{\sum_{i=1}^n \ln X_i}$ is the MLE estimator of θ.

2. Show that $\hat{\theta}$ is gamma distributed.

3. Show that $\hat{\theta}$ is asymptotic unbiased estimator of θ.

4. Is $\hat{\theta}$ an efficient estimator of θ?

5.38. Let $X_1, \dots, X_n$ be iid $\mathcal{N}(\theta, 1)$. Show that the best unbiased estimator of θ^2 is $\overline{X}_n^2 - \frac{1}{n}$. Calculate its variance and show that it is greater thatn the Cramer-Rao lower bound.

Chapter 6

Hypothesis Testing

6.1 Introduction

Point estimation and confidence intervals are useful statistical inference procedures. Another type of inference that is frequently used concerns tests of hypotheses. As in the last section, suppose our interest centers on a random variable X which has density function $f(x;\theta)$ where $\theta \in \Theta$. Suppose we think, due to theory or a preliminary experiment, that $\theta \in \Theta_0$ or $\theta \in \Theta_1$ where Θ_0 and Θ_1 are subsets of Θ and $\Theta_0 \cup \Theta_1 = \Theta$. We label the hypothesis as

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{6.1}$$

We call H_0 the *null hypothesis* and H_1 the *alternative hypothesis.* A hypothesis of the form $\theta = \theta_0$ is called a *simple hypothesis* while a hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a *composite hypothesis.* A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called a *two-sided test.* A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a *one-sided test.*

Often the null hypothesis represents no change or no difference from the past while the alternative represents change or difference. The alternative is often referred to as the research worker's hypothesis. The decision rule to take H_0 or H_1 is based on a sample $X_1, \ldots, X_n$ from the distribution of X and hence, the decision could be right or wrong. There are only two types of statistical errors we may commit: rejecting H_0 when H_0 is true (called the Type I error) and accepting H_0 when H_0 is wrong (called the Type II error).

Let $\mathcal{D}$ denote the sample space. A test of H_0 versus H_1 is based on a subset C of $\mathcal{D}$. This set C is called the *critical region* and its corresponding decision rule is:

Table 6.1: Decision Table for a Test of Hypothesis

Decision	H_0 is True	H_1 is True
Reject H_0	Type I Error	Correct Decision
Accept H_0	Correct Decision	Type II Error

- Reject H_0 (Accept H_1) if $(X_1, \dots, X_n) \in C$;
- Retain H_0 (Reject H_0) if $(X_1, \dots, X_n) \notin C$.

Our goal is to select a critical region which minimizes the probability of making error. In general, it is not possible to simultaneously reduce Type I and Type II errors because of a see-saw effect: if one takes $C = \emptyset$ then H_0 would be never rejected so the probability of Type I error would be 0, but the Type II error occurs with probability 1. Type I error is usually considered to be worse than Type II. Therefore, we will choose a critical regions which, on one hand, bound the probability of Type I error at a certain level, and on the other hand, minimizes the probability of Type II error.

Definition 6.1.1. A critical region C is called *of size* α if

$$\alpha = \max_{\theta \in \Theta_0} \mathbb{P}_\theta[(X_1, \dots, X_n) \in C].$$

α is also called the *significance level* of the test associated with critical region C.

Over all critical regions of size α, we will look for the one whom has the lowest probability of Type II error. It also means that for $\theta \in \Theta_1$, we want to maximize

$$1 - \mathbb{P}_\theta[\text{Type II Error}] = \mathbb{P}_\theta[(X_1, \dots, X_n) \in C].$$

We call the probability on the right side of this equation the *power of the test at* θ. So our task is to find among all the critical region of size α the one with highest power.

We define the *power function* of a critical region by

$$\gamma_C(\theta) = \mathbb{P}_\theta[(X_1, \dots, X_n) \in C], \quad \theta \in \Theta_1.$$

Example 6.1.2. Suppose $X_1, \dots, X_n$ is a random sample from a $N(\mu, 1)$ distribution. Consider the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

where $\mu_0 < \mu_1$ are specified. Let's consider a critical region C of the form $C = \{\bar{X}_n > k\}$. Since $\bar{X}_n$ has the $N(\mu, \frac{1}{n})$ distribution, the size of critical regions is

$$\alpha = \mathbb{P}_{\mu_0}[\bar{X}_n > k] = 1 - \Phi(\sqrt{n}(k - \mu_0)).$$

The power function of the critical region C is

$$\gamma_C(\mu_1) = \mathbb{P}_{\mu_1}[\bar{X}_n > k] = 1 - \Phi(\sqrt{n}(k - \mu_1)).$$

In particular, if we have $\mu_0 = 0, \mu_1 = 1, n = 100$ then at the significant level 5%, we would reject H_0 in favor of H_1 if $\bar{X}_n > 0.1965$ and the power of the test is $1 - \Phi(-8.135) = 0.9999$.

Example 6.1.3 (Large Sample Test for the Mean)**.** Let $X_1, \dots, X_n$ be a random sample from the distribution of X with mean μ and finite variance σ^2. We want to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$

where μ_0 is specified. To illustrate, suppose μ_0 is the mean level on a standardized test of students who have been taught a course by a standard method of teaching. Suppose it is hoped that a new method which incorporates computers will have a mean level $\mu > \mu_0$, where $\mu = \mathbb{E}[X]$ and X is the score of a student taught by the new method. This conjecture will be tested by having n students (randomly selected) to be taught under this new method.

Because $\bar{X}_n \to \mu$ in probability, an intuitive decision rule is given by

Reject H_0 in favor of H_1 if $\bar{X}_n$ is much large than μ_0.

In general, the distribution of the sample mean cannot be obtained in closed form. So we will use the Central Limit Theorem to find the critical region. Indeed, since

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \xrightarrow{w} N(0,1),$$

we obtain a test with an approximate size α:

Reject H_0 in favor of H_1 if $\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}} \geq x_\alpha$.

The power of the test is also approximated by using the Central Limit Theorem

$$\gamma(\mu) = \mathbb{P}[\bar{X}_n \geq \mu_0 + x_\alpha \sigma/\sqrt{n}] \approx \Phi\Big(- x_\alpha - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\Big).$$

So if we have some reasonable idea of what σ equals, we can compute the approximate power function.

Finally, note that if X has normal distribution then $\frac{\bar{X}_n - \mu}{S/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom. Thus we can establish a rejection rule having exact level α:

Reject H_0 in favor of H_1 if $T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} \geq t_{\alpha,n-1}$,

where $t_{\alpha,n-1}$ is the upper α critical point of a t distribution with $n-1$ degrees of freedom.

One way to report the results of a hypothesis test is to state that the null hypothesis was or was not rejected at a specified α-value or level of significance. For example, we can say that H_0: $\mu = 0$ was rejected at the 0.05 level of significance. This statement of conclusions is often inadequate because it gives the decision maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

To avoid these difficulties the *P-value approach* has been adopted widely in practice. The P-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true. Thus, a P-value conveys much information about the weight of evidence against H_0, and so a decision maker can draw a conclusion at any specified level of significance. We now give a formal definition of a P-value.

Definition 6.1.4. The *P-value* is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

This mean that if $\alpha > P$-value, we would reject H_0 while if $\alpha < P$-value, we would not reject H_0.

6.2 Method of finding test

6.2.1 Likelihood Ratio Tests

Let $L(\mathbf{x};\theta)$ be the likelihood function of the sample $(X_1,\dots,X_n)$ from a distribution with density $p(x;\theta)$.

Definition 6.2.1. The *likelihood test statistic* for testing $H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\mathbf{x};\theta)}{\sup_{\theta\in\Theta} L(\mathbf{x};\theta)}.$$

A likelihood ratio test is any test that has a rejection region of the form $C = \{\mathbf{x} : \lambda(\mathbf{x}) \le c\}$ for some $c \in [0,1]$.

The motivation of likelihood ratio test comes from the fact that if θ_0 is the true value of θ then, asymptotically, $L(\theta_0)$ is the maximum value of $L(\theta)$. Therefore, if H_0 is true, λ should be close to 1; while if H_1 is true, λ should be smaller.

Example 6.2.2 (Likelihood Ratio Test for the Exponential Distribution)**.** Suppose $X_1,\dots,X_n$ are iid with pdf $f(x;\theta) = \theta^{-1}e^{-x/\theta}\mathbb{I}_{\{x>0\}}$ and $\theta > 0$. Let's consider the hypotheses

$$H_0 :\ \theta = \theta_0 \text{ versus } H_1 :\ \theta \ne \theta_0,$$

where $\theta_0 > 0$ is a specified value. The likelihood ratio test statistic simplifies to

$$\lambda(\mathbf{X}) = e^n\Big(\frac{\bar{X}_n}{\theta_0}\Big)^n e^{-n\bar{X}_n/\theta_0}.$$

The decision rule is to reject H_0 if $\lambda \le c$. Using differential calculus, it is easy to show that $\lambda \le c$ iff $\bar{X} \le c_1\theta_0$ or $\bar{X} \ge c_2\theta_0$ for some positive constants c_1, c_2. Note that under the null hypothesis, H_0, the statistic $\frac{2}{\theta_0}\sum_{i=1}^n X_i$ has a χ^2 distribution with $2n$ degrees of freedom. Therefore, the following decision rule results in a level α test:

$$\text{Reject } H_0 \text{ if } \tfrac{2}{\theta_0}\textstyle\sum_{i=1}^n X_i \le \chi^2_{1-\alpha/2}(2n) \text{ or } \tfrac{2}{\theta_0}\textstyle\sum_{i=1}^n X_i \ge \chi^2_{\alpha/2}(2n),$$

where $\chi^2_{1-\alpha/2}(2n)$ is the lower $\alpha/2$ quantile of a χ^2 distribution with $2n$ degrees of freedom and $\chi^2_{\alpha/2}(2n)$ is the upper $\alpha/2$ quantile of a χ^2 distribution with $2n$ degrees of freedom.

If $\varphi(\mathbf{X})$ is a sufficient statistic for θ with pdf or pmf $g(t;\theta)$, then we might consider constructing an likelihood ratio test based on φ and its likelihood function $L^*(t;\theta) = g(t;\theta)$ rather than on the sample $\mathbf{X}$ and its likelihood function $L(\mathbf{x};\theta)$.

Theorem 6.2.3. *If $\varphi(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the likelihood ratio test statistics based on φ and $\mathbf{X}$, respectively, then $\lambda^*(\varphi(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.*

Proof. From the Factorization Theorem, the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$, where $g(t;\theta)$ is the pdf or pmf of T and $h(\mathbf{x})$ does not depend on θ. Thus

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0} L(\mathbf{x};\theta)}{\sup_{\Theta} L(\mathbf{x};\theta)} = \frac{\sup_{\Theta_0} f(\mathbf{x};\theta)}{\sup_{\Theta} f(\mathbf{x};\theta)} = \frac{\sup_{\Theta_0} g(T(\mathbf{x});\theta)h(\mathbf{x})}{\sup_{\Theta} g(T(\mathbf{x});\theta)h(\mathbf{x})} \\ &= \frac{\sup_{\Theta_0} L^*(T(\mathbf{x});\theta)}{\sup_{\Theta} L^*(\mathbf{x};\theta)} = \lambda^*(T(\mathbf{x})).\end{aligned}$$

□

6.3 Method of evaluating test

6.3.1 Most powerful test

Now we consider a test of a simple hypothesis H_0 versus a simple alternative H_1. Let $f(x;\theta)$ denote the density of a random variable X where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from the distribution of X.

Definition 6.3.1. A subset C of the sample space is called a *best critical region* of size α for testing the simple hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1,$$

if $\mathbb{P}_{\theta_0}[X \in C] = \alpha$ and for every subset A of the sample space

$$\mathbb{P}_{\theta_0}[X \in A] = \alpha \ \text{ implies } \ \mathbb{P}_{\theta_1}[X \in C] \geq \mathbb{P}_{\theta_1}[X \in A].$$

The following theorem of Neyman and Pearson provides a systematic method of determining a best critical region.

Theorem 6.3.2. *Let $(X_1, \ldots, X_n)$ be a sample from a distribution that has density $f(x;\theta)$. Then the likelihood of $X_1, X_2, \ldots, X_n$ is*

$$L(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \textit{for } \mathbf{x} = (x_1, \ldots, x_n).$$

Let θ_0 and θ_1 be distinct fixed values of θ so that $\Theta = \{\theta_0, \theta_1\}$, and let k be a positive number. Let C be a subset of the sample space such that

(a) $\frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} \leq k$ *for each* $\mathbf{x} \in C$;

(b) $\frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} \geq k$ *for each* $\mathbf{x} \in \mathcal{D} \backslash C$;

(c) $\alpha = \mathbb{P}_{\theta_0}[\mathbf{X} \in C]$.

Then C is a best critical region of size α for testing the simple hypothesis

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1.$$

Proof. We prove the theorem when the random variables are of the continuous type. If A is another critical region of size α, we will show that

$$\int_C L(\mathbf{x}; \theta_1) d\mathbf{x} \geq \int_A L(\mathbf{x}; \theta_1) d\mathbf{x}.$$

Write C as the disjoint union of $C \cap A$ and $C \cap A^c$ and A as the disjoint union of $A \cap C$ and $A \cap C^c$, we have

$$\begin{aligned}\int_C L(\mathbf{x}; \theta_1) d\mathbf{x} - \int_A L(\mathbf{x}; \theta_1) d\mathbf{x} &= \int_{C\cap A^c} L(\mathbf{x}; \theta_1) d\mathbf{x} - \int_{A\cap C^c} L(\mathbf{x}; \theta_1) d\mathbf{x} \\ &\geq \frac{1}{k} \int_{C\cap A^c} L(\mathbf{x}; \theta_0) d\mathbf{x} - \frac{1}{k} \int_{A\cap C^c} L(\mathbf{x}; \theta_0) d\mathbf{x},\end{aligned}$$

where the last inequality follows from conditions (a) and (b). Moreover, we have

$$\int_{C\cap A^c} L(\mathbf{x}; \theta_0) d\mathbf{x} - \int_{A\cap C^c} L(\mathbf{x}; \theta_0) d\mathbf{x} = \int_C L(\mathbf{x}; \theta_0) d\mathbf{x} - \int_A L(\mathbf{x}; \theta_0) d\mathbf{x} = \alpha - \alpha = 0.$$

This implies the desired result. □

Example 6.3.3. Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a random sample from the distribution $N(\theta, 1)$. It is desired to test the simple hypothesis

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta = 1.$$

We have

$$\frac{L(0; \mathbf{x})}{L(1; \mathbf{x})} = \frac{\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2} \sum_{i=1}^n x_i^2 \Big)}{\frac{1}{(2\pi)^{n/2}} \exp\Big(-\frac{1}{2} \sum_{i=1}^n (x_i - 1)^2 \Big)} = \exp\Big(-\sum_{i=1}^n x_i + \frac{n}{2} \Big).$$

If $k > 0$, the set of all points $(x_1, \ldots, x_n)$ such that

$$\exp\Big(-\sum_{i=1}^n x_i + \frac{n}{2} \Big) \leq k \Leftrightarrow \frac{1}{n} \sum_{i=1}^n x_i \geq \frac{1}{2} - \frac{\ln k}{n} = c$$

is a best critical region, where c is a constant that can be determined so that the size of the critical region is α. Since $\bar{X}_n \sim N(0, 1/n)$,

$$\mathbb{P}_{\theta_0}(\bar{X}_n \geq c) = \alpha \Leftrightarrow c = \Phi^{-1}(1 - \alpha),$$

where Φ^{-1} is the reverse function of $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$.

If $\bar{X}_n \geq c$, the simple hypothesis $H_0 : \theta = 0$ would be rejected at the significance level α; if $\bar{X}_n < c$, the hypothesis H_0 would be accepted.

The probability of rejecting H_0 when H_0 is true is α; the probability of rejecting H_0, when H_0 is false, is the value of the power of the test at $\theta = 1$. That is

$$\mathbb{P}_{\theta_1}[\bar{X}_n \geq c] = \int_c^\infty \frac{1}{\sqrt{2\pi/n}} \exp\Big(- \frac{(x-1)^2}{2/n}\Big) dx.$$

For example, if $n = 25,\ \alpha = 0.05$ then $c = 0.329$. Thus the power of this best test of H_0 against H_1 is 0.05 at $\theta = 1$ is

$$\int_{0.329}^\infty \frac{1}{\sqrt{2\pi/25}} \exp\Big(- \frac{(x-1)^2}{2/25}\Big) dx = 1 - \Phi(-3.355) = 0.999.$$

6.3.2 Uniformly most powerful test

We now define a critical region when it exists, which is a best critical region for testing a simple hypothesis H_0 against an alternative composite hypothesis H_1.

Definition 6.3.4. The critical region C is a *uniformly most powerful* (UMP) critical region of size α for testing the simple hypothesis H_0 against an alternative composite hypothesis H_1 if the set C is a best critical region of size a for testing H_0 against each simple hypothesis in H_1. A test defined by this critical region C is called a *uniformly most powerful* (UMP) test, with significance level α, for testing the simple hypothesis H_0 against the alternative composite hypothesis H_1.

It is well-known that uniformly most powerful tests do not always exist. However, when they do exist, the Neyman-Pearson theorem provides a technique for finding them.

Example 6.3.5. Let $(X_1, X_2, \ldots, X_n)$ be a random sample from the distribution $N(0, \theta)$, where the variance θ is an unknown positive number. We will show that there exists a uniformly most powerful test with significance level α for testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

The joint density of $X_1, \ldots, X_n$ is

$$L(\theta; x_1, \ldots, x_n) = \frac{1}{(2n\theta)^{n/2}} \exp\Big(- \frac{1}{2\theta} \sum_{i=1}^n x_i^2 \Big).$$

Let θ' be a number greater than θ_0 an let k denote a positive number. Let C be the set of points where

$$\frac{L(\theta_0; \mathbf{x})}{L(\theta'; \mathbf{x})} \leq k \Leftrightarrow \sum_{i=1}^n x_i^2 \geq \frac{2\theta_0\theta'}{\theta' - \theta_0} \Big(\frac{n}{2} \ln \frac{\theta'}{\theta_0} - \ln k \Big) = c.$$

The set $C = \Big\{ (x_1, \ldots, x_n) : \sum_{i=1}^n x_i^2 \geq c \Big\}$ is then a best critical region for our testing problem. It remains to determine c so that this critical region has size α, i.e.,

$$\alpha = \mathbb{P}_{\theta_0}\Big[\sum_{i=1}^n X_i^2 \geq c \Big].$$

This can be done using the observation that $\frac{1}{\theta_0}\sum_{i=1}^n X_i^2$ has a χ^2-distribution with n degrees of freedom. Note that for each number $\theta' > \theta_0$, the foregoing argument holds. It means that C is a uniformly most powerful critical region of size α.

In conclusion, if $\sum_{i=1}^n X_i^2 \geq c$, H_0 is rejected at the significance level α and H_1 is accepted; otherwise, H_0 is accepted.

Example 6.3.6. Let $(X_1, \ldots, X_n)$ be a sample from the normal distribution $N(a, 1)$, where a is unknown. We will show that there is no uniformly most powerful test of the simple hypothesis

$$H_0 : a = a_0 \quad \text{versus} \quad H_1 : a \neq a_0.$$

However, if the alternative composite hypothesis is either $H_1 : a > a_0$ or $H_1 : a < a_0$, a uniformly most powerful test will exist in each instance.

Let a_1 be a number not equal to a_0. Let k be a positive number and consider

$$\frac{\frac{1}{(2\pi)^{n/2}}\exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i - a_0)^2\Big)}{\frac{1}{(2\pi)^{n/2}}\exp\Big(-\frac{1}{2}\sum_{i=1}^n (x_i - a_1)^2\Big)} \leq k \Leftrightarrow (a_1 - a_0)\sum_{i=1}^n x_i \geq \frac{n}{2}(a_1^2 - a_0^2) - \ln k.$$

This last inequality is equivalent to

$$\sum_{i=1}^n x_i \geq \frac{n}{2}(a_1 - a_0) - \frac{\ln k}{a_1 - a_0},$$

provided that $a_1 > a_0$, and it is equivalent to

$$\sum_{i=1}^n x_i \leq \frac{n}{2}(a_1 - a_0) - \frac{\ln k}{a_1 - a_0},$$

if $a_1 < a_0$. The first of these two expressions defines a best critical region for testing $H_0 : a = a_0$ against the hypothesis $a = a_1$ provided that $a_1 > a_0$, while the second expression defines a best critical region for testing $H_0 : a = a_0$ against the hypothesis $a = a_1$ provided that $a_1 < a_0$. That is, a best critical region for testing the simple hypothesis against an alternative simple hypothesis, say $a = a_0 + 1$, will not serve as a best critical region for testing $H_0 : a = a_0$ against the alternative simple hypothesis $a = a_0 - 1$. By definition, then, there is no uniformly most powerful test in the case under consideration. However, if the alternative composite hypothesis is either $H_1 : a > a_0$ or $H_1 : a < a_0$, a uniformly most powerful test will exist in each instance.

Remark 5. The sufficiency is importance for finding a test. Indeed, let $X_1, \ldots, X_n$ be a random sample from a distribution that has pdf $f(x, \theta)$, $\theta \in \Theta$. Suppose that $Y = u(X_1, \ldots, X_n)$ is a sufficient statistic for θ. It follows from the factorization theorem that the joint pdf of $X_1, \ldots, X_n$ may be written

$$L(x_1, \ldots, x_n; \theta) = k_1(u(x_1, \ldots, x_n); \theta)k_2(x_1, \ldots, x_n),$$

where $k_2(x_1, \ldots, x_n)$ does not depend upon θ. It implies that the ratio

$$\frac{L(x_1, \ldots, x_n; \theta')}{L(x_1, \ldots, x_n; \theta'')} = \frac{k_1(u(x_1, \ldots, x_n); \theta')}{k_1(u(x_1, \ldots, x_n), \theta'')}$$

depends upon $x_1, \ldots, x_n$ only through $u(x_1, \ldots, x_n)$. Consequently, if there is a sufficient statistic $Y = u(X_1, \ldots, X_n)$ for θ and if a best test or a uniformly most powerful test is desired, there is no need to consider tests which are based upon any statistic other than the sufficient statistic.

6.3.3 Monotone likelihood ratio

Consider the general one-sided hypotheses of the form

$$H_0 : \theta \leq \theta_0 \quad versus \quad H_1 : \theta > \theta_0. \tag{6.2}$$

In this section we introduce general forms of uniformly most powerful tests for these hypotheses when the sample has a so called monotone likelihood ratio.

Definition 6.3.7. Let $\mathbf{X}' = (X_1, \ldots, X_n)$ be a random sample with common pdf (or pmf) $f(x;\theta), \theta \in \Theta$. We say that its likelihood function $L(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ has *monotone likelihood ratio* in the statistic $y = u(\mathbf{x})$ if for $\theta_1 < \theta_2$, the ratio

$$\frac{L(\mathbf{x};\theta_1)}{L(\mathbf{x};\theta_2)}$$

is a monotone function of $y = u(\mathbf{x})$.

Theorem 6.3.8. *Assume that $L(\mathbf{x};\theta)$ has a monotone decreasing likelihood ratio in the statistic $y = u(\mathbf{x})$. The following test is uniformly most powerful of level α for the hypotheses* (6.2)*:*

$$\textit{Reject } H_0 \textit{ if } u(\mathbf{X}) \geq c, \tag{6.3}$$

where c is determined by $\alpha = \mathbb{P}_{\theta_0}[u(\mathbf{X}) \geq c]$.

In case $L(\mathbf{x};\theta)$ has a monotone increasing likelihood ratio in the statistic $y = u(\mathbf{x})$ we can construct a uniformly most powerful test in a similar way.

Proof. We first consider the simple null hypothesis: $H_0' : \theta = \theta_0$. Let $\theta_1 > \theta_0$ be arbitrary but fixed. Let C denote the most powerful critical region for θ_0 versus θ_1. By the Neyman-Pearson Theorem, C is defined by,

$$\frac{L(\mathbf{X};\theta_0)}{L(\mathbf{X};\theta_1)} \leq k, \quad \text{if and only if } \mathbf{X} \in C,$$

where k is determined by $\alpha = \mathbb{P}_{\theta_0}[\mathbf{X} \in C]$. However, since $\theta_1 > \theta_0$,

$$\frac{L(X;\theta_0)}{L(X;\theta_1)} = g(u(\mathbf{X})) \leq k \Leftrightarrow u(\mathbf{X}) > g^{-1}(k),$$

where $g(u(x)) = \frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)}$. Since $\alpha = \mathbb{P}_{\theta_0}[u(\mathbf{X}) \geq g^{-1}(k)$, we have $c = g^{-1}(k)$. Hence, the Neyman-Pearson test is equivalent to the test defined by (6.3). Moreover, the test is uniformly most powerful for θ_0 versus $\theta_1 > \theta_0$ because the test only depends on $\theta_1 > \theta_0$ and $g^{-1}(k)$ is uniquely determined under θ_0.

Let $\gamma(\theta)$ denote the power function of the test (6.3). For any $\theta' < \theta''$, the test (6.3) is the most powerful test for testing θ' versu θ'' with the level $\gamma(\theta')$, we have $\gamma(\theta'') > \gamma(\theta')$. Hence $\gamma(\theta)$ is a nondecreasing function. This implies $\max_{\theta<\theta_0} \gamma(\theta) = \alpha$. □

Example 6.3.9. Let $X_1, \dots, X_n$ be a random sample from a Bernoulli distribution with parameter $p = \theta$, where $0 < \theta < 1$. Let $\theta_0 < \theta_1$. Consider the ratio of likelihood,

$$\frac{L(x_1, \dots, x_n; \theta_0)}{L(x_1, \dots, x_n; \theta_1)} = \Big(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\Big)^{\sum x_i} \Big(\frac{1-\theta_0}{1-\theta_1}\Big)^n.$$

Since $\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} < 1$, the ratio is an decreasing function of $y = \sum x_i$. Thus we have a monotone likelihood ratio in the statistic $Y = \sum X_i$.

Consider the hypotheses

$$H_0 : \theta < \theta_0 \quad versus \quad H_1 : \theta > \theta_0.$$

By Theorem 6.3.8, the uniformly most powerful level α decision rule is given by

$$\text{Reject } H_0 \text{ if } Y = \sum_{i=1}^n X_i \geq c,$$

where c is such that $\alpha = \mathbb{P}_{\theta_0}[Y \geq c]$.

6.4 Some well-known tests for a single sample

6.4.1 Hypothesis test on the mean of a normal distribution, variance σ^2 known

In this section, we will assume that a random sample $X_1, X_2, \dots, X_n$ has been taken from a normal $N(\mu, \sigma^2)$ population. It is known that $\bar{X}$ is an unbiased point estimator of μ.

Hypothesis tests on the mean

Null hypothesis: $H_0 : \mu = \mu_0$
Test statistic: $Z_0 = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

Alternative hypothesis	Rejection criteria	P value
$H_1 : \mu \neq \mu_0$	$\lvert Z_0\rvert > z_{\alpha/2}$	$2[1 - \Phi(\lvert Z_0\rvert)]$
$H_1 : \mu > \mu_0$	$Z_0 > z_\alpha$	$1 - \Phi(Z_0)$
$H_1 : \mu < \mu_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Example 6.4.1. The following data give the score of 10 students in a certain exam.

$$75 \quad 64 \quad 75 \quad 65 \quad 72 \quad 80 \quad 71 \quad 68 \quad 78 \quad 62.$$

Assume that the score is normally distributed with mean μ and known variance $\sigma^2 = 36$, test the following hypotheses at the 0.05 level of significance and find the P-value of each test.

(a) $H_0 : \mu = 70$ against $H_1 : \mu \neq 70$.

(b) $H_0 : \mu = 68$ against $H_1 : \mu > 68$.

(c) $H_0 : \mu = 75$ against $H_1 : \mu < 75$.

Solution: (a) We may solve the problem by following the six-step procedure as follows.

1. The parameter of interest is μ, the score.
2. We are going to test: $H_0 : \mu = 70, \quad H_1 : \mu \neq 70$.
3. Sample size $n = 10$,
   and sample mean $\overline{X} = \frac{1}{10}(75 + 64 + 75 + 65 + 72 + 80 + 71 + 68 + 78 + 62) = 71$.
4. Significance level $\alpha = 0.05$ so $z_{\alpha/2} = 1.96$.
5. The test statistic is
   $$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 70}{6/\sqrt{10}} = 0.5270.$$
6. Since $|Z_0| < z_{\alpha/2}$ we do not reject $H_0 : \mu = 70$ in favour of $H_1 : \mu \neq 70$ at the 0.05 level of significance. More precisely, we conclude that the mean score is 70 based on a sample of 10 measurements.

The P-value of this test is $2(1 - \Phi(|Z_0|)) = 2(1 - \Phi(0.5270)) = 0.598$.

(b)

1. The parameter of interest is μ, the score.
2. We are going to test: $H_0 : \mu = 68, \quad H_1 : \mu > 68$.
3. Sample size $n = 10$,
   and sample mean $\overline{X} = \frac{1}{10}(75 + 64 + 75 + 65 + 72 + 80 + 71 + 68 + 78 + 62) = 71$.
4. Significance level $\alpha = 0.05$ so $z_{\alpha} = 1.645$.
5. The test statistic is
   $$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 68}{6/\sqrt{10}} = 1.581.$$
6. Since $Z_0 < z_\alpha$ we do not reject $H_0 : \mu = 68$ in favour of $H_1 : \mu > 68$ at the 0.05 level of significance. More precisely, we conclude that the mean score is 68 based on a sample of 10 measurements.

The P-value of this test is $1 - \Phi(Z_0) = 1 - \Phi(1.581) = 0.057$.

(c)

1. The parameter of interest is μ, the score.
2. We are going to test: $H_0 : \mu = 75, \quad H_1 : \mu < 75$.
3. Sample size $n = 10$,
   and sample mean $\overline{X} = \frac{1}{10}(75 + 64 + 75 + 65 + 72 + 80 + 71 + 68 + 78 + 62) = 71$.

4. Significance level $\alpha = 0.05$ so $z_\alpha = 1.645$.

5. The test statistic is
$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 75}{6/\sqrt{10}} = -2.108.$$

6. Since $Z_0 < -z_\alpha$ we reject $H_0 : \mu = 75$ in favour of $H_1 : \mu < 75$ at the 0.05 level of significance. More precisely, we conclude that the mean score is less than 75 based on a sample of 10 measurements.

The P-value of this test is $\Phi(Z_0) = \Phi(-2.108) = 0.018$.

Connection between hypothesis tests and confidence intervals

There is a close relationship between the test of a hypothesis about any parameter, say θ, and the confidence interval for θ. If $[l, u]$ is a $100(1-\alpha)\%$ confidence interval for the parameter θ, the test of size α of the hypothesis
$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$
will lead to rejection of H_0 if and only if θ_0 is not in the $100(1-\alpha)\%$ confidence interval $[l, u]$.

Although hypothesis tests and CIs are equivalent procedures insofar as decision making or inference about μ is concerned, each provides somewhat different insights. For instance, the confidence interval provides a range of likely values for μ at a stated confidence level, whereas hypothesis testing is an easy framework for displaying the risk levels such as the P-value associated with a specific decision.

Type II error and choice of sample size

In testing hypotheses, the analyst directly selects the type I error probability. However, the probability of type II error β depends on the choice of sample size. In this section, we will show how to calculate the probability of type II error β. We will also show how to select the sample size to obtain a specified value of β.

In the following we will derive the formula for β of the two-sided test. The ones for one-sided tests can be derived in a similar way and we leave it as an exercise for the reader.

Finding the probability of type II error β: Consider the two-sided hypothesis
$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Suppose the null hypothesis is false and that the true value of the mean is $\mu = \mu_0 + \delta$ for some δ. The test statistic Z_0 is
$$Z_0 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \frac{\delta\sqrt{n}}{\sigma} \sim \mathcal{N}\Big(\frac{\delta\sqrt{n}}{\sigma}, 1\Big).$$

Therefore, the probability of type II error is $\beta = \mathbb{P}_{\mu_0+\delta}(|Z_0| \leq z_{\alpha/2})$, i.e., *Type II error for two-sided test*

$$\beta = \Phi\Big(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\Big) - \Phi\Big(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\Big). \tag{6.4}$$

Sample size formula There are no closed form for n from equation (6.4). However, we can estimate n as follows.

Case 1 If $\delta > 0$, then $\Phi(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}) \approx 0$ then

$$\beta \approx \Phi\Big(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\Big) \Leftrightarrow n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Case 2 If $\delta < 0$, then $\Phi(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}) \approx 1$ then

$$\beta \approx 1 - \Phi\Big(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\Big) \Leftrightarrow n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Therefore, the sample size formula is defined by

Sample size formula for two-sided test

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}$$

Large-sample test

We have developed the test procedure for the null hypothesis $H_0 : \mu = \mu_0$ assuming that the population is normally distributed and that σ^2 is known. In many if not most practical situations σ^2 will be unknown. Furthermore, we may not be certain that the population is well modeled by a normal distribution. In these situations if n is large (say $n > 40$) the sample variance s^2 can be substituted for σ^2 in the test procedures with little effect. Thus, while we have given a test for the mean of a normal distribution with known σ^2, it can be easily converted into a large-sample test procedure for unknown σ^2 that is valid regardless of the form of the distribution of the population. This large-sample test relies on the central limit theorem just as the large-sample confidence interval on σ^2 that was presented in the previous chapter did. Exact treatment of the case where the population is normal, σ^2 is unknown, and n is small involves use of the t distribution and will be deferred in the next section.

6.4.2 Hypothesis test on the mean of a normal distribution, variance σ^2 unknown

Hypothesis test on the mean

We assume again that a random sample $X_1, X_2, \ldots, X_n$ has been taken from a normal $N(\mu, \sigma^2)$ population. Recall that $\overline{X}$ and $s(X)^2$ are sample mean and sample variance of the random sample $X_1, X_2, \ldots, X_n$, respectively. It is known that

$$t_{n-1} = \frac{\overline{X} - \mu}{s(X)/\sqrt{n}}$$

has a t distribution with $n-1$ degree of freedom. This fact leads to the following test on the mean μ.

Null hypothesis: $H_0 : \mu = \mu_0$
Test statistic: $T_0 = \frac{\bar{X}-\mu_0}{s(X)/\sqrt{n}}$

Alternative hypothesis	Rejection criteria	P-value
$H_1 : \mu \neq \mu_0$	$\lvert T_0\rvert > t_{\alpha/2,n-1}$	$2\mathbb{P}(t_{n-1} > \lvert T_0\rvert)$
$H_1 : \mu > \mu_0$	$T_0 > t_{\alpha,n-1}$	$\mathbb{P}(t_{n-1} > T_0)$
$H_1 : \mu < \mu_0$	$T_0 < -t_{\alpha,n-1}$	$\mathbb{P}(t_{n-1} < -T_0)$

Where $t_{a,n-1}$ satisfies $\mathbb{P}[t_{n-1} > t_{a,n-1}] = a$.

Because the t-table in the Appendix contains a few critical values for each t distribution, computation of the exact P-value directly from the table is usually impossible. However, it is easy to find upper and lower bounds on the P-value from this table.

Example 6.4.2. The following data give the IQ score of 10 students.

$$112 \quad 116 \quad 115 \quad 120 \quad 118 \quad 125 \quad 118 \quad 113 \quad 117 \quad 121.$$

Suppose that the IQ score is normally distributed $\mathcal{N}(\mu, \sigma^2)$, test the following hypotheses at the 0.05 level of significance and estimate the P-value of each test.

(a) $H_0 : \mu = 115$ against $H_1 : \mu \neq 115$.

(b) $H_0 : \mu = 115$ against $H_1 : \mu > 115$.

(c) $H_0 : \mu = 120$ against $H_1 : \mu < 120$.

Solution (a)

1. The parameter of interest is the mean IQ score μ.

2. We are going to test $H_0 : \mu = 115$ against $H_1 : \mu \neq 115$.

3. Sample size $n = 10$,
sample mean $\overline{X} = 117.5$,
sample variance $s^2(X) = 14.944$.

4. Significance level $\alpha = 0.05$ so $t_{\alpha/2,9} = 2.262$.

5. The test statistic is
$$T_0 = \frac{\bar{X} - \mu_0}{s(X)/\sqrt{n}} = \frac{117.5 - 115}{\sqrt{14.944}/\sqrt{10}} = 2.04.$$

6. Since $|T_0| < t_{\alpha/2,9}$ we do not reject $H_0 : \mu = 115$ in favour of $H_1 : \mu \neq 115$ at the 0.05 level of significance. More precisely, we conclude that the average IQ score is 115 based on a sample of 10 measurements.

Based on the table of Student distribution, we know that the P-value of this test is $2\mathbb{P}(t_9 > 2.04) \in (0.05; 0.1)$. The actual value of the P-value is 0.072.

Type II error and choice of sample size

When the true value of the mean is $\mu = \mu_0 + \delta$, the distribution for T_0 is called the *non-central t distribution* with $n-1$ degrees of freedom and non-centrality parameter $\delta\sqrt{n}/\sigma$. Therefore, the type II error of the two-sided alternative would be

$$\beta = \mathbb{P}(|T_0'| \le t_{\alpha/2,n-1})$$

where T_0' denotes the non-central t random variable.

6.4.3 Hypothesis test on the variance of a normal distribution

The hypothesis testing procedures

We assume that a random sample $X_1, X_2, \ldots, X_n$ has been taken from a normal $N(\mu, \sigma^2)$ population. Since $(n-1)s^2(X)/\sigma^2$ follows the chi-square distribution with $n-1$ degrees of freedom, we obtain the following test for value of σ^2.

Null hypothesis: $H_0 : \sigma = \sigma_0$
Test statistic: $\chi_0^2 = \frac{(n-1)s^2(X)}{\sigma_0^2}$

Alternative hypothesis	Rejection criteria	P-value
$H_1 : \sigma \neq \sigma_0$	$\chi_0^2 > c_{\alpha/2,n-1}$ or $T_0 < c_{1-\alpha/2,n-1}$	$1 - \lvert 2\mathbb{P}(\chi_{n-1}^2 > \chi_0^2) - 1\rvert$
$H_1 : \sigma > \sigma_0$	$\chi_0^2 > c_{\alpha,n-1}$	$\mathbb{P}(\chi_{n-1}^2 > \chi_0^2)$
$H_1 : \sigma < \sigma_0$	$\chi_0^2 < c_{1-\alpha,n-1}$	$\mathbb{P}(\chi_{n-1}^2 < \chi_0^2)$

Where $c_{a,n-1}$ satisfy $\mathbb{P}[\chi_{n-1}^2 > c_{a,n-1}] = a$.

Example 6.4.3. An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ (fluid ounces)^2. If the variance of fill volume exceeds 0.01 (fluid ounces)^2, an unacceptable proportion of bottles will be underfilled or overfilled. Is there evidence in the sample data to suggest that the manufacturer has a problem with underfilled or overfilled bottles? Use $\alpha = 0.05$, and assume that fill volume has a normal distribution.

Solution

1. The parameter of interest is the population variance σ^2.

2. We are going to test $H_0 : \sigma^2 = 0.01$ against $H_1 : \sigma^2 > 0.01$.

3. Sample size $n = 20$,
   sample variance $s^2(X) = 0.0153$.

4. Significance level $\alpha = 0.05$ so $c_{\alpha,19} = 30.14$.

5. The test statistic is
$$\chi_0^2 = \frac{(n-1)s^2(X)}{\sigma_0^2} = \frac{19 \times 0.0153}{0.01} = 29.07.$$

6. Since $\chi_0^2 < c_{\alpha,19}$, we conclude that there is no strong evidence that the variance of fill volume exceeds 0.01 (fluid ounces)^2.

Since $\mathbb{P}(\chi_1^29 > 27.2) = 0.10$ and $\mathbb{P}(\chi_1^29 > 30.4) = 0.05$, we conclude that the P-value of the test is in the interval $(0.05, 0.10)$. Note that the actual P-value is 0.0649.

6.4.4 Test on a population proportion

Large-Sample tests on a proportion

Let $(X_1, \ldots, X_n)$ be a random sample observing from a random variable X with $B(1,p)$ distribution. Then $\hat{p} = \overline{X}$ is a point estimator of p. By the Central limit theorem, when n is large, $\hat{p}$ is approximately normal with mean p and variance $p(1-p)/n$. We thus obtain the following test for value of p.

Null hypothesis: $H_0 : p = p_0$
Test statistic: $Z_0 = \frac{\sqrt{n}(\bar{X}-p_0)}{\sqrt{p_0(1-p_0)}}$

Alternative hypothesis	Rejection criteria	P-value
$H_1 : p \neq p_0$	$\lvert Z_0\rvert > z_{\alpha/2}$	$2(1-\Phi(\lvert Z_0\rvert)$
$H_1 : p > p_0$	$Z_0 > z_\alpha$	$1-\Phi(Z_0)$
$H_1 : p < p_0$	$Z_0 < -z_\alpha$	$\Phi(Z_0)$

Example 6.4.4. A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step not exceed 0.05 and that the manufacturer demonstrate process capability at this level of quality using $\alpha = 0.05$. The semiconductor manufacturer takes a random sample of 200 devices and finds that four of them are defective. Can the manufacturer demonstrate process capability for the customer?

Solution

1. The parameter of interest is the process fraction defective p.

2. $H_0 : p = 0.05$ against $H_1 : p < 0.05$.

3. The sample size $n = 200$, and sample proportion $\overline{X} = \frac{4}{200} = 0.02$.

4. Significance level $\alpha = 0.05$ so $z_\alpha = 1.645$.

5. The test statistic is
$$Z_0 = \frac{\sqrt{n}(\overline{X} - p_0)}{\sqrt{p_0(1-p_0)}} = \frac{\sqrt{200}(0.02 - 0.05)}{\sqrt{0.05 \times 0.95}} = -1.947.$$

6. Since $Z_0 < -z_\alpha$, we reject H_0 and conclude that the process fraction defective p is less than 0.05. The P-value for this value of the test statistic is $\Phi(-1.947)) = 0.0256$, which is less than $\alpha = 0.05$. We conclude that the process is capable.

Type II error and choice of sample size

Suppose that p is the true value of the population proportion. The approximate β-error is defined as follows

- the two-sided alternative $H_1 : p \neq p_0$

$$\beta \approx \Phi\Big(\frac{p_0 - p + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\Big) - \Phi\Big(\frac{p_0 - p - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\Big)$$

- the one-sided alternative $H_1 : p < p_0$

$$\beta \approx 1 - \Phi\Big(\frac{p_0 - p - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\Big)$$

- the one-sided alternative $H_1 : p > p_0$

$$\beta \approx \Phi\Big(\frac{p_0 - p + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\Big)$$

These equations can be solved to find the approximate sample size n that gives a test of level α that has a specified β risk. The sample size is defined as follows.

- the two-sided alternative $H_1 : p \neq p_0$

$$n = \Big[\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0}\Big]^2$$

- the one-sided alternative

$$n = \Big[\frac{z_\alpha\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p(1-p)}}{p - p_0}\Big]^2$$

6.5 Some well-known tests for two samples

6.5.1 Inference for a difference in means of two normal distributions, variances known

In this section we consider statistical inferences on the difference in means $\mu_1 - \mu_2$ of two normal distributions, where the variances σ_1^2 and σ_2 are known. The assumptions for this section are summarized as follows.

$X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
$X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2. (6.5)
The two populations represented by X_1 and X_2 are independent.
Both populations are normal.

The inference for $\mu_1 - \mu_2$ is based on the following result.

Theorem 6.5.1. *Under the assumptions stated above, the quantity*

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic: $Z_0 = \dfrac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Alternative hypothesis	Rejection criteria
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$\lvert Z_0 \rvert > z_{\alpha/2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$Z_0 > z_\alpha$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$Z_0 < z_\alpha$

When the population variances are unknown, the sample variances s_1^2 and s_2^2 can be substituted into the test statistic Z_0 to produce a large-sample test for the difference in means. This procedure will also work well when the populations are not necessarily normally distributed. However, both n_1 and n_2 should exceed 40 for this large-sample test to be valid.

Example 6.5.2. A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\overline{X}_1 = 121$ minutes and $\overline{X}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Solution:

1. The quantity of interest is the difference in mean drying time, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.

2. We are going to test: $H_0 : \mu_1 - \mu_2 = 0$ **vs** $H_1 : \mu_1 > \mu_2$.

3. The sample means $n_1 = n_2 = 10$.

4. The significance level $\alpha = 0.05$ so $z_\alpha = 1.645$.

5. The test statistic is
$$Z_0 = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{10}}} = 2.52.$$

6. Since $\Phi^{-1}(\alpha) = \Phi^{-1}(0.05) = 1.645 < Z_0$, we reject H_0 at the $\alpha = 0.05$ level and conclude that adding the new ingredient to the paint significantly reduces the drying time.

Alternatively, we can find the P-value for this test as

$$P\text{-value} = 1 - \Phi(2.52) = 0.0059.$$

Therefore $H_0 : \mu_1 = \mu_2$ would be rejected at any significance level $\alpha \geq 0.0059$.

Type 2 error and choice of sample size

6.5.2 Inference for the difference in means of two normal distributions, variances unknown

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Suppose we have two independent normal populations with unknown means μ_1 and μ_2, and unknown but equal variances σ^2. Assume that assumptions (6.5) hold.

The *pooled estimator* of σ^2, denoted by S_p^2 is defined by

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The inference for $\mu_1 - \mu_2$ is based on the following result.

Theorem 6.5.3. *Under all the assumption mentioned above, the quantity*

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test statistic: $T_0 = \dfrac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Alternative hypothesis	Rejection criteria
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$\lvert T_0\rvert > t_{\alpha/2,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$T_0 > t_{\alpha,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$T_0 < -t_{\alpha,n_1+n_2-2}$

Example 6.5.4. The IQ's of 9 children in a district of a large city have empirical mean 107 and standard deviation 10. The IQs of 12 children in another district have empirical mean 112 and standard deviation 9. Test the equality of means at the 0.05 significance of level.

Example 6.5.5. Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown in the following table.

Observation Num.	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75

Is there any difference between the mean yields? Use $\alpha = 0.05$, and assume equal variances.

Cases 2: $\sigma_1^2 \neq \sigma_2^2$

In some situations, we cannot reasonably assume that the unknown variances σ_1^2 andσ_2^2 are equal. There is not an exact t-statistic available for testing $H_0 : \mu_1 - \mu_2 = \Delta_0$ in this case. However, if H_0 is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

is distributed approximately as t with degrees of freedom given by

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Therefore, if $\sigma_1^2 \neq \sigma_2^2$, the hypotheses on differences in the means of two normal distribution are tested as in the equal variances case, except that T_0^* is used as the test statistic and $n_1 + n_2 - 2$ is replaced by ν in determining the degrees of freedom for the test.

6.5.3 Paired t-test

A special case of the two-sample t-tests of previous section occurs when the observations on the two populations of interest are collected in pairs. Each pair of observations, say (X_j, X_j), is taken under homogeneous conditions, but these conditions may change from one pair to another. For example, suppose that we are interested in comparing two different types of tips for a hardness-testing machine. This machine presses the tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen can be determined. If several specimens were selected at random, half tested with tip 1, half tested with tip 2, and the pooled or independent t-test in the previous was applied, the results of the test could be erroneous. The metal specimens could have been cut from bar stock that was produced in different heats, or they might not be homogeneous in some other way that might affect hardness. Then the observed difference between mean hardness readings for the two tip types also includes hardness differences between specimens.

A more powerful experimental procedure is to collect the data in pairs - that is, to make two hardness readings on each specimen, one with each tip. The test procedure would then consist of analyzing the differences between hardness readings on each specimen. If there is no difference between tips, the mean of the differences should be zero. This test procedure is called the paired t-test.

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be a set of n paired observations where we assume that the mean and variance of the population represented by X are μ_X and σ_X^2, and the mean and variance of the population represented by Y are μ_Y and σ_Y^2. Define the differences between each pair of observations as $D_j = X_j - Y_j, j = 1, 2, \dots, n$. The D_js are assumed to be normally distributed with mean $\mu_D = \mu_X - \mu_Y$ and variance σ_D^2, so testing hypotheses about the difference between μ_X and μ_Y can be accomplished by performing a one-sample t-test on μ_D.

Null hypothesis: $H_0 : \mu_D = \Delta_0$

Test statistic: $T_0 = \dfrac{\overline{D} - \Delta_0}{S_D/\sqrt{n}}$

Alternative hypothesis	Rejection criteria
$H_1 : \mu_D \neq \Delta_0$	$\lvert T_0\rvert > t_{\alpha/2,n-1}$
$H_1 : \mu_D > \Delta_0$	$T_0 > t_{\alpha,n-1}$
$H_1 : \mu_D < \Delta_0$	$T_0 < -t_{\alpha,n-1}$

Example 6.5.6. An article in the Journal of Strain Analysis (1983, Vol. 18, No. 2) compares several methods for predicting the shear strength for steel plate girders. Data for two of these methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown in the following table.

Karlsruhe Method	1.186	1.151	1.322	1.339	1.200	1.402	1.365	1.537	1.559
Lehigh Method	1.061	0.992	1.063	1.062	1.065	1.178	1.037	1.086	1.052
Difference D_j	0.119	0.159	0.259	0.277	0.138	0.224	0.328	0.451	0.507

Test whether there is any difference (on the average) between the two methods?

Solution:

$$\overline{D} = 0.2736, \quad S_D^2 = 0.018349, \quad T_0 = 6.05939, \quad t_{0.025,8} = 2,306.$$

We conclude that there is difference between the two method at the 0.05 level of significance.

6.5.4 Inference on the variance of two normal populations

A hypothesis-testing procedure for the equality of two variances is based on the following result.

Theorem 6.5.7. *Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample from a normal population with mean μ_1 and variance σ_1^2 and let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample from a second normal population with mean μ_2 and variance σ_2^2. Assume that both normal populations are independent. Let s_1^2 and s_2^2 be the sample variances. Then the ratio*

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

This result is based on the fact that $(n_1-1)s_1^2/\sigma_1^2$ is a chi-square random variable with $n_1 - 1$ degrees of freedom, that $(n_2-1)s_1^2/\sigma_2^2$ is a chi-square random variable with n_2-1 degrees of freedom, and that the two normal populations are independent. Clearly under the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, the ratio $F_0 = s_1^2/s_2^2$ has an F_{n_1-1,n_2-1} distribution. Let f_{α,n_1-1,n_2-1} be a constant satisfying

$$\mathbb{P}[F_0 > f_{\alpha,n_1-1,n_2-1}] = \alpha.$$

It follows from the property of F distribution that

$$f_{1-\alpha,n_1-1,n_2-1} = \frac{1}{f_{\alpha,n_1-1,n_2-1}}.$$

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic: $F_0 = \dfrac{s_1^2}{s_2^2}$

Alternative hypothesis	Rejection criteria
$H_1 : \sigma_1^2 = \sigma_2^2$	$F_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $F_0 < f_{1-\alpha/2,n_1-1,n_2-1}$
$H_1 : \sigma_1^2 > \sigma_2^2$	$F_0 > f_{\alpha,n_1-1,n_2-1}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$F_0 < f_{1-\alpha,n_1-1,n_2-1}$

Example 6.5.8. Oxide layers on semiconductor wafers are etched in a mixture of gases to achieve the proper thickness. The variability in the thickness of these oxide layers is a critical characteristic of the wafer, and low variability is desirable for subsequent processing steps. Two different mixtures of gases are being studied to determine whether one is superior in reducing the variability of the oxide thickness. Twenty wafers are etched in each gas. The sample standard deviations of oxide thickness are $s_1 = 1.96$ angstroms and $s_2 = 2.13$ angstroms, respectively. Is there any evidence to indicate that either gas is preferable? Use $\alpha = 0.05$.

Solution: At the $\alpha = 0.05$ level of significance we need to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \textbf{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Since $n_1 = n_2 = 20$, we will reject H_0 if $F_0 = \frac{s_1^2}{s_2^2} > f_{0.025,19,19} = 2.53$ or $F_0 < f_{0.975,19,19} = \frac{1}{2.53} = 0.40$.

Computation: $F_0 = \frac{1.96^2}{2.13^2} = 0.85$. Hence we cannot reject the null hypothesis H_0 at the 0.05 level of significance. Therefore, there is no strong evidence to indicate that either gas results in a smaller variance of oxide thickness.

6.5.5 Inference on two population proportions

We now consider the case where there are two binomial parameters of interest, say, p_1 and p_2, and we wish to draw inferences about these proportions. We will present large-sample hypothesis testing based on the normal approximation to the binomial.

Suppose that two independent random samples of sizes n_1 and n_2 are taken from two populations, and let X_1 and X_2 represent the number of observations that belong to the class of interest in samples 1 and 2, respectively. Furthermore, suppose that the normal approximation to the binomial is applied to each population, so the estimators of the population proportions $\hat{P}_1 = X_1/n_1$ and $\hat{P}_2 = X_2/n_2$ have approximate normal distribution. Moreover, under the null hypothesis $H_0 : p_1 = p_2 = p$, the random variable

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\dfrac{n_1+n_2}{n_1 n_2}}}$$

is distributed approximately $N(0,1)$.

This leads to the test procedures described below.

Null hypothesis: $H_0 : p_1 = p_2$

Test statistic: $Z_0 = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}(1-\hat{p})\dfrac{n_1+n_2}{n_1 n_2}}}$ with $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$.

Alternative hypothesis	Rejection criteria
$H_1 : p_1 \neq p_2$	$\|Z_0\| > z_{\alpha/2}$
$H_1 : p_1 > p_2$	$Z_0 > z_\alpha$
$H_1 : p_1 < p_2$	$Z_0 < -z_\alpha$

Example 6.5.9. Two different types of polishing solution are being evaluated for possible use in a tumble-polish operation for manufacturing interocular lenses used in the human eye following cataract surgery. Three hundred lenses were tumble- polished using the first polishing solution, and of this number 253 had no polishing-induced defects. Another 300 lenses were tumble-polished using the second polishing solution, and 196 lenses were satisfactory upon completion. Is there any reason to believe that the two polishing solutions differ? Use $\alpha = 0.01$.

6.6 The chi-square test

6.6.1 Goodness-of-fit test

Suppose that a large population consists of items of k different types, and let p_i denote the probability that an item selected at random will be of type $i(i = 1, \dots, k)$. Suppose that the following hypothesis is to be tested

$$H_0 : p_i = p_i^0 \quad \text{for } i = 1, \dots k \quad \textbf{vs} \quad H_1 : p_i = p_i^0 \quad \text{for at least one value of } i$$

We shall assume that a random sample of size n is to be taken from the given population. That is, n independent observations are to be taken, and there is probability p_i that each observation will be of type $i(i = 1, ..., k)$. On the basis of these n observations, the hypothesis is to be tested.

For each i, we denote N_i the number of observations in the random sample that are of type i.

Theorem 6.6.1 (Pearson's theorem)**.** *The following statistic*

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$$

has the property that if H_0 is true and the sample size $n \to \infty$, then Q converges in distribution to the χ^2 distribution with $k - 1$ degrees of freedom.

Chi-squared goodness-of-fit test for simple hypothesis

Suppose that we observer an i.i.d. sample $X_1, \dots, X_n$ of random variables that take a finite number of values $B_1, \dots, B_k$ with unknown probability $p_1, \dots, p_k$. Consider the hypothesis

$$H_0 : p_i = p_i^0 \quad \text{for } i = 1, \dots k \quad \textbf{vs} \quad H_1 : p_i = p_i^0 \quad \text{for at least one value of } i$$

If the null hypothesis H_0 is true then by Pearson's theorem,

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi^2_{k-1}$$

where N_i is number of X_j equal to B_j. On the other hand, if H_1 holds, then for some index i^*, $p_{i^*} \neq p_i^0$. We write

$$\frac{\nu_{i^*} - np_{i^*}^0}{\sqrt{np_{i^*}^0}} = \sqrt{\frac{p_{i^*}}{p_{i^*}^0}} \frac{\nu_{i^*} - np_{i^*}}{\sqrt{np_{i^*}}} + \sqrt{n}\frac{p_{i^*} - p_{i^*}^0}{\sqrt{p_{i^*}^0}}.$$

the first term converges to $\mathcal{N}(0, (1 - p_{i^*})p_{i^*}/p_{i^*}^0)$ by the central limit theorem while the second term diverges to plus or minus infinity. It means that if H_1 holds then $Q \to \infty$. Therefore, we will reject H_0 if $Q \geq c_{\alpha,k-1}$ where $c_{\alpha,k-1}$ is chosen such that the error of type 1 is equal to the level of significance α:

$$\alpha = \mathbb{P}_0(Q > c_{\alpha,k-1}) \approx \mathbb{P}(\chi^2_{k-1} > c_{\alpha,k-1}).$$

This test is called *chi-squared goodness-of-fit test*

Null hypothesis: $H_0 : p_i = p_i^0 \quad$ for $i = 1, \dots k$

Test statistic

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0}$$

Alternative hypothesis	Rejection criteria	P-value
$H_1 : p_i = p_i^0$ for at least one value of i	$Q \geq c_{\alpha,k-1}$	$\mathbb{P}(\chi^2_{k-1} > Q)$

Example 6.6.2. A study of blood types among a sample of 6004 people gives the following result

Blood type	A	B	AB	O
Number of people	2162	738	228	2876.

Table 6.2: Blood types

A previous study claimed that the proportions of people whose blood of types A, B, AB and O are $33.33\%, 12.5\%, 4.17\%$ and 50%, respectively.

We can use the data in Table 6.2 to test the null hypothesis H_0 that the probabilities (p_1, p_2, p_3, p_4) of the four blood type equal $(\frac{1}{3}, \frac{1}{8}, \frac{1}{24}, \frac{1}{2})$. The χ^2 test statistic is then

$$Q = \frac{(2162 - 6004 \times \frac{1}{3})^2}{6004 \times \frac{1}{3}} + \frac{(738 - 6004 \times \frac{1}{8})^2}{6004 \times \frac{1}{8}} + \frac{(228 - 6004 \times \frac{1}{24})^2}{6004 \times \frac{1}{24}} + \frac{(2876 - 6004 \times \frac{1}{2})^2}{6004 \times \frac{1}{2}} = 20.37$$

To test H_0 at the level α_0, we would compare Q to the $1 - \alpha_0$ quantile of the χ^2 distribution with three degrees of freedom. Alternatively, we can compute the P-value, which would be the smallest α_0 at which we could reject H_0. In general, the P-value is $1 - F(Q)$ where F is the cumulative distribution function of the χ^2 distribution with $k - 1$ degrees of freedom. In this example $k = 4$ and $Q = 20.37$ then the p-value is 1.42×10^{-4}.

Goodness-of-fit for continuous distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from unknown distribution P and consider the following hypotheses

$$H_0 : P = P_0 \quad \textbf{vs} \quad H_1 : P \neq P_0$$

for some particular, possibly continuous distribution P_0. To do this, we will split a set of all possible outcomes of X_i, say $\mathcal{X}$, into a finite number of intervals $I_1, \ldots, I_k$. The null hypothesis H_0 implies that for all intervals

$$\mathbb{P}(X \in I_j) = P_0(X \in I_j) = p_j^0.$$

Therefore, we can do a chi-squared test for

$$H_0' : \mathbb{P}(X \in I_j) = p_j^0 \quad \text{for all } j \leq k \quad \textbf{vs} \quad H_1' : \text{ otherwise}.$$

It is clear that H_0 implies H_0'. However, the fact that H_0' holds does not guarantee that H_0 hold. There are many distribution different from P that have the same distribution on the intervals $I_1, \ldots, I_k$ as P. On one hand, if we group into more and more intervals, our discrete approximation of P will get closer and closer to P, so in some sense H_0' will get closer to H_0. However, we can not split into too many intervals either, because the χ^2_{k1}-distribution approximation for statistic T in Pearsons theorem is asymptotic. The rule of thumb is to group the data in such a way that the expected count in each interval np_i^0 is at least 5.

Example 6.6.3. Suppose that we wish to test the null hypothesis that the logarithms of the lifetime of ball bearings are an i.i.d. sample from the normal distribution with mean $ln(50) = 3.912$ and variance 0.25. The observed logarithms are

2.88	3.36	3.95	3.99	4.53	4.59	3.50	3.73
4.02	4.22	4.66	4.66	3.74	3.82	4.23	4.23
4.85	4.85	5.16	3.88	3.95	4.23	4.43	

In order to have the expected count in each interval be at least 5, we can use at most $k = 4$ intervals. We shall make these intervals each have probability 0.25 under the null hypothesis. That is, we shall divide the intervals at the 0.25, 0.5, and 0.75 quantiles of the hypothesized normal distribution. These quantiles are

$$\begin{aligned} 3.912 + 0.5\Phi^{-1}(0.25) &= 3.575 \\ 3.912 + 0.5\Phi^{-1}(0.5) &= 3.912 \\ 3.912 + 0.5\Phi^{-1}(0.75) &= 4.249. \end{aligned}$$

The number of observation in each of the four intervals are then $3, 4, 8$ and 8. We then calculate

$$Q = 3.609.$$

Our table of the χ^2 distribution with three degrees of freedom indicates that 3.609 is between the 0.6 and 0.7 quantiles, so we would not reject the null hypothesis at levels less 0.3 and reject the null hypothesis at levels greater than 0.4. (Actually, the P-value is 0.307.)

Goodness-of-fit for composite hypotheses

We can extend the goodness-of-fit test to deal with the case in which the null hypothesis is that the distribution of our data belongs to a particular parametric family. The alternative hypothesis is that the data have a distribution that is not a member of that parametric family. There are two changes to the test procedure in going from the case of a simple null hypothesis to the case of a composite null hypothesis. First, in the test statistic Q, the probabilities p_i^0 are replaced by estimated probabilities based on the parametric family. Second, the degrees of freedom are reduced by the number of parameters.

Let us start with a discrete case when a random variable takes a finite number of values $B_1, \ldots, B_k$ and

$$p_i = \mathbb{P}(X = B_i), \quad i = 1, \ldots, k.$$

We would like to test a hypothesis that this distribution comes from a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In other words, if we denote

$$p_j(\theta) = \mathbb{P}_\theta(X = B_j),$$

we want to test

$$H_0 : p_j = p_j(\theta) \text{ for all } j \leq r \text{ for some } \theta \in \Theta \quad \textbf{vs} \quad H_1 : \text{otherwise}.$$

The situation now is complicated since we want to test if $p_j = p_j(\theta), j \leq r$ at least for some $\theta \in \Theta$ which means that we may have many candidates for θ. One way to approach this problem is as follows.

Step 1: Assuming that hypothesis H_0 holds, we can find an estimator θ^* of this unknown θ.

Step 2: Try to test if, indeed, the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistics

$$Q^* = \sum_{i=1}^{k} \frac{(N_i - np_i(\theta^*))^2}{np_i(\theta^*)}$$

in chi-squared goodness-of-fit test.

This approach looks natural, the only question is what estimate θ^* to use and how the fact that θ^* also depends on the data will affect the convergence of Q. It turns out that if we let θ^* be the maximum likelihood estimate, i.e. θ that maximizes the likelihood function

$$\phi(\theta) = p_1(\theta)^{N_1} \ldots p_k(\theta)^{N_k},$$

then the statistics Q^* converges in distribution to a χ^2_{r-s-1} distribution with $r - s - 1$ degrees of freedom, where s is the dimension of the parameter set Θ. Note that we must assume $s \leq r - 2$ so that we have at least one degree of freedom. Very informally, by dimension we understand the number of free parameters that describe the set

$$\{(p_1(\theta), \ldots, p_k(\theta)) : \theta \in \Theta\}.$$

The we will reject H_0 if $T \leq c$ where the threshold c is determined from the condition

$$\mathbb{P}(T > c|H_0) = \alpha$$

where $\alpha \in [0, 1]$ is the level of significance.

Example 6.6.4. Suppose that a gene has two possible alleles A_1 and A_2 and the combinations of these alleles define three genotypes A_1A_1, A_1A_2 and A_2A_2. We want to test a theory that

$$\begin{cases} \text{Probability to pass } A_1 \text{ to a child } & = \theta \\ \text{Probability to pass } A_2 \text{ to a child } & = 1 - \theta \end{cases}$$

and that the probabilities of genotypes are given by

$$\begin{aligned} p_1(\theta) &= \mathbb{P}(A_1A_1) = \theta^2 \\ p_2(\theta) &= \mathbb{P}(A_1A_2) = 2\theta(1-\theta) \\ p_3(\theta) &= \mathbb{P}(A_2A_2) = (1-\theta)^2. \end{aligned}$$

Suppose that given a random sample $X_1, \ldots, X_n$ from the population the counts of each genotype are N_1, N_2 and N_3. To test the theory we want to test the hypothesis

$$H_0 : p_i = p_i(\theta), i = 1, 2, 3 \quad \textbf{vs} \quad H_1 : \text{otherwise}.$$

First of all, the dimension of the parameter set is $s = 1$ since the distributions are determined by one parameter θ. To find the MLE θ^* we have to maximize the likelihood function

$$p_1(\theta)^{N_1} p_2(\theta)^{N_2} p_3(\theta)^{N_3}.$$

After computing the critical point by setting the derivative equal to 0, we get

$$\theta^* = \frac{2N_1 + N_2}{2n}.$$

Therefore, under the null hypothesis H_0 the statistics

$$Q^* = \sum_{i=1}^{3} \frac{(N_i - np_i(\theta^*))^2}{np_i(\theta^*)}$$

converges to χ^2 distribution with 1 degree of freedom. Therefore, if $\alpha = 0.05$ we will reject H_0 if $Q^* > 3.841$.

In the case when the distributions $\mathbb{P}_\theta$ are continuous or, more generally, have infinite number of values that must be grouped in order to use chi-squared test (for example, normal or Poisson distribution), it can be a difficult numerical problem to maximize the grouped likelihood function

$$\mathbb{P}_\theta(I_1)^{N_1} \cdots \mathbb{P}_\theta(I_k)^{N_k}.$$

It is tempting to use a usual non-grouped MLE $\hat{\theta}$ of θ instead of the above θ^* because it is often easier to compute, in fact, for many distributions we know explicit formulas for these MLEs. However, if we use $\hat{\theta}$ in the statistic

$$\hat{Q} = \sum_{i=1}^{k} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

then it will no longer converges to χ^2_{r-s-1} distribution. It has been shown that typically this $\hat{Q}$ will converge to a distribution "in between" χ^2_{k-s-1} and χ^2_{k-s}[1]. Thus, a conservative decision rule is to reject H_0 whether $\hat{Q} > c$ where c is chosen such that $\mathbb{P}(\chi^2_{k-1} > c) = \alpha$.

Example 6.6.5 (Testing Whether a Distribution Is Normal)**.** Consider now a problem in which a random sample $X_1, ..., X_n$ is taken from some continuous distribution for which the p.d.f. is unknown, and it is desired to test the null hypothesis H_0 that this distribution is a normal distribution against the alternative hypothesis H_1 that the distribution is not normal. To perform a χ^2 test of goodness- of-fit in this problem, divide the real line into k subintervals and count the number N_i of observations in the random sample that fall into the ith subinterval $(i = 1, ..., k)$.

If H_0 is true, and if μ and σ^2 denote the unknown mean and variance of the normal distribution, then the parameter vector θ is the two-dimensional vector $\theta(\mu, \sigma^2)$. The probability $\pi_i(\theta)$, or $\pi_i(\mu, \sigma^2)$, that an observation will fall within the ith subinterval, is the probability assigned to that subinterval by the normal distribution with mean μ and variance σ^2. In other words, if the ith subinterval is the interval from a_i to b_i, then

$$\pi_i(\mu, \sigma^2) = \Phi\Big(\frac{b_i - \mu}{\sigma}\Big) - \Phi\Big(\frac{a_i - \mu}{\sigma}\Big).$$

It is important to note that in order to calculate the value of the statistic Q^*, the M.L.E.s μ^* and σ^{2*} must be found by using the numbers $N_1, ..., N_k$ of observations in the different subintervals. The M.L.E.s should not be found by using the observed values of $X_1, ..., X_n$ themselves. In other words, μ^* and σ^{2*} will be the values of μ and σ^2 that maximize the likelihood function

$$L(\mu, \sigma^2) = [\pi_1(\mu, \sigma^2)]^{N_1} \cdots [\pi_k(\mu, \sigma^2)]^{N_k}. \tag{6.6}$$

Because of the complicated nature of the function $\pi_i(\mu, \sigma^2)$ a lengthy numerical computation would usually be required to determine the values of μ and σ^2 that maximize $L(\mu, \sigma^2)$. On the other hand, we know that the M.L.E.s of μ and σ^2 based on the n observed values $X_1, ..., X_n$ in the original sample are simply the sample mean $\overline{X}_n$ and the sample variance s_n^2. Furthermore, if the estimators that maximize the likelihood function $L(\mu, \sigma^2)$ are used to calculate the statistic Q^*, then we know that when H_0 is true, the distribution of Q^* will be approximately the χ^2 distribution with $k-3$ degrees of freedom. On the other hand, if the M.L.E.s $\overline{X}_n$ and s_n^2, which are based on the observed values in the original sample, are used to calculate $\hat{Q}$, then this χ^2 approximation to the distribution of $\hat{Q}$ will not be appropriate. Indeed, the distribution of $\hat{Q}$ is asymptotically "in between" χ^2_{k-3} and χ^2_{k-1}.

[1] Chernoff, Herman; Lehmann, E. L. (1954) The use of maximum likelihood estimates in χ^2 tests for goodness of fit. Ann. Math. Statistics 25, pp. 579-586.

Return to Example 6.6.3. We are now in a position to try to test the composite null hypothesis that the logarithms of ball bearing lifetimes have some normal distribution. We shall divide the real line into the subintervals $(\infty, 3.575]$, $(3.575, 3.912]$, $(3.912, 4.249]$, and $(4.249, +\infty)$. The counts for the four intervals are $3, 4, 8$, and 8. The M.L.E.s based on the original data gives $\hat{\mu} = 4.150$ and $\hat{\sigma}^2 = 0.2722$. The probabilities of the four intervals are $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.1350, 0.1888, 0.2511, 0.4251)$. This make the value of $\hat{Q}$ equal to 1.211.

The tail area corresponding to 1.211 needs to be computed for χ^2 distributions with $k-1=3$ and $k-3=1$ degrees of freedom. For one degree of freedom, the p-value is 0.2711, and for three degrees of freedom the p-value is 0.7504. So, our actual p-value lies in the interval $[0.2711, 0.7504]$. Although this interval is wide, it tells not to reject H_0 at level α if $\alpha < 0.2711$.

6.6.2 Tests of independence

In this section we will consider a situation when our observations are classified by two different features and we would like to test if these features are independent. For example, we can ask if the number of children in a family and family income are independent. Our sample space $\mathcal{X}$ will consist of $a \times b$ pairs.

$$\mathcal{X} = \{(i,j) : i = 1, \ldots, a,\ j = 1, \ldots, b\}$$

where the first coordinate represents the first feature that belongs to one of a categories and the second coordinate represents the second feature that belongs to one of b categories. An i.i.d. sample $X_1, ..., X_n$ can be represented by a contingency table below where N_{ij} is the number all observations in a cell (i,j).

	Feature 2			
Feature 1	1	2	$\cdots$	b
1	N_{11}	N_{12}	$\cdots$	N_{1b}
2	N_{21}	N_{22}	$\cdots$	N_{2b}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	N_{a1}	N_{a2}	$\cdots$	N_{ab}

We would like to test the independence of two features which means that

$$\mathbb{P}[X = (i,j)] = \mathbb{P}[X^1 = i]\mathbb{P}[X^2 = j].$$

Denote $\theta_{ij} = \mathbb{P}[X = (i,j)]; p_i = \mathbb{P}[X^1 = i]; q_j\mathbb{P}[X^2 = j]$. Then we want to test

$$\begin{aligned} &H_0 : \theta_{ij} = p_i q_j \text{ for all } (i,j) \text{ for some } (p_1, \ldots, p_a) \text{ and } (q_1, \ldots, q_b) \\ &H_1 : \text{otherwise.} \end{aligned}$$

We can see that this null hypothesis H_0 is a special case of the composite hypotheses from previous lecture and it can be tested using the chi-squared goodness-of-fit test. The total number of groups is $k = a \times b$. Since p_is and q_js should add up to one, one parameter in each sequence, for example p_a and q_b, can be computed in terms of other probabilities and we can take $(p_1, ..., p_{a-1})$

and $(q_1, ..., q_{b-1})$ as free parameters of the model. This means that the dimension of the parameter set is

$$s = (a-1) \times (b-1).$$

Therefore, if we find the maximum likelihood estimates for the parameters of this model then the chi-squared statistic satisfies

$$Q = \sum_{ij} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \xrightarrow{w} \chi^2_{k-s-1} = \chi^2_{(a-1)(b-1)}$$

To formulate the test it remains to find the maximum likelihood estimates of the parameters. We need to maximize the likelihood function

$$\prod_{ij} (p_i q_j)^{N_{ij}} = \Big(\prod_i p_i^{N_{i+}}\Big)\Big(\prod_j p_j^{N_{+j}}\Big),$$

where $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$. Since p_is and q_js are not related to each other, maximizing the likelihood function above is equivalent to maximizing $\prod_i p_i^{N_{i+}}$ and $\prod_j p_j^{N_{+j}}$ separately. We have

$$\ln\Big(\prod_i p_i^{N_{i+}}\Big) = \sum_{i=1}^{a-1} N_{i+} \ln p_i + N_{a+} \ln(1 - p_1 - \cdots - p_{a-1}).$$

An elementary computation shows that

$$p_i^* = \frac{N_{i+}}{n}, \quad i = 1, \ldots, a.$$

Similarly, the MLE for q_j is

$$q_j^* = \frac{N_{+j}}{n}, \quad ij = 1, \ldots, b.$$

Therefore, chi-square statistic Q in this case can be written as

$$Q = \sum_{ij} \frac{\left(N_{ij} - \frac{N_{i+}N_{+j}}{n}\right)^2}{\frac{N_{i+}N_{+j}}{n}}.$$

We reject H_0 if $Q > c_{\alpha,(a-1)(b-1)}$ where the threshold $c_{\alpha,(a-1)(b-1)}$ is determined from the condition

$$\mathbb{P}[\chi^2_{(a-1)(b-1)} > c_{\alpha,(a-1)(b-1)}] = \alpha.$$

6.6.3 Test of homogeneity

Suppose that the population is divided into R groups and each group (or the entire population) is divided into C categories. We would like to test whether the distribution of categories in each group is the same.

	Category 1	Category 2	$\cdots$	Category C	Σ
Group 1	N_{11}	N_{12}	$\cdots$	N_{1C}	N_{1+}
Group 2	N_{21}	N_{22}	$\cdots$	N_{2b}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group R	N_{R1}	N_{R2}	$\cdots$	N_{RC}	N_{R+}
Σ	N_{+1}	N_{+2}	$\cdots$	N_{+C}	n

If we denote

$$p_{ij} = \mathbb{P}(Category_j | Group_i)$$

so that for each group i we have

$$\sum_{j=1}^{C} p_{ij} = 1,$$

then we want to test the following hypotheses

$$\begin{aligned} &H_0 : p_{ij} = p_j \text{ for all groups} i \leq R \\ &H_1 : \text{ otherwise}. \end{aligned}$$

If observations $X_1, ..., X_n$ are sampled independently from the entire population then homogeneity over groups is the same as independence of groups and categories. Indeed, if have homogeneity

$$\mathbb{P}(Category_j | Group_i) = \mathbb{P}(Category_j),$$

then we have

$$\mathbb{P}(Category_j, Group_i) = \mathbb{P}(Category_j)\mathbb{P}(Group_i).$$

This means that to test homogeneity we can use the test of independence above. Denote

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(N_{ij} - \frac{N_{i+}N_{+j}}{n}\right)^2}{\frac{N_{i+}N_{+j}}{n}} \xrightarrow{w} \chi^2_{(C-1)(R-1)}.$$

We reject H_0 at the significance level α if $Q > c_{\alpha,(C-1)(R-1)}$ where the threshold $c_{\alpha,(C-1)(R-1)}$ is determined from the condtion

$$\mathbb{P}[\chi^2_{(C-1)(R-1)} > c_{\alpha,(C-1)(R-1)}] = \alpha.$$

Example 6.6.6. In this example, 100 people were asked whether the service provided by the fire department in the city was satisfactory. Shortly after the survey, a large fire occured in the city. Suppose that the same 100 people were asked whether they thought that the service provided by the fire department was satisfactory. The result are in the following table:

	Satisfactory	Unsatisfactory
Before fire	80	20
After fire	72	28

Suppose that we would like to test whether the opinions changed after the fire by using a chi-squared test. However, the i.i.d. sample consisted of pairs of opinions of 100 people (X_i^1, X_i^2), $i = 1, \ldots, 100$ where the first coordinate/feature is a persons opinion before the fire and it belongs to one of two categories

$$\{\text{“}Satisfactory\text{”}, \text{“}Unsatisfactory\text{”}\},$$

and the second coordinate/feature is a persons opinion after the fire and it also belongs to one of two categories

$$\{\text{“}Satisfactory\text{”}, \text{“}Unsatisfactory\text{”}\},$$

So the correct contingency table corresponding to the above data and satisfying the assumption of the chi-squared test would be the following:

	Satisfactory	Unsatisfactory
Satisfactory	70	10
Unsatisfactory	2	18

In order to use the first contingency table, we would have to poll 100 people after the fire independently of the 100 people polled before the fire.

6.7 Exercises

6.7.1 Significance level and power function

6.1. Suppose that X has a pdf of the form $f(x;\theta) = \theta x^{\theta-1}\mathbb{I}_{\{0<x<1\}}$ where $\theta \in \{1,2\}$. To test the simple hypotheses $H_0 : \theta = 1$ against $H_1 : \theta = 2\}$, one uses a random sample X_1, X_2 of size $n = 2$ and define the critical region to be $C = \{(x_1, x_2) : x_1 x_2 \geq \frac{3}{4}\}$. Find the power function of the test.

6.2. Suppose that X has a binomial distribution with the number of trials $n = 10$ and with $p \in \{\frac{1}{4}, \frac{1}{2}\}$. The simple hypothesis $H_0 : p = \frac{1}{2}$ is rejected, and the alternative simple hypothesis $H_1 : p = \frac{1}{4}$ is accepted, if the observed value of X_1, a random sample of size 1, is less than or equal to 3. Find the significance level and the power of the test.

6.3. Let us say the life of a light bulb, say X, is normally distributed with mean θ and standard deviation 5000. Past experience indicates that $\theta = 30000$. The manufacturer claims that the light bulb made by a new process have mean $\theta > 30000$. It is possible that $\theta = 35000$. Check his claim by testing $H_0 : \theta = 30000$ against $H_1 : \theta > 30000$. We shall observe n independent values of X, say $X_1, \ldots, X_n$, and we shall reject H_0 (thus accept H_1) if and only if $\bar{x} \geq c$. Determine n and c so that the power function $\gamma(\theta)$ of the test has the values $\gamma(30000) = 0,01$ and $\gamma(35000) = 0,98$.

6.4. Suppose that X has a Poisson distribution with mean λ. Consider the simple hypothesis $H_0 : \lambda = \frac{1}{2}$ and the alternative composite hypothesis $H_1 : \lambda < \frac{1}{2}$. Let $X_1, \ldots, X_{12}$ denote a random sample of size 12 from this distribution. One rejects H_0 if and only if the observed value of $Y = X_1 + \ldots + X_{12} \leq 2.$. Find $\gamma(\lambda)$ for $\lambda \in (0, \frac{1}{2}]$ and the significance level of the test.

6.5. Let $Y_1 < Y_2 < Y_2 < Y_4$ be the order statistics of a random sample of size $n = 4$ from a distribution with pdf $f(x;\theta) = 1/\theta, 0 < x < \theta$, zero elsewhere, where $\theta > 0$. The hypothesis $H_0 : \theta = 1$ is rejected and $H_1 : \theta > 1$ is accepted if the observed $Y_4 \geq c$.

1. Find the constant c so that the significance level is $\alpha = 0.05$.
2. Determine the power function of the test.

6.7.2 Null distribution

6.6. Let $X_1, \ldots, X_n$ be a random sample from a $N(a_0, \sigma^2)$ distribution where $0 < \sigma^2 < \infty$ and a_0 is known. Show that the likelihood ratio test of $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$ can be based upon the statistics $W = \sum_{i=1}^n (X_i - a_0)^2/\sigma_0^2$. Determine the null distribution of W and give the rejection rule for a level α test.

6.7. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with mean $\lambda > 0$.

1. Show that the likelihood ratio test of $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ is based upon the statistic $Y = X_1 + \ldots + X_n$. Obtain the null distribution of Y.

2. For $\lambda_0 = 2$ and $n = 5$, find the significance level of the test that rejects H_0 if $Y \leq 4$ or $Y \geq 17$.

6.8. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli $B(1, \theta)$ distribution, where $0 < \theta < 1$.

1. Show that the likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ is based upon the statistic $Y = X_1 + \ldots + X_n$. Obtain the null distribution of Y.

2. For $n = 100$ and $\theta_0 = 1/2$, find c_1 so that the test reject H_0 when $Y \leq c_1$ or $Y \geq c_2 = 100 - c_1$ has the approximate significance level $\alpha = 0.05$.

6.9. Let $X_1, \ldots, X_n$ be a random sample from a $\Gamma(\alpha = 3, \beta = \theta)$ distribution, where $0 < \theta < \infty$.

1. Show that the likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ is based upon the statistic $Y = X_1 + \ldots + X_n$. Obtain the null distribution of $2Y/\theta_0$.

2. For $\theta_0 = 3$ and $n = 5$, find c_1 and c_2 so that the test that rejects H_0 when $Y \leq c_1$ or $Y \geq c_2$ has significance level 0.05.

6.7.3 Best critical region

6.10. Let X_1, X_2 be a random sample of size 2 from a random variable X having the pdf $f(x;\theta) = \frac{e^{-x/\theta}}{\theta}\mathbb{I}_{\{0<x<\infty\}}$. Consider the simple hypothesis $H_0 : \theta = \theta' = 2$ and the alternative hypothesis $H_1 : \theta = \theta'' = 4$. Show that the best test of H_0 against H_1 may be carried out by use of the statistics $X_1 + X_2$.

6.11. Let $X_1, \ldots, X_n$ be a random sample of size 10 from a normal distribution $N(0, \sigma^2)$. Find a best critical region of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$. Is this a best critical region of size 0.05 for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 4$? Against $H_1 : \sigma^2 = \sigma_1^2 > 1$.

6.12. If $X_1, \ldots, X_n$ is a random sample from a distribution having pdf of the form $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, zero elsewhere, show that a best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$ is $C = \Big\{(x_1, \ldots, x_n) : c \leq x_1 x_2 \ldots x_n\Big\}$.

6.13. Let $X_1, \ldots, X_n$ denote a random sample from a normal distribution $N(\theta, 100)$. Show that $C = \big\{(x_1, \ldots, x_n) : \bar{x} \geq c\big\}$ is a best critical region for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$. Find n and c so that

$$\mathbb{P}_{H_0}[(X_1, \ldots, X_n) \in C] = \mathbb{P}_{H_0}[\bar{X} \geq c] = 0.05$$

and

$$\mathbb{P}_{H_1}[(X_1, \dots, X_n) \in C] = \mathbb{P}_{H_1}[\bar{X} \geq c] = 0.90,$$

approximately.

6.14. Let $X_1, \dots, X_n$ be iid with pmf $f(x;p) = p^x(1-p)^{1-x}, x = 0, 1$, zero elsewhere. Show that $C = \{(x_1, \dots, x_n) : \sum x_i \leq c\}$ is a best critical region for testing $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{1}{3}$. Use the Central Limit Theorem to find n and c so that approximately $\mathbb{P}_{H_0}[\sum X_i \leq c] = 0.10$ and $\mathbb{P}_{H_1}[\sum X_i \leq c] = 0.80$.

6.15. Let $X_1, \dots, X_{10}$ denote a random sample of size 10 from a Poisson distribution with mean λ. Show that the critical region C defined by $\sum_{i=1}^{10} x_i \geq 3$ is a best critical region for testing $H_0 : \lambda = 0.1$ against $H_1 : \lambda = 0.5$. Determine, for this test, the significance level α and the power at $\theta = 0.5$.

6.16. Let X have the pmf $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$, zero elsewhere. We test the simple hypothesis $H_0 : \lambda = \frac{1}{4}$ against the alternative composite hypothesis $H_1 : \theta < \frac{1}{4}$ by taking a random sample of size 10 and rejecting $H_0 : \theta = \frac{1}{4}$ iff the observed values $x_1, \dots, x_10$ of the sample observations are such that $\sum_{i=1}^{10} x_i \leq 1$. Find the power function $\gamma(\theta)$, $0 < \theta \leq \frac{1}{4}$, of this test.

6.7.4 Some tests for single sample

Tests on mean

6.17. (a) The sample mean and standard deviation from a random sample of 10 observations from a normal population were computed as $\overline{x} = 23$ and $\sigma = 9$. Calculate the value of the test statistic of the test required to determine whether there is enough evidence to infer at the 5% significance level that the population mean is greater than 20.
(b) Repeat part (a) with $n = 30$.
(c) Repeat part (b) with $n = 40$.

6.18. (a) A statistics practitioner is in the process of testing to determine whether there is enough evidence to infer that the population mean is different from 180. She calculated the mean and standard deviation of a sample of 200 observations as $\overline{x} = 175$ and $\sigma = 22$. Calculate the value of the test statistic of the test required to determine whether there is enough evidence at the 5% significance level.
(b) Repeat part (a) with $s = 45$.
(c) Repeat part (a) with $s = 60$.

6.19. A courier service advertises that its average delivery time is less than 6 hours for local deliveries. A random sample of times for 12 deliveries to an address across town was recorded. These data are shown here. Is this sufficient evidence to support the couriers advertisement, at the 5% level of significance?

$$3.03,\ 6.33,\ 7.98,\ 4.82,\ 6.50,\ 5.22,\ 3.56,\ 6.76,\ 7.96,\ 4.54,\ 5.09,\ 6.46.$$

$\overline{X} = 5,6875;\ s^2 = 2,1325;\ T_0 = -0.7413.$

6.20. Aircrew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 centimeters per second. We know that the standard deviation of burning rate is $\sigma = 2$ centimeters per second. The experimenter decides to specify a type I error probability or significance level of $\alpha = 0.05$ and selects a random sample of $n = 25$ and obtains a sample average burning rate of $\overline{X} = 51.3$ centimeters per second. What conclusions should be drawn?

6.21. The mean water temperature downstream from a power plant cooling tower discharge pipe should be no more than $100°F$. Past experience has indicated that the standard deviation of temperature is $2°F$. The water temperature is measured on nine randomly chosen days, and the average temperature is found to be $98°$F.
(a) Should the water temperature be judged acceptable with $\alpha = 0.05$?
(b) What is the P-value for this test?
(c) What is the probability of accepting the null hypothesis at $\alpha = 0.05$ if the water has a true mean temperature of $104°F$?

6.22. A study reported body temperatures $(°F)$ for 25 female subjects follow:
97.8, 97.2, 97.4, 97.6, 97.8, 97.9, 98.0, 98.0, 98.0, 98.1, 98.2, 98.3,
98.3, 98.4, 98.4, 98.4, 98.5, 98.6, 98.6, 98.7, 98.8, 98.8, 98.9, 98.9, and 99.0.
(a) Test the hypotheses $H_0 : \mu = 98.6$ versus $H_1 : \mu \neq 98.6$, using $\alpha = 0.05$. Find the P-value.
(b) Compute the power of the test if the true mean female body temperature is as low as 98.0.
(c) What sample size would be required to detect a true mean female body temperature as low as 98.2 if we wanted the power of the test to be at least 0.9?

6.23. Cloud seeding has been studied for many decades as a weather modification procedure. The rainfall in acre-feet from 20 clouds that were selected at random and seeded with silver nitrate follows:
18.0, 30.7, 19.8, 27.1, 22.3, 18.8, 31.8, 23.4, 21.2, 27.9,
31.9, 27.1, 25.0, 24.7, 26.9, 21.8, 29.2, 34.8, 26.7, 31.6.
(a) Can you support a claim that mean rainfall from seeded clouds exceeds 25 acre-feet? Use $\alpha = 0.01$.
(b) Compute the power of the test if the true mean rainfall is 27 acre-feet.
(c) What sample size would be required to detect a true mean rainfall of 27.5 acre-feet if we wanted the power of the test to be at least 0.9?

6.24. The life in hours of a battery is known to be approximately normally distributed, with standard deviation $\sigma = 1.25$ hours. A random sample of 10 batteries has a mean life of $\overline{x} = 40.5$ hours.
(a) Is there evidence to support the claim that battery life exceeds 40 hours? Use $\alpha = 0.05$.
(b) What is the P-value for the test in part (a)?
(c) What is the power for the test in part (a) if the true mean life is 42 hours?
(d) What sample size would be required to ensure that the probability of making type II error

does not exceed 0.10 if the true mean life is 44 hours?
(e) Explain how you could answer the question in part (a) by calculating an appropriate confidence bound on life.

6.25. Medical researchers have developed a new artificial heart constructed primarily of titanium and plastic. The heart will last and operate almost indefinitely once it is implanted in the patients body, but the battery pack needs to be recharged about every four hours. A random sample of 50 battery packs is selected and subjected to a life test. The average life of these batteries is 4.05 hours. Assume that battery life is normally distributed with standard deviation $\sigma = 0.2$ hour.
(a) Is there evidence to support the claim that mean battery life exceeds 4 hours? Use $\alpha = 0.05$.
(b) Compute the power of the test if the true mean battery life is 4.5 hours.
(c) What sample size would be required to detect a true mean battery life of 4.5 hours if we wanted the power of the test to be at least 0.9?
(d) Explain how the question in part (a) could be answered by constructing a one-sided confidence bound on the mean life.

Tests on population variance

6.26. After many years of teaching, a statistics professor computed the variance of the marks on her final exam and found it to be $\sigma^2 = 250$. She recently made changes to the way in which the final exam is marked and wondered whether this would result in a reduction in the variance. A random sample of this years final exam marks are listed here. Can the professor infer at the 10% significance level that the variance has decreased?

$$57 \quad 92 \quad 99 \quad 73 \quad 62 \quad 64 \quad 75 \quad 70 \quad 88 \quad 60.$$

6.27. With gasoline prices increasing, drivers are more concerned with their cars' gasoline consumption. For the past 5 years, a driver has tracked the gas mileage of his car and found that the variance from fill-up to fill-up was $\sigma^2 = 23$ mpg^2. Now that his car is 5 years old, he would like to know whether the variability of gas mileage has changed. He recorded the gas mileage from his last eight fill-ups; these are listed here. Conduct a test at a 10% significance level to infer whether the variability has changed.

$$28 \quad 25 \quad 29 \quad 25 \quad 32 \quad 36 \quad 27 \quad 24.$$

Tests on proportion

6.28. (a) Calculate the P-value of the test of the following hypotheses given that $\hat{p} = 0.63$ and $n = 100$:

$$H_0 : p = 0.6 \quad \textbf{vs} \quad H_1 : p > 0.6.$$

(b) Repeat part (a) with $n = 200$.
(c) Repeat part (a) with $n = 400$.
(d) Describe the effect on P-value of increasing sample size.

6.29. Has the recent drop in airplane passengers resulted in better on-time performance? Before the recent economic downturn, one airline bragged that 92% of its flights were on time. A random sample of 165 flights completed this year reveals that 153 were on time. Can we conclude at the 5% significance level that the airlines on-time performance has improved?

6.30. In a random sample of 85 automobile engine crank- shaft bearings, 10 have a surface finish roughness that exceeds the specifications. Does this data present strong evidence that the proportion of crankshaft bearings exhibiting excess surface roughness exceeds 0.10? State and test the appropriate hypotheses using $\alpha = 0.05$.

6.31. An study claimed that nearly one-half of all engineers continue academic studies beyond the B.S. degree, ultimately receiving either an M.S. or a Ph.D. degree. Data from an article in Engineering Horizons (Spring 1990) indicated that 117 of 484 new engineering graduates were planning graduate study.
(a) Are the data from Engineering Horizons consistent with the claim reported by Fortune? Use $\alpha = 0.05$ in reaching your conclusions.
(b) Find the P-value for this test.
(c) Discuss how you could have answered the question in part (a) by constructing a two-sided confidence interval on p.

6.32. A researcher claims that at least 10% of all football helmets have manufacturing flaws that could potentially cause injury to the wearer. A sample of 200 helmets revealed that 16 helmets contained such defects.
(a) Does this finding support the researchers claim? Use $\alpha = 0.01$.
(b) Find the P-value for this test.

6.7.5 Some tests for two samples

Compare two means

6.33. In random samples 12 from each of two normal populations, we found the following statistics: $\overline{x}_1 = 74, s_1 = 18$ and $\overline{x}_2 = 71, s_2 = 16$.
(a) Test with $\alpha = 0.05$ to determine whether we can infer that the population means differ.
(b) Repeat part (a) increasing the standard deviation to $s_1 = 210$ and $s_2 = 198$.
(c) Describe what happens when the sample stan- dard deviations get larger.
(d) Repeat part (a) with sample size 150.
(e) Discuss the effects of increasing the sample size.

6.34. Random sampling from two normal populations produced the following results

$$\begin{array}{lll} \overline{x}_1 = 412 & s_1 = 128 & n_1 = 150 \\ \overline{x}_2 = 405 & s_2 = 54 & n_2 = 150. \end{array}$$

(a) Can we infer that at the 5% significance level that μ_1 is greater than μ_2.
(b) Repeat part (a) decreasing the standard deviation to $s_1 = 31$, $s_2 = 16$.

(c) Describe what happens when the sample stan- dard deviations get smaller.
(d) Repeat part (a) with samples of size 20.
(e) Discuss the effects of decreasing the sample size.
(f) Repeat part (a) changing the mean of sample 1 to $\overline{x}_1 = 409$.

6.35. Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The fill volume can be assumed normal, with standard deviation $\sigma_1 = 0.020$ and $\sigma_2 = 0.025$ ounces. A member of the quality engineering staff suspects that both machines fill to the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from the output of each machine.

Machine 1		Machine 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

(a) Do you think the engineer is correct? Use $\alpha = 0.05$.
(b) What is the P-value for this test?
(c) What is the power of the test in part (a) for a true difference in means of 0.04?
(d) Find a 95% confidence interval on the difference in means. Provide a practical interpretation of this interval.
(e) Assuming equal sample sizes, what sample size should be used to assure that the probability of making type II error is 0.05 if the true difference in means is 0.04? Assume that $\alpha = 0.05$.

6.36. Every month a clothing store conducts an inventory and calculates losses from theft. The store would like to reduce these losses and is considering two methods. The first is to hire a security guard, and the second is to install cameras. To help decide which method to choose, the manager hired a security guard for 6 months. During the next 6-month period, the store installed cameras. The monthly losses were recorded and are listed here. The manager decided that because the cameras were cheaper than the guard, he would install the cameras unless there was enough evidence to infer that the guard was better. What should the manager do?

Security guard	355	284	401	398	477	254
Cameras	486	303	270	386	411	435

Pair t-test

6.37. Many people use scanners to read documents and store them in a Word (or some other software) file. To help determine which brand of scanner to buy, a student conducts an experiment in which eight documents are scanned by each of the two scanners he is interested in. He records the number of errors made by each. These data are listed here. Can he infer that brand A (the more expensive scanner) is better than brand B?

Document	1	2	3	4	5	6	7	8
BrandA	17	29	18	14	21	25	22	29
BrandB	21	38	15	19	22	30	31	37

6.38. In an effort to determine whether a new type of fertilizer is more effective than the type currently in use, researchers took 12 two-acre plots of land scattered throughout the county. Each plot was divided into two equal-sized subplots, one of which was treated with the current fertilizer and the other with the new fertilizer. Wheat was planted, and the crop yields were measured.

Plot	1	2	3	4	5	6	7	8	9	10	11	12
Current fertilizer	56	45	68	72	61	69	57	55	60	72	75	66
New fertilizer	60	49	66	73	59	67	61	60	58	75	72	68

(a) Can we conclude at the 5% significance level that the new fertilizer is more effective than the current one?
(b) Estimate with 95% confidence the difference in mean crop yields between the two fertilizers.
(c) What is the required condition(s) for the validity of the results obtained in parts (a) and (b)?

Compare two variances

6.39. Random samples from two normal population produced the following statistics

$$s_1^2 = 350, \quad n_1 = 30, \quad s_2^2 = 700, \quad n_2 = 30.$$

(a) Can we infer at the 10% significance level that the two population variances differ?
(b) Repeat part (a) changing the sample sizes to $n_1 = 15$ and $n_2 = 15$.
(c) Describe what happens to the test statistics and the conclusion when the sample sizes decrease.

6.40. A statistics professor hypothesized that not only would the means vary but also so would the variances if the business statistics course was taught in two different ways but had the same final exam. He organized an experiment wherein one section of the course was taught using detailed PowerPoint slides whereas the other required students to read the book and answer questions in class discussions. A sample of the marks was recorded and listed next. Can we infer that the variances of the marks differ between the two sections?

Class 1	64	85	80	64	48	62	75	77	50	81	90
Class 2	73	78	66	69	79	81	74	59	83	79	84

6.41. An operations manager who supervises an assembly line has been experiencing problems with the sequencing of jobs. The problem is that bottle- necks are occurring because of the inconsistency of sequential operations. He decides to conduct an experiment wherein two different methods are used to complete the same task. He measures the times (in seconds). The data are listed here. Can he infer that the second method is more consistent than the first method?

Method 1	8.8	9.6	8.4	9.0	8.3	9.2	9.0	8.7	8.5	9.4
Method 2	9.2	9.4	8.9	9.6	9.7	8.4	8.8	8.9	9.0	9.7

Compare two proportions

6.42. Random samples from two binomial populations yielded the following statistics:

$$\hat{p}_1 = 0.45 \quad n_1 = 100 \quad \hat{p}_2 = 0.40 \quad n_2 = 100.$$

(a) Calculate the P-value of a test to determine whether we can infer that the population proportions differ.
(b) Repeat part (a) increasing the sample sizes to 400.
(c) Describe what happens to the p-value when the sample sizes increase.

6.43. Random samples from two binomial populations yielded the following statistics:

$$\hat{p}_1 = 0.60 \quad n_1 = 225 \quad \hat{p}_2 = 0.55 \quad n_2 = 225.$$

(a) Calculate the P-value of a test to determine whether we there is evidence to infer that the population proportions differ.
(b) Repeat part (a) $\hat{p}_1 = 0.95$ and $\hat{p}_2 = 0.90$.
(c) Describe the effect on the P-value of increasing the sample proportions.
(d) Repeat part (a) $\hat{p}_1 = 0.10$ and $\hat{p}_2 = 0.05$.
(e) Describe the effect on the P-value of decreasing the sample proportions.

6.44. Many stores sell extended warranties for products they sell. These are very lucrative for store owners. To learn more about who buys these warranties, a random sample was drawn of a stores customers who recently purchased a product for which an extended warranty was available. Among other vari- ables, each respondent reported whether he or she paid the regular price or a sale price and whether he or she purchased an extended warranty.

	Regular Price	Sale Price
Sample size	229	178
Number who bought extended warranty	47	25

Can we conclude at the 10% significance level that those who paid the regular price are more likely to buy an extended warranty?

6.45. Surveys have been widely used by politicians around the world as a way of monitoring the opinions of the electorate. Six months ago, a survey was undertaken to determine the degree of support for a national party leader. Of a sample of 1100, 56% indicated that they would vote for this politician. This month, another survey of 800 voters revealed that 46% now support the leader.
(a) At the 5% significance level, can we infer that the national leaders popularity has decreased?
(b) At the 5% significance level, can we infer that the national leaders popularity has decreased by more than 5%?

6.46. A random sample of 500 adult residents of Maricopa County found that 385 were in favour of increasing the highway speed limit to 75 mph, while another sample of 400 adult residents of Pima County found that 267 were in favour of the increased speed limit. Do these data indicate that there is a difference in the support for increasing the speed limit between the residents of the two counties? Use $\alpha = 0.05$. What is the P-value for this test?

6.47. Two different types of injection-molding machines are used to form plastic parts. A part is considered defective if it has excessive shrinkage or is discolored. Two random samples, each of size 300, are selected, and 15 defective parts are found in the sample from machine 1 while 8 defective parts are found in the sample from machine 2. Is it reasonable to conclude that both machines produce the same fraction of defective parts, using $\alpha = 0.05$? Find the P-value for this test.

6.7.6 Chi-squared tests

Chi-squared test on distribution

6.48. A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of Sixes	0	1	2	3
Number of Rolls	48	35	15	3

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. What do they conclude?

6.49. Suppose that the distribution of the heights of men who reside in a certain large city is the normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. Suppose also that when the heights of 500 men who reside in a certain neighbourhood of the city were measured, the distribution in the following table was obtained. Test the hypothesis that, with regard to height, these 500 men form a random sample from all the men who reside in the city.

Height (in inch)	<66	66-67.5	67.7-68.5	68.5-70	>70
Number of men	18	177	198	102	5

6.50. The 50 values in the following table are intended to be a random sample from the standard normal distribution.

1.28	1.22	0.32	0.80	1.38	1.26	2.33	0.34	1.14	0.64
0.41	0.01	0.49	0.36	1.05	0.04	0.35	2.82	0.64	0.56
0.45	1.66	0.49	1.96	3.44	0.67	1.24	0.76	0.46	0.11
0.35	1.39	0.14	0.64	1.67	1.13	0.04	0.61	0.63	0.13
0.72	0.38	0.85	1.32	0.85	0.41	0.11	2.04	1.61	1.81

a) Carry out a χ^2 test of goodness-of-fit by dividing the real line into five intervals, each of which has probability 0.2 under the standard normal distribution.
b) Carry out a χ^2 test of goodness-of-fit by dividing the real line into ten intervals, each of which has probability 0.1 under the standard normal distribution.

Chapter 7

Regression

7.1 Simple linear regression

7.1.1 Simple linear regression model

Suppose that we have a pair of variables (X, Y) and a variable Y is a linear function of X plus random noise:

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon,$$

where a random noise ϵ is assumed to have normal distribution $\mathcal{N}(0, \sigma^2)$. A variable X is called a predictor variable, Y - a response variable and a function $f(x) = \beta_0 + \beta_1 x$ - a linear regeression function.

Suppose that we are given a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are described by the above model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. We have three unknown parameters β_0, β_1 and σ^2 and we want to estimate them using a given sample. The points $X_1, \ldots, X_n$ can be either random or non random, but from the point of view of estimating linear regression function the nature of Xs is in some sense irrelevant so we will think of them as fixed and non random and assume that the randomness comes from the noise variables ϵ_i. For a fixed X_i, the distribution of Y_i is equal to $\mathcal{N}(f(X_i), \sigma^2)$. The likelihood function of the sequence $(Y_1, \ldots, Y_n)$ is

$$L(Y_1, \ldots, Y_n; \beta_0, \beta_1, \sigma^2) = (2\pi\sigma)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f(X_i))^2}.$$

Let us find the m.l.e. of $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ that maximize this likelihood function L. First of all, it is obvious that $(\hat{\beta}_0, \hat{\beta}_1)$ is also minimized

$$L^*(\beta_0, \beta_1) := \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

so $\hat{\beta}_0, \hat{\beta}_1$ are solution to

$$\begin{cases} \dfrac{\partial L^*}{\partial \beta_0} = -\sum_{i=1}^n 2(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\ \dfrac{\partial L^*}{\partial \beta_1} = -\sum_{i=1}^n 2(Y_i - (\beta_0 + \beta_1 X_i))X_i = 0. \end{cases}$$

Denote

$$\bar{X} = \frac{1}{n}\sum X_i, \quad \bar{Y} = \frac{1}{n}\sum Y_i, \quad \overline{X^2} = \frac{1}{n}\sum X_i^2, \quad \overline{XY} = \frac{1}{n}\sum_i X_i Y_i,$$

we obtain

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \end{aligned}$$

Denote $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

The numerator in the last sum is the sum of squares of the residuals and the numerator is the variance of Y and R^2 is usually interpreted as the proportion of variability in the data explained by the linear model. The higher R^2 the better our model explains the data. Next, we would like to do statistical inference about the linear model.

1. Construct confidence intervals for parameters of the model β_0, β_1 and σ^2.
2. Construct prediction intervals for Y given any point X.
3. Test hypotheses about parameters of the model.

The distribution of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$ are defined by the following result.

Proposition 7.1.1. *1. Vector* $(\hat{\beta}_0, \hat{\beta}_1)$ *has a normal distribution with mean* (β_0, β_1) *and covariance matrix*

$$\Sigma = \frac{\sigma^2}{n\sigma_x^2}\begin{pmatrix} \overline{X^2} & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}, \quad \textit{where } \sigma_x^2 = \overline{X^2} - \bar{X}^2.$$

2. $\hat{\sigma}^2$ *is independent of* $\hat{\beta}_0, \hat{\beta}_1$.

3. $\frac{n\hat{\sigma}^2}{\sigma^2}$ *has* χ^2_{n-2} *distribution with* $n-2$ *degrees of freedom.*

7.1.2 Confidence interval for σ^2

It follows from Proposition 7.1.1 that $\frac{n\hat{\sigma}^2}{\sigma^2}$ is χ^2_{n-2} distributed, so if we choose $c_{1-\alpha/2,n-1}, c_{\alpha/2,n-1}$ such that

$$\mathbb{P}[\chi^2_{n-1} > c_{1-\alpha/2,n-1}] = 1 - \frac{\alpha}{2}, \quad \mathbb{P}[\chi^2_{n-1} > c_{\alpha/2,n-1}] = \frac{\alpha}{2},$$

then

$$\mathbb{P}\Big[\frac{n\hat{\sigma}^2}{c_{\alpha/2,n-1}} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_{1-\alpha/2,n-1}}\Big] = 1-\alpha.$$

Therefore the $(1-\alpha)$ CI for σ^2 is

$$\frac{n\hat{\sigma}^2}{c_{\alpha/2,n-1}} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_{1-\alpha/2,n-1}}.$$

7.1.3 Confidence interval for β_1

It follows from Proposition 7.1.1 that

$$\sqrt{\frac{n\sigma_x^2}{\sigma^2}}(\hat{\beta}_1 - \beta_1) \sim \mathcal{N}(0,1) \quad \text{and } \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

and $\hat{\beta}_1$ is independent of $\hat{\sigma}^2$, so

$$\frac{\sqrt{\frac{n\sigma_x^2}{\sigma^2}}(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}}} = (\hat{\beta}_1 - \beta_1)\sqrt{\frac{(n-2)\sigma_x^2}{\hat{\sigma}^2}}$$

has a t_{n-2} distribution with $n-2$ degrees of freedom. Choose x_α such that

$$\mathbb{P}[|t_{n-2}| < x_\alpha] = 1-\alpha$$

we obtain the $(1-\alpha)$ CI for β_1 as follows

$$\hat{\beta}_1 - x_\alpha\sqrt{\frac{\hat{\sigma}^2}{(n-2)\sigma_x^2}} \le \beta_1 \le \hat{\beta}_1 + x_\alpha\sqrt{\frac{\hat{\sigma}^2}{(n-2)\sigma_x^2}}.$$

7.1.4 Confidence interval for β_0

A similar argument as above yields

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\Big(\frac{1}{n} + \frac{\bar{X}^2}{n\sigma_x^2}\Big)\sigma^2}} : \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(1 + \frac{\bar{X}^2}{\sigma_x^2}\Big)}}$$

has a Student's t distribution with $n-2$ degrees of freedom. Thus the $(1-\alpha)$ CI for β_0 is

$$\hat{\beta}_0 - x_\alpha\sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(1 + \frac{\bar{X}^2}{\sigma_x^2}\Big)} \le \beta_0 \le \hat{\beta}_0 + x_\alpha\sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(1 + \frac{\bar{X}^2}{\sigma_x^2}\Big)}.$$

7.1.5 Prediction intervals

Suppose now that we have a new observation X for which Y is unknown and we want to predict Y or find the confidence interval for Y. According to simple regression model,

$$Y = \beta_0 + \beta_1 X + \epsilon$$

and it is natural to take $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ as the prediction of Y. Let us find the distribution of their difference $\hat{Y} - Y$.

Proposition 7.1.2. *The random variable*

$$\frac{\hat{Y} - Y}{\sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(n+1+\frac{(\bar{X}-X)^2}{\sigma_x^2}\Big)}}$$

has a Student's t distribution with $n-2$ degrees of freedom.

Choose x_α such that $\mathbb{P}[|t_{n-2}| < x_\alpha] = 1-\alpha$ we obtain the $(1-\alpha)$ CI for Y as follows

$$\hat{Y} - x_\alpha \sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(n+1+\frac{(\bar{X}-X)^2}{\sigma_x^2}\Big)} \le Y \le \hat{Y} + x_\alpha \sqrt{\frac{\hat{\sigma}^2}{n-2}\Big(n+1+\frac{(\bar{X}-X)^2}{\sigma_x^2}\Big)}.$$

Appendies

Table of Normal distribution $\Phi(z) = \int_{-\infty}^{z} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.5398	0.5438	0.54776	0.55172	0.55567	0.55966	0.5636	0.56749	0.57142	0.57535
0.2	0.5793	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

Table of Student distribution[1]

1 side	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
2 side	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.08	1.386	1.886	2.92	4.303	6.965	9.925	14.09	22.33	31.6
3	0.765	0.978	1.25	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.19	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.61
5	0.727	0.92	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.44	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.86	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.1	1.383	1.833	2.262	2.821	3.25	3.69	4.297	4.781
10	0.7	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.93	4.318
13	0.694	0.87	1.079	1.35	1.771	2.16	2.65	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.14
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.69	0.865	1.071	1.337	1.746	2.12	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.74	2.11	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.33	1.734	2.101	2.552	2.878	3.197	3.61	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.86	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.85
21	0.686	0.859	1.063	1.323	1.721	2.08	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.06	1.319	1.714	2.069	2.5	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.06	2.485	2.787	3.078	3.45	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.69
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.31	1.697	2.042	2.457	2.75	3.03	3.385	3.646
40	0.681	0.851	1.05	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2	2.39	2.66	2.915	3.232	3.46
80	0.678	0.846	1.043	1.292	1.664	1.99	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.29	1.66	1.984	2.364	2.626	2.871	3.174	3.39
120	0.677	0.845	1.041	1.289	1.658	1.98	2.358	2.617	2.86	3.16	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.96	2.326	2.576	2.807	3.09	3.291

[1] $\mathbb{P}[T_1 < 1.376] = 0.8 \,\mathrm{v}\, \mathbb{P}[|T_1| < 1.376] = 0.6$

Table of χ^2-distribution $\mathbb{P}[\chi_n^2 > \alpha]$

DF: n	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.00004	0.001	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797
22	8.643	10.982	27.301	30.813	33.924	36.781	37.659	40.289	42.796	45.962	48.268
23	9.26	11.689	28.429	32.007	35.172	38.076	38.968	41.638	44.181	47.391	49.728
24	9.886	12.401	29.553	33.196	36.415	39.364	40.27	42.98	45.559	48.812	51.179
25	10.52	13.12	30.675	34.382	37.652	40.646	41.566	44.314	46.928	50.223	52.62
26	11.16	13.844	31.795	35.563	38.885	41.923	42.856	45.642	48.29	51.627	54.052
27	11.808	14.573	32.912	36.741	40.113	43.195	44.14	46.963	49.645	53.023	55.476
28	12.461	15.308	34.027	37.916	41.337	44.461	45.419	48.278	50.993	54.411	56.892
29	13.121	16.047	35.139	39.087	42.557	45.722	46.693	49.588	52.336	55.792	58.301
30	13.787	16.791	36.25	40.256	43.773	46.979	47.962	50.892	53.672	57.167	59.703

Bibliography

[1] Casella, George, and Roger L. Berger. Statistical inference. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[2] Cacoullos, T. (1989) Exercises in probability, Springer-Verlag New York Inc.

[3] DeGroot, M., & Mark J. Schervish. Probability and Statistics. 3rd ed. Boston, MA: Addison-Wesley, 2002.

[4] Hogg, R., McKean, J.W., & Craig, A.T. (2005) Introduction to Mathematical Statistics, 6th Edition. Pearson Education International.

[5] Jacod, J., Protter, P. (2003) Probability Essential. Springer.

[6] Montgomery, D. C., & Runger, G. C. (2010). Applied statistics and probability for engineers. John Wiley & Sons.

[7] Panchenko, D. (2006) Lecture note "Statistics for Applications". http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/readings/

[8] Rahman N. A. (1983) Theoretical exercises in probability and statistics, second edition. Macmillan Publishing.

[9] Rice, John. Mathematical statistics and data analysis. Nelson Education, 2006.

[10] Shao, J. (2005) Mathematical Statistics: Exercises and Solutions. Springer

[11] Yuri, S. & Kelbert, M. (2008) Probability and Statistics by Example: Volume 1 and 2. Cambridge University Press.

[12] Shevtsova, I. (2011). On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. arXiv preprint arXiv:1111.6554.

[13] Nguyen Duy Tien, Vu Viet Yen (2001) Probability Theory (in Vietnamese). Educational Publishing House.